Stata 9 Supplement for: Biostatistics for the Health Sciences

R. Clifford Blair

and

Richard A Taylor

May 29, 2007
Preface

This manual is not an introduction to Stata for Windows® and contains only a small portion of the material that would be found therein.¹ Rather, this manual teaches the reader how to use Stata Release 9 for Windows² to analyze data related to the problems addressed in Biostatistics For The Health Sciences by R. Clifford Blair and Richard A. Taylor.

To this end, Stata is used to analyze various of the data sets and problems contained in that text. The reader will be required, especially in the early part of the manual, to input selected data to the system. Additional data sets accompany this manual so that the reader will not be burdened with data entry once this skill is acquired.

It is assumed that the reader is familiar with the Microsoft Windows® operating system so that little instruction in this regard will be provided. It is further assumed that the reader is using Stata Release 9 rather than some earlier version. It is strongly suggested that the user avail herself/himself of some external source material on Stata while working through this manual. A good starting point would be Getting Started With Stata For Windows.[2]

Chapter 1 familiarizes the reader with some basic Stata concepts. Key among these is the entry and saving of data which is prerequisite to all that follows. Chapters 2 through 10 address the corresponding chapters in the text. That is, Chapter 2 in the manual addresses problems in Chapter 2 in the text and so on. Section headings in these latter chapters reflect the sections/pages in the text addressed in that section of the manual. Thus, while studying the text readers can quickly locate the comparable area in the manual by referring to the manual’s table of contents. In addition, a manual appendix relates specific examples, data tables etc. in the text to specific pages in the manual. These provisions allow easy integration of Stata analyses as study of the text progresses.

Finally, we would caution that data analysis, whether by hand calculations or software is no substitute for a clear understanding of the underlying statistical concept. Such analysis can, however, be an invaluable aid to mastery of such concepts and is indispensable for research purposes.

¹ Getting Started With Stata For Windows published by StataCorp of College Station TX provides just such an introduction.
² While this manual is specific to Stata for Windows, with minor exceptions it also applies to Mac OS and the various Unix environments including Linux. Data sets provided with this manual will also be appropriate for these systems. See [2] for details.
# Contents

1 Preliminaries .............................. 1  
1.1 Introduction .............................. 1  
1.2 Getting Started ............................ 1  
1.3 Inputting Data ............................. 3  
1.4 Performing Analyses ........................ 5  
1.5 Editing And Printing Your Output ....... 5  
1.6 Creating New Variables .................... 7  
1.7 Graphs ................................. 9  
Exercises .................................. 10  

2 Descriptive Methods ......................... 11  
2.1 Introduction .............................. 11  
2.4 Distributions ............................. 11  
2.5 Graphs ................................. 15  
2.6 Numerical Methods ........................ 21  
Exercises .................................. 25  

3 Probability .................................. 27  
3.1 Introduction .............................. 27  
3.3 Contingency Tables ........................ 27  
3.4 The Normal Curve ......................... 29  
Exercises .................................. 30  

4 Introduction To Inference And One Sample Methods 31  
4.1 Introduction .............................. 31  
4.2 Sampling Distributions ..................... 31  
4.3 Hypothesis Testing ........................ 32  
4.4 Confidence Intervals ....................... 39  
Exercises .................................. 41  

5 Paired Samples Methods ...................... 43  
5.1 Introduction .............................. 43  
5.2 Methods Related to Mean Difference ....... 43  
5.3 Methods Related to Proportions .......... 45  

iii
CONTENTS

5.4 Methods Related to Paired Samples Risk Ratios . . . . . . . . . 47
5.5 Methods Related to Paired Samples Odds Ratios . . . . . . . . . 48
Exercises .................................. 50

6 Two Independent Samples Methods 51
6.1 Introduction .................................. 51
6.2 Methods Related to Differences Between Means ............... 51
6.3 Methods Related to Proportions ......................... 53
6.4 Methods Related to Independent Samples Risk Ratios ...... 54
6.5 Independent Samples Odds Ratios ........................ 56
Exercises .................................. 58

7 Multi-Sample Methods 59
7.1 Introduction .................................. 59
7.2 The Oneway Analysis of Variance (ANOVA) F Test .......... 59
7.3 The 2 By \( k \) Chi-Square Test .......................... 60
7.4 Multiple Comparison Procedures ........................... 61
Exercises .................................. 62

8 The Assessment of Relationships 63
8.1 Introduction .................................. 63
8.2 The Pearson Product-Moment Correlation Coefficient .. 63
8.3 The Chi-Square Test For Independence ..................... 64
Exercises .................................. 65

9 Linear Regression 67
9.1 Introduction .................................. 67
9.2 Simple Linear Regression ............................. 67
9.3 Multiple Linear Regression ............................. 68
Exercises .................................. 70

10 Methods Based on the Permutation Principle 71
10.1 Introduction .................................. 71
10.3 Applications .................................. 71
Exercises .................................. 74

A Table Relating Text Items to Tasks 77
Chapter 1

Preliminaries

1.1 Introduction

The hand calculations you perform in conjunction with your studies of the statistical concepts in *Biostatistics For The Health Sciences* by R. Clifford Blair and Richard A. Taylor\(^1\) are designed to enhance your understanding of the concepts under study. Modern statistical practice only rarely resorts to such methods for the analysis of data collected for research purposes. This is because (1) data sets collected in applied research contexts are often too large for hand calculations and (2) such calculations are prone to errors. Instead, researchers resort to tried and true statistical software that can be relied upon to provide quick, accurate results. The Stata system is one of the most popular and powerful of these.

In this manual, you will learn to use Stata to analyze data associated with the problems outlined in the text. You will learn some basic Stata concepts in this chapter. Chapters 2 through 10 will address problems in the like numbered chapters in the text. No guarantee is provided that the methods outlined here are the best or most efficient way of accomplishing the task at hand. Indeed, it is our hope that as you experiment and gain experience with Stata, you will improve on these methods.

1.2 Getting Started

When you installed Stata you were asked to designate a working directory where your datasets, graphs and other Stata-related files will be kept. The suggested default is `c:\data`. If you don’t have such a directory on your system you should create one then copy all the files that accompany this manual into your working directory.

There are a number of ways to start Stata. We will use the simple method of double clicking on the shortcut icon \(\square\) we’ve placed on the desktop. If you don’t

\(^1\)Henceforth referred to as “the text.”
have the shortcut on your desktop you can select the sequence Start⇒Programs⇒Stata 9⇒Stata/SE. The notation Start⇒Programs⇒Stata 9⇒Stata/SE means that you first select the Start menu, then Programs, then Stata 9 then Stata/SE. The last entry in this sequence will either be Stata/SE, Intercooled Stata or Small Stata depending on which version you have installed on your system.

Unless some of the default settings for Stata have been altered, the windows shown in Figure 1.1 will appear. Other windows are available but we will concentrate on these basic four. The window in the lower right hand corner is the Command window. This is where you will enter instructions or commands to Stata to perform various tasks. For example, you can tell Stata to load a particular dataset into memory, compute various statistics on the data thus loaded or perform a myriad of other tasks. As an example, if you look at the bottom left portion of our screen you will see C:\aamystuff\data. This is the default directory we created earlier as discussed above for storing datasets, graphs etc.\footnote{No, we didn’t use the suggested default C:\data because that would make our lives too simple.} But suppose this is the wrong directory or you simply want to use a different directory for the project at hand. You can change the default directory by typing

\begin{figure}
\centering
\includegraphics[width=\textwidth]{stata_screen.png}
\caption{Screen appearance when Stata is started.}
\end{figure}
1.3. INPUTTING DATA

```
cd C:\aamystuff\books\basic\stata9
```

in the Command window and pushing Enter. This is the directory we are currently using to write this manual. As you will see, many of the commands that can be typed in at the Command window can also be entered by clicking various of the buttons above the screen.

The window above the Command window is the Results window. This is where results will be displayed. So if you ask Stata to compute the mean and standard deviation of a set of variables, the result will appear here.

At the bottom left is the Variables window. When a dataset is loaded into memory all the variables defined in that dataset will be displayed here. The upper left and final window is the Review window. Previously executed commands appear here. As we begin to use Stata the roles of these windows will become clearer.

1.3 Inputting Data

Logically, before you can analyze data, you must input it to the analysis system. This is done by either typing the data into a worksheet in the Data Editor or accessing a preexisting worksheet stored on the harddisk.

Open the Data Editor by clicking on the Data Editor button that lies just below the word Window at the top of the screen, or alternatively, type `edit` in the Command Window and press enter. As you can see, this opens a blank worksheet. Each row of the worksheet is an observation while each column is a variable. For example, if we had the heights and weights of five subjects, the two values for the first subject would go in columns one and two of the first row, those for the second subject would go in columns one and two of the second row and so on. By default, the variables are named `Var1`, `Var2` etc. but you can change these as needed. As you enter data the rows will be numbered 1, 3, 3 etc. The intersection of a row and column is termed a cell. Stata uses the default notation `Var[i]` to designate the cell defined by the intersection of the variable `Var` and the ith observation. Thus, `Var2[3]` is the cell defined by the intersection of the variable named `Var2` and the third observation.

Let’s use the variable labeled `x` in Table 2.6 on page 34 of the text to demonstrate data entry. To begin, click cell `Var1[1]`. Notice that the cell is darkened. The dark background indicates the cell in which data will be entered when you type and press Enter or Tab. Now type the number 3, which is the first value in Table 2.6, in cell `Var1[1]`. Pressing Enter at this point moves the cursor to the next observation, while pressing Tab moves the cursor to the next variable. Go ahead and press Enter and enter the second observation (also a 3). Continue until all observations for `Var1` have been entered.

It is a good idea to name the variables in your worksheet. We will name our variable `x` because that was the only designation used in table 2.6 but we would usually want to use a more informative name such as “sdp” for systolic blood pressure or “age.” As you will see later, variable labels can be employed to make variable names such as sdp clearer.
To change the variable name from Var1 to \( x \), double click on any cell in the Var1 column. The Variable Properties window opens. Type \( x \) in the Name: window then click OK. When you are finished entering the data for variable \( x \) and changing Var1 to \( x \), click the Preserve button. The Preserve button stores a temporary copy of your worksheet in the computer’s memory. If at a later point you decide you want to have things the way they were, you could click the Restore button to bring back the data as it existed at the time of the last Preserve. It is important to understand that Preserve does not make a permanent copy on your hard disk. This is done with the Save or Save As... commands. Your Data Editor window should now appear as shown in Figure 1.2.

To save our work to the hard disk thereby making a permanent copy, we must first close the Data Editor window. Then choose File⇒Save As... Be sure you are saving into the directory you created earlier for this purpose. Let’s name this data file table 2.6 data.dta You do not need to add the .dta extension as Stata will automatically do that for you. After clicking Save, this worksheet will now be available to you at any time you choose to load it into the Data Editor.

To see that this is true, be sure the Data Editor is closed, then type clear at the Command window and press Enter. This removes the current data from the computer’s memory. You can re-open the Data Editor to assure yourself that no data is available to the editor as the blank worksheet attests. Close the editor. Now click the open folder icon or click File⇒Open.... Double click the table 2.6 data.dta file (Note that Stata has appended the suffix .dta to the file) or type the file name in the File name: window and click Open. There is no obvious result but in fact the table 2.6 data.dta has been placed back in memory as you can verify by opening the Data Editor. Also notice that the variable \( x \) appears in the Variables window.
1.4 Performing Analyses

Now that we have entered our data, we can use Stata to perform analyses on these data. We will provide a simple example in this section but will give numerous examples in the chapters that follow.

Stata uses two methods for performing analyses and other tasks. The first is menu driven while the second makes use of entries in the Command window. Experienced users tend to prefer the Command window approach and we will favor this method in our subsequent discussions. We will also provide a few examples of the menu driven approach for comparison purposes. As you gain experience with Stata you will develop your own preferences.

Before beginning the analysis, be sure your data are loaded into memory by opening the Data Editor and verifying their presence. If they are not displayed, load the table 2.6 data.dta file as described above. Close the editor.

Let’s begin by calculating a few descriptive statistics on the variable we’ve labeled \( x \). Type

\[
\text{summarize } x
\]

in the Command window and press Enter. As may be seen in the Results window this produces (1) the number of observations in \( x \), (2) the mean of \( x \), (3) the standard deviation of \( x \) and (4) the minimum and maximum values of \( x \). Had we entered \text{summarize} without the \( x \), Stata would provide results for all variables in the data set. In this case, because there is only one variable, the result would have been the same. Now type

\[
\text{summarize } x, \text{ detail}
\]

and press Enter. Detail is an option to the summarize command that causes Stata to produce more statistics as you see. Options are always separated from commands by a comma as was done here. You will learn about these and many other statistics as you proceed through the text. Your results should appear as shown in Figure 1.3.

The same result can be obtained by selecting Statistics⇒Summaries, tables & tests⇒Summary statistics⇒Summary statistics. This opens the summarize - Summary statistics dialog box. Enter \( x \) in the Variables: (leave empty for all variables) box. Because we want the , detail option appended, select the Display additional statistics option then click OK. The same result appears. In fact \text{summarize } x, \text{ detail} appears in the Review window indicating that this sequence was executed.

1.5 Editing And Printing Your Output

Once you have produced output in the Results window, you will likely want to edit and/or print it. For example, you may wish to add your name, date and a homework assignment number. You may then want to print out the
result to turn in to your instructor. The easiest way to do this is to select **File⇒Print⇒Results**. The usual print screen appears. Click **Print**. At this point the **Printer Settings** box appears. Several options are provided in this box. You may choose to have lines numbered, a header and/or logo added to the output, as well as your name and a project descriptor. After filling in these blanks just click **OK** to print. Should you wish to do more extensive editing e.g. deleting certain portions of the output or adding comments at various places, you will have to create a **log** file.

There are two types of log files depending on their format. Log files with the suffix .smcl are in native Stata format. They preserve all formatting in the **Results** window such as bolding, underlining etc. If you want to preserve your output as is for possible future printing, this is probably the best choice. In order to edit the log file you must preserve the log in plain ASCII format. These files have the suffix .log and can be accessed by any common editor or word processor.\(^3\) To demonstrate, lets regenerate the statistics we produced earlier.

Be sure the **table 2.6 data.dta** are loaded into memory. Now select the **Begin Log** icon which is next to the printer icon. The **Begin logging Stata output** window opens. Choose **Log (*.log)** in the **Save as type**: window then type **edited-output1** in the **File name**: window and save it into your working directory. To this point you have created a .log file that will capture all output into the **Results** window until the log is closed.

Recreate the statistics you produced earlier by typing `summarize x` then

\(^3\)Files saved in .smcl format can be converted to the .log format with the command `translate file.smcl file.log`
press **Enter** and then **summarize x, detail** followed by **Enter**. A copy of this output is going into the log file you created. This is enough for our little demonstration so let’s close the log file by typing **log close**. We can now edit our log file with our favorite editor or use one of Stata’s built in editors. We will use the built in editor.

Now click the **New Do-file Editor** icon just beneath the word “User.” The Do-file editor opens. From the editor open the **edited-output1.log** file. Edit and save the file then close the editor. Our edited output is depicted in Figure 1.4

### 1.6 Creating New Variables

From time to time you may wish to use the variables in your data set to create new variables. This can be done quite easily in Stata with the **generate** command which may be abbreviated as **g**, **ge** or **gen**. The form of the generate command is

\[
generate \textit{newvar}=\textit{expr} \]

where **newvar** is the new variable to be created and **expr** is an algebraic expression that defines the new variable.
To demonstrate, let’s create the second column in Table 2.6 which, as you can see, is the square of the first column. Be sure that table 2.6 data.dta is still loaded into memory. Then type

\[ \text{gen xsquare}=x^2 \]

The symbol \(^2\) (shift 6) means “raise to the power of.” Notice that a new variable named \(xsquare\) appears in the variables window. Now type and enter

\[ \text{list x xsquare} \]

The new variable is indeed the square of \(x\).

As a second example, let’s generate the values in the column headed \(x - \bar{x}\) in Table 2.6. These values are termed “deviation scores” and will be explained in text Chapter 2. To calculate these deviation scores, we must subtract the average (i.e. mean) of \(x\) from each value of \(x\). We previously found that the average of \(x\) was 4.000 so we could write an expression to subtract this value from each of the \(x\) scores. But let’s do this in a slightly more sophisticated way so that we may demonstrate yet another Stata command.

We first calculate the mean of \(x\) by use of the \texttt{summariz} command as follows

\[ \text{summarize x} \]

The \texttt{summarize} command not only prints results to the screen but also stores these results in a function \(r()\). One such result is the mean of \(x\) so we can enter

\[ \text{gen deviation}=x-r(\text{mean}) \]

Now list \(deviation\). These are the deviation scores we set out to calculate. To see all of the results stored in \(r()\) type

\[ \text{return list} \]

Our commands from the \texttt{Review} window are shown in Figure 1.5

Let’s save our work by selecting the \texttt{Save} icon. When asked if it’s ok to write over the old file say yes.
1.7 Graphs

Stata can be used to create a wide variety of graph types ranging from the simple to the complex. You will employ some of these capabilities in Chapter 2. For the moment, we will provide only a simple example of one graph type by constructing a histogram of the $x$ variable in Table 2.6.

Type and enter

```
histogram x, discrete
```

A histogram of $x$ appears as shown in Figure 1.6. Close the window without saving changes. You might want to run this again while leaving off the “discrete” option. What difference did it make?


Figure 1.7: Reconstruction of Table 2.6.

Exercises

1.1 Use Stata to construct the remaining variables in Table 2.6. Then list the variables that make up Table 2.6. The result should appear as in Figure 1.7. (Hint: abs() takes the absolute value of the variable in the parentheses.

1.2 Find the means of all the variables in Table 2.6.

1.3 Label the output from Exercise 1.2 with your name, the current date and the names of the authors of your favorite biostatistics text (You may use ours if you can’t think of any.). Print out the result.

1.4 Enter the data from text Table 7.1 into a new worksheet. Save the worksheet with the name table 7.1 data.

1.5 Find the means of the three diet groups in Table 7.1. Label them mean one, mean two and mean three. Save them in the table 7.1 data worksheet.

1.6 See if you can figure out how to recreate the histogram formed in Section 1.7 without using the Command window.
Chapter 2

Descriptive Methods

2.1 Introduction

In this chapter you will learn to use Stata to construct distributions and graphs as well as to calculate most of the statistics presented in text Chapter 2. You will also learn that in a few instances Stata and the text do not agree on the calculated values of certain descriptive statistics because different definitions of these statistics are used by the software and text. You will also be afforded an opportunity to practice data entry and data transformations.

2.4 Distributions (page 15)

Task 2.1

Use the data in Table 2.1 (table 2.1 data.dta) to construct frequency, relative frequency, and cumulative relative frequency distributions as shown in Table 2.2.

Solution

Load the Table 2.1 data which is in the table 2.1 data.dta file. There are two variables in this file, pt_num (Patient Number) and pain_level (Pain Level). We will use the tabulate command to form the desired distributions so enter

```
tabulate pain_level
```

Because variable names may be abbreviated to the shortest string of characters that uniquely defines them given the variables currently in memory, we could just as well enter

```
tabulate pa
```
to obtain the same result. Notice we could not abbreviate `pain_level` to `p` because Stata would not know if we meant `pain_level` or `pt_num`.

Either command gives us the desired distributions but pain assessments are not in order. They are listed as mild, moderate, none and severe. This means that the cumulative relative frequencies are not meaningful. This is not surprising because Stata has no way of knowing which of these is more or less severe than any other. We will rectify this as follows. First enter:

```
label define pain_label 1 none 2 mild 3 moderate 4 severe
```

This command creates a label named `pain_label` that associates the numbers 1 through 4 with the four pain levels. Now enter the command:

```
encode pain_level, gen(pain_level_num) label(pain_label)
```

The `encode` command encodes string (i.e. nonnumeric) values into numeric values. With this command we have said ”create a new numeric variable called `pain_level_num` from the old string variable `pain_level`. Stata will see that there are four values of `pain_level` and so will automatically assign the numbers 1 through 4 to these values. But left to its own devices, Stata will associate the numbers alphabetically (with capitals preceding lower case). By adding the `pain_label` label, we tell Stata which numbers to associate with which pain level. Now enter

```
list pain_level pain_level_num
```

The listings look exactly the same but they aren’t.

Stata has no idea as to the low to high ordering of `pain_level` but completely understands the ordering of `pain_level_num`. We can now form our distributions with the command:

```
tabulate pain_level_num
```

This yields the desired result.

One other small change would make our output appear a little nicer. Rather than having `pain_level_num` as the column heading, let’s change this to Pain Level. To do this, open the editor and double click in any column. The Variable Properties window opens. In the Name: window type `pain_level_num`. In the Label: window type `Pain Level`. Click OK and close the editor. Once again enter

```
list pain_level pain_level_num
```

Your output should appear as in Figure 2.1. Save the `table 2.1 data.dta` file as we will need it later.

Several points regarding this output should be noted. First, what your text refers to as Frequency, Relative Frequency, and Cumulative Relative Frequency are labeled Freq., Percent and Cum. by Stata. Second, with the exception that Stata uses percents where the text uses proportions and the number of decimal
places reported, the distributions provided by Stata are the same as those in text Table 2.2. Third, Stata orders the distributions, beginning at the top, from the lowest value of the encoded variable (1 in this case) to the highest value (4 in this case) while the text uses the opposite order. The style used by Stata is the more common. The text authors prefer the ascending ordering but either is correct and provides the same information.

**Task 2.2**

Use the data in dataset **table 2.3 data.dta** to construct the simple frequency distribution depicted in the text Table 2.3.

**Solution**

Because the blood pressure data are numeric as opposed to text, we need not go through the process of creating a numeric counterpart variable as was necessary with Task 2.1. Thus, we can construct the simple frequency distribution without any preliminaries.

After loading **table 2.3 data.dta** type and enter

```
tabulate bp
```

Results are shown in Figure 2.2. Notice that we have edited this output so as to have it fit in the figure. Again, the order of presentation is the opposite of that provided in Table 2.3. Also note that values with zero counts are not provided in the Stata output. This is common practice as it saves space.

**Task 2.3**

Use the data in dataset **table 2.3 data.dta** to construct the grouped distributions depicted in the text Table 2.4.

**Solution**

We will take this opportunity to introduce you to do-files and the Do-file Editor. Do-files are particularly useful when you know you may want to execute the same series of commands or an edited version of the same commands at a later point in time. The idea is to place your commands in a do-file, save them, then
CHAPTER 2. DESCRIPTIVE METHODS

Figure 2.2: Simple frequency distribution of BP scores as shown in Table 2.3.

<table>
<thead>
<tr>
<th>BP</th>
<th>Freq</th>
<th>SFreq</th>
<th>SF</th>
<th>FSFreq</th>
</tr>
</thead>
<tbody>
<tr>
<td>86</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>90</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>92</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>93</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>95</td>
<td>3</td>
<td>3</td>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td>96</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>97</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>98</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>99</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>100</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td>102</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>103</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>105</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>106</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>107</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>3</td>
</tr>
</tbody>
</table>

Total 199

run them or an edited version thereof whenever you wish. This saves having to re-type them in the Command window.

To begin, load the table 2.3 data.dta data into memory if it isn’t already there. Now click the New Do-file Editor button just beneath the word User to open the Do-file Editor. Type in the following:

```stata
generate bp_cat=recode(bp, 93, 101, 109, 117, 125, 133, 141, 149)
label define bp_label 93 "86 to 93" 101 "94 to 101" 109 "102 to 109" 117 "110 to 117" 125 "118 to 125" 133 "126 to 133" 141 "134 to 141" 149 "142 to 149"
label values bp_cat bp_label
label variable bp_cat "Interval"
tabulate bp_cat
```

Be sure to press Enter after each command.

When all the commands are entered click Do current file. This executes the commands in the do-file.\(^1\) If you’ve made no errors the output shown in Figure 2.3 should appear. If you’ve made an error, correct it and run the file again. When an error occurs, it is sometimes wise to clear memory and re-load the data in order to avoid conflicts with previously executed commands. Note that if you’ve made no error, running the commands a second time will cause an error because the variable bp_cat already exists which will cause the generate command to produce an error. When the commands execute correctly save the do-file in your working directory. We named ours group_freq.do. The Do-file Editor will add the .do extension.

\(^1\)Don’t confuse the Do current file command with the run command. We won’t be using the latter.
2.5. GRAPHS

Figure 2.3: Grouped frequency distributions for BP variable in 12 categories.

Let’s go over the individual commands in the do-file so that you might become more familiar with Stata commands. (1) The *generate* command creates a new variable named *bp_cat*. The *recode* portion provides the values of *bp_cat* as follows. The value 93 is assigned to *bp_cat* anytime *bp* is less than or equal to 93. The value 101 is assigned when *bp* is greater than 93 but less than or equal to 101 and so on. So after the first command is executed we have a variable named *bp_cat* that is made up of the numbers 93, 101, 109, 117, 125, 133, 141, 149. (2) The second command creates labels to be associated with each of these numbers. Thus, the label “86 to 93” is associated with the number 93, “94 to 101” is associated with 94 and so on. Collectively, these labels are named *bp_label*. (3) The third command applies these labels to the numbers in *bp_cat*. (4) This command applies the label “Interval” to the *bp_cat* variable. You did something similar to this earlier by using the data editor and double clicking on a column to open the Variable Properties window so you could attach the label “Pain Level” to the variable *pain_level_num*. (5) The last command uses *tabulate* to produce the desired distributions. It is important to understand that you did not need the Do-file Editor to produce this output. You could have just entered each command singly into the Command window.

2.5 Graphs (page 19)

**Task 2.4**

Use the data in Table 2.1 (table 2.1 data.dta) to construct the relative frequency bar graph depicted in text Figure 2.1.
Solution

Load table 2.1 data.dta into memory. Don’t forget, you enhanced this dataset in connection with Task 2.1. Then enter, either from the Command window or a do-file, the following commands.

```stata
  gen freq=1 if pain_level!=""
  count if !missing(freq)
  gen rel_freq=freq/r(N)
  graph bar (sum) rel_freq,over(pain_level_num) ytitle("Relative Frequency")
```

If no errors occur, your output should appear as in Figure 2.4. Let’s go over these commands. (1) The first command used `generate`, which can be abbreviated as `gen`, to create a variable named `freq` whose values are all ones so long as `pain_level` is not blank. Stata uses `!` to mean “not equal to” and `""` indicates a blank. So a one is generated in `freq` so long as `pain_level` is not missing. You’ll see why shortly. (2) The second line uses the command `count` to count the number of (non-missing) observations in the data set. This number is printed to the screen but is also stored as `r(N)` which we will use shortly. The `if !missing(freq)` component tells `count` to count only those observations where `freq` is not missing. We don’t have any missing values but you may want to use `count` in a situation where there are missing numbers so you may as well learn how this is done. (3) This command uses `generate` to create the new variable `rel_freq` which is defined as `freq/r(N)` which is `1/r(N)` or one divided by the total number of pain level observations. Notice that if we summed `rel_freq` for any one of the pain levels, “none” for example, the result would be the relative frequency of “none.” (4) The `graph bar` command produces the bar graph. This command indicates that it is the sum of `rel_freq` whose graph is to be constructed. Remember, the sum of `rel_freq` is the relative frequency for a given pain level. The `over(pain_level) portion indicates that a bar will be constructed for each pain level i.e. none, mild moderate and severe. The last portion `ytitle("Relative Frequency")` labels the y-axis.
2.5. GRAPHS

**Figure 2.5:** Stata rendering of the histogram depicted in text Figure 2.2.

![Histogram Image]

**Task 2.5**

Use the data in Table 2.3 (table 2.3 data.dta) to construct the relative frequency histogram depicted in text Figure 2.2.

**Solution**

Load the table 2.3 data.dta data. Then enter

```
    histogram bp, fraction start(84.5) width(5) xlabel(84.5(5)144.5) ytitle("Relative Frequency") color(gs10)
```

If there are no errors the output should appear as in Figure 2.5. As the name implies, the histogram command forms histograms. The histogram bp, portion instructs Stata to construct a histogram of the bp variable. The fraction portion indicates that the bars are to represent fractions (rather than frequencies or percents etc.) because we want a relative frequency histogram. The start(84.5) width(5) component tells Stata to form intervals (or bins as they are referred to technically) beginning with a blood pressure score of 84.5 and with each such interval being five points in length. The x axis is to be labeled from 84.5 to 144.5 in increments of five as attested to by the entry xlabel(84.5(5)144.5). The ytitle component gives the title for the y axis while color gives the color of the bars. The entry gs10 gives the bars a medium gray color.

**Task 2.6**

Use the data in Table 2.3 (table 2.3 data.dta) to construct the relative frequency polygon depicted in text Figure 2.4.

**Solution**

Unfortunately, Stata does not provide a simple, straightforward means to deal
with this problem. In fact, some rather sophisticated Stata code is involved.\footnote{Indeed, the commands that follow are based on code written by Nick Cox and modified by Bill Rising for this problem.} For this reason, we provide the following do-file with little comment. We would note, however, that the portion in the first line that states `start(79.5)` `width(5)` defines the intervals and the "82(5)147" portion of the third line controls the placement of the dots. With this information you may be able to modify this code to construct polygons for other data.

Load table 2.3 `data.dta` then construct a do-file (use `fig_2dot4.do` if you wish) containing the following commands. When you run the do-file you should obtain the output seen in manual Figure 2.6 (i.e. text Figure 2.4).

```stata

twoway histogram gen bp, start(79.5) width(5) fraction gen(bpfrac bpbin) levelsof bpbin, local(have) numlist "82(5)147"
local should "r(numlist)"
local need : list should - have local i = N qui foreach v of local need {
replace bpbin = 'v' in 'i'
replace bpfrac = 0 in 'i'
local -i}
twoway /*
*/ (scatter bpfrac bpbin, connect(l) sort), legend(off) ytitle("Relative Frequency")
```

**Task 2.7**

Use the data in Table 2.3 (`table 2.3 data.sta`) to construct the cumulative relative frequency polygon depicted in text Figure 2.6.
2.5. \textit{GRAPHS}

Solution

Unfortunately, Stata does not provide a simple, straightforward means to deal with this problem. In fact, some rather sophisticated Stata code is involved.\footnote{Indeed, the commands that follow are based on code written by Nick Cox and modified by Bill Rising for this problem.} For this reason, we provide the following do-file (fig_2dot6.do) without comment.

\begin{verbatim}
local lowleft 84.5
local width 5
local extralow = 'lowleft' - 'width'
twoway_histogram_gen bp, start('lowleft') width('width') fraction
   gen(bpfrac bpbin)
   /* next line makes up for a bug in twoway_histogram_gen */
   replace bpbin = . if bpfrac >= .
   gen rightend = bpbin+width'/2
   quietly summarize rightend
   local lowleft = 'r(min)' - 'width' /* in case lowleft is too low */
   local hiright = 'r(max)' /* easier to see computing from dataset */
   levelsof rightend, local(have)
   numlist "'lowleft'('width')'hiright"
   local should = "'r(numlist)'"
   local need : list should - have
   local i = N
   qui foreach v of local need {
      replace rightend = 'v' in 'i'
      replace bpfrac = 0 in 'i'
      local –i
   }
   by rightend, sort: gen bpcum = (n==1)*bpfrac
   replace bpcum = sum(bpcum)
twoway /*
   */(scatter bpcum rightend, connect(l) sort xlab('should')ytitle("Cumulative Relative Frequency")), legend(off)
\end{verbatim}

Output from these commands is shown in manual Figure 2.7

\textbf{Task 2.8}

Construct a stem-and-leaf-display for the blood pressure data in text Table 2.3.

Solution

Be sure that \texttt{table 2.3 data.dta} is loaded into memory. Then enter

\begin{verbatim}
stem bp, lines(1)
\end{verbatim}

The output is as shown in manual Figure 2.8. The , \texttt{lines(1)} component of the command forces Stata to put all leaves belonging to a single stem on a single line. Run the command again without this component to see what happens when this is omitted.
Figure 2.7: Stata rendering of cumulative relative frequency polygon shown in text Figure 2.6.

```
Figure 2.8: Stata rendering of a stem-and-leaf display for the BP variable in text Table 2.3.
```

```
```

```
```
2.6 Numerical Methods (page 23)

In text Example 2.4 on page 27 you are asked to find the median of the numbers 3, 2, 0, 2, 1, 1, 2, and 2. Two different methods are used that produce two different answers, namely, 2.00 and 1.75. It is argued that perhaps the more useful answer is provided by the method that produces the 1.75. What answer is provided by Stata?

Solution

Enter the numbers into a new worksheet. Name the variable \( x \) and save the worksheet as example 2.4 data.dta. The command `centile x` calculates specified percentiles of the \( x \) variable. When no percentiles are specified the default is the 50th percentile which is defined as the median. Therefore, enter

\[ \text{centile } x \]

The result declares the 50th percentile to be 2 which agrees with the score based (as opposed to the point based) definition of the median.

Find the mean, median, and standard deviation of the data in the table associated with Example 2.7 on page 30. Is the median reported by Stata the same as that reported in the text? Why do you think this is the case?

Solution

Because no data are provided, we must create a new worksheet containing the data. Fortunately we can save ourselves considerable effort by taking advantage of the fact that the data in the table are formed into a simple frequency distribution. To do so, open a new worksheet. Enter the scores into the first column and the frequencies into the second. Name the scores “score” and the frequencies “frequency” then save the data as example 2.7 data.dta.

The `summarize` command when used in this context takes the form

\[ \text{summarize [varlist] [weight] [, options]} \]

From this point on, when outlining the general form of a command such as the above, we will follow the following convention. Optional arguments will be placed in brackets, entries that are to be typed exactly as specified will use this font while those that are to be replaced with specifically named entries will use this font.

In the above expression, `varlist`, `weight` and `, options` are optional and may, therefore, be omitted while `summarize` must be present and typed exactly as shown. If `varlist` is omitted, `summarize` will operate on all variables in the dataset. If `weight` is omitted, no weights will be applied to the observations in
CHAPTER 2. DESCRIPTIVE METHODS

varlist. If any options are specified, they must be preceded by a comma. With these details in mind enter

```
summarize score [fweight=frequency], detail
```

The fweight weighttype indicates that the weights in frequency are frequencies. The option detail is used because it causes additional statistics to be printed some of which we need. (Try running `summarize score [weighttype=weight]` to see which statistics are omitted.)

The output gives the mean as 2.12, the standard deviation as 1.812763 and the median (i.e. 50th percentile) as 2.2. As may be seen, the median reported here is 2.2 which contrasts with the value of 2.17 reported in the text as the answer for Example 2.7. Thus, Stata calculates the median as the middle score (or average of the two middle scores) rather than by the scale point based method given by Equation 2.5.

**Task 2.11**

Find the variance of the $x$ variable in Table 2.6 (dataset `Table 2.6 data.dta`). Is the result provided by Stata the same as that given in the text for example 2.15 on page 36?

**Solution**

Be sure `Table 2.6 data.dta` is loaded. Then enter

```
summarize x, detail
```

The variance is reported as .6 which agrees with the value calculated in the text.

**Task 2.12**

Find the twenty-fifth, sixtieth, and seventy-fifth percentiles of the blood pressure data in dataset `Table 2.3 data.dta`. Do the answers provided by Stata match those given in the solution to text Example 2.17 on page 38?

**Solution**

Load the `Table 2.3 data.dta` data. A general form of the `centile` command is

```
centile [varlist][, options]
```

where `varlist` is the list of variables whose percentiles are to be computed and `, options` are various options that may be specified. For the problem at hand enter

```
centile bp, centile(25 60 75)
```

The option “centile(list of centiles)” was used to specify centiles to be calculated. If no options are specified the 50th percentile (median) will be returned.
2.6. NUMERICAL METHODS

The results are reported for the twenty-fifth, sixtieth, and seventy-fifth percentiles respectively as, 115, 130, and 135. The answers provided in the text are respectively, 114.67, 130.38 and 134.64. The discrepancy appears to result from a score based definition of percentile used by Stata and a scale point based definition used by the text.

**Task 2.13**

Use Stata to perform the analyses outlined in Example 2.21 on page 43.

**Solution**

Enter the numbers 1, 3, 3, 9 into a new worksheet. Name this variable x and close the worksheet. Enter

```
gen z=(x-4)/3.46
```

This places a z score for each value of x into a variable named z. If you list z you will see these z scores. Now enter

```
summarize z, detail
```

The output gives the mean of z as 0 with standard deviation 1.001185 which, given rounding, is the expected answer.

**Task 2.14**

Use Stata to conduct the analyses alluded to in Example 2.23 on page 44.

**Solution**

Load table 2.8 data.sta into memory. Then enter the command

```
summarize A B C, detail
```

Stata reports the skew for distributions A, B and C as 0, −.7059569 and .7059569 respectively. The values reported in the text are 0, −.648 and .648. The discrepancy comes about because when calculating the Z scores used in the calculation of skew, we used the n − 1 divisor for the standard deviation while Stata used a method that is equivalent to dividing by n. This difference becomes smaller as sample size increases. What would happen if A B and C were omitted from the summarize command given above? Why?

**Task 2.15**

Use Stata to perform the analyses outlined in Example 2.24 on page 46.

**Solution**

Load the table 2.9 data.dta dataset into memory. Then enter the command

```
summarize A B, detail
```

Kurtosis for distributions A and B are reported as 6 and 1.5 respectively. This contrasts with the values 5.04 and 1.26 given in the text. The discrepancy comes about because when calculating the Z scores used in the calculation of kurtosis, we used the $n - 1$ divisor for the standard deviation while Stata used a method that is equivalent to dividing by $n$. This difference becomes smaller as sample size increases.
Exercises

2.1 Use the data in dataset **table 2.3 data.dta** to construct the grouped distributions depicted in the text Table 2.5.

2.2 Construct the histogram depicted in text Figure 2.3.

2.3 Use Stata to find the mean and median of the following data.

<table>
<thead>
<tr>
<th>Score</th>
<th>frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>9</td>
<td>134</td>
</tr>
<tr>
<td>8</td>
<td>442</td>
</tr>
<tr>
<td>7</td>
<td>391</td>
</tr>
<tr>
<td>6</td>
<td>777</td>
</tr>
</tbody>
</table>

2.4 Calculate z scores for the bp variable in **table 2.3 data.dta**. Find the mean, standard deviation, skew and kurtosis of the resultant z scores. Hint: First find the mean and standard deviation of bp. Then use a command of the form `gen z=(you fill this in)` to find the z scores. Then find the mean, standard deviation skew and kurtosis of the resultant z scores.

2.5 Use Stata to find the skew and kurtosis of the bp data in **table 2.3 data.dta**. How do these values compare to the values you found for the z scores in the previous exercise?
Chapter 3

Probability

3.1 Introduction

In this chapter you will learn to use Stata to form contingency tables as well as to find areas under the normal curve.

3.3 Contingency Tables (page 52)

Task 3.1

Use dataset table 3.1 data.dta to construct text Table 3.1 located on page 54.

Solution

Forming these data into a table of the form shown in Table 3.1 is quite simple and may be accomplished with a variety of commands. For example, after loading the table 3.1 data.dta into memory, you can enter

```
  tab smoking_status disease_status
```

to obtain the table shown in Figure 3.1.

Figure 3.1: A Stata rendering of Table 3.1.
We could improve the appearance of this table by attaching labels to the values “S”, “SN”, “D” and “DN.” Unfortunately this task is made unduly complex because we recorded the data in this form rather than as numerals, e.g. 1 and 0. This stems from the fact that labels cannot be attached to nonnumeric values. This means that before attaching such labels, we must associate numbers with each nonnumeric value. We will discuss how this is done in conjunction with the following do-file (table 3.1.do).

```
label define relate_smoke_num 0 "S" 1 "SN"
encode smoking_status, gen(smoke_num) label(relate_smoke_num)
label define smoke_label 0 "Smoker" 1 "Not a Smoker"
label value smoke_num smoke_label
label define relate_disease_num 0 "D" 1 "DN"
encode disease_status, gen(disease_num) label(relate_disease_num)
label define disease_label 0 "Disease" 1 "No Disease"
label value disease_num disease_label
```

The first four lines pertain to the smoking variable while the second four accomplish exactly the same things for the disease variable. For this reason, we’ll only discuss the first four.

Our goal is to attach labels to the values of the two variables in order to make the table a bit more presentable. Our first problem is that labels can only be attached to numeric variables and our values have been recorded as nonnumeric. So, we will, (1) associate numbers with the various nonnumeric values, then (2) attach labels to these numeric values. The first two lines address (1) while the second two accomplish (2).

Line one defines a label `relate_smoke_num` that contains the rules for associating numbers with the nonnumeric values. The second line creates a new variable `smoke_num` that contains the same values as does `smoking_status` but has numeric values associated with each value as governed by `relate_smoke_num`. We now have numeric values which we can label.

The third line defines a label `smoke_label` that contains the rules for associating each numeric value with a narrative descriptor. That is, 0 will be referred to as “Smoker” while 1 will be referred to as “Not a Smoker.” Line four associates the narrative descriptors in `smoke_label` with the values in `smoke_num`. Notice that lines one and two would not have been necessary had we entered numeric values into our data set rather than the nonnumeric values we used. The next four lines perform the same tasks for the disease status variable. After running the do-file the output is as shown in Figure 3.2.

**Task 3.2**

Use dataset table 3.1 data.dta to construct text Table 3.2 located on page 55.

**Solution**

Use the commands from Task 3.1 with the exception that the last line is changed
3.4. THE NORMAL CURVE

**Figure 3.2**: A Stata rendering of Table 3.1 with variable labeling.

<table>
<thead>
<tr>
<th>Smoking Status</th>
<th>Disease Status</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Smoker</td>
<td>Disease</td>
<td>3</td>
</tr>
<tr>
<td>Not a Smoker</td>
<td>No Disease</td>
<td>6</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>12</td>
</tr>
<tr>
<td></td>
<td></td>
<td>8</td>
</tr>
</tbody>
</table>

**Table 3.1**

To: `tabulate smoke_num disease_num, nofreq cell`. The `nofreq cell` tells Stata not to print frequencies but to provide percentages instead. Note that these are percentages rather than proportions but the same information is conveyed.

### 3.4 The Normal Curve (page 62)

**Task 3.3**

Use Stata to find the area of the normal curve indicated in Example 3.9 on page 64.

**Solution**

Stata provides the `normal(z)` function to find the area below any point on the normal curve. We can display this area with the `display` command. For example, if we enter `display normal(-1.65)` the result 0.04947147 is displayed on the screen. We can obtain the solution to the problem at hand by entering

```
display normal((220-250)/25)
```

which produces 0.11506967 which, except for rounding, is the result reported in the text.
Exercises

3.1 Construct two different worksheets to represent the data in the accompanying contingency table. Use only nonnumeric variables for the first and only numeric variables for the second. Then use these worksheets to construct the contingency table.

<table>
<thead>
<tr>
<th></th>
<th>positive</th>
<th>negative</th>
</tr>
</thead>
<tbody>
<tr>
<td>male</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>female</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>6</td>
</tr>
</tbody>
</table>

3.2 Use Stata in lieu of the normal curve table (text Appendix A) to find the areas of the normal curve indicated in Examples 3.10, 3.11 and 3.12.
Chapter 4

Introduction To Inference And One Sample Methods

4.1 Introduction

In this chapter you will use Stata to find binomial probabilities as well as probabilities associated with the normal curve. You will also learn to perform various hypothesis tests, including equivalence tests, and form certain confidence intervals. In addition, you will perform power calculations and compute required sample sizes to attain specified levels of power.

4.2 Sampling Distributions (page 75)

Task 4.1

Generate the binomial probabilities in Table 4.1 on page 83.

Solution

The function \texttt{Binomial}(n, k, p), where \( n \) is the number of trials, \( k \) is the number of successes and \( p \) is the probability of success for a single trial, returns the probability of attaining \( k \) or more successes in \( n \) trials. Therefore, the probability of obtaining \textit{exactly} \( k \) successes is:

\[ \text{Binomial}(n, k, p) - \text{Binomial}(n, k+1, p) \]

Therefore the following commands will produce the probabilities seen in Table 4.1.

\[
\begin{align*}
\text{display Binomial(5, 0, .10)-Binomial(5, 1, .10)} \\
\text{display Binomial(5, 1, .10)-Binomial(5, 2, .10)} \\
\text{display Binomial(5, 2, .10)-Binomial(5, 3, .10)}
\end{align*}
\]
CHAPTER 4. INFERENCE AND ONE SAMPLE METHODS

\[
\text{display } \text{Binomial(5, 3, .10)} - \text{Binomial(5, 4, .10)} \\
\text{display } \text{Binomial(5, 5, .10)}
\]

**Task 4.2**

The Z value calculated in connection with Example 4.7 on page 86 was \(-5.33\). Use Stata to find the associated probability.

**Solution**

The function `normal(z)` returns the area of the normal curve below \(z\). So we enter

\[
\text{display normal(-5.33)}
\]

which returns 4.911e-08 which can be expressed as .00000004911.

**4.3 Hypothesis Testing (page 86)**

**Task 4.3**

Use Stata to perform the one mean Z test outlined in Example 4.9 on page 92.

**Solution**

Because the one mean Z test is based on the normal curve and we are provided summary information, perhaps the simplest way to perform this test is by use of the function `normal(z)` which will provide us with the \(p\)-value. To this end we enter

\[
\text{display 1-normal((128.2-120)/(40/sqrt(150)))}
\]

which returns .00602414. Because this value is less than \(\alpha = .01\), the null hypothesis is rejected. The expression 1-normal(\(z\)) was used because normal(\(z\)) returns the area below \(z\) and, because of the form of the alternative hypothesis, we need the area above \(z\).

Normally, we would not bother performing the critical \(Z\) versus obtained \(Z\) version of the test but in the spirit of learning more about Stata we will next find critical \(Z\). To this end we note that the function `invnormal(p)` returns the \(Z\) value that has the proportion \(p\) of the normal curve below it. Because we are dealing with a one-tailed test with alternative of the form \(H_A : \mu > 120\) and \(\alpha = .01\), the desired \(Z\) value cuts off .01 in the upper tail of the curve which means that .99 of the curve lies below it. So we enter

\[
\text{display invnormal(.99)}
\]

which returns critical \(Z\) of 2.3263479 which agrees with that given in the solution to Example 4.9.
4.3. HYPOTHESIS TESTING

**Task 4.4**

Use Stata to perform the one mean Z test outlined in Example 4.12 on page 97.

**Solution**

We will again use the `normal(z)` function to find the p-value for the test. Because the test is two-tailed and 87.1 is above the distribution mean of 80, the p-value will be double the area above obtained Z. The area above obtained Z is given by 1-normal(z) so that to double this value we enter

\[
\text{display 2-2*normal((87.1-80)/(21/sqrt(40)))}
\]

which produces a p-value of .03249224 which agrees with the rounded value reported in the text. Because this value is less than \( \alpha = .10 \), the null hypothesis is rejected.

Critical Z can be obtained by entering

\[
\text{display invnormal(.95)}
\]

and

\[
\text{display invnormal(.05)}
\]

which produce, respectively, 1.6448536 and -1.6448536.\(^1\)

**Task 4.5**

Use Stata to perform the one mean t test outlined in Example 4.14 on page 104.

**Solution**

Enter the data from Example 4.14 on page 104 into the first column of a new worksheet. Name this variable \( x \). Save this worksheet as `example 4.14 data.dta`. With this worksheet loaded into memory enter

\[
ttest x==8
\]

This command performs a one mean t test of the null hypothesis \( H_0 : \mu = 8 \) and produces the output seen in Figure 4.1.

Notice that in addition to other information, obtained \( t \) of -3.1278 is reported as well as \( p \)-values for alternatives \( H_A : \mu < 8 \) (0.0102), \( H_A : \mu \neq 8 \) (0.0204), and \( H_A : \mu > 8 \) (0.9898). Because the alternative for Example 4.14 is of the form \( H_A : \mu < 8 \), the appropriate \( p \)-value for this test is 0.0102. Because this value is greater than \( \alpha = .01 \), the null hypothesis is not rejected.

The function `invttail(df, p)` where `df` is degrees of freedom and `p` is the right tail area of the t distribution with `df` degrees of freedom, can be used to obtain critical t. To this end we enter

\[^1\text{Note that 1.6448536 rounds to 1.64 when rounded to two decimal places. However, the value 1.65 is used by convention because it is slightly more conservative.}\]
CHAPTER 4. INFERENCE AND ONE SAMPLE METHODS

Figure 4.1: Stata analysis of Example 4.14.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Obs</th>
<th>Mean</th>
<th>Std. Err.</th>
<th>Std. Dev.</th>
<th>[95% Conf. Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x$</td>
<td>8</td>
<td>0.1426684</td>
<td>0.0070763</td>
<td>0.00993375</td>
<td>0.092397</td>
</tr>
</tbody>
</table>

mean = mean(0)

$\hat{\mu} = 0.1426684$
degrees of freedom = 6

$\text{Pr}(\hat{x} < 0.14) = 0.3192$
$\text{Pr}(\hat{x} \geq 0.14) = 0.6808$

display invttail(6, .99)

which returns $-3.1426684$. Again, because obtained $t$ of $-3.1278$ is greater than critical of $-3.1426684$, the null hypothesis is not rejected.

Task 4.6

Use Stata to perform the one sample test for a proportion outlined under the heading one-tailed test on page 108.

Solution

The `bitesti` command with general form

```
bitesti nt ns p[, detail]
```

can be used for exact tests of hypotheses based on the binomial distribution. In this command, $nt$ is the number of trials, $ns$ is the number of successes and $p$ is the probability of success on any given trial. the `detail` option is used to produce additional output. To perform the test outlined on page 108, we enter

```
bitesti 10 6 0.38
```

which produces the output seen in Figure 4.2. As may be seen, the $p$-value associated with the alternative $H_0 : \pi > .38$ is 0.134760 which agrees with the rounded value calculated in the text.
4.3. HYPOTHESIS TESTING

**Task 4.7**

Use Stata to perform the one sample test for a proportion outlined in Example 4.19 on page 113.

**Solution**

Enter the command

\[
\text{bitesti 8 2 .35}
\]

As may be seen in the output, Stata reports the two-tailed \( p \)-value as 0.721414 which contrasts with the value of 0.85562 reported in the text. As discussed in the text, there are a number of different methods for finding two-tailed \( p \)-values for exact tests based on the binomial distribution. Obviously, Stata uses a different method from the one outlined in the text.

**Task 4.8**

Use Stata to carry out the two-tailed equivalence test outlined in Example 4.25 on page 120.

**Solution**

Begin by entering the sample data into the first column of a new worksheet. Name this variable \( x \). We can now use Stata to conduct two one-tailed \( t \) tests as is done in the text. However, there is a simpler way to establish two-tailed equivalence which we will also demonstrate.\(^2\) We begin by conducting the tests as was done in the text.

For Test One enter

\[
\text{ttest x==2}
\]

This command conducts a one mean \( t \) test of the null hypothesis \( H_0 : \mu = 2 \). The \( p \)-value associated with the alternative \( H_A : \mu < 2 \) is 0.0027 which is less than .05 so that the null hypothesis is rejected.

For Test Two enter

\[
\text{ttest x==-2}
\]

This command conducts a one mean \( t \) test of the null hypothesis \( H_0 : \mu = -2 \). The \( p \)-value associated with the alternative \( H_A : \mu > -2 \) is 0.0285 which is less than .05 so that the null hypothesis is rejected. Because both Test One and Test Two were significant, we reject the equivalence null hypothesis and declare equivalence.

A much simpler way of conducting a two-tailed equivalence test\(^3\) is by forming a 100 \((1 - 2\alpha)\) percent confidence interval and noting the relationship of

\(^2\)The authors of your text presented two-tailed equivalence testing in the manner they did because they believed that, while more tedious, this method led to a better understanding of the equivalence concept.

\(^3\)You may wish to delay learning this form of testing until you’ve studied Sections 4.4 and 4.5.
CHAPTER 4. INFERENCE AND ONE SAMPLE METHODS

$L$ and $U$ to $EI_L$ and $EI_U$. If both $L$ and $U$ are between $EI_L$ and $EI_U$, the two-tailed equivalence null hypothesis is rejected otherwise, it is not rejected. You should be able to determine why this is true after you study Section 4.4 and especially Section 4.5. Let us demonstrate this technique on the current example.

Enter the command

\texttt{ci x, level(90)}

The \texttt{level(90)} option specifies a two-sided 90 percent confidence interval. This figure is arrived at by noting that $100(1 - 2\alpha) = 100(1 - (2)(.05)) = 90$. Stata returns the interval $-1.759686$ to $.7596864$. Because both ends of the confidence interval are within the equivalence interval $-2$ to $2$, the null equivalence hypothesis is rejected.

\textbf{Task 4.9}

Use Stata to carry out the exact two-tailed equivalence test outlined in Example 4.26 on page 122.

\textbf{Solution}

To conduct Test One enter

\texttt{bitesti 8 4 .7}

This command conducts an exact binomial test with 8 trials, 4 successes and $\pi = .7$. The resultant $p$-value for the alternative $H_A: \pi < .7$ is 0.194104 which agrees with the rounded value calculated in the text.

For Test Two, enter

\texttt{bitesti 8 4 .3}

This command conducts an exact binomial test with 8 trials, 4 successes and $\pi = .3$. The resultant $p$-value for the alternative $H_A: \pi > .3$ is 0.194104 which agrees with the rounded value calculated in the text. Because both $p$-values are less than $\alpha = .20$, the equivalence null hypothesis is rejected.

\textbf{Task 4.10}

Compute power for the situation described in Example 4.32.

\textbf{Solution}

The command \texttt{sampsi} returns sample size or power calculations for selected tests of significance. For the one mean $Z$ test the form is

\texttt{sampsi $\mu_0$ $\mu$, sd($\sigma$) n(n) [onesample onesided alpha($\alpha$)]}

where $\mu_0$ is the hypothesized population mean, $\mu$ is the actual population mean, sd($\sigma$) is the population standard deviation and n(n) is the sample size. If
the options `onesample`, `onesided` or `alpha` ($\alpha$) are omitted, the assumption respectively is that calculations are for a two sample test that is two-tailed with $\alpha = .05$.

For the problem at hand enter

```
sampsi 90 88, sd(20) n(25) onesample
```

The output is shown in Figure 4.3. As may be seen Stata gives the power calculation as 0.0791 which is slightly different from the value of 0.0721 calculated in the text.

The discrepancy of $0.0791 - 0.0721 = 0.007$ can be explained by reference to Panel B of Figure 4.25 on page 134 of the text. In this figure power is depicted as the darkly shaded portion of the alternative curve that lies in the lower tail critical region of the null curve. Stata adds to this area the portion of the alternative curve that extends into the right hand critical region of the null curve the reasoning being that this too would bring about rejection of the false null hypothesis albeit for the wrong reason. That is, while $\mu < \mu_0$, rejection in the right hand tail would lead one to believe that $\mu > \mu_0$. If you enter

```
sampsi 90 88, sd(20) n(25) onesample onesided alpha(.025)
```

Stata will return the result given in the text.

### Task 4.11

Use Stata to find the power alluded to in Example 4.34 on page 135.

\footnote{Some authors refer to this as a Type III error while others simply treat it as part of the power calculation as does Stata.}
Solution

The command \texttt{sampsi} returns sample size or power calculations for selected tests of significance. For the test of a single proportion the form is

\begin{verbatim}
sampsi \pi_0 \pi, n(n) [onesample onesided alpha(\alpha)]
\end{verbatim}

where $\pi_0$ is the hypothesized population proportion, $\pi$ is the actual population proportion and $n(n)$ is the sample size. If the options \texttt{onesample}, \texttt{onesided} or \texttt{alpha (\alpha)} are omitted, the result respectively will be a calculation for a two sample test that is two-tailed with $\alpha = .05$.

For the problem at hand enter

\begin{verbatim}
sampsi .35 .50, n(8) onesample onesided
\end{verbatim}

The result is a power calculation of 0.2356 which is dramatically different from the value of .14454 calculated in the text. Stata issues the following warning in this regard:

\begin{quote}
Note: For the above sample size(s) and proportion(s), the normal approximation to the binomial may not be very accurate. Thus, power calculations are questionable.
\end{quote}

As Stata cautions, the problem arises because Stata uses an approximate method based on the normal curve for its calculation while the calculation in the text is exact. Stata will give us a better approximation if we replace the .05 default with .02533 which is the exact level of significance calculated in the text. To do this enter

\begin{verbatim}
sampsi .35 .50, n(8) onesample onesided alpha(.02533)
\end{verbatim}

Stata now returns power of 0.1549 which is a much better approximation. The lesson to be learned here is that if you are using an exact test for a proportion, you should calculate and use the exact level of significance rather than simply inputting the intended level. Better yet, use a method that will return exact power.

\textbf{Task 4.12}

Use Stata to calculate the sample size described in Example 4.36 on page 136.

\textbf{Solution}

To this point we have used \texttt{sampsi} to calculate power but by adding \texttt{power()} as an option to the command, sample size will be returned.

For the problem at hand enter

\begin{verbatim}
sampsi 10 8, sd(4) power(.8) alpha(.01) onesample
\end{verbatim}

The sample size estimate is 47 which is the same result reported in the text.
4.4 Confidence Intervals (page 137)

Task 4.13

Use Stata to construct the 95 percent confidence interval outlined in Example 4.37 on page 143.

Solution

The `cii` command produces confidence intervals for means, proportions and counts. For a normally distributed variable the general form is

\[ \text{cii } n \text{ sample\_mean sd[, level(lev\_conf.)]} \]

where \( n \) is the sample size, \( \text{sample\_mean} \) is the sample mean and \( sd \) is the (sample) standard deviation. The option \( \text{level()} \) specifies the level of confidence. If \( \text{level()} \) is omitted, 95 is assumed. For the problem at hand enter

\[ \text{cii 60 90.1 16, level(95)} \]

The output produces \( L = 8.596676 \) and \( U = 94.23324 \). These values contrast with the interval \( L = 86.05 \) and \( U = 94.15 \). The discrepancy lies in the fact that \( \text{cii} \) assumes that \( sd \) represents the sample standard deviation while Example 4.37 gave the population standard deviation. That is, \( \text{cii} \) is constructing an interval for the case \( \sigma \) not known while the problem dealt with a case where \( \sigma \) was known. The discrepancy is small and gets smaller as sample size increases.

Task 4.14

Use Stata to construct the confidence interval described in Example 4.41 on page 147.

Solution

Enter the blood glucose data from Example 4.41 into a new worksheet. Name the variable \( x \). The general form of the `ci` command is

\[ \text{ci varlist[, level(lev\_conf.)]} \]

where \( \text{varlist} \) is the list of variables for which confidence intervals are to be constructed and \( \text{level()} \) indicates the level of confidence. For the problem at hand enter

\[ \text{ci x, level(99)} \]

The output provides the interval \( L = 94.65409 \) and \( U = 120.1459 \) which agrees with the results calculated in the text.

Task 4.15

Use Stata to form the exact two-sided 95 percent confidence interval constructed in Example 4.42 on page 150.
Solution
The *cii* command can be used to form exact confidence intervals for a proportion based on the binomial distribution. The general form of this command when used for this purpose is

\[ cii \text{ nt ns[,level(lev) exact]} \]

where *nt* is the number of trials, *ns* is the number of successes and the option *level()* specifies the level of confidence. To construct the exact interval described in Example 4.42 enter

\[ cii \text{ 10 4,level(95) exact} \]

which returns the interval \( L = .1215523 \) and \( U = .7376219 \) both of which agree with the rounded interval given in the text.
Exercises

4.1 Use Stata to reproduce the binomial probabilities in Table 4.5 on text page 109.

4.2 In Example 4.8 on page 86, the solution to the problem involves finding the area of the normal curve that lies above a $Z$ value of 1.36. The solution gives this area as .0895. Use Stata to find this area.

4.3 Use Stata to perform the one mean $Z$ test outlined in Example 4.11 on page 95 as well as to find critical $Z$ for the test.

4.4 Use Stata to perform the one mean $Z$ test outlined in Example 4.13 on page 98 as well as to find critical $Z$ values for the test.

4.5 Use Stata to perform the one mean $t$ test outlined in Example 4.16 on page 106.

4.6 Use Stata to perform the test outlined in Example 4.18 on page 111.

4.7 Use Stata to perform the test outlined in Example 4.20 on page 114.

4.8 Use Stata to perform the following test of hypothesis using both the method outlined in the text (i.e. Test One and Test Two) and the shortcut (i.e. confidence interval) method.

\[ H_{0E}: \pi \leq .45 \text{ or } \pi \geq .55 \]
\[ H_{AE}: .45 < \pi < .55 \]

where $\hat{p} = .47$ and $n = 500$. Use $\alpha = .05$.

4.9 In reference to Example 4.32 on page 133, calculate the portion of the alternative curve that lies in the right hand critical region of the null curve thereby reconciling the difference of approximately .0069975 between the power calculation provided by Stata and that calculated in the text.

4.10 Use Stata to obtain the exact power calculated in Example 4.35.

4.11 Use Stata to form the confidence interval (or an approximation thereto) described in Example 4.38 on page 144.

4.12 Use Stata to form the confidence intervals (or approximations thereto) described in Example 4.39 on page 144.

4.13 Use Stata to construct the confidence interval outlined in Exercise 4.22 on page 157.

4.14 Use Stata to construct the two-sided confidence interval alluded to in Example 4.43 on page 151.
CHAPTER 4. INFERENCE AND ONE SAMPLE METHODS
Chapter 5

Paired Samples Methods

5.1 Introduction

In this chapter you will use Stata to perform tests of hypotheses and form confidence intervals for various paired samples problems. You will conduct paired samples $t$ tests, McNemar’s test, as well as tests on risk and odds ratios. You will also form confidence intervals for each of these statistics and perform equivalence tests.

5.2 Methods Related to Mean Difference (page 160)

**Task 5.1**

Use Stata to perform the paired samples $t$ test described in Example 5.1 on page 162.

**Solution**

Load the dataset containing the Table 5.1 data which is in file table 5.1 data.dta. Obtain difference scores by entering

```
ttest posttreatment==pretreatment
```

Stata tests the null hypothesis $H_0: \mu_d = 0$. The (partial) output is shown in Figure 5.1.

As may be seen, obtained $t$ is given as 2.930 which agrees with the value calculated in the text. The $p$-value associated with the alternative $H_A: \mu_d > 0$ is 0.0055 which is less than $\alpha = .05$ so that the null hypothesis is rejected.
Figure 5.1: Stata analysis of Example 5.1.

```
. mean(diff) = mean(postreatment - pretreatment)  t = 2.030
Ho: mean(diff) = 0  degrees of freedom = 14
Hay mean(diff) < 0  Hat mean(diff) = 0  Hat mean(diff) > 0
P(T ≤ |t|) = 0.0045  P(T > |t|) = 0.0010  P(T > 0) = 0.0055
```

**Task 5.2**

Use Stata to perform the equivalence test described in Example 5.4 on page 168.

**Solution**

Load **table 5.5 data.dta** into memory. The one-tailed equivalence test may be carried out by conducting a one-tailed paired samples *t* test as follows. To obtain difference scores enter

```
gen dif=lower_dose-higher_dose
```

Then test the null hypothesis $H_0: \mu_d = 6$ by entering

```
ttest dif==6
```

Obtained $t$ is reported as $-2.6099$ which agrees with the value calculated in the text. The *p*-value associated with the alternative $H_A: \mu_d < 6$ is given as $0.0091$ which is less than $\alpha = .05$ so that the null hypothesis is rejected.

**Task 5.3**

Use the paired data in Table 5.1 to construct a one-sided 95 percent confidence interval for the lower bound of $\mu_d$.

**Solution**

Load the Table 5.1 data which is in file **table 5.1 data.dta** into memory. Generate difference scores by entering

```
gen dif=posttreatment-pretreatment
```

We now take advantage of the fact that the lower end of a two-sided $100(1-2\alpha)$ percent confidence interval will be identical to the lower end of a $100(1-\alpha)$ percent one-sided interval. This means that we can find the lower end of a two-sided $100(1-(2)(.05)) = 90$ percent interval to solve the problem.\(^1\) To this end enter

```
ci dif, level(90)
```

\(^1\)If the logic here is not clear, you may wish to review Section 4.5.2 on text page 154.
5.3. METHODS RELATED TO PROPORTIONS

Figure 5.2: Stata analysis of the data in Table 5.7

<table>
<thead>
<tr>
<th></th>
<th>Controls Exposed</th>
<th>Unexposed</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exposed</td>
<td>2</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>Unexposed</td>
<td>6</td>
<td>6</td>
<td>12</td>
</tr>
<tr>
<td>Total</td>
<td>8</td>
<td>8</td>
<td>16</td>
</tr>
</tbody>
</table>

McNemar’s chi2(1) = 2.00  Prob > chi2 = 0.1573
Exact McNemar significance probability = 0.2001

Proportion with factor

<table>
<thead>
<tr>
<th>Cases</th>
<th>Controls</th>
<th>.25 [95% Conf. Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td>diff.</td>
<td>-.25</td>
<td>-1.3835606 - 0.1890998</td>
</tr>
<tr>
<td>ratio</td>
<td>.2</td>
<td>1.1876259 1.332204</td>
</tr>
<tr>
<td>rel. diff.</td>
<td>-.5</td>
<td>-1.348659 1.948693</td>
</tr>
<tr>
<td>odds ratio</td>
<td>.3333333</td>
<td>.0329021 1.664446 (exact)</td>
</tr>
</tbody>
</table>

which returns a confidence interval whose lower end is 2.553433 which agrees with the value calculated in the text.

5.3 Methods Related to Proportions (page 174)

**Task 5.4**

Use both approximate and exact methods to perform a two-tailed McNemar’s test on the data in Table 5.7 on page 175 at the $\alpha = .05$ level.

**Solution**

Load the 5.7 data.dta data into memory. Notice that the vaccine one and vaccine two data have been entered as numeric values with “no disease” being coded 0 and “disease” coded 1. The **mcc** (mcc stands for matched case control) command performs both approximate and exact versions of McNemar’s test. The general form is

```
mcc var1 var2
```

where **var1** and **var2** represent the outcomes for the two paired groups.

For the problem at hand enter

```
mcc vac2 vac1
```

The output is shown in Figure 5.2.²

²What do you believe the result would be if the command **mcc vac1 vac2** were entered.

As may be seen from the line “McNemar’s chi2(1) = 2.00 Prob > chi2 = 0.1573,” the chi-square statistic for the approximate result is 2.0 which is the
same value calculated in the text. The exact result is given in the line “Exact McNemar significance probability = 0.2891.” which is the value calculated in conjunction with Example 5.11 on page 179 of the text.

**Task 5.5**

Use the Outcome variable in Table 5.14 to perform the exact test outlined in Example 5.13 on page 183.

**Solution**

The Outcome variable, minus the noninformative observations, from Table 5.14 are located in **table 5.14 data.dta**. Load this dataset into memory. The `bitest` command can be used to perform exact hypothesis tests based on the binomial distribution. Its general form is

```
bitest var==p[, detail]
```

where `var` is the variable upon which the test is to be conducted and `p` is the probability of success on a given trial. For the problem at hand enter

```
bitest outcome==.4
```

The output shows that the $p$-value associated with the alternative $H_A: \pi > .4$ is 0.514145 which is the value calculated in the text.

**Task 5.6**

Construct the confidence intervals described in Example 5.15 on page 186.

**Solution**

The `prtesti` command can be used to form the approximate confidence interval. The general form is

```
prtesti n \hat{p} \pi_0[, level(lev.conf.)]
```

where $n$ is the sample size, $\hat{p}$ is the sample proportion and $\pi_0$ is the hypothesized population proportion. If, `level()` is omitted, 95 is assumed.

Because we are using `prtesti` to form a confidence interval, we can specify any legitimate value for $\pi_0$. For the problem at hand enter

```
prtesti 10 .4 .6
```

The output specifies the 95 percent confidence interval as 0.0963637 to 0.7036363 which agrees with the values calculated in the text.

The `cii` command can be used to form exact confidence intervals for a proportion based on the binomial distribution. The general form of this command when used for this purpose is

```
cii nt ns[,level(lev.conf.) exact]
```
5.4. METHODS RELATED TO PAIRED SAMPLES RISK RATIOS

where \( nt \) is the number of trials, \( ns \) is the number of successes and the option \texttt{level()} specifies the level of confidence. To construct the exact interval described in Example 5.15 enter

\[
cii 10 4
\]

The 95 percent confidence interval is given as \( 0.1215523 \) to \( 0.7376219 \) which agrees with the calculations performed in the text. Notice that it was not necessary to specify the level nor to use the \texttt{exact} option because the defaults produce exact 95 percent intervals.

5.4 Methods Related to Paired Samples Risk Ratios (page 190)

\textbf{Task 5.7}

Carry out the tasks outlined in Example 5.17 on page 191.

\textbf{Solution}

Refer to Task 5.4 and Figure 5.2 in this manual. The statistic termed “ratio” in the figure is in fact the risk ratio. It’s value is reported as \( 0.5 \) which is the result obtained in the text.

The chi-square statistic calculated via Equation 5.5 is reported as 2.0 which agrees with the value calculated in the text. The calculation based on a \( Z \) test is not reported but can be easily hand calculated. Also, while the labeling of the output is not ideal for our purpose, the statistics are correct and meet our needs.

\textbf{Task 5.8}

Use Stata to perform the equivalence test as outlined in Example 5.20 on page 195.

\textbf{Solution}

Among other things, the command \texttt{mcci} calculates confidence intervals for paired samples risk ratios. We can use this facility to conduct the equivalence test. The general form of this command is

\[
mcci a b c d[,level(lev,conf.)]
\]

where \( a, b, c \) and \( d \) are as shown in Table 5.8 on page 176 of the text and \texttt{level()} sets the level of confidence for confidence intervals.

We note that if the upper end of a one-sided 95 percent confidence interval is less than or equal to 1.1, the equivalence null hypothesis will be rejected and the researchers will conclude that the level of risk is acceptable. Because Stata
does not construct one-sided intervals of this sort, we will form a two-sided 90 percent interval and compare the value of $U$ to 1.1. To this end enter

\texttt{mcci 126 414 390 999,level(90)}

The output gives $U$ for the paired samples risk ratio as 1.143185 which is greater than 1.1 so that the equivalence null hypothesis is not rejected.

\textbf{Task 5.9}

Use Stata to form the confidence interval described in Example 5.22 on page 197.

\textbf{Solution}

Again, we will employ the command \texttt{mcci} to form the confidence interval. To this end enter

\texttt{mcci 226 546 378 0,level(95)}

Notice that it makes no difference what value we specify for cell $d$ because this cell plays no role in the calculation of $\hat{R}R$.

The output gives $L = 1.171356$ and $U = 1.394671$ which agrees with the rounded values calculated in the text. We see that 1.0 is not in this interval so that the null hypothesis $H_0: \hat{R}R = 1$ is rejected.

\section*{5.5 Methods Related to Paired Samples Odds Ratios (page 199)}

\textbf{Task 5.10}

Use Stata to perform an exact test of the hypothesis $H_0: OR = 1$ for the situation described in Example 5.24 on page 201.

\textbf{Solution}

We will take advantage of the fact that the command \texttt{mcci} constructs exact confidence intervals for the paired samples odds ratio to perform the test. To this end enter

\texttt{mcci 13 25 5 55,level(95)}

The sample odds ratio is reported as 5.0 with the associated exact 95 percent confidence interval for the odds ratio being 1.880087 to 16.72368. Because 1.0 is not in this interval, the null hypothesis is rejected.

\textbf{Task 5.11}

Perform the exact two-tailed equivalence test described in Example 5.27 on page 206.
5.5. METHODS RELATED TO PAIRED SAMPLES ODDS RATIOS

Solution

As discussed on page 35 of this manual, a simple way of conducting a two-tailed equivalence test is by forming a 100 \((1 - 2\alpha)\) percent confidence interval and noting the relationship of \(L\) and \(U\) to \(E_I L\) and \(E_I U\). If both \(L\) and \(U\) are between \(E_I L\) and \(E_I U\), the two-tailed equivalence null hypothesis is rejected otherwise, it is not rejected. We will use \texttt{mcci} to form this interval as follows

\[
\texttt{mcci 13 8 7 19,level(90)}
\]

Stata reports \(\widehat{OR}\) as 1.142857 which agrees with the calculation in the text. The 90 percent confidence interval is reported as \(L = 0.4285438\) and \(U = 3.102944\). The equivalence null hypothesis is rejected only if both of these values are in the equivalence interval \(.883\) to \(1.2\). Neither \(L\) nor \(U\) is in this interval so that the equivalence null hypothesis is not rejected. This is also the conclusion reached in the text.

\[
\text{Task 5.12}
\]

Construct the exact confidence interval described in Example 5.30 on page 211.

Solution

Once again calling on \texttt{mcci} we enter

\[
\texttt{mcci 19 13 11 19,level(95)}
\]

Stata computes the 95 percent confidence interval as \(.4885561\) to \(2.913432\) which agrees with the interval calculated in the text.
CHAPTER 5. PAIRED SAMPLES METHODS

Exercises

5.1 Use Stata to perform the test of hypothesis outlined in Example 5.2 on page 163.

5.2 Use Stata to perform the equivalence test outlined in Example 5.5 on page 169.

5.3 Use the paired data in Table 5.2 on page 163 to form a 95 percent confidence interval for the estimation of $\mu_1 - \mu_2$ where $\mu_1$ represents the mean for Treatment One and $\mu_2$ represents the mean for Treatment Two.

5.4 Use Stata to perform the test outlined in Example 5.12 on page 180.

5.5 Use the Outcome variable in Table 5.16 to perform the tests alluded to in Example 5.14 on page 185.

5.6 Use Stata to construct the confidence intervals discussed in Example 5.16 on page 188.

5.7 Conduct the test outlined in Example 5.18 on page 192.

5.8 Conduct the test of significance alluded to in Example 5.19 on page 192.

5.9 Use Stata to perform the exact two-tailed test of significance outlined in Example 5.25 on page 202.

5.10 Use Stata to construct the approximate confidence interval discussed in Example 5.28 on page 208.
Chapter 6

Two Independent Samples Methods

6.1 Introduction

In this chapter you will use Stata to perform tests of hypotheses and form confidence intervals for various two (independent) sample methods. You will conduct independent samples $t$ tests and tests for differences between proportions. You will also form confidence intervals for each of these statistics and perform equivalence tests.

6.2 Methods Related to Differences Between Means (page 215)

Task 6.1

Conduct the independent samples $t$ test described in Example 6.1 on page 220.

Solution

Load the table 6.1 data.dta data into memory. We will use the `ttest` command with general form

```
  ttest var1 == var2, unpaired level(lev\text{conf.})
```

to perform the test. Here, $var1$ and $var2$ represent the outcome variables for the two groups, `unpaired` specifies that the test is not for paired data and `level()` specifies the level of the associated confidence interval. For the problem at hand enter

```
  ttest grp\_one == grp\_two, unpaired level(95)
```
Figure 6.1: Stata output for an independent samples t test conducted on data from Table 6.1.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Obs</th>
<th>Mean</th>
<th>Std. Err.</th>
<th>Std. Dev.</th>
<th>[95% Conf. Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td>grp_one</td>
<td>25</td>
<td>171.8</td>
<td>2.235517</td>
<td>11.5921</td>
<td>124.6065 - 219.205</td>
</tr>
<tr>
<td>grp_two</td>
<td>25</td>
<td>176.2</td>
<td>2.690726</td>
<td>10.14223</td>
<td>129.4259 - 222.972</td>
</tr>
<tr>
<td>combined</td>
<td>50</td>
<td>174.5</td>
<td>2.056444</td>
<td>11.41531</td>
<td>129.2236 - 219.775</td>
</tr>
</tbody>
</table>

The output from this command is shown in Figure 6.1. As may be seen, obtained \( t \) is given as \( t = -0.8080 \) which agrees with the result provided in the text. The associated \( p \)-value is 0.4259 so that the null hypothesis is not rejected at the \( \alpha = 0.05 \) level of significance.

**Task 6.2**

Conduct the equivalence test described in Example 6.3 on page 225.

**Solution**

Unfortunately, `ttest` does not provide a direct means to test \( H_0: \mu_1 - \mu_2 = -5 \) which is required for Test Two. Rather, it insists on testing \( H_0: \mu_1 - \mu_2 = 0 \). We can circumvent this difficulty in several ways. We will demonstrate one of these here and leave a second as an exercise.

Suppose that \( \mu_1 - \mu_2 \) is exactly equal to \(-5\). If we add five points to each score in group one, we can now conceive of this group as having been sampled from a population whose mean is five points higher than previously so that \( \mu_1 - \mu_2 \) is now zero. Rejecting the null hypothesis \( H_0: \mu_1 - \mu_2 = 0 \) after adding the points is the same thing as rejecting \( H_0: \mu_1 - \mu_2 = -5 \) for the original data. So this will be our strategy. To add the five points and create a new variable enter

```
gen grp_one5=grp_one+5
```

To conduct the t test enter

```
ttest grp_one5==grp_two, unpaired
```

Obtained \( t \) is \( t = 0.3802 \) which is the value calculated in the text. The \( p \)-value associated with the alternative \( H_A: \mu_1 - \mu_2 > 0 \) (or \( H_A: \mu_1 - \mu_2 > -5 \)) is given as 0.3533 so that the null hypothesis is not rejected.

**Task 6.3**

Form the confidence interval described in Example 6.5 on page 228.
6.3 Methods Related to Proportions (page 230)

**Task 6.4**

Perform the two-sample test for a difference between proportions addressed in Example 6.7 on page 233.

**Solution**

The `prtesti` command can be used to perform this test. The general form is

\[
\text{prtesti } n_1 \ s_1 \ n_2 \ s_2[, \ \text{count level(lev_con.f.)}]
\]

where \( n_1 \) and \( n_2 \) are the sample sizes of groups one and two respectively, \( s_1 \) and \( s_2 \) are the numbers of successes in the two groups and \text{level()}\ is the level of confidence for a confidence interval. The term \text{count} indicates that \( s_1 \) and \( s_2 \) are the number of successes in each group. If \text{count} is omitted, \( s_1 \) and \( s_2 \) are the proportion of successes in the two groups. For the problem at hand, enter

\[
\text{prtesti } 314 \ 23 \ 316 \ 39, \ \text{count level(95)}
\]

which produces obtained \( Z \) of \( z = -2.1137 \) which is the value calculated in the text. The \( p \)-value for the alternative \( H_A : \pi_1 < \pi_2 \) is given as 0.0173 so that the null hypothesis is rejected.

**Task 6.5**

Perform the two-tailed equivalence test described in Example 6.9 on page 235.

**Solution**

As discussed on page 35 of this manual, a simple way of conducting a two-tailed equivalence test is by forming a \( 100(1 - 2\alpha) \) percent confidence interval and noting the relationship of \( L \) and \( U \) to \( EIL \) and \( EIU \). If both \( L \) and \( U \) are between \( EIL \) and \( EIU \), the two-tailed equivalence null hypothesis is rejected, otherwise, it is not rejected. We will use `prtesti` to form this interval as follows.\(^1\)

\[
\text{prtesti } 100 \ .07 \ 100 \ .06, \ \text{level(90)}
\]

Notice that \text{count} was not specified as an option because the proportion rather than the number of successes was entered for each sample. Notice also that the

\(^1\)See Task 6.4 for the general form of `prtesti`.
level of confidence is specified as 90 for the two-tailed equivalence test to be conducted at $\alpha = .05$.

The output gives the confidence interval as $-.0473344$ to $.0673344$. In order to reject the two-tailed equivalence null hypothesis, both $L$ and $U$ would have to lie between $-.04$ and $.04$. Because this is not the case, the null hypothesis is not rejected.

**Task 6.6**

Construct the confidence interval described in Example 6.11 on page 236.

**Solution**

Enter

```
prtesti 299 106 313 66, count level(95)
```

which produces a 95 percent confidence interval of $.0730676$ to $.2142373$. This differs only slightly from the $.070$ to $.218$ interval calculated in the text. The difference most likely stems from a small correction factor designed to enhance accuracy that is employed in the text.²

### 6.4 Methods Related to Independent Samples

**Risk Ratios (page 238)**

The `csi` command produces various results related to the independent samples risk ratio. The general form is

```
csi a c b d[, level(lev.conf.)]
```

where $a$, $b$, $c$, and $d$ are cell frequencies as shown in text Table 6.7 on page 239 and `level` is the optional level of confidence for an associated confidence interval.

**Task 6.7**

Calculate the risk ratio and perform the hypothesis test outlined in Example 6.13 on page 240.

**Solution**

We use the frequencies from the table on page 240 with the `csi` command as follows.

```
csi 42 21 2981 4088
```

The output is shown in Figure 6.2.

As may be seen, the reported risk ratio is 2.718492 which agrees with the rounded result reported in the text. The test of hypothesis is carried out by

²See Equations 6.6 and 6.7 in the text.
Figure 6.2: Stata output associated with the analysis of Example 6.13.

The sample risk ratio is given as 0.5341923 which agrees with the rounded calculation in the text. We note that $U = 1.028568$ is greater than one so that the null hypothesis is not rejected. This agrees with the result given in the text. (If you understand the logic of this analysis, you are doing quite well.)

---

**Task 6.8**

Calculate the risk ratio and conduct the hypothesis test outlined in Example 6.14 on page 240.

**Solution**

Stata does not provide a one-tailed test of the null hypothesis $H_0 : RR = 1$. However, because you paid close attention to the material in Section 4.5.2 on page 154 of the text, you know that one-sided confidence intervals can be used to conduct one-tailed hypothesis tests. For the problem at hand we need only form a one-sided 95 percent CI for estimation of the upper bound of $RR$. If $U \leq 1$ the null hypothesis will be rejected.

Unfortunately, Stata also does not provide a direct means for constructing one-sided confidence intervals so we will take advantage of the fact that the upper end of a two-sided 100 $(1 - \alpha) = 100 (1 - 2(.05)) = 90$ percent confidence interval will be identical to the upper end of a 100 $(1 - \alpha) = 100 (1 - .05) = 95$ percent one-sided interval. With this in mind, we enter

```
csi 10 17 12344 11202, level(90)
```

The sample risk ratio is given as 0.5341923 which agrees with the rounded calculation in the text. We note that $U = 1.028568$ is greater than one so that the null hypothesis is not rejected. This agrees with the result given in the text. (If you understand the logic of this analysis, you are doing quite well.)

---

**Task 6.9**

Carry out the tasks outlined in Example 6.17 on page 245.

---

Notice that the form of the $Z$ statistic used in the text does not square to equal the chi-square statistic as is usually the case.
CHAPTER 6. TWO INDEPENDENT SAMPLES METHODS

Solution

From Figure 6.2 we see that the 95 percent confidence interval is 1.613433 to 4.580417. Because one is not in this interval, the null hypothesis is rejected.

6.5 Methods Related to Independent Samples
Odds Ratios (page 247)

Task 6.10

Calculate the odds ratio and conduct the equivalence test outlined in Example 6.21 on page 252.

Solution

Because Stata does not provide a direct means of conducting Test One, we will conduct the test by constructing a one-sided confidence interval. Unfortunately, Stata also does not provide a direct means for constructing one-sided confidence intervals so we will take advantage of the fact that the upper end of a two-sided 100(1 − 2α) percent confidence interval will be identical to the upper end of a 100(1 − α) percent one-sided interval. This means that we can find the upper end of a two-sided 100(1 − (2)(.05)) = 90 percent interval to solve the problem. We will use the cci command for this purpose. The general form of the cci command is

cci a b c d[, level(lev.conf.)]

where a, b, c and d are as shown in Table 6.9 on text page 248 and level() specifies the confidence level for a confidence interval estimation of OR. For the purpose at hand enter

cci 126 819 119 811, level(90)

The sample odds ratio is given as 1.048481 which agrees with the text calculation. The upper end of the 90 percent confidence interval is given as 1.326051. Because this value is not less than or equal to 1.1, the equivalence null hypothesis is not rejected.4

Task 6.11

Construct the confidence interval described in Example 6.23 on page 254.

Solution

Enter

cci 13 56 4 65, level(95)

4If the logic here is not clear, you may wish to review Section 4.5.2 on text page 154.
The odds ratio is given as 3.772321 which is the value calculated in the text. However, the 95 percent confidence interval of 1.07494 to 16.64271 does not agree with the 1.164 to 12.231 interval given in the text. The discrepancy is attributable to the fact that an exact procedure is provided by Stata while the text uses an approximation based on the normal curve. The approximation does not appear to serve well for samples of this size.
Exercises

6.1 Form a confidence interval to carry out the equivalence test described in Task 6.2.

6.2 Use Stata to conduct the test of hypothesis described in Example 6.2.

6.3 Perform the equivalence test described in Example 6.4 on page 226.

6.4 Find the confidence interval discussed in Example 6.6 on page 229.

6.5 Perform the test described in Example 6.8 on page 233.

6.6 Perform the equivalence test outlined in Example 6.10 on page 235.

6.7 Construct the one-sided confidence interval described in Example 6.12 on page 237.

6.8 Conduct the equivalence test outlined in Example 6.15 on page 243. HINT: Notice how Task 6.8 was carried out.

6.9 Conduct the two-tailed equivalence test discussed in Example 6.16 on page 243.

6.10 Perform the equivalence test discussed in Example 6.22 on page 253.

6.11 Conduct the test discussed in Example 6.24 on page 255.
Chapter 7

Multi-Sample Methods

7.1 Introduction

In this chapter you will use Stata to perform a One-way ANOVA and 2 by $k$ chi-square test.

7.2 The One-way Analysis of Variance (ANOVA) F Test (page 264)

Task 7.1

Use the data in Table 7.1 to compute $MS_w$, $MS_b$, obtained $F$ and to produce an ANOVA table.

Solution

Load the table 7.1 data.dta data into memory. Notice that the first column termed factor contains the group membership designation for each response in the second column which is termed response. We will use the oneway command to perform the analysis. The general form is

\[
\text{oneway outcome\_var group\_var}
\]

where outcome\_var is the response variable and group\_var is the group designation variable. For the problem at hand enter

\[
\text{oneway response factor}
\]

The Stata output is shown in Figure 7.1.

As may be seen in this table, $MS_w$ and $MS_b$ are given as 625 and 934.86667 respectively which agrees with the rounded values calculated in the text. Obtained $F$ is given as 1.50 which is the value given in the text. The associated $p$-value is .2630.
Figure 7.1: Stata rendering of ANOVA table for data in text Table 7.1.

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>F</th>
<th>Prob &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between groups</td>
<td>1669.74</td>
<td>2</td>
<td>934.87</td>
<td>1.50</td>
<td>0.260</td>
</tr>
<tr>
<td>Within groups</td>
<td>583</td>
<td>12</td>
<td>48.58</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>2252.74</td>
<td>14</td>
<td>160.91</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Gartlott's test for equal variances: $\chi^2(20) = 0.2000$, Prob=chi2 = 0.884

Figure 7.2: Chi-square analysis of nutritional data.

<table>
<thead>
<tr>
<th>Instruction</th>
<th>1</th>
<th>2</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Birth Weight</td>
<td>22</td>
<td>39</td>
<td>62</td>
</tr>
<tr>
<td></td>
<td>204</td>
<td>277</td>
<td>581</td>
</tr>
<tr>
<td>Total</td>
<td>214</td>
<td>316</td>
<td>630</td>
</tr>
</tbody>
</table>

Pearson $\chi^2(1) = 4.4678$, Pr = 0.035

7.3 The 2 By $k$ Chi-Square Test (page 276)

Task 7.2

Perform the chi-square test outlined in Example 7.6 on page 280.

Solution

Load the example 7.6 data.dta data into memory. Notice that the combinations of the first and second columns define the four cells while the third column contains the observed frequency for each cell. We will use the `tabulate` command to perform the analysis. When used in this context, its general form is

```
. tab birth_weight instruction [fweight=frequency], chi2
```

where `var1` and `var2` are the row and column variables defining the cells, `fweight=weight` specifies the variable containing cell frequencies and `chi2` indicates that a chi-square test is to be performed. For the problem at hand enter

```
. tab birth_weight instruction [fweight=frequency], chi2
```

The output is shown in Figure 7.2. Obtained chi-square is reported as 4.4678 which agrees with the value calculated in the text. The associated $p$-value is 0.035.
7.4 Multiple Comparison Procedures (page 282)

**Task 7.3**
Perform the Tukey’s HSD tests outlined in Example 7.9 on page 289.

**Solution**
Oddly, Stata does not provide an internal means to perform this test. Third party code can be attached to the system for this purpose but this will not be addressed in this manual.
Exercises

7.1 Carry out the ANOVA analysis outlined in Example 7.3 on page 269.

7.2 Carry out the ANOVA analysis outlined in Example 7.4 on page 271.

7.3 Perform the chi-square analysis outlined in Example 7.5 on page 279.
Chapter 8

The Assessment of Relationships

8.1 Introduction

In this chapter you will use Stata to calculate the Pearson product-moment correlation coefficient and to test the hypothesis $H_0 : \rho = 0$. Additionally, you will conduct the chi-square test for independence.

8.2 The Pearson Product-Moment Correlation Coefficient (page 295)

Task 8.1

Calculate the Pearson product-moment correlation coefficient for the data in Table 8.1 on page 297. Perform a two-tailed test of the hypothesis $H_0 : \rho = 0$ at the .05 level.

Solution

Load the table 8.1 data.dta data into memory. We will use the \texttt{pwcorr} command to obtain the desired correlation and test of significance. The general form of this command is

\begin{verbatim}
pwcorr var_list[, sig]
\end{verbatim}

where \texttt{var_list} is a list of variables whose correlations are sought and \texttt{sig} requests that the $p$-value for a test of the hypothesis $H_0 : \rho = 0$ be printed. For the problem at hand enter

\begin{verbatim}
pwcorr x y, sig
\end{verbatim}
Figure 8.1: Chi-square analysis of data described in Example 8.7.

<table>
<thead>
<tr>
<th>County</th>
<th>Vaccinated</th>
<th>Not Vaccinated</th>
<th>Unknown</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>County One</td>
<td>419</td>
<td>136</td>
<td>452</td>
<td>617</td>
</tr>
<tr>
<td>County Two</td>
<td>224</td>
<td>210</td>
<td>440</td>
<td>874</td>
</tr>
<tr>
<td>County Three</td>
<td>330</td>
<td>624</td>
<td>600</td>
<td>1,554</td>
</tr>
<tr>
<td>Total</td>
<td>973</td>
<td>950</td>
<td>1,672</td>
<td>3,605</td>
</tr>
</tbody>
</table>

Pearson chi2(4) = 206.8111 Pr = 0.000

Stata gives the correlation coefficient as 0.9635 which agrees with the value calculated in the text. The p-value for the test of significance is given as 0.0000 so that the null hypothesis is rejected at the $\alpha = 0.05$ level.

8.3 The Chi-Square Test For Independence (page 312)

Task 8.2

Perform the chi-square test described in Example 8.7 on page 314.

Solution

Load the example 8.7 data.dta data into memory. Notice that the combinations of the first and second columns define the nine cells while the third column contains the observed frequency for each cell. We will use the `tabulate` command to perform the analysis. When used in this context, its general form is

```
tabulate var1 var2 [fweight = weight], chi2
```

where `var1` and `var2` are the row and column variables defining the cells, `fweight=weight` specifies the variable containing cell frequencies and `chi2` indicates that a chi-square test is to be performed. For the problem at hand enter

```
tabulate county vaccination [fweight = frequency], chi2
```

The output is shown in Figure 8.1. As may be seen, obtained chi-square is reported as $\text{Pearson chi2}(4) = 206.8111$ which is the value reported in the text. The notation $(4)$ indicates that the chi-square statistic is based on 4 degrees of freedom. The associated $p$-value is 0.000.
Exercises

8.1 Calculate the correlation coefficient discussed in Example 8.2 on page 298.

8.2 Paice et al. [1] conducted a survey among doctor trainees in the U.K. to answer various questions related to bullying that they may have experienced in their current position. Of particular interest was the relationship between the trainees rank in her/his present position and their propensity to complain when such behavior was experienced. Results obtained from one survey question were as follows.

When those who indicated that they had experienced such behavior were asked if they had complained to anyone, the responses were as follows. Among trainees in Rank One, 30 answered yes, 60 answered no and 0 said they don’t know. In Rank Two, 66 answered yes, 152 answered no and 5 indicated they don’t know. Among Rank Three trainees, 57 answered yes, 107 answered no and 3 indicated they don’t know. Perform a chi-square test at $\alpha = .05$ to determine whether there is a relationship between trainee rank and propensity to complain. Interpret the result.
CHAPTER 8. THE ASSESSMENT OF RELATIONSHIPS
Chapter 9

Linear Regression

9.1 Introduction

In this chapter, you will use Stata to construct simple and multiple regression models, generate \( \hat{R}^2 \) values and conduct tests of significance.

9.2 Simple Linear Regression (page 320)

**Task 9.1**

Construct the simple linear regression model alluded to in Example 9.1 on page 321. Solution

Open the table 8.1 data.dta dataset. The `regress` command is one of several regression analysis commands implemented by Stata and is the one we will use for this and the following problems. The general form of the `regress` command is as follows

```
regress y x1 x2 · · · xk
```

where \( y \) is the dependent variable and \( x1 \ x2 \ · · · \ xk \) are a series of predictors. For the problem at hand where we have but a single predictor enter

```
regress y x
```

The results are shown in Figure 9.1 In this table the model terms a and b are under the heading coef. and are termed cons and x respectively. These values are reported as .8741497 and .7059369 which agree with the rounded values calculated in the text.

**Task 9.2**

Find \( SS_{reg} \), \( SS_{res} \), and \( \hat{R}^2 \) for the data in text Table 9.1 on page 323.
CHAPTER 9. LINEAR REGRESSION

Figure 9.1: Regression analysis of Table 8.1 data.

The table below shows the regression analysis of Table 8.1 data.

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>Number of obs = 12</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>107.443599</td>
<td>1</td>
<td>107.443599</td>
<td>F(1, 11) = 168.86</td>
</tr>
<tr>
<td>Residual</td>
<td>8.28973408</td>
<td>11</td>
<td>0.75339458</td>
<td>Prob &gt; F = 0.0000</td>
</tr>
<tr>
<td>Total</td>
<td>115.733333</td>
<td>12</td>
<td>9.64444444</td>
<td>R-squared = 0.9284</td>
</tr>
</tbody>
</table>

| y   | Coef.  | Std. Err. | t     | P>|t|  | [95% Conf. Interval] |
|-----|--------|-----------|-------|-----|-----------------|
| _cons | 0.6871497 | 0.0649112 | 10.67 | 0.000 | 0.5597446 - 0.814554 |

Solution

Load the table 9.1 data.dta worksheet into memory. Enter

```
regress y x
```

Stata gives the Regression sum of squares ($SS_{reg}$) which it terms the “Model” sum of squares as 107.443599 and the residual sum of squares ($SS_{res}$) which it labels “Residual” as 8.28973408. Both values agree with the text calculations. $R^2$ is given as “R-squared = 0.9284” which is the value calculated in the solution to Example 9.2 on page 324.

Task 9.3

Test the null hypothesis $R^2 = 0$ for a model used to predict $y$ from $x$ for the data in Table 9.1 on page 323.

Solution

Using the output from the solution to Task 9.2, we see that obtained $F$ is reported as 168.49 with an associated $p$-value of 0.0000. The value of 168.49 reported by Stata differs slightly from the value of 167.56 calculated in the text. This small difference appears to be the result of rounding in the text calculations.

9.3 Multiple Linear Regression (page 329)

Task 9.4

Use the data in Table 9.2 on page 330 to construct a two predictor model with $x_1$ and $x_2$ being used to predict $y$.

Solution

Load the table 9.2 data.dta dataset into memory. Enter the command
9.3. **MULTIPLE LINEAR REGRESSION**

```
regress y x1 x2
```

Stata gives the regression model as \( \hat{y} = 2.685874 + 0.4036254x1 + 0.9998021x2 \) which agrees with the results calculated in the text.

**Task 9.5**

Test the null hypothesis \( H_0 : \beta_1 = \beta_2 = 0 \) for the model you constructed in Task 9.4.

**Solution**

From the output for the solution to Task 9.4, we note that obtained \( F \) for the regression model is 4.73 which is the value calculated in the text. The associated \( p \)-value is 0.0307 which results in rejection of the null hypothesis for a test conducted at \( \alpha = .05 \). This is the same conclusion reached in the text.

**Task 9.6**

Use the data in Table 9.2 on page 330 to determine whether adding \( x_2 \) to a model that contains \( x_1 \) adds significantly to \( \hat{R}^2 \).

**Solution**

This problem can be approached in a number of different ways. For example, we could use Stata to find \( \hat{R}^2_{y,1} \) and \( \hat{R}^2_{y,12} \) then use a calculator to apply Equation 9.24. We will use a simpler more direct, if somewhat less obvious, approach.

Referring to the output obtained from Task 9.4 we see that a \( t \) value of 1.88 is reported for a test of \( H_0 : \beta_2 = 0 \) which has an associated \( p \)-value of 0.084. This test is equivalent to a test of \( \hat{R}^2_{y,12} - \hat{R}^2_{y,1} = 0 \) which more directly expresses the added contribution of \( x_2 \) to a model initially made up of \( x_1 \). The associated \( p \)-value of 0.084 means that we would fail to reject the null hypothesis \( \hat{R}^2_{y,12} - \hat{R}^2_{y,1} = 0 \) at the .05 level. This is the conclusion reached on page 337 of the text. Therefore, we are unable to demonstrate that adding \( x_2 \) to a model that contains \( x_1 \) adds to \( \hat{R}^2 \).
Exercises

9.1 Find the regression equation for predicting $y$ from $x$ for the data in Table 9.1 on page 323.

9.2 Use Stata to construct the model described in Exercise 9.7 on page 341 of the text.

9.3 Test the null hypothesis $H_0 : R^2 = 0$ for the model you constructed in Exercise 9.2.

9.4 Use the data in Table 9.2 on page 330 to determine whether adding $x_1$ to a model that contains $x_2$ adds significantly to $\hat{R}^2$. 
Chapter 10

Methods Based on the Permutation Principle

10.1 Introduction

In this chapter you will use Stata to perform Wilcoxon’s signed-ranks test, Wilcoxon’s rank-sum (Mann-Whitney) test, the Kruskal-Wallis test and Fisher’s exact test.

10.3 Applications (page 348)

Task 10.1

Perform the Wilcoxon’s signed-ranks test described in Example 10.15 on page 369.

Solution

Load the table 10.7 data.dta data into memory. Notice that this file contains pre- and post-treatment values with variable names pre and post. We will use the signrank command for the analysis whose general form is

\[ \text{signrank entry1} = \text{entry2} \]

When matched scores are to be analyzed entry1 and entry2 are the names of the two variables. When a set of difference scores are to be analyzed entry1 is the name of the difference score variable and entry2 is usually zero which represents the median of the difference score distribution. For the problem at hand enter

\[ \text{signrank post} = \text{pre} \]
The output reports an obtained $Z$ value of $-1.112$ with an associated $p$-value of $0.2659$. This results in failure to reject the null hypothesis at $\alpha = 0.05$. This is the same conclusion reached in the text. It should be noted, however, that $p$-values reported by Stata for the Wilcoxon’s signed-ranks test are approximate rather than exact.

Let’s now create a set of difference scores by entering

```
gen d=post-pre
```

Suppose table 10.7 data.dta had contained these difference scores rather than the pre-treatment and post-treatment scores themselves. In this case the analysis would be conducted by entering

```
signrank d=0
```

As you can see, the same result is obtained.

**Task 10.2**

Perform the Wilcoxon’s rank-sum test described in Example 10.20 on page 382.

**Solution**

Load the example 10.20 data.dta data into memory. Notice that the first column contains an indicator of group membership while column two contains the outcome variable. We will use the `ranksum` command to perform the test. Its general form is

```
ranksum outcome, by(group)
```

where `outcome` is the outcome variable and `group` is the indicator of group membership. Because the variables in example 10.20 data.dta are named `outcome` and `group`, just enter the above command.

Stata reports a $Z$ value of $-0.655$ with a two-tailed $p$-value of $0.5127$. Taking half this value gives a one-tailed $p$-value of $0.2564$. As noted in the solution to Example 10.20, the exact $p$-value is $0.35$ so that the approximation provided by Stata is not very good but this is not surprising for such small sample sizes.

**Task 10.3**

Perform a Kruskal-Wallis test on the raw data in the table on page 393.

**Solution**

Enter the group number, i.e. 1, 2, or 3, of each of the 6 subjects in the first column and their outcome measure in the second column of a new worksheet. Name the first variable `group` and the second `outcome`. The entries should appear as in Figure 10.1.

The command for performing the Kruskal-Wallis test is `kwallis` with general form
Figure 10.1: Data arranged for Kruskal-Wallis test analysis.

<table>
<thead>
<tr>
<th>group</th>
<th>outcome</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>2</td>
<td>9</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>7</td>
</tr>
<tr>
<td>6</td>
<td>12</td>
</tr>
</tbody>
</table>

kwallis outcome, by(group)

where outcome is the outcome variable and group is an indicator of group membership. Because we named the variables outcome and group, just enter the above command.

Stata provides a chi-squared statistic with a $p$-value of 0.1561. As is shown in the solution to this problem (page 395), the exact $p$-value is 0.20 so that the Stata approximation is not too good. But this is to be expected with such small sample sizes.

Task 10.4

Perform a two-tailed Fisher's exact test on the data in Table 10.23 on page 405.

Solution

We will use the tabi command to perform the analysis. The general form is

\[ \text{tabi } f_{011} f_{012} \backslash f_{021} f_{022}, \text{exact} \]

where $f_{011}$ represents the observed frequency of the cell in the first row and first column, $f_{012}$ is the observed frequency of the cell in the first row second column etc. The option exact causes Stata to perform Fisher's exact test. For the problem at hand enter

\[ \text{tabi } 4 \ 2 \ \ 1 \ 8, \text{exact} \]

Stata gives the two-tailed $p$-value as 0.089 which agrees with the value calculated in the solution to Example 10.30 on page 405. A one-tailed $p$-value of 0.047 is also provided.
Exercises

10.1 Conduct the test described in Example 10.16 on page 371.

10.2 Perform the test alluded to in Example 10.21 on page 383.

10.3 Conduct the test described in Example 10.22 on page 385.

10.4 Conduct the Kruskal-Wallis test described in Example 10.27 on page 395.

10.5 Conduct the test described in Example 10.31 on page 407.
Bibliography


## Appendix A

### Table Relating Text Items to Tasks

This table relates various portions of the text to specific tasks and the task page number.

<table>
<thead>
<tr>
<th>Text</th>
<th>Task</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Table 2.2</td>
<td>2.1</td>
<td>11</td>
</tr>
<tr>
<td>Table 2.3*</td>
<td>2.2</td>
<td>13</td>
</tr>
<tr>
<td>Table 2.4</td>
<td>2.3</td>
<td>13</td>
</tr>
<tr>
<td>Figure 2.1</td>
<td>2.4</td>
<td>15</td>
</tr>
<tr>
<td>Table 2.3*</td>
<td>2.5</td>
<td>17</td>
</tr>
<tr>
<td>Figure 2.4</td>
<td>2.6</td>
<td>17</td>
</tr>
<tr>
<td>Figure 2.6</td>
<td>2.7</td>
<td>18</td>
</tr>
<tr>
<td>Example 2.4</td>
<td>2.9</td>
<td>21</td>
</tr>
<tr>
<td>Example 2.7</td>
<td>2.10</td>
<td>21</td>
</tr>
<tr>
<td>Table 2.6</td>
<td>2.11</td>
<td>22</td>
</tr>
<tr>
<td>Example 2.17</td>
<td>2.12</td>
<td>22</td>
</tr>
<tr>
<td>Example 2.21</td>
<td>2.13</td>
<td>23</td>
</tr>
<tr>
<td>Example 2.23</td>
<td>2.14</td>
<td>23</td>
</tr>
<tr>
<td>Example 2.24</td>
<td>2.15</td>
<td>23</td>
</tr>
<tr>
<td>Table 3.1</td>
<td>3.1</td>
<td>27</td>
</tr>
<tr>
<td>Example 3.9</td>
<td>3.3</td>
<td>29</td>
</tr>
<tr>
<td>Table 4.1</td>
<td>4.1</td>
<td>31</td>
</tr>
<tr>
<td>Example 4.7</td>
<td>4.2</td>
<td>32</td>
</tr>
<tr>
<td>Example 4.9</td>
<td>4.3</td>
<td>32</td>
</tr>
<tr>
<td>Example 4.12</td>
<td>4.4</td>
<td>33</td>
</tr>
</tbody>
</table>

77
<table>
<thead>
<tr>
<th>Text</th>
<th>Task</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Example 4.14</td>
<td>4.5</td>
<td>33</td>
</tr>
<tr>
<td>page 108</td>
<td>4.6</td>
<td>34</td>
</tr>
<tr>
<td>Example 4.19</td>
<td>4.7</td>
<td>35</td>
</tr>
<tr>
<td>Example 4.25</td>
<td>4.8</td>
<td>35</td>
</tr>
<tr>
<td>Example 4.26</td>
<td>4.9</td>
<td>36</td>
</tr>
<tr>
<td>Example 4.32</td>
<td>4.10</td>
<td>36</td>
</tr>
<tr>
<td>Example 4.34</td>
<td>4.11</td>
<td>37</td>
</tr>
<tr>
<td>Example 4.36</td>
<td>4.12</td>
<td>38</td>
</tr>
<tr>
<td>Example 4.37</td>
<td>4.13</td>
<td>39</td>
</tr>
<tr>
<td>Example 4.41</td>
<td>4.14</td>
<td>39</td>
</tr>
<tr>
<td>Example 4.42</td>
<td>4.15</td>
<td>39</td>
</tr>
<tr>
<td>Example 5.1</td>
<td>5.1</td>
<td>43</td>
</tr>
<tr>
<td>Example 5.4</td>
<td>5.2</td>
<td>44</td>
</tr>
<tr>
<td>Table 5.1</td>
<td>5.3</td>
<td>44</td>
</tr>
<tr>
<td>Table 5.7</td>
<td>5.4</td>
<td>45</td>
</tr>
<tr>
<td>Example 5.13</td>
<td>5.5</td>
<td>46</td>
</tr>
<tr>
<td>Example 5.15</td>
<td>5.6</td>
<td>46</td>
</tr>
<tr>
<td>Example 5.17</td>
<td>5.7</td>
<td>47</td>
</tr>
<tr>
<td>Example 5.20</td>
<td>5.8</td>
<td>47</td>
</tr>
<tr>
<td>Example 5.22</td>
<td>5.9</td>
<td>48</td>
</tr>
<tr>
<td>Example 5.24</td>
<td>5.10</td>
<td>48</td>
</tr>
<tr>
<td>Example 5.27</td>
<td>5.11</td>
<td>48</td>
</tr>
<tr>
<td>Example 5.30</td>
<td>5.12</td>
<td>49</td>
</tr>
<tr>
<td>Example 6.1</td>
<td>6.1</td>
<td>51</td>
</tr>
<tr>
<td>Example 6.3</td>
<td>6.2</td>
<td>52</td>
</tr>
<tr>
<td>Example 6.5</td>
<td>6.3</td>
<td>52</td>
</tr>
<tr>
<td>Example 6.7</td>
<td>6.4</td>
<td>53</td>
</tr>
<tr>
<td>Example 6.9</td>
<td>6.5</td>
<td>53</td>
</tr>
<tr>
<td>Example 6.11</td>
<td>6.6</td>
<td>54</td>
</tr>
<tr>
<td>Table 7.1</td>
<td>7.1</td>
<td>59</td>
</tr>
<tr>
<td>Example 7.6</td>
<td>7.2</td>
<td>60</td>
</tr>
<tr>
<td>Example 7.9</td>
<td>7.3</td>
<td>61</td>
</tr>
<tr>
<td>Table 8.1</td>
<td>8.1</td>
<td>63</td>
</tr>
<tr>
<td>Example 8.7</td>
<td>8.2</td>
<td>64</td>
</tr>
<tr>
<td>Example 9.1</td>
<td>9.1</td>
<td>67</td>
</tr>
<tr>
<td>Table 9.1*</td>
<td>9.2</td>
<td>67</td>
</tr>
<tr>
<td>Table 9.1*</td>
<td>9.3</td>
<td>68</td>
</tr>
<tr>
<td>Table 9.2*</td>
<td>9.4</td>
<td>68</td>
</tr>
<tr>
<td>Table 9.2*</td>
<td>9.5</td>
<td>69</td>
</tr>
<tr>
<td>Text</td>
<td>Task</td>
<td>Page</td>
</tr>
<tr>
<td>-----------------------</td>
<td>------</td>
<td>------</td>
</tr>
<tr>
<td>Table 9.2*</td>
<td>9.6</td>
<td>69</td>
</tr>
<tr>
<td>Example 10.15</td>
<td>10.1</td>
<td>71</td>
</tr>
<tr>
<td>Example 10.20</td>
<td>10.2</td>
<td>72</td>
</tr>
<tr>
<td>page 393</td>
<td>10.3</td>
<td>72</td>
</tr>
<tr>
<td>Table 10.23</td>
<td>10.4</td>
<td>73</td>
</tr>
</tbody>
</table>

* Appears more than once in list.