

**SPSS MANUAL**  
for Moore's

**The Basic Practice of Statistics**  
**Fifth Edition**

Patricia Humphrey  
*Georgia Southern University*

W.H. Freeman and Company  
New York

Copyright © 2010 by W.H. Freeman and Company

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

ISBN-10: 1-4292-2785-0

# Contents

<b>Preface</b>	<b>iii</b>
<b>CHAPTER 0</b>	
<b>Introduction to SPSS</b>	<b>1</b>
0.1 Accessing SPSS	2
0.2 Opening and Saving Data Files	3
0.3 Defining Variables and Entering Data	5
0.4 Opening Excel Files	7
0.5 Recoding Variables	8
0.6 Deleting/Inserting a Case or a Column	10
0.7 Selecting Cases	11
0.8 Using SPSS Help	12
<b>CHAPTER 1</b>	
<b>Looking at Data – Distributions</b>	<b>13</b>
1.1 Displaying Distributions with Graphs	14
1.2 Describing Distributions with Numbers	25
1.3 Normal Distributions	30
<b>CHAPTER 2</b>	
<b>Exploring/Looking at Data – Relationships</b>	<b>34</b>
2.1 Scatterplots	35
2.2 Correlation	40
2.3 Least-Squares Regression	42
2.4 Cautions about Correlation and Regression	44
2.5 Relations in Categorical Variables	48
<b>CHAPTER 3</b>	
<b>Producing Data</b>	<b>52</b>
3.1 First Steps	53
3.2 Design of Experiments	55
3.3 Sampling Design	57
3.4 Toward Statistical Inference	58

<b>CHAPTER 4</b>	
<b>Probability: The Study of Randomness</b>	<b>60</b>
4.1 Randomness	61
4.2 Probability Models	64
4.3 Random Variables	66
4.4 Means of Random Variables	68
4.5 General Probability	69
<b>CHAPTER 5</b>	
<b>Sampling Distributions</b>	<b>71</b>
5.1 <b>Sampling Distributions for Counts and Proportions</b>	<b>72</b>
Binomial Probabilities	72
Probabilities for $\hat{p}$	75
Normal Approximations	75
5.2 <b>Poisson Random Variables</b>	<b>76</b>
5.3 <b>The Sampling Distribution of a Sample Mean</b>	<b>77</b>
Sum of Independent Normal Measurements	79
Sum and Difference of Sample Means	79
<b>CHAPTER 6</b>	
<b>Introduction to Inference</b>	<b>80</b>
6.1 <b>Confidence Intervals with <math>\sigma</math> Known</b>	<b>81</b>
Choosing the Sample Size	82
6.2 <b>Tests of Significance</b>	<b>83</b>
6.3 <b>Use and Abuse of Tests</b>	<b>84</b>
6.4 <b>Power and Inference as a Decision</b>	<b>86</b>
<b>CHAPTER 7</b>	
<b>Inference for Distributions</b>	<b>88</b>
7.1 <b>Inference for the Mean of a Population</b>	<b>89</b>
One-sample $t$ Confidence Interval	89
One-sample $t$ test	92
Matched Pair $t$ Procedure	94
The Sign Test	97
7.2 <b>Comparing Two Means</b>	<b>97</b>
7.3 <b>Optional Topics in Comparing Distributions</b>	<b>100</b>

<b>CHAPTER 8</b>	
<b>Inference for Proportions</b>	<b>101</b>
<b>8.1 Inference for a Single Proportion</b>	<b>102</b>
A Large-Sample Confidence Interval	102
A Plus-Four Confidence Interval	102
Choosing a Sample Size	103
Significance Tests	103
<b>8.2 Comparing Two Proportions</b>	<b>105</b>
A Large-Sample Confidence Interval for Difference of Proportions	105
A Plus-Four Confidence Interval for Difference of Proportions	106
Significance Tests for Difference of Proportions	106
<b>CHAPTER 9</b>	
<b>Inference for Two-Way Tables</b>	<b>108</b>
<b>9.1 Data Analysis for Two-Way Tables</b>	<b>109</b>
<b>9.2 Inference for Two-Way Tables</b>	<b>110</b>
Comparison with the 2-PropZTest	113
<b>9.3 Goodness of Fit</b>	<b>117</b>
<b>CHAPTER 10</b>	
<b>Inference for Regression</b>	<b>119</b>
<b>10.1 Simple Linear Regression</b>	<b>120</b>
Hypotheses other than 0	123
Confidence Intervals in Regression Inference	124
Confidence Intervals for a Mean or Individual Response	127
<b>10.2 More Detail about Simple Linear Regression</b>	<b>128</b>
Sample Correlation and the $t$ test	130
<b>CHAPTER 11</b>	
<b>Multiple Regression</b>	<b>131</b>
<b>11.1 Inference for Multiple Regression</b>	<b>132</b>
<b>11.2 A Case Study</b>	<b>135</b>
<b>11.3 Another Type of Plot</b>	<b>138</b>

<b>CHAPTER 12</b>	
<b>One-Way Analysis of Variance</b>	<b>140</b>
12.1 Inference for One-Way Analysis of Variance	141
<b>CHAPTER 13</b>	
<b>Two-Way Analysis of Variance</b>	<b>145</b>
13.1 Plotting Means	146
13.2 Inference for Two-Way ANOVA	147
<b>CHAPTER 14</b>	
<b>Bootstrap Methods and Permutation Tests</b>	<b>151</b>
<b>CHAPTER 15</b>	
<b>Nonparametric Tests</b>	<b>152</b>
15.1 The Wilcoxon Rank Sum Test	153
15.2 The Wilcoxon Signed Rank Test	158
15.3 The Kruskal-Wallis Test	162
<b>CHAPTER 16</b>	
<b>Logistic Regression</b>	<b>165</b>
16.1 The Logistic Regression Model	166
Model for Logistic Regression	167
<b>CHAPTER 17</b>	
<b>Statistics for Quality: Control and Capability</b>	<b>171</b>
17.1 Statistical Process Control	172
17.2 Process Capability Indices	175
17.3 Control Charts for Sample Proportions	177
<b>CHAPTER 18</b>	
<b>Time Series Forecasting</b>	<b>180</b>
18.1 Trends and Seasons	181
18.2 Time Series Models	184

<b>Problem Statements</b>	<b>191</b>
<b>Solutions</b>	<b>288</b>

## Preface

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

With its powerful computation abilities SPSS has been a statistical staple for many years; I first encountered it as a Master's student many years ago in the days of punch cards. Today, students and teachers can have instant access to many statistical procedures on their desktop or laptop. SPSS is not, however, a panacea. It will not tell you what analysis or test is appropriate for a given set of data; that is the realm of the practicing statistician, as is interpretation of the output. Just as any computer program will have its drawbacks, SPSS does not function well with data that have already been summarized, nor will its base or student versions perform all the calculations or tests a practicing statistician (or even a student) might want without additional add-on modules. That said, when the data are suited, SPSS is an extremely useful aid.

This manual serves as a companion to your W. H. Freeman Introductory Statistics text by David Moore and others. Examples either taken from the text or similar to those in the text are worked using SPSS. The tremendous capabilities and usefulness of this computer package, as well as its limitations, are demonstrated throughout. It is hoped that students, teachers, and practitioners of statistics will continue to make use of these capabilities, and that readers will find this manual to be helpful.

## Acknowledgments

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

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

email: [phumphre@georgiasouthern.edu](mailto:phumphre@georgiasouthern.edu)  
homepage: <http://math.georgiasouthern.edu/~phumphre/>

## CHAPTER

# 0

# Introduction to SPSS

0.1	Accessing SPSS
0.2	Opening and Saving Data Files
0.3	Defining Variables and Entering Data
0.4	Opening Excel Files
0.5	Recoding Variables
0.6	Deleting/Inserting a Case or a Column
0.7	Selecting Cases
0.8	Using SPSS Help

## **Introduction**

In this chapter we introduce SPSS, the Statistical Package for the Social Sciences. This manual is intended to help a student perform the statistical procedures presented in your W. H. Freeman text: *Introduction to the Practice of Statistics*, *The Basic Practice of Statistics*, *The Practice of Business Statistics*, or *The Practice of Statistics in the Life Sciences*. This supplement is intended to use SPSS for Windows version 16 (current at this writing). However, the instructions included here will work for most versions and for most basic statistical procedures.

Throughout this manual, the following convention is used: commands you click or text you type are in boldface and underlined (e.g., go to **File**) and, most of the time, variables are in boldface (e.g., **Count**).

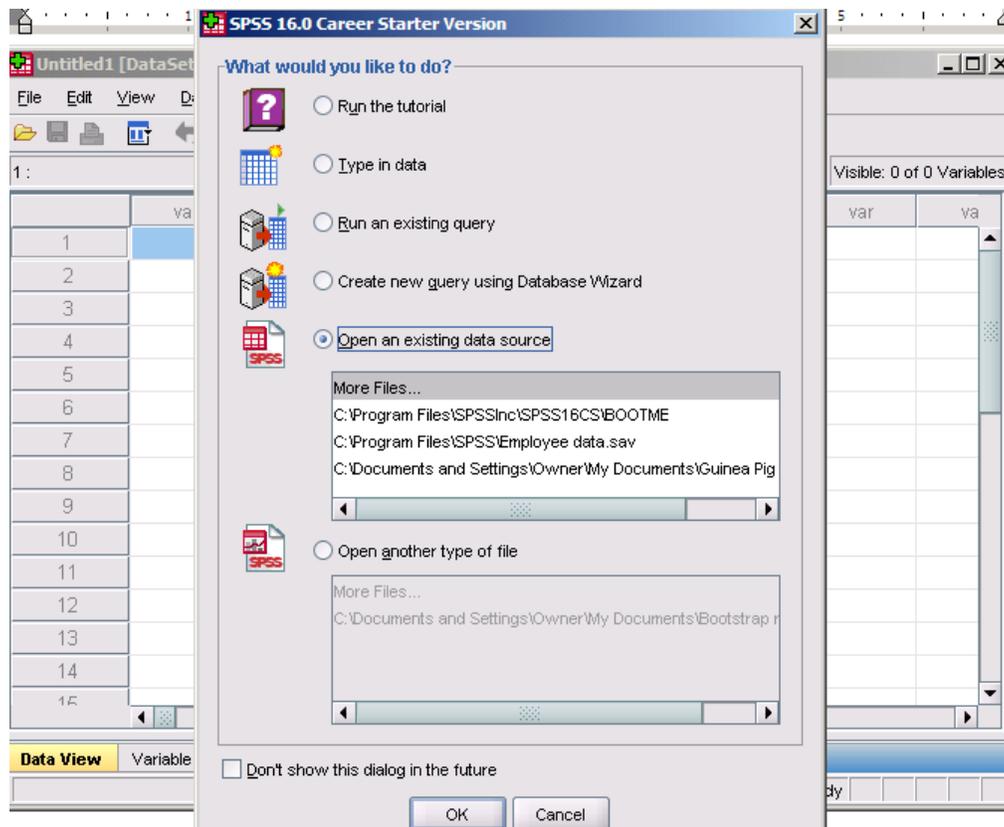
## 0.1 Accessing SPSS

If you work in a lab, locate SPSS in the computer. You should look for the following icon on the desktop:



SPSS 16.0.Ink

Your computer may have a similar icon with an earlier version number. Double click on the icon to start SPSS running. If there is no desktop icon, use the **Start** Menu on your computer and open All Programs, locate the SPSS Inc folder, and follow it to find the program. Once the program has been started, you will briefly see an introductory screen (similar to an Excel or Word start-up screen) followed (most likely, unless this has been disabled by checking the box at the bottom) by this screen.

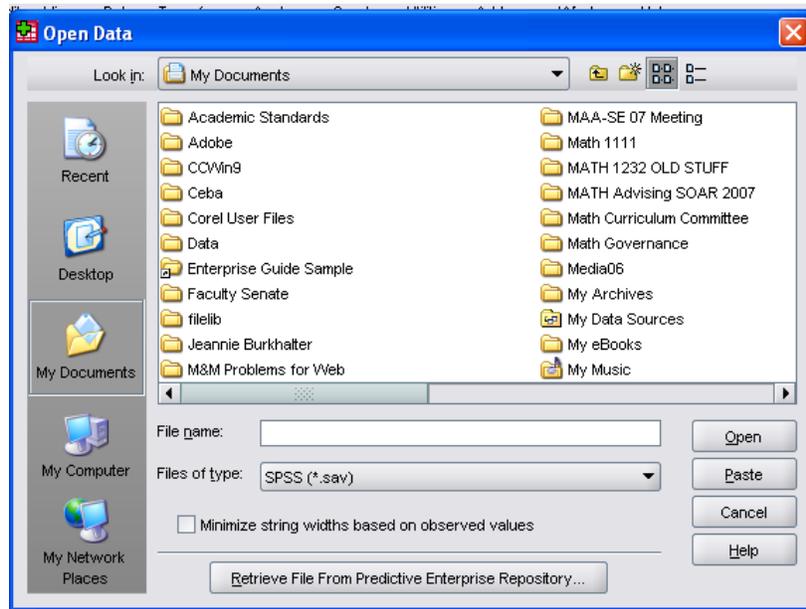


Two SPSS windows will actually be open at this point — the output viewer and the Data Editor. Before doing any statistics or graphs, we must have data. The purpose of the introductory screen is for the program to determine your data source. Behind the data source selection box is a blank spreadsheet — the Data View screen.

Since all data sets used in your text are on the included CD-Rom (and on the Companion Website), here you will most likely click **OK** (or press **Enter**) to select the default option which is to **Open** an existing data source.

## 0.2 Opening and Saving Data Files

If the introductory data source selection menu is not presented (or you want to proceed to another data set within an SPSS session), click **File**, **Open**, **Data**. Initially, you will see the screen below.



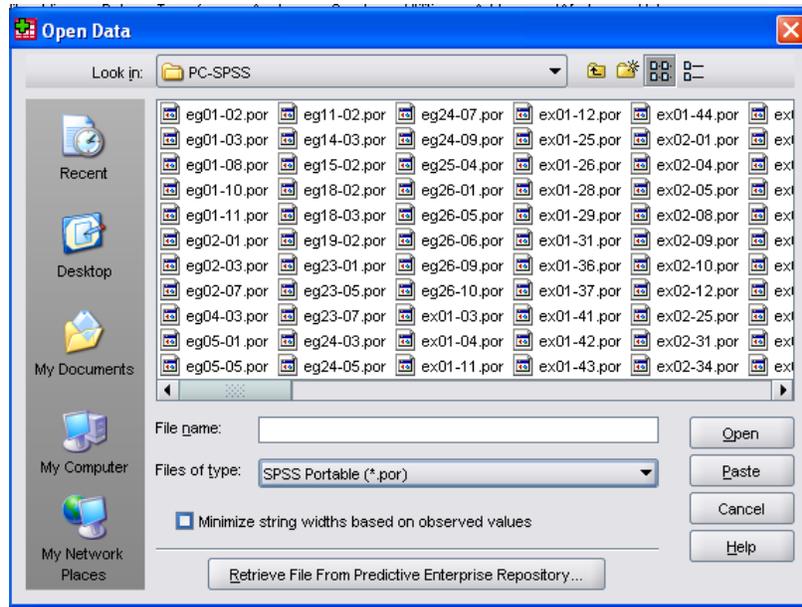
File selection works much the same as any other Windows program. In the **Look in** box, select the location of your data set (drive and folder). The SPSS default data file extension is .sav. Data files on the CD-Rom are saved as SPSS portable worksheets, so change the box labeled **Files of type** to SPSS Portable (.por) as in the screen following.

File naming conventions used on the CD and Website are the following:

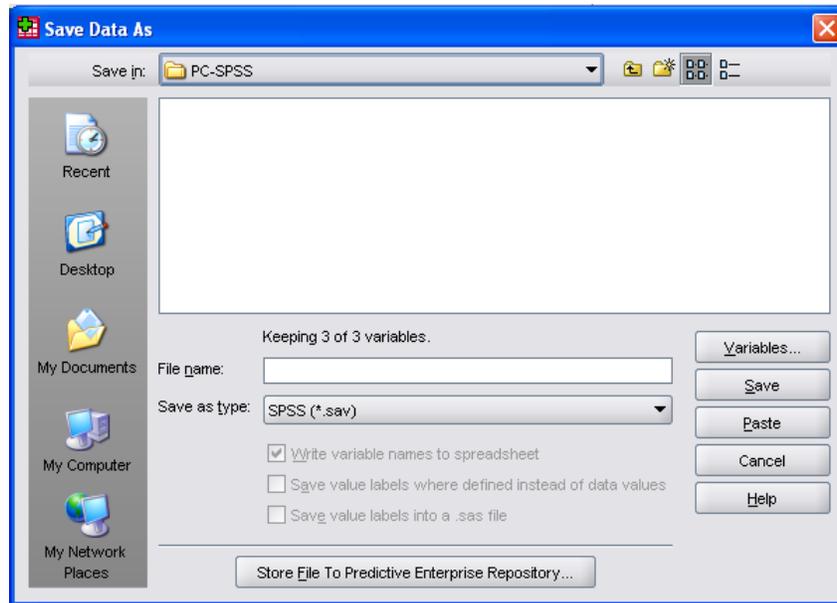
1. The first two characters describe the type of data set. Examples are eg, exercises are ex, figures are fg, and tables are ta.
2. The second two characters indicate the chapter number.
3. Numbers after the dash correspond to the example, exercise, figure, or table number within the chapter.

Once you have located the file you want, click **Open**.

#### 4 Chapter 0 – Introduction to SPSS



To save an SPSS data file, click **File, Save As** or **File, Save all Data** depending on your SPSS version. In the File Name box, type the name you wish to give your data. The default folder is PC-SPSS. Be sure to change that if you want a different location, such as diskette or flash drive.



### 0.3 Defining Variables and Entering Data

In the event you need to enter your own data for a project, on the first (opening) screen select the **Type in data** button and click **OK**. You will see the blank SPSS Data Editor spreadsheet window.

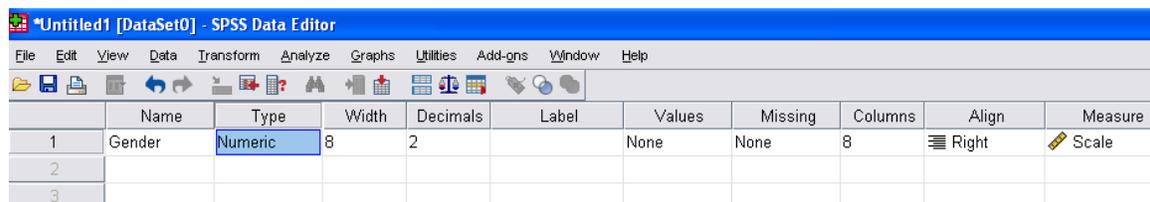
**Example 0.1: Creating an SPSS Data File by Entering Data.** The following data set contains 10 randomly selected scores in the final exam of a basic statistics course at XYZ College. Along with the final exam scores, the number of classes missed during the semester and the gender of the students were also recorded. The data set is given below:

<u>Gender</u>	<u>Number Classes Missed</u>	<u>Final Score</u>
Male	2	83
Female	0	93
Male	6	61
Female	1	73
Female	0	95
Female	4	75
Male	3	77
Male	4	71
Female	5	68
Female	4	59

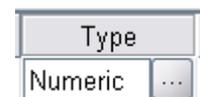
One can simply begin either by typing the data into the spreadsheet, or by defining the variables. For completeness sake, both steps should be completed, but order is unimportant.

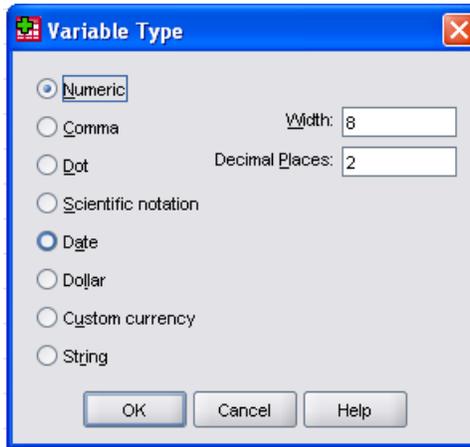
To define the variables, click on the **Variable View** tab at the bottom of the Data Editor window.

Under **Name**, type the name of the first variable (eight characters or fewer, beginning with a letter or the underscore sign). In this case, the name of the first variable is **Gender**. Press the **Tab** key to advance to the Type box. Notice what SPSS defaults variables to: Numeric, with two decimal places, occupying eight columns. Gender is a categorical variable. We need to change this.



To change the variable type, click in the highlighted box, then click on the small button that appears at the right.





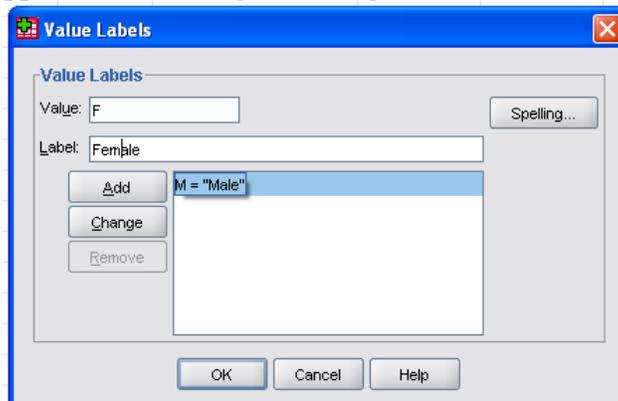
To change this variable to a categorical one, click on **String**. You will be allowed to change the maximum number of characters (the default is eight) if desired. When finished, click **OK**.

Enter the name of the second variable **NumMiss**. This variable name could stand for some explanation on output, rather than just this cryptic name. Press the **Tab** key to move to the **Label** Column. Type in a more appropriate “long” variable name, such as **Number Classes Missed**.

Enter the third variable name **Final** and label it as Final Exam Score as just detailed. At this point, our variables definition should look like that below.

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measu
1	Gender	String	8	0		None	None	8	≡ Left	Nominal
2	NumMiss	Numeric	8	2	Number Classes Missed	None	None	8	≡ Right	Scale
3	Final	Numeric	8	2	Final Exam Score	None	None	8	≡ Right	Scale
4										

Lastly, consider the gender variable. We’d like our data entry to be as easy as possible, but have SPSS print out the full word for Male and Female students. If we just want to enter M (or F) we can define value labels that will print the full descriptor. Click the cursor in the **Values** field of the **Gender** variable. A small box like that shown on the previous page will appear. Click it to get a dialog box.



Here, I have already added the label for Males and input both the value and label for Females. To add the value label, click **Add**. When finished adding labels, click **OK**.

To start entering the data, click on **Data View** and enter the values, pressing **Tab** after each entry. The program will automatically advance to the next row after the third variable for an individual has been entered. For capital F and M, engage **Caps Lock**. If you make a typographical error, simply click on the cell and type in the correct value.

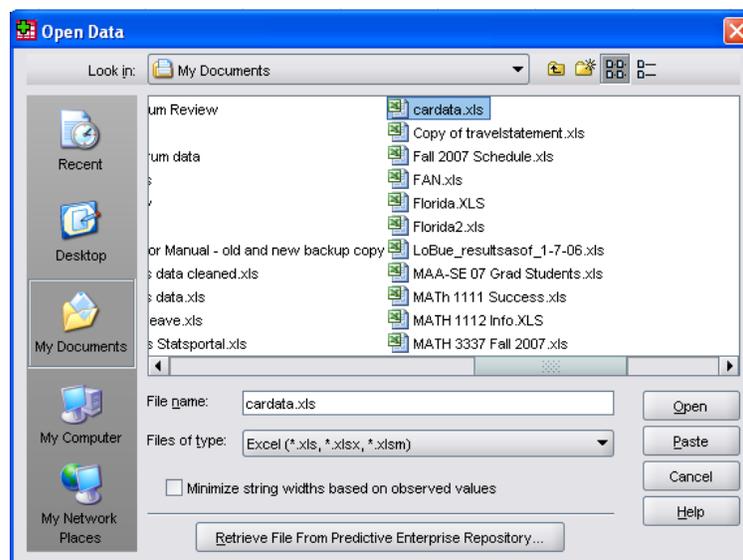
	Gender	NumMiss	Final
1	M	2.00	83.00
2	F	0.00	93.00
3	M	6.00	61.00
4	F	1.00	73.00
5	F	0.00	95.00
6	F	4.00	75.00
7	M	3.00	77.00
8	M	4.00	71.00
9	F	5.00	68.00
10	F	4.00	59.00
11			

To save your data in an SPSS formatted file, follow the instructions in the preceding section.

## 0.4 Opening Excel Files

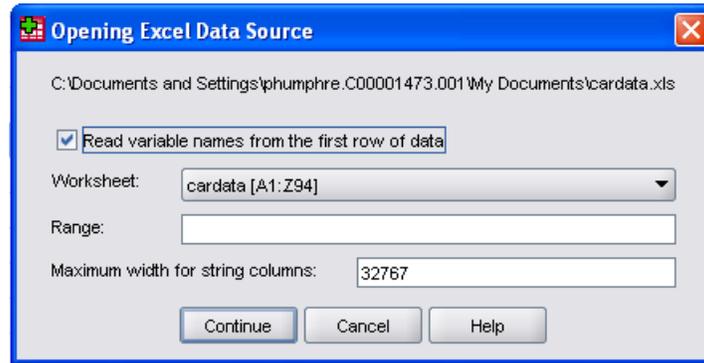
To open an Excel data file, follow these steps:

1. Click on **File, Open, Data**, then the Open File window will appear as already shown.
2. Choose the directory or location where the desired file is located. In our case, the file is stored in the My Documents folder.
3. Change the **Files of Type** box from the default .sav to Excel (.xls) or all files.



4. Click on the desired file name. In our case, it is **cardata.xls** as shown above.

Depending on the Excel file, one needs to know whether the names of the variables are located in the first row or not, i.e., where do the actual data start? Is it in the first or second row? In our case, the first row does have names, so leave the **Read variable names from the first row of data** box checked. Click **Continue** to open the file.



SPSS will take most variable attributes from the information in the Excel file. You probably will want to give more meaningful “long” variable names. Click on **Variable View** and add **Labels** as shown above.

## 0.5 Recoding Variables

One can change a categorical (string) variable to numeric. Also, one can transform a quantitative variable from one form to another by categorizing or by recoding the variable. The following example shows how to categorize a numeric variable in SPSS.

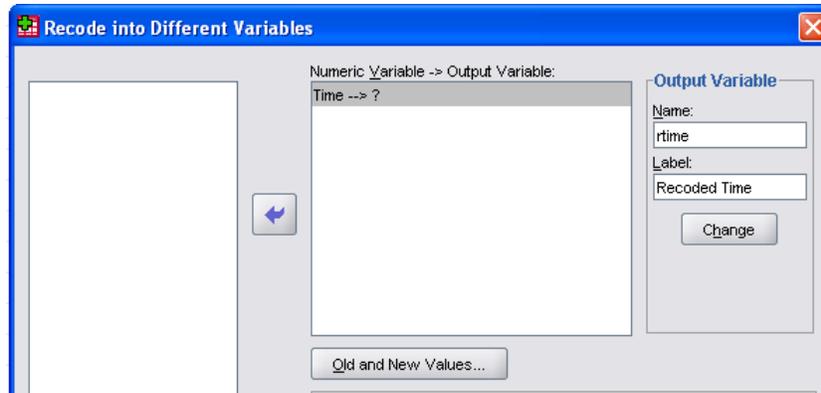
**Example 0.2: Recoding Variables.** The following data represent the waiting time (in seconds) for a random sample of 30 customers at a local bank.

49, 160, 80, 220, 170, 92, 178, 66, 124, 144, 71, 183, 248, 191, 155, 166, 256, 300, 180, 166, 171, 280, 144, 110, 267, 188, 160, 90, 205, 136

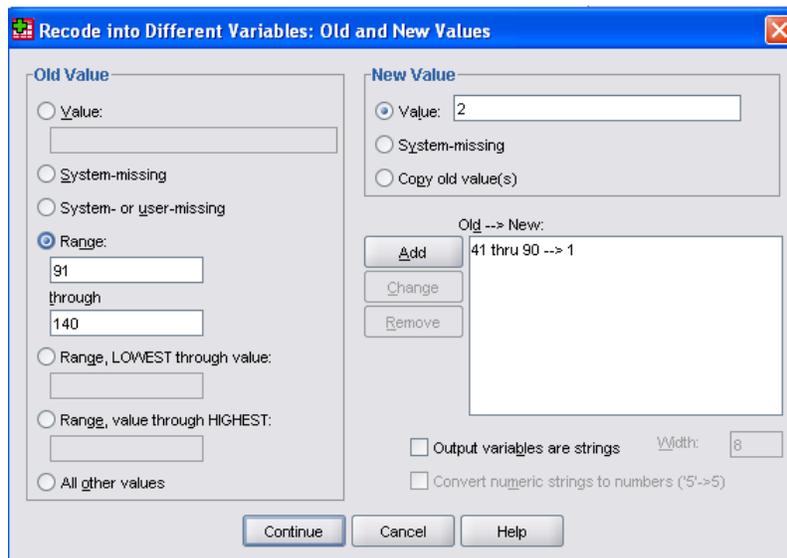
It may be more useful to group these data into non-overlapping classes (i.e., to create frequency tables). Let us recode these data to six equal width classes. The number of classes is usually determined by sample size and should fall between 5 and 20 intervals. A good rule of thumb is to use the square root of the sample size as a rough estimate for the number of classes.

The names of these classes will be 1, 2, 3, 4, 5, and 6. All observations between 41 and 90 seconds (inclusive) will be assigned to class 1, all observations between 91 and 140 will be assigned to class 2, all observations between 141 and 190 to class 3, and so on. Here is how to do it in SPSS.

Click on **Transform, Recode into Different Variable**. Time has already been highlighted as the input variable (since it's the only one in this spreadsheet). If there are more variables in your sheet, click to select the one of interest. Click the arrow to move Time (or your selected variable) to the working area at right.



Name the new output variable in the box at right, and give it a “long” name or label if desired. Click the **Change** button to record the new variable name. Click on the **Old and New Values** button. I have already defined the first category displayed in the box at right. Here, I am defining the second category as including the range 91-140 with new value 2. Click the **Add** button to complete this category definition, and define the others. When all categories have been defined, click **Continue** to return to the first Recode box. Click **OK** at the bottom of the box to create the new variable.



It will be useful to tell anyone who looks at your output what these recoded values represent. To do this, click on the **Variable View** tab at the bottom of the worksheet, then click in the **Values** box and add value labels as discussed previously.

## 0.6 Deleting/Inserting a Case or a Column

The data presented in the previous example will be used to illustrate the points of this section. To delete a case (an entire row of data), follow these steps:

1. Locate the case to be deleted by scrolling through the data.
2. Click on the case number at the left. The entire row will be highlighted.  
Suppose observation 14 should be deleted because it was an extra.
3. Press Delete on the keyboard.

The screenshot shows the SPSS Data Editor window titled "Untitled1 [DataSet0] - SPSS Data Editor". The menu bar includes File, Edit, View, Data, Transform, Analyze, Graphs, Utilities, Add-ons, and Win. The toolbar contains various icons for file operations and data manipulation. The status bar shows "14 : Time" and "191". The data table has the following structure:

	Time	rtime	VAR00001	VAR00002	VAI
8	66.00	1.00	.	.	.
9	124.00	2.00	.	.	.
10	144.00	3.00	.	.	.
11	71.00	1.00	.	.	.
12	183.00	3.00	.	.	.
13	248.00	5.00	.	.	.
14	191.00	4.00	.	.	.
15	155.00	3.00	.	.	.
16	166.00	3.00	.	.	.

To insert a case, follow these steps:

1. In the Data View window, click on the case number *below* where the new case should be.
2. Click **Edit, Insert Cases**. A blank row will be inserted.
3. Type in the desired data values for that observation.

The screenshot shows the SPSS Data Editor window titled "Untitled1 [DataSet0] - SPSS Data Editor". The menu bar includes File, Edit, View, Data, Transform, Analyze, Graphs, and Utilit. The status bar shows "10 : Time" and "144". The data table has the following structure:

	Time	rtime	VAR00001
5	170.00	3.00	.
6	92.00	2.00	.
7	178.00	3.00	.
8	66.00	1.00	.
9	124.00	2.00	.
10	144.00	3.00	.
11	71.00	1.00	.
12	183.00	3.00	.
13	248.00	5.00	.
14	155.00	3.00	.

The screenshot shows the SPSS Data Editor window titled "Untitled1 [DataSet0] - SPSS Data Editor". The menu bar includes File, Edit, View, Data, Transform, Analyze, Graphs, and U. The status bar shows "10 : Time". The data table has the following structure:

	Time	rtime	VAR00001
5	170.00	3.00	.
6	92.00	2.00	.
7	178.00	3.00	.
8	66.00	1.00	.
9	124.00	2.00	.
10	.	.	.
11	144.00	3.00	.
12	71.00	1.00	.
13	183.00	3.00	.

To insert a new variable it is easiest to define one in the Variable View as previously described after those already in the data set. If you want to insert one in the middle, follow these steps:

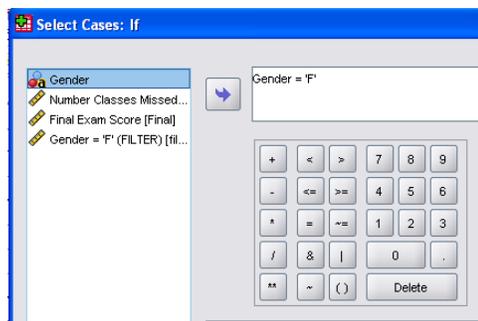
1. In the Variable View, click on the variable below where the new one is to be inserted.
2. Click on **Edit > Insert Variables**. A new variable with name of the form **VAR00xx** will be inserted. Change this name to the desired name, and also change any of the default characteristics (variable type, number of decimal places, etc.) as needed.

To delete a variable within the Data View, click on the variable name and press the Delete key on the keyboard.

## 0.7 Selecting Cases

Statistical analyses are sometimes needed for part of the data rather than for the entire data set. For example, it may be desired to compare the Females against the Males for the data on absences and final exam scores used in Example 0.1 (page 5). We might also want to do a regression with and without outliers to examine their impact.

1. Click **Data > Select Cases**.
2. Move the button highlight to **If condition is satisfied** and click the **If** button.
3. Highlight the variable name to be used and press the right arrow box to transfer this into the condition box. Complete the condition (in this case we want to select Females). Click **Continue** to return to the main Select Cases box.
4. Click **OK** to perform the selection. We see in the screen at right that Males will now be ignored, and a new variable named filter\_\$ has been created. This variable has values 0 or 1 according to whether the case has been excluded or not.

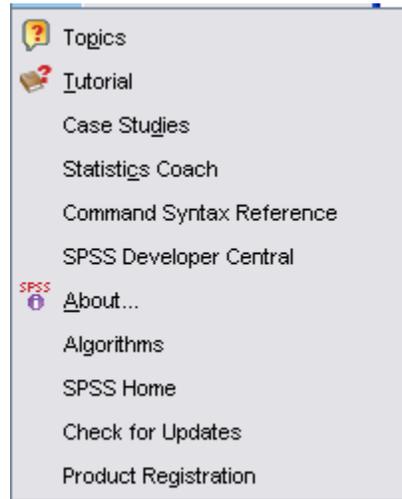


	Gender	Nummiss	Final	filter_ \$
1	M	2.00	83.00	0
2	F	0.00	93.00	1
3	M	6.00	61.00	0
4	F	1.00	73.00	1
5	F	0.00	95.00	1
6	F	4.00	75.00	1
7	M	3.00	77.00	0
8	M	4.00	71.00	0
9	F	5.00	68.00	1

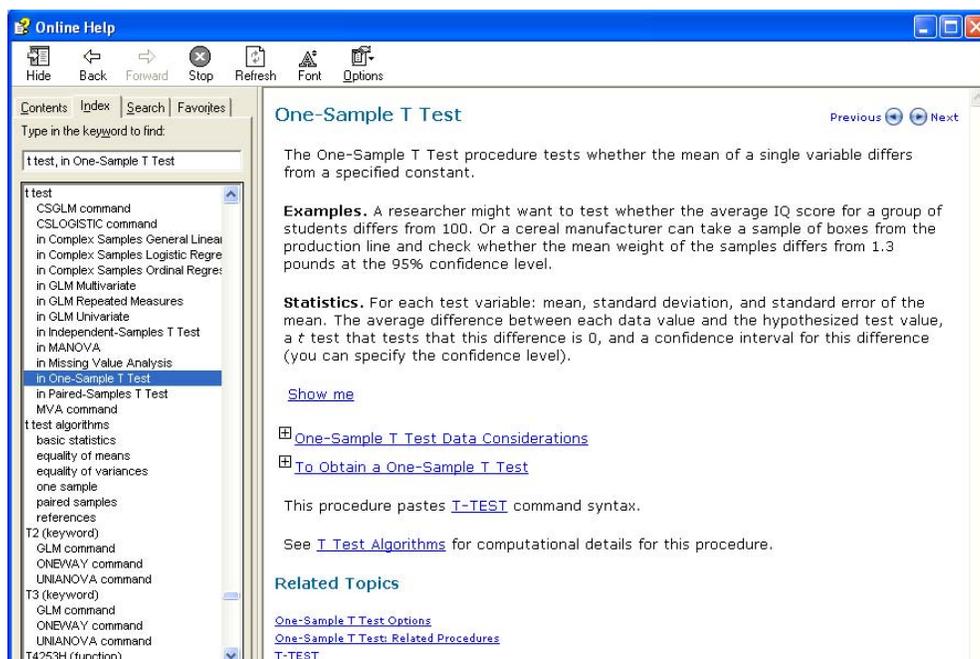
To return to using all cases and remove the filter, go back to **Data > Select Cases** and select the **All Cases** button, then click **OK**.

## 0.8 Using SPSS Help

Suppose you were looking for information on how to do something in SPSS and you can't find it in this manual (heaven, forbid). Help is available in several forms by clicking **Help** on the right-hand side of the top menu bar. Context specific help is available in every dialog box simply by clicking the button.

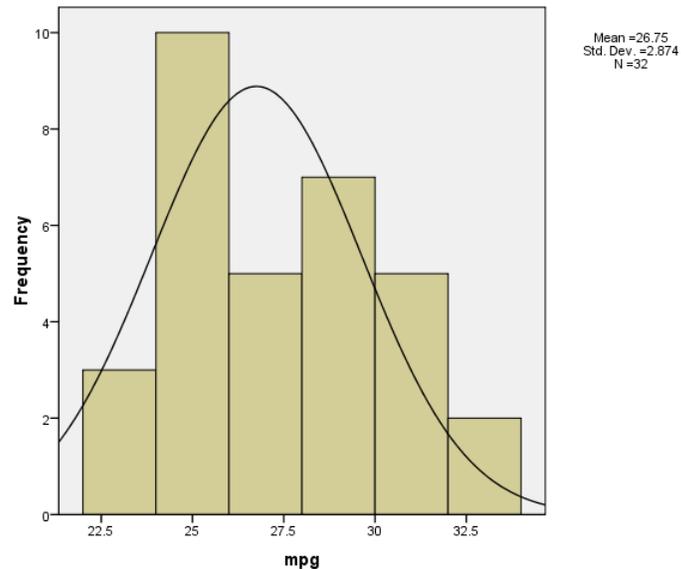


The **Tutorial** offers basic information on certain topics, in much the same manner as this manual. The **Statistics Coach** presents a series of screens to narrow down the search of topic and presents sample output as well. Lastly, one can search for help by **Topics**. This author recommends that if searching by topic, select **Index** after the initial Topics selection. Enter the topic name in the search box. As you type more characters, the index at left will move to try to “zero in” on the topic of interest. When you see it, highlight the topic name and click **Display**. The screen below illustrates the initial results from a search for *t*-tests.



## CHAPTER

# 1



# Looking at Data— Distributions

- |     |                                       |
|-----|---------------------------------------|
| 1.1 | Displaying Distributions with Graphs  |
| 1.2 | Describing Distributions with Numbers |
| 1.3 | Normal Distributions                  |

## Introduction

In this chapter, we use SPSS to view data sets. We first show how to make bar graphs, pie charts, histograms, and time plots. Then we compute basic statistics, such as the mean, median, and standard deviation, and show how to view data further with boxplots. Lastly, we use show how to perform calculations involving normal distributions.

## 1.1 Displaying Distributions with Graphs

**Example 1.1 Radio Station Formats–Bar Graph and Pie Chart.** The radio audience rating service Arbitron places the country’s 13,838 radio stations into categories that describe the kind of programs they broadcast. Their categorizations appear in the table below.

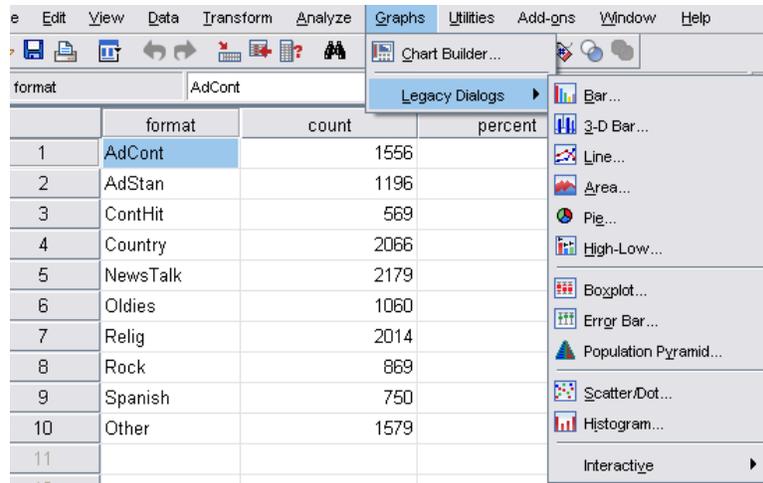
Format	Count of Stations	Percent of Stations
Adult Contemporary	1,556	11.2
Adult Standards	1,196	8.6
Contemporary Hit	569	4.1
Country	2,066	14.9
News/Talk/Information	2,179	15.7
Oldies	1,060	7.7
Religious	2,014	14.6
Rock	869	6.3
Spanish Language	750	5.4
Other formats	1,579	11.4
<b>Total</b>	<b>13,838</b>	<b>99.9</b>

We’d like to create graphics to display this information. Since station format is a categorical variable, we can use a bar graph or pie chart to display these data.

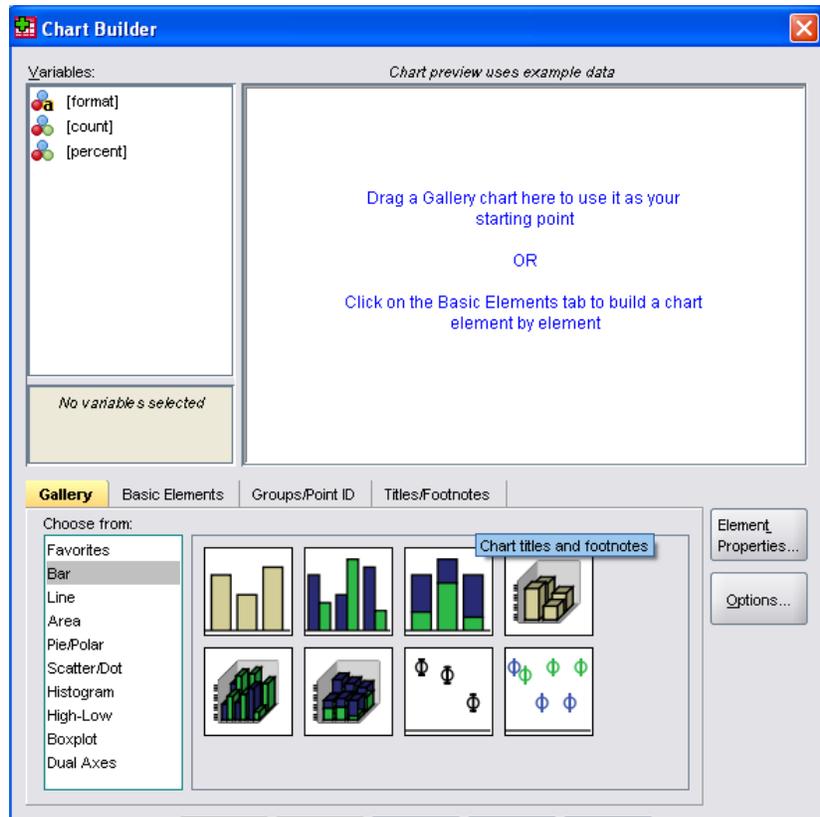
The data were entered using abbreviations for the formats as below.

	format	count	percent
1	AdCont	1556	11.2
2	AdStan	1196	8.6
3	ContHit	569	4.1
4	Country	2066	14.9
5	NewsTalk	2179	15.7
6	Oldies	1060	7.7
7	Relig	2014	14.6
8	Rock	869	6.3
9	Spanish	750	5.4
10	Other	1579	11.4

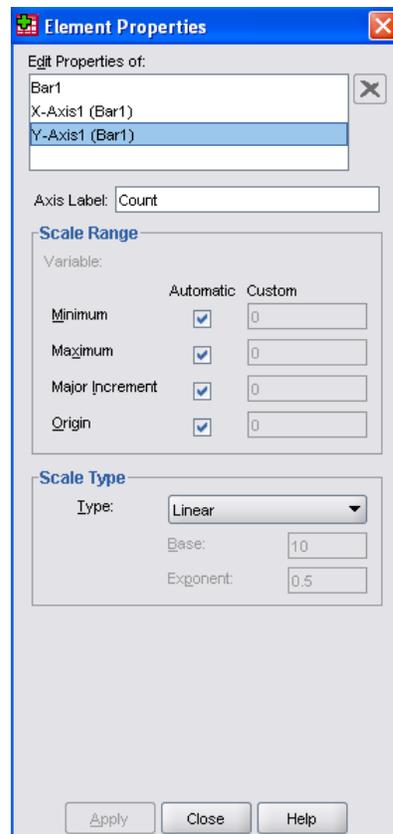
Click **Graphs**.



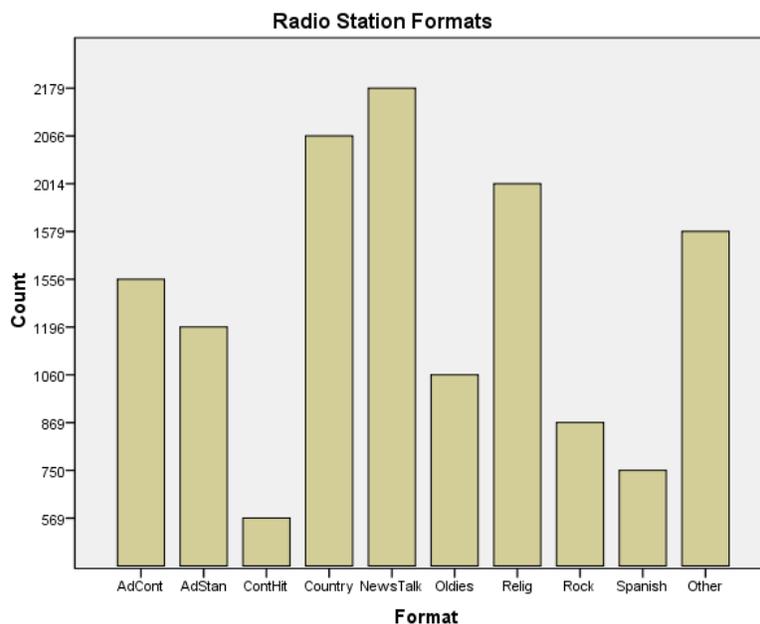
The Chart Builder is an intuitive way to build a graph that prompts you through the process much as Excel's Chart Wizard. However, in many cases using the **Legacy Dialogs** is easier.



As indicated, locate the type of graph you want (here, we want a bar graph) and drag it into the display box. I will drag the first bar graph type at the upper left into the box. You will next be asked to define the properties for each axis of your plot.



Notice that you can use default scaling (let SPSS figure it out), or define your own. We also give each axis a label through these boxes. Drag the variables into the desired locations (we've used **format** as x and **count** as y). If you want to give your graph a title, click on the **Titles/Footnotes** tab and enter it. Click **OK** to generate the graph into the Output window.

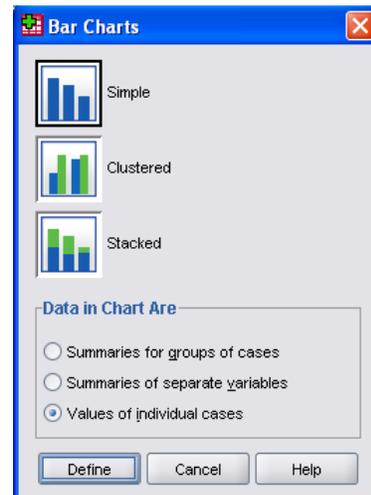


Notice that the scaling on the y axis produced by the automatic scaling choice is rather strange. To change this, go back to the Chart Builder and change the y axis properties. Uncheck the box for automatic increments and enter one of your own choice (500 might be a good value here).

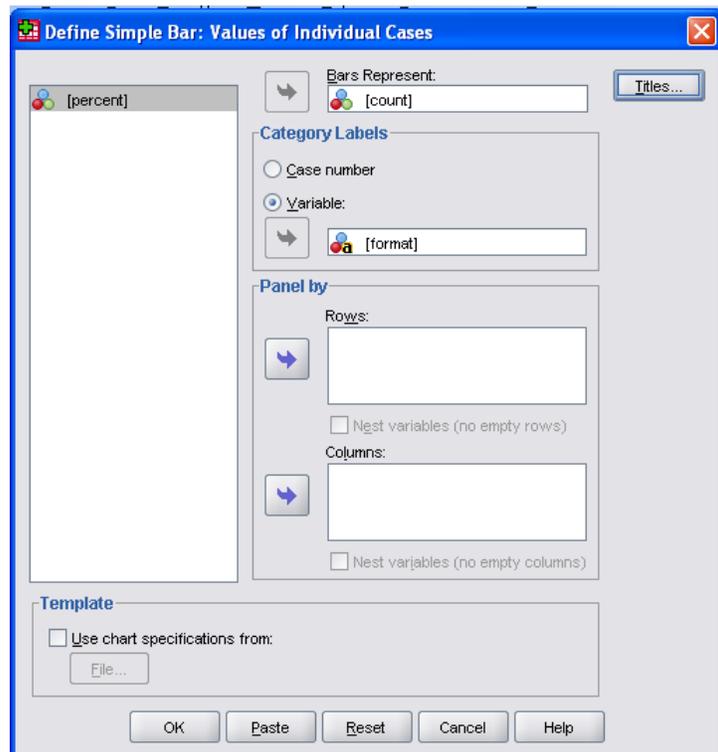
To copy the graph into another program, such as Word, click in the graph and use Ctrl-c to copy it to the Pasteboard and Ctrl-v to paste it into the document. To modify the graphic size or other properties, right-click on the pasted picture and select **Format Picture**.

An alternative way to generate this same graphic is to select **Legacy Dialog, Bar**. The first Dialog box asks what type of bar graph you want.

In this box, I have indicate that I want a simple bar graph. Since the data are already summarized in our table that was entered, I have said that the bars will be created using Values of individual cases. If the data, in this case for example, were in a spreadsheet where each row represented a single radio station and one of the variables was format, I would have selected to create bars as Summaries of separate variables. Press **Define** to continue.



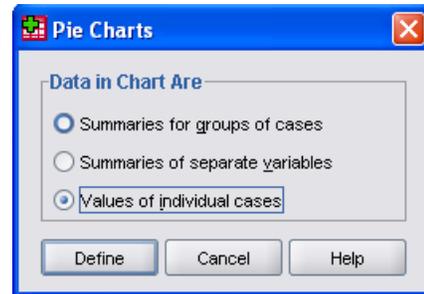
In this dialog box, we define the roles of our variables. We have indicated that the bars represent the variable **count** and the categories are from the variable **format**. Click the **Titles** button to add a title for your graph. Click **OK** to generate the graph into the Output window.



To create a pie chart for the same data, we'll use the percent data, since pie charts always represent the fraction of the whole. These should add up to 100% (to within rounding error).

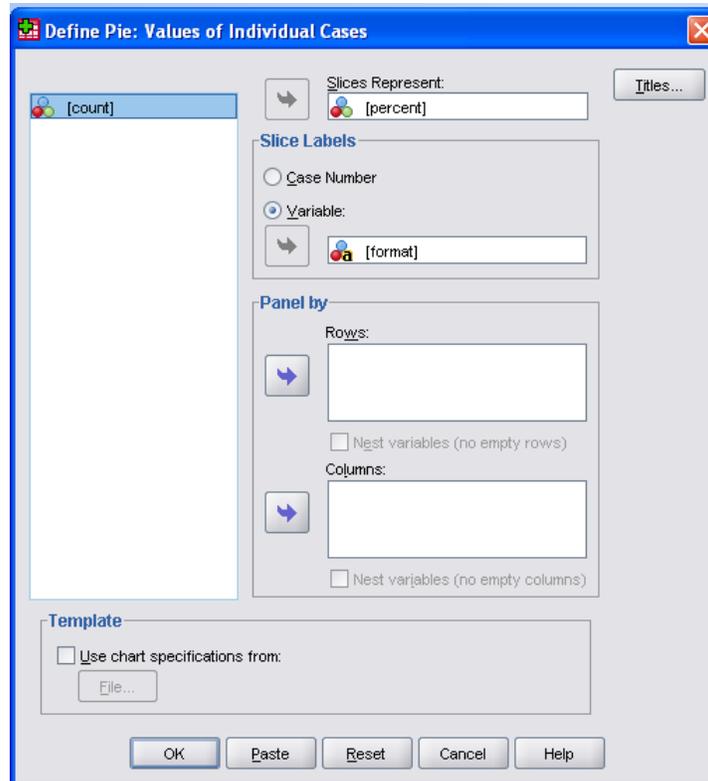
Since our data are already summarized, using **Legacy Dialogs, Pies** is recommended. You will be first asked (similarly to creating a bar graph previously) how SPSS should view the data. Again, since our data are already summarized, we've selected that we will be using Values of individual cases, rather than having SPSS compute the summaries for us from raw data.

Click **Define** to continue.

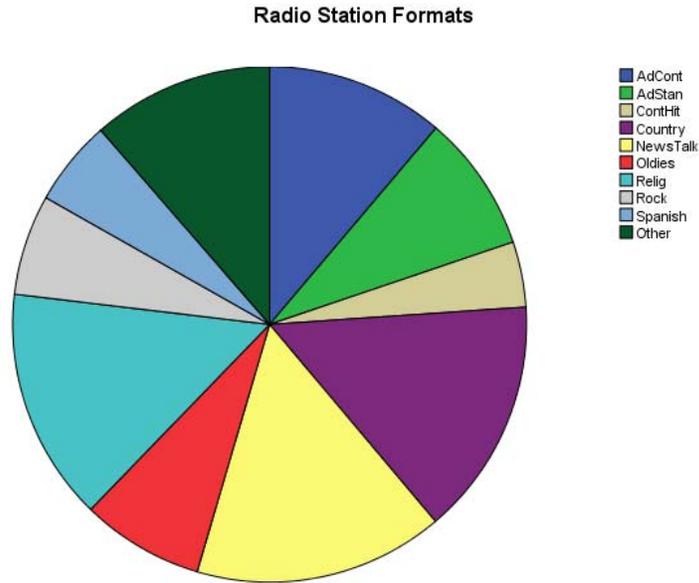


This dialog box should be self-explanatory. Our slice labels are the radio station formats, and the slices represent the percent of all U.S. radio stations with that format. Click the **Titles** button to add an appropriate title for your graph.

Click **OK** to generate the graph.



The graphic on the following page is the default style. The different station formats are represented by different colors. If you want to change the display, there are many options. If you right click in the graphic, you can select to **Edit Content in a separate window**. Here, you can change the fill from different colors to patterns (good for black-and-white printers), add a title if you've forgotten, **add labels** that represent the actual percent in each slice, along with many other options.



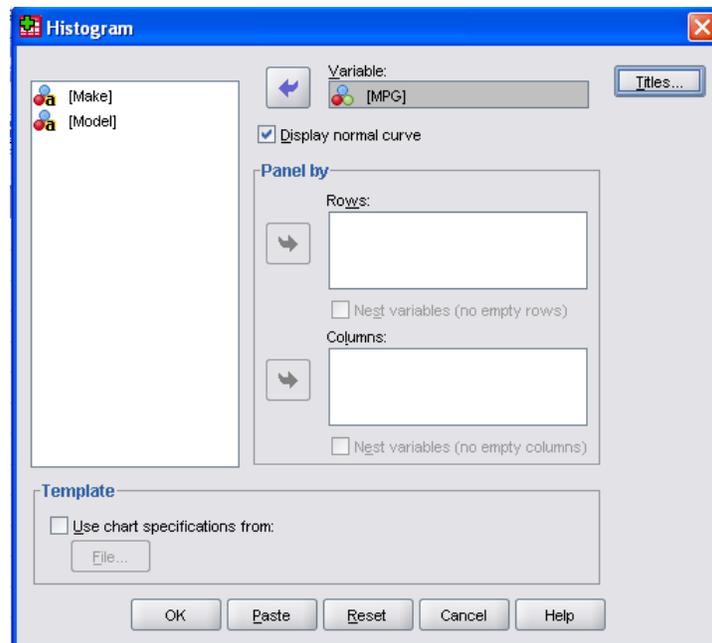
**Example 1.2 Gas Mileage—a Histogram.** Histograms are connected bar graphs for numeric data. Bars in bar graphs are not connected, as their order is arbitrary. The table below gives highway gas mileage for several 2001 model cars. We’d like to make a histogram to represent this data.

Model	MPG	Model	MPG
Acura 3.5 RL	24	Lexus GS300	24
Audi A6 Quattro	24	Lincoln-Mercury LS	25
BMW 740I	23	Lincoln-Mercury Sable	27
Buick Regal	30	Mazda 626	28
Cadillac Catera	24	Mercedes-Benz E320	28
Cadillac Eldorado	27	Mercedes-Benz E430	25
Chevrolet Lumina	29	Mercedes-Benz E55 AMG	24
Chrysler Sebring	30	Mitsubishi Diamante	25
Dodge Stratus	30	Mitsubishi Galant	28
Honda Accord	30	Nissan Maxima	26
Hyundai Sonata	28	Oldsmobile Intrigue	28
Infiniti I30	26	Saab 9-3	28
Infiniti Q45	23	Saturn L100	33
Jaguar S/C	22	Toyota Camry	32
Jaguar Vanden Plus	24	Volkswagen Passat	31
Jaguar XJ8L	24	Volvo S80	26

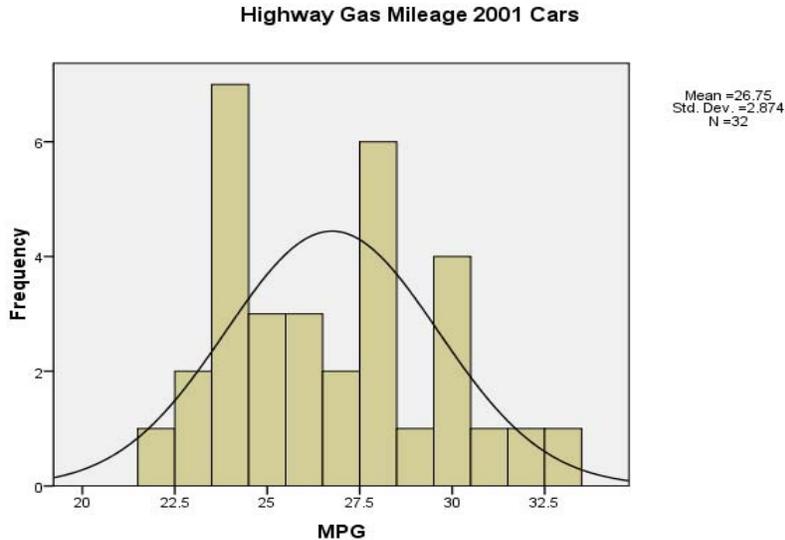
The data have been entered and a portion is shown below. Since we are really interested in the **MPG**, it really is unnecessary to enter the Makes and Models.

	Make	Model	MPG
1	ACURA	3.5RL	24
2	AUDI	A6QUATTRO	24
3	BMW	740I	23
4	BUICK	REGAL	30
5	CADILLAC	CATERA	24
6	CADILLAC	ELDORADO	27
7	CHEVROLET	LUMINA	29
8	CHRYSLER	SEBRING	30
9	DODGE	STRATUS	30

The **Chart Builder** does *not* work for true histograms, since it tries to treat our (numeric) variable as categories. For a histogram, using **Graph, Legacy Dialogs, Histogram** is the easiest to define the plot. **MPG** is the variable for which we want the graph, so move it into the Variable box. Click to check the box **Display normal curve** if desired (this is useful in trying to determine if data are at least approximately normal) and click the **Titles** button to add an appropriate title for your graph. Finally, click **OK** to generate the graph.

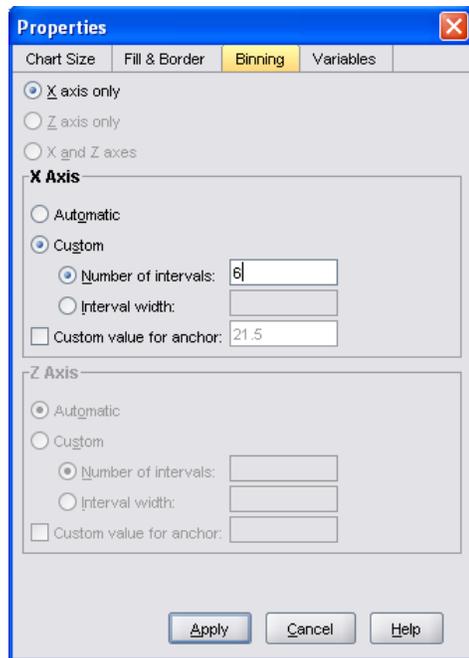


My graph is below. Notice that we also get the sample mean and standard deviation. These data are not approximately normally distributed; the bell curve is not close to smoothing the histogram.

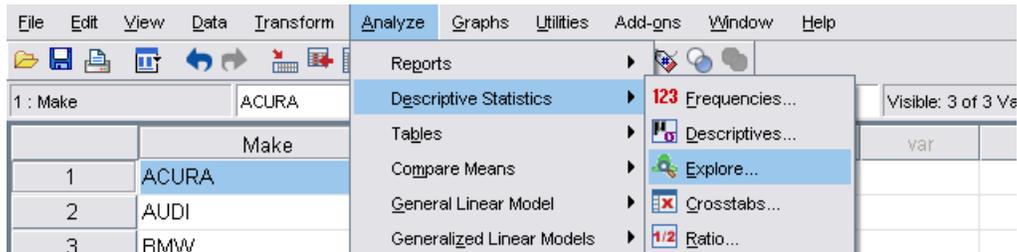


Notice that the intervals used in this graph seem to be 1 mpg each. This should be changed to something more logical (we had data on 32 cars and this histogram has 12 bars, which is too many). Right-click on the graph in the output window and select **Edit Content in separate window**. Right-click on one of the bars and select **Properties Window** from the menu.

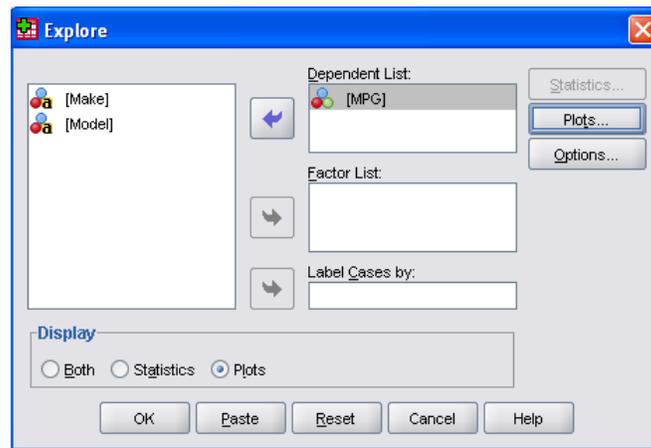
Using this dialog box we can change the binning for the  $x$  axis of our histogram. The button should be on **Automatic**. Click to move it to **Custom**. You have your choice of either specifying the number of intervals, or the width. Here, I have chosen to use six intervals. If desired, you can also specify a minimum  $X$  value to display by clicking to put a check in the box next to **Custom value for anchor** and entering a value in the box beside it. Click **Apply** to change the settings and the graph in the Chart Editor window will change to what has been specified.



**Example 1.3 Gas Mileage—a Stem-and-Leaf Plot.** Histograms show the number of observations in a given bar, but they lose the actual values. Stem-and-leaf plots display not only the shape of a distribution, but they preserve the actual data values (at least to rounding). For the same data on gas mileage used in our histogram, we'd like to create a stem-and-leaf plot. Click on **Analyze** on the menu bar, then **Descriptive Statistics > Explore**.



The window below should open. Click to highlight **MPG** and use the arrow button in the box to move it into the **Dependent List** box. Click to select **Plots** (the default is to display both the plots and summary statistics). Click the **Plots** button at the top right to make sure that **stem-and-leaf** has been selected (the button beside it should be filled in). Click **Continue** to return to the Explore box and **OK** to create the plot.



Stem-and-Leaf Plot

Frequency	Stem &	Leaf
.00	2 .	
3.00	2 .	233
10.00	2 .	4444444555
5.00	2 .	66677
7.00	2 .	8888889
5.00	3 .	00001
2.00	3 .	23

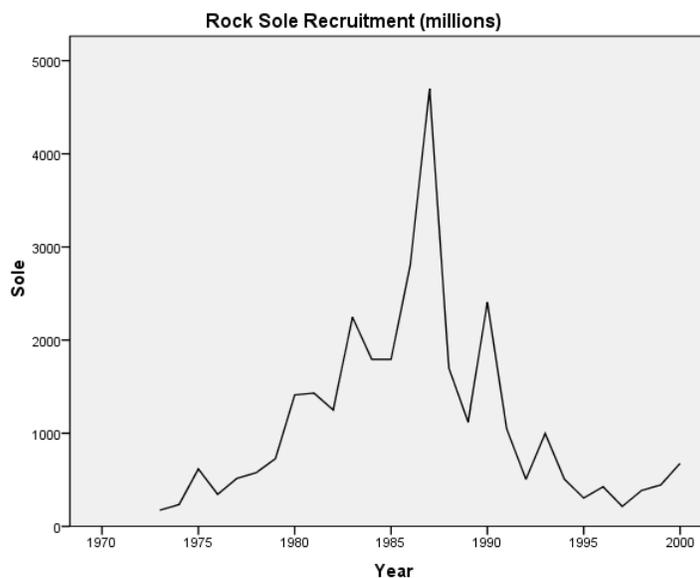
Stem width: 10  
Each leaf: 1 case(s)

The stem-and-leaf plot on the preceding page has had stems split into five lines per stem. Note that the stem value (width) is labeled as 10 implying the leaves represent the one's place in this data set. The Frequency column indicates how many observations are in each line and is given as an aid in locating the median of the distribution.

**Example 1.4 Rock Sole Recruitment—a Time Plot.** Here are data on the recruitment (in millions) of new fish to the Rock Sole population in the Bering Sea between 1973 and 2000. Make a time plot of the recruitment.

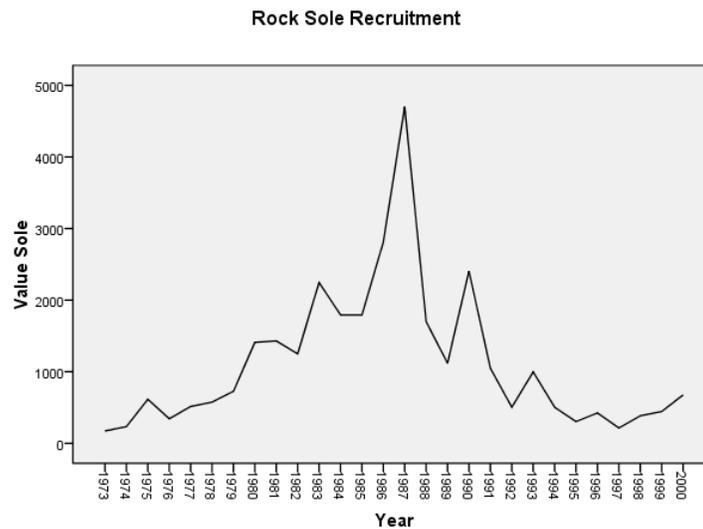
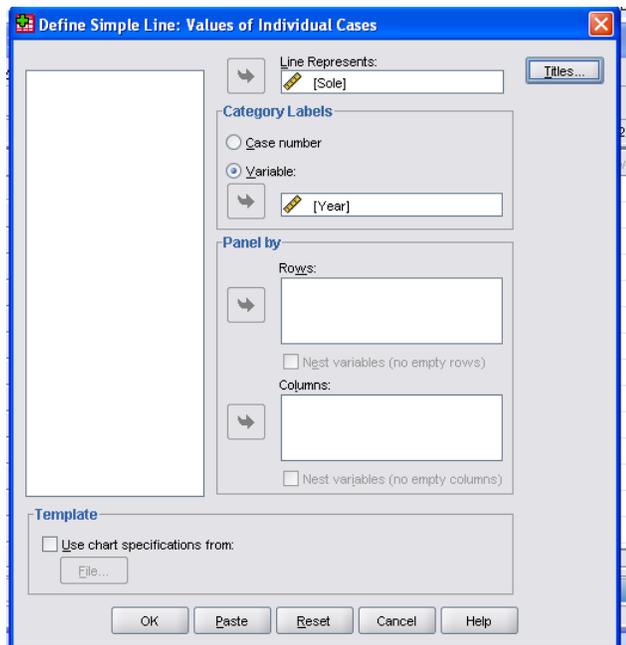
Year	Recruitment	Year	Recruitment	Year	Recruitment	Year	Recruitment
1973	173	1980	1411	1987	4700	1994	505
1974	234	1981	1431	1988	1702	1995	304
1975	616	1982	1250	1989	1119	1996	425
1976	344	1983	2246	1990	2407	1997	214
1977	515	1984	1793	1991	1049	1998	385
1978	576	1985	1793	1992	505	1999	445
1979	727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

With the data entered into two columns, we select **Graphs** > **Chart Builder**. We will first create a scatterplot of the data and add the connecting lines using the Chart Editor. Select the first (simple) line plot and drag it into the window. Drag **Year** to the  $x$  axis and **Sole** to the  $y$  axis. Using the Element Properties Box, define axis labels for each variable in the Element Properties Box. Click **Apply** for each label definition. Click on the **Titles** tab, give your graph a title, and click **Apply**. Close the Element Properties Box and click **OK** to display the graph.



This can also be easily done with the **Line** option under **Legacy Dialogs**. Select the **Simple** type and that Data are **Values of individual cases**.

My definition is at right. Notice that our Year is treated as a category variable. Be sure to add a plot title by clicking the button at upper right. When finished, click **OK** to display the plot.



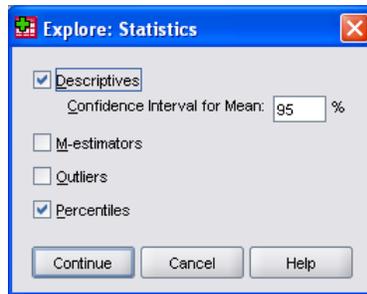
This plot is just a little different from the one created using the Chart Builder. Each year is labeled.

## 1.2 Describing Distributions with Numbers

Your text describes many summary statistics that can be calculated from a set of data. Among these are the mean, median, range, IQR, and standard deviation. There are others as well.

**Example 1.5 Gas Mileage Descriptive Statistics.** Recall our data used in Example 1.2 (page 19) on highway gas mileage for some 2001 cars. The histogram we created also told us the mean and standard deviation of the data. However, since these data were not very symmetric, these summary statistics are probably not the best to describe these data. Perhaps the five-number summary (minimum,  $Q_1$ , median,  $Q_3$ , maximum) would be a better descriptor.

Click **Analyze, Descriptive Statistics, Explore**. This was used before in Example 1.3 (page 22) for the stem-and-leaf plot of these data. Click to highlight **MPG** and click the right arrow to move it into the **Dependent List**. Click the **Statistics** button.



Since we want to see the quartiles, click the box next to **Percentiles** to check mark it. If not checked, the IQR will be shown on the output, but not the actual quartiles. Click **Continue** to return to the Explore box and then **OK**.

The first part of the output is common to all SPSS statistics computations. It merely tells us the number and percent of valid observations and the number of missing values (observations without data).

**Case Processing Summary**

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
	32	100.0%	0	.0%	32	100.0%

The next part of the output gives most of the statistics we are interested in.

Descriptives		Statistic	Std. Error
Mean		26.75	.508
95% Confidence Interval for Mean	Lower Bound	25.71	
	Upper Bound	27.79	
5% Trimmed Mean		26.67	
Median		26.50	
Variance		8.258	
Std. Deviation		2.874	
Minimum		22	
Maximum		33	
Range		11	
Interquartile Range		5	
Skewness		.363	.414
Kurtosis		-.769	.809

We see three of the five values needed: the median highway gas mileage for these cars is 26.50, with minimum 22 and maximum 33. The interquartile range is given as 5 but we don't know where it is located. Also given are the mean, standard deviation ( $s$ ), variance ( $s^2$ ), and the range (max – min).

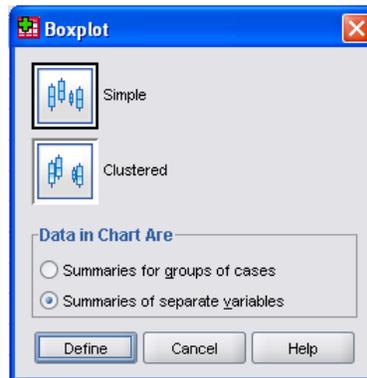
Some of these statistics are most likely unfamiliar to you. The 5% Trimmed mean calculates the mean after the top and bottom 5% have been removed. It is an attempt to make a mean more resistant to outliers (since they are, by definition, at the extremes). Skewness is a measure of how skewed a set of data is. Values near 0 indicate that a data set is relatively symmetric. Kurtosis is a measure of how flat (or peaked) a set of data is. If you look at the histogram for these data on the first page of this chapter (or the stem-and-leaf plot on page 22) we can see that the center of this distribution is lower than expected for a “bell-shaped” curve. That is why this is negative.

The next box of output gives percentiles of the distribution. Tukey's Hinges (24 and 28.5) are the quartiles as described in your text (median of each half of the data). Notice the difference of these is not the same as the IQR given above. This is partly because statisticians have yet to come to full agreement on how quartiles should be found.

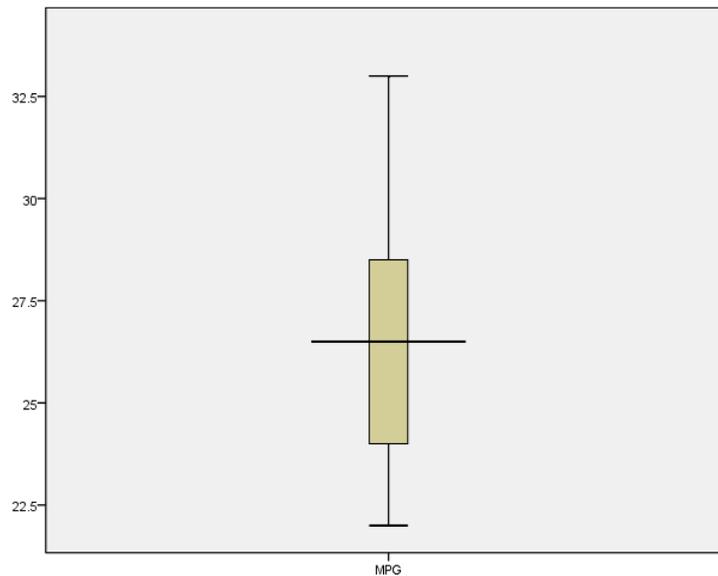
	Percentiles						
	5	10	25	50	75	90	95
Weighted Average(Definition 1)	22.65	23.30	24.00	26.50	28.75	30.70	32.35
Tukey's Hinges			24.00	26.50	28.50		

**Example 1.6 Gas Mileage—a Boxplot.** Boxplots are yet another way to picture a distribution of data. They have several advantages over stem-and-leaf plots and histograms: they are objective, being based on the five-number summary, and add an objective criterion for whether or not an observation is an outlier. They are also good for visually comparing two or more distributions. Since they are based solely on the five-number summary, however, they cannot identify such potential features in a distribution as bimodality (having two peaks). Since time involved in generating plots is minimal, it's always a good idea to look at several different plots of a distribution.

To create a boxplot, click **Graphs**, **Legacy Dialogs**, **Boxplots**. (The Chart Builder is unreliable in identifying outliers.)



**Simple** should be selected as the default. Click that we are using **Summaries of separate variables** (even though we have only one). Then select **Define**. Drag **MPG** into **Boxes Represent** and say **OK**. You should see the plot below. Notice that there are no outliers in this distribution; if there were, they would be identified with circles above (or below) the “whiskers.”



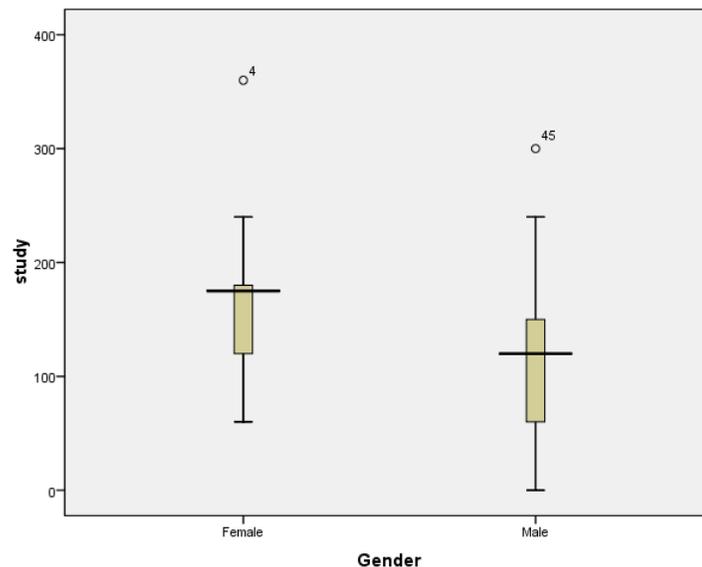
**Example 1.7 Study Time–Side-by-Side Boxplots.** As mentioned above, boxplots are a good way to visually compare two distributions. Data presented on the next page are reported nightly study time claimed by samples of first-year college men and women. We will make side-by-side boxplots to compare these distributions. We will also compute summary statistics to compare the two distributions.

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

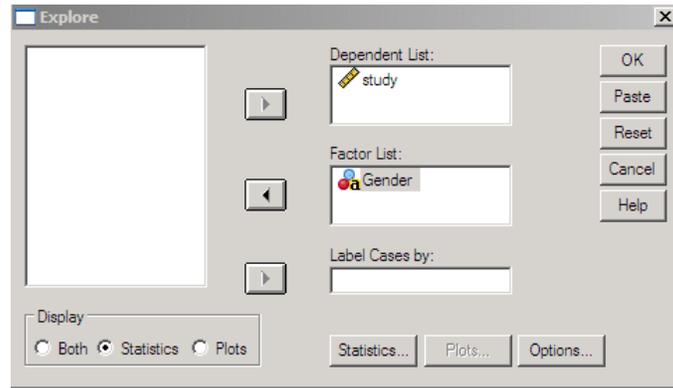
Data have been entered into SPSS with a column for **Gender** and one for **study**. Value labels were defined (see Chapter 0) so that F (for Female) and M will display the whole word on our plot.

	Gender	study
1	F	180
2	F	120
3	F	180
4	F	360
5	F	240
6	F	120

Define the boxplot similarly to Example 1.6 earlier, but in this case we want to select **Summaries for Groups of Cases** in the first dialog box. Drag **Gender** to the **Category axis** and **study** to the **Variable** box. **OK** will generate the graph below. We see one high outlier in each distribution. These are identified with their case (row) numbers so that you can locate them easily in the data window.



To compute the summary statistics, use **Analyze, Descriptive Statistics, Explore**. Since we have one column for gender and one for the study time, we enter **study** into the **Dependent List** (click to highlight it then click the right arrow) and **Gender** into the **Factor List**. If we had had two separate lists for Males and Females, these would have both been entered into the **Dependent List** box. Click the button to ask for statistics. Press **OK** to obtain the results.



	Gender		Statistic	Std. Error	
study	Female	Mean	165.1667	10.31817	
		95% Confidence Interval for Mean	Lower Bound		144.0636
			Upper Bound		186.2697
		5% Trimmed Mean			161.8519
		Median			175.0000
		Variance			3193.937
		Std. Deviation			56.51493
		Minimum			60.00
	Maximum		360.00		
	Range		300.00		
	Interquartile Range		60.00		
	Skewness		1.301	.427	
	Kurtosis		3.844	.833	
	Male	Mean	117.1667	13.55424	
		95% Confidence Interval for Mean	Lower Bound		89.4451
			Upper Bound		144.8882
5% Trimmed Mean			114.0741		
Median			120.0000		
Variance			5511.523		
Std. Deviation			74.23963		
Minimum			.00		
Maximum		300.00			
Range		300.00			
Interquartile Range		97.50			
Skewness		.666	.427		
Kurtosis		-.083	.833		

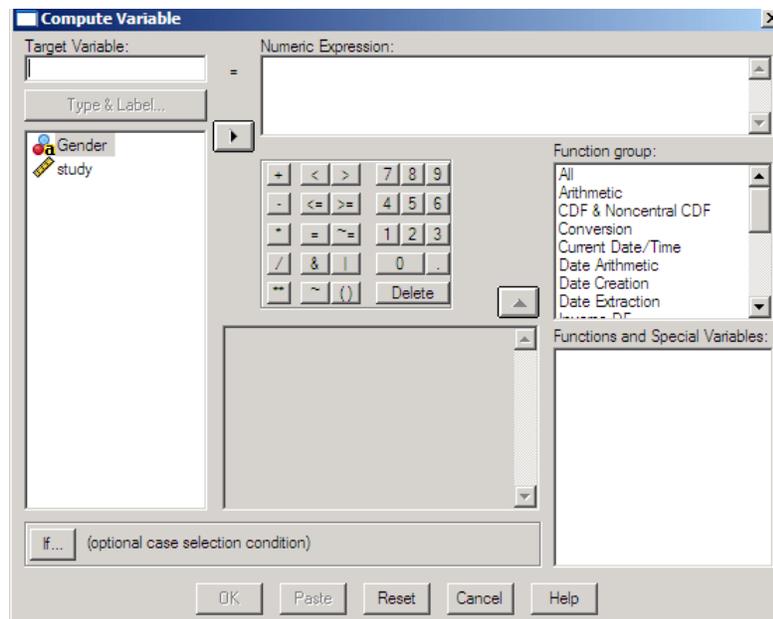
In the table of descriptive statistics, Females appear first. Their mean is almost 50 minutes higher than the Males (165.2 minutes per night for studying compared to 117.2); Males' study time is also more variable with a standard deviation of 74.3 compared to 56.5 for Females.

### 1.3 Normal Distributions

We have already seen that SPSS can impose a Normal “bell curve” onto a histogram of data. This is an aid in determining if a set of data might have come from a Normal distribution. Most Normal calculations are actually easier to do with a calculator or table, but SPSS has a built-in calculator that can be used as well.

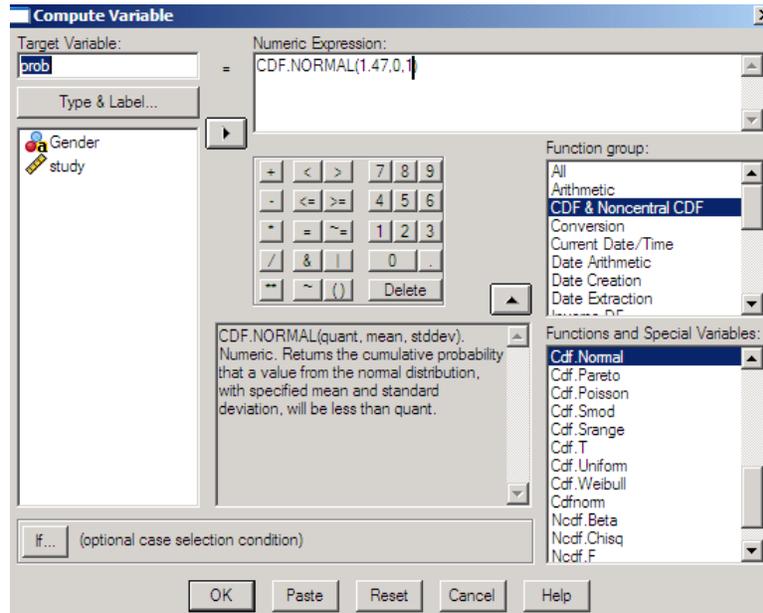
**Example 1.8—Finding the Area in a Standard Normal Curve** What proportion of observations for a standard Normal distribution are less than 1.47?

We will use **Transform, Compute Variable** to answer this question. The window below should open.



In the **Function group** window, click to highlight the **CDF and Noncentral CDF** option. CDF (cumulative distribution functions) for many different variable types will appear in the lower box. Scroll down to find **Cdf.Normal**. Double-click it and more information on how to use the command will appear in the central gray box, as well as a command shell in the Numeric Expression box at the top. Define a new variable to receive the results of the calculation in the Target Variable box at top left. In the screen below, the results of our calculation will be stored in the variable **prob**. We have indicated that we want area less than 1.47 with mean 0 and standard deviation 1. **OK** will perform the computation. There is also a Cdfnorm option that could be used here,

since we were working with a standard Normal distribution, but this option is more general, as most problems of this type are *not* standard Normal.



The new variable has been added into the worksheet. We can see the value is 0.93 (or for more accuracy, 0.929219123008314). Usually convention is to report these to four decimal places, the same number of significant digits as are in the table, so we'd report that the proportion of observations in a standard Normal distribution less than  $z = 1.47$  is 0.9292.

45 : prob	0.929219123008314		
	Gender	study	prob
1	F	180.00	.93
2	F	120.00	.93

The Normal calculations in SPSS work much like the tables—that is, one can only obtain probabilities below a desired value. If we wanted to know, for example, what proportion of observations in a standard Normal distribution are above 1.47, we would have done the same calculation as we just did. We subtract the result from 1 and obtain  $1 - 0.9292 = 0.0708$ .

**Example 1.9 Women's Heights – Finding Area Between Two Values.** American women have heights that are normally distributed with mean 64 inches and standard deviation 2.7 inches. What proportion of American women are between 63 and 66 inches tall?

Completion of this question requires two of the normal calculations as done in the previous example. We will find the proportion having heights less than 63 inches and subtract that from the proportion having heights less than 66 inches.

Target Variable: prob	=	Numeric Expression: CDF.NORMAL(63,64,2.7)	gives	0.355553273797402	and
Target Variable: prob	=	Numeric Expression: CDF.NORMAL(66,64,2.7)	gives	0.770574674035266	

Don't worry about intermediate messages that the variable **prob** is being redefined. Since this is just a place-holder for our results, it can be safely over-written. To four decimal places, the desired probability is  $.7706 - .3556 = .4150$ . About 41.5% of American women are between 63" (5'3") and 66" (5'6") tall.

**Example 1.10 Women's Heights—Finding Percentiles.** Given the above distribution of heights of American women, how tall are the tallest 10%?

We'll use the same **Transform, Compute Variables** dialog as above, but a different **Function Group**. Here we want the Function Group **Inverse DF**. Click to highlight it, and similarly to the CDF group, inverse possibilities for many different distribution types will be displayed in the lower box. Locate and double-click **Idf.Normal**. Just as with the cdf, some information about the function will appear in the central box and the command shell is transferred into the **Numeric Expression** box. It will paste at the beginning of the box, so you may need to clear any remnants of an old expression before continuing.

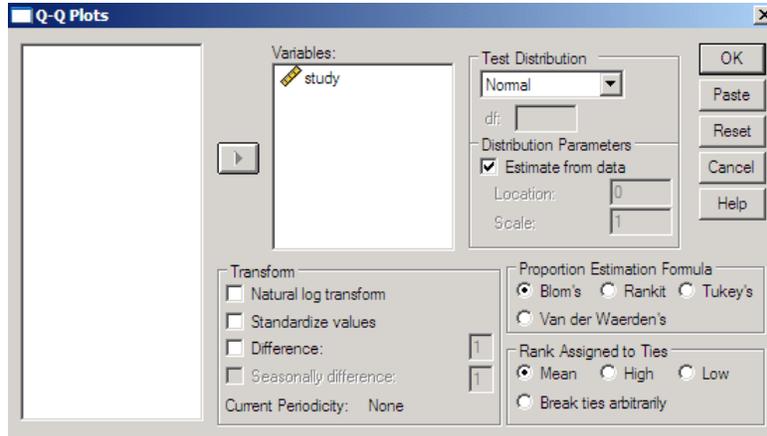
My expression and answer look like this:

Target Variable: height	=	Numeric Expression: IDF.NORMAL(.9,64,2.7)	1 : height	67.4601892269704
----------------------------	---	--	------------	------------------

Notice that the expression was entered using .9 and not .1. We were interested in the *tallest* 10% of American women. The command shell works with the accumulated area to the left of the point specified. Finding the tallest 10%, that means 90% are that tall or shorter. The tallest 10% of American women are at least 67.46 inches tall (about 5'7.5").

**Example 1.11 Study Time—Are the Data Normal?** In Example 1.7 (page 28) we looked at data on study times for male and female freshmen students. Both genders had high outliers. If we considered combining these as representing freshmen students in general, might these be considered as having come from a Normal distribution?

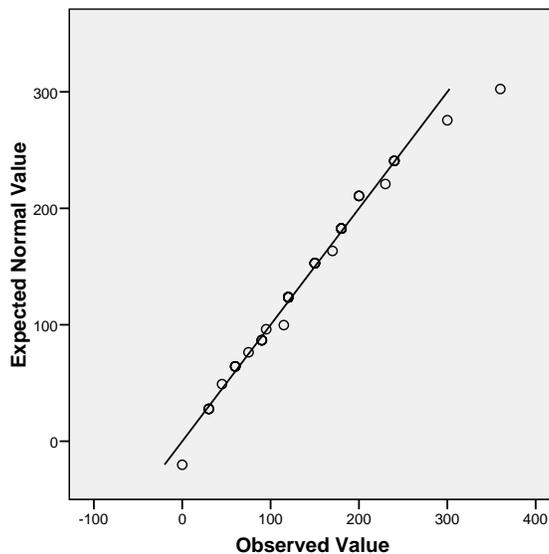
We use **Analyze, Descriptive Statistics, Q-Q Plots** to create a normal quantile plot of the entire distribution of study times.



Variable **study** has been entered into the **Variables** box and the **Test Distribution** is Normal (the default). Click **OK** to generate the plot of interest (and a lot of other stuff).

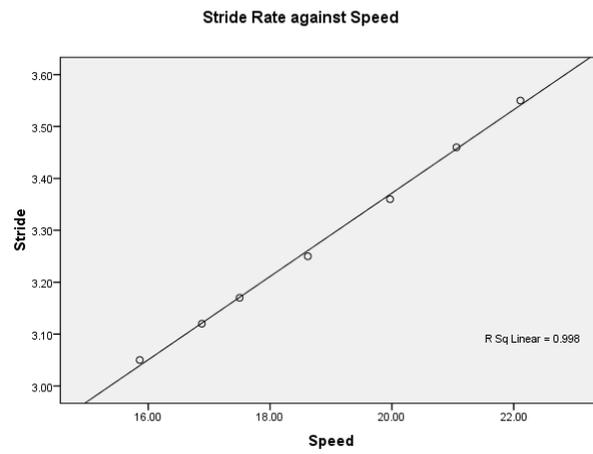
The plot of interest is below. On plots like these, we look for whether or not the plot is approximately a straight line. To help us, SPSS adds a straight line. Note that, except for the two high outliers that we already knew about, this plot is reasonably along the line. We can believe that freshmen study time is an approximately Normal random variable, based on these data.

Normal Q-Q Plot of study



## CHAPTER

# 2



# Looking at Data— Exploring Relationships

	2.1	Scatterplots
	2.2	Correlation
	2.3	Least-Squares Regression
	2.4	Cautions about Correlation and Regression
	2.5	Relations for Categorical Variables

## Introduction

In this chapter, we use SPSS to graph the relationship between two quantitative variables using a scatterplot. We then show how to compute the correlation and find the least-squares regression line through the data. Lastly, we show how to work with the residuals of the regression line. We will also examine stacked bar graphs—a way to visualize relationships in categorical variables.

## 2.1 Scatterplots

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

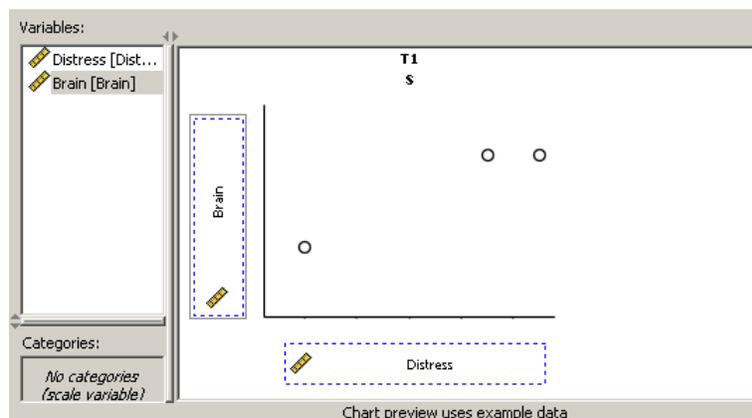
**Example 2.1 Brain Activity and Stress.** It has been suggested that emotional stress leads to increased brain activity. Make a scatterplot of brain activity level against social distress score.

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

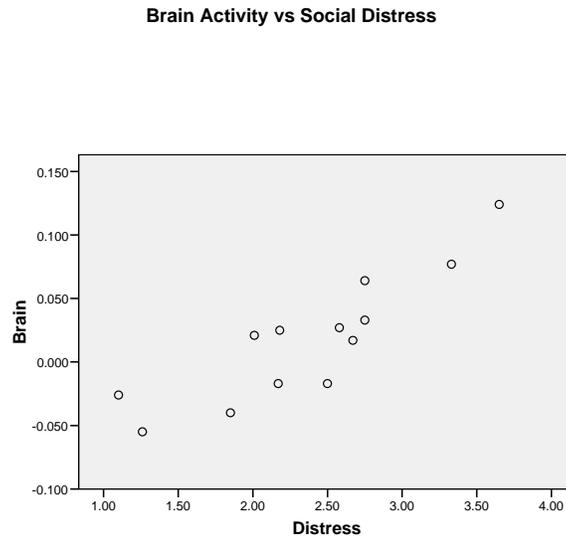
*Solution.* We entered the data into SPSS after defining the variables **Distress** and **Brain**. Since the Brain activity values have three decimal places, be sure to change that on the Variable View screen, since the SPSS default is two decimal places. The first few values are shown below.

	Distress	Brain	
1	1.26	-.055	
2	1.85	-.040	
3	1.10	-.026	
4	2.50	-.017	
5	2.17	-.017	

This plot is one that the **Chart Builder** in the **Graphs** menu is good for. Select the **Simple Scatter** from the available plot types and drag it to the preview window. Define axis labels in the Property dialog boxes as described in Chapter 1 of this manual, and click the **Titles** tab to give the plot a meaningful title. Be sure to **Apply** each attribute. The preview window should look like the one below.



Click OK to generate the graph into the output window.



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

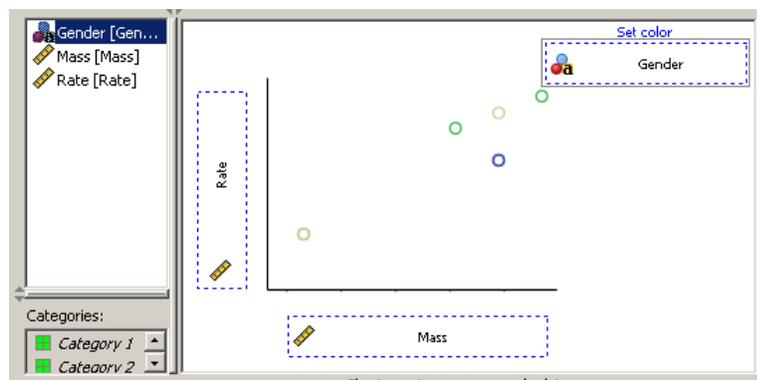
**Example 2.2 Body Mass and Gender.** We would like to examine whether or not there seems to be a difference in metabolic rate versus body mass for males and females. We will use a grouped scatterplot with different symbols for males and females, and display both at once.

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

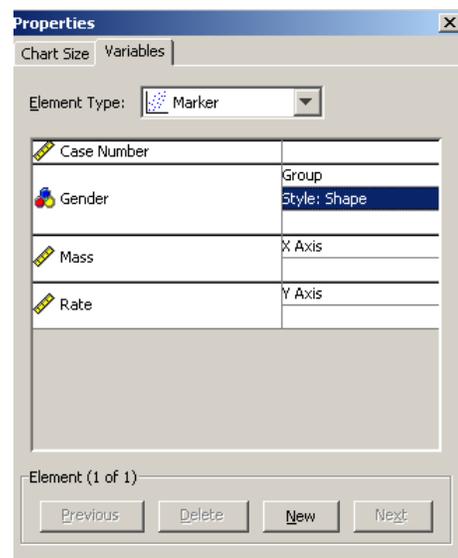
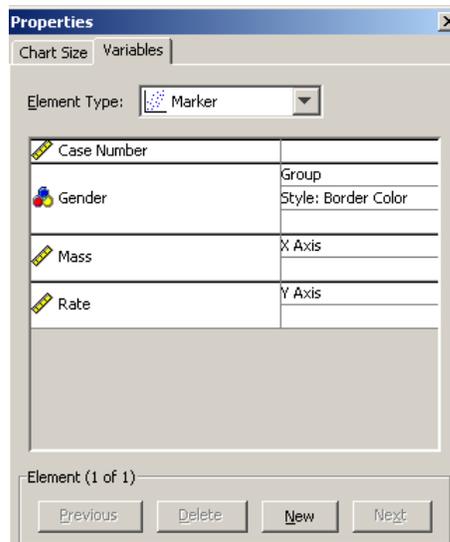
*Solution.* We first entered the data for the males and then the females after having defined Gender as a string variable. A portion of the data worksheet is shown below.

	Gender	Mass	Rate
1	M	62.00	1792.00
2	M	62.90	1666.00
3	M	47.40	1362.00
4	M	48.70	1614.00
5	M	51.90	1460.00
6	M	51.90	1867.00
7	M	46.90	1439.00
8	F	40.30	1189.00
9	F	33.10	913.00

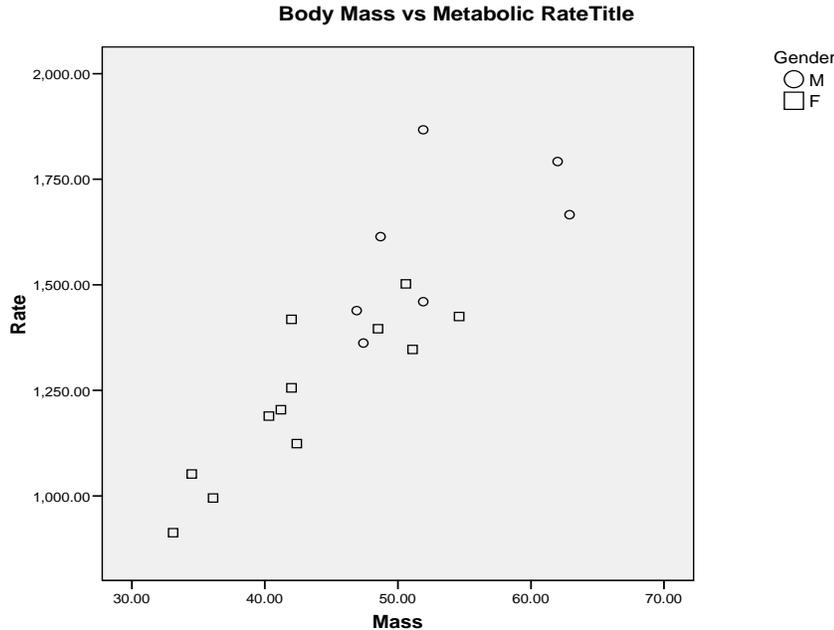
Plot definition is similar to the one in the above example; drag Gender to the grouping variable in the plot preview window. Click **OK** to generate the graph.



The initial plot uses color to distinguish the males from the females. Males are shown as blue circles in the upper right of the graph, and females are green circles that tend to be at the lower left. This might be an acceptable graph if you have a color printer; if not, we need to change the plotting symbols to something that can be distinguished in black and white. Double-click one of the data point circles in the graph. This will start the Chart Editor and bring up a Variables Properties window, as at the left below.



Click on **Style: Border Color** for **Gender**. A subwindow will appear with possible style options. Select **Style: Border Shape** and **Apply**. **Close** the Properties box. Now we have different shapes to mark the data points for the different genders. My final graph is below.



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

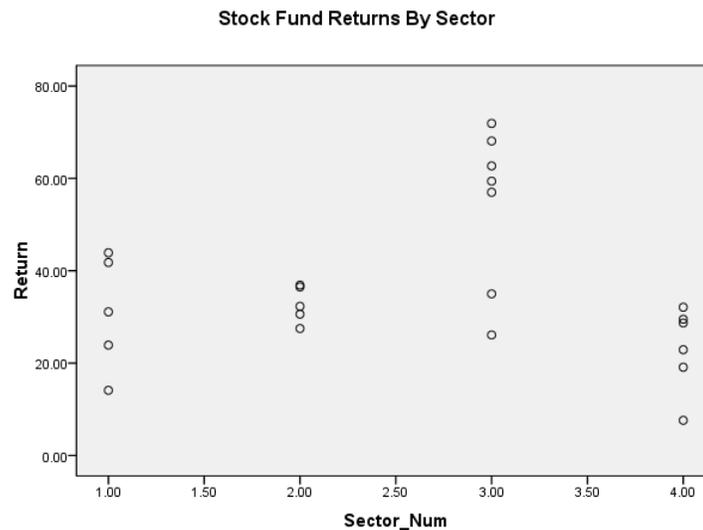
**Example 2.3 Mutual Fund Returns.** How does the return on investment vary among sector mutual funds? Data in the table below are annual total returns from several sector funds. We will make a plot of the total return against market sector. We’ll also compute the mean return for each sector, add the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

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

*Solution.* We have entered the data into two columns: one for the actual sector name, and another for the returns. We will plot the market sectors on the x axis as the values 1, 2, 3, and 4. Because there are multiple returns for each sector, we create a new variable called Sector\_Num and enter each of the values 1 through 4 as many times into a new column as there are returns for that sector.

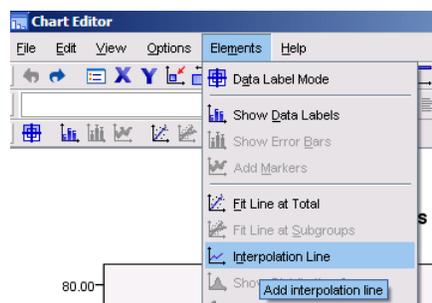
	Sector	Return	Sector_Num
1	Consumer	23.90	1.00
2	Consumer	14.10	1.00
3	Consumer	41.80	1.00
4	Consumer	43.90	1.00
5	Consumer	31.10	1.00
6	Finance	32.30	2.00
7	Finance	36.50	2.00

Then we define a **Simple Scatterplot** with **Sector\_Num** on the  $x$ -axis and **Return** on the  $y$ -axis. Be sure to give the plot a title. Click OK at the bottom of the dialog box to generate a plot like the one below.

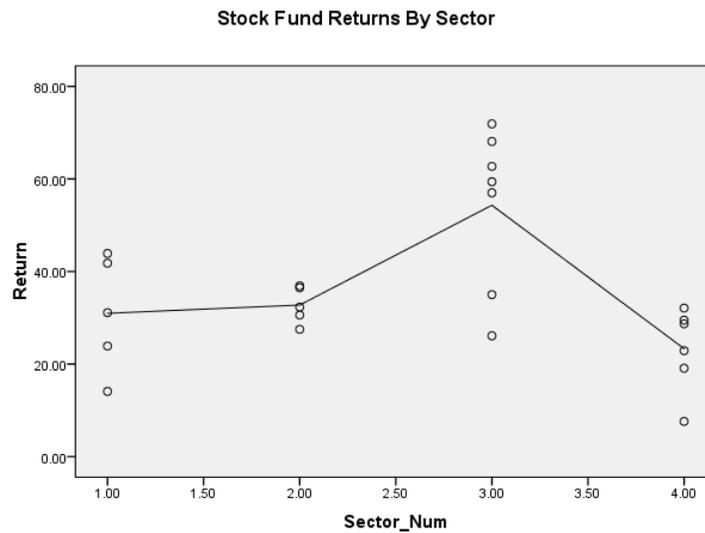


So far, we see that the returns in sector 3 (Technology) are much more variable than the others—there is potential for greater return, but also less. Natural Resources seems to have the lowest returns and Financial Services the most consistent (least variability).

Adding a line to interpolate the means is easy. Click in the graph and select **Edit Content in Separate Window**. Then click **Elements, Interpolation line**.



The interpolation line for the means will be added into the graph, as seen on the next page.



## 2.2 Correlation

In this section, we compute the correlation coefficient  $r$  between paired quantitative variables.

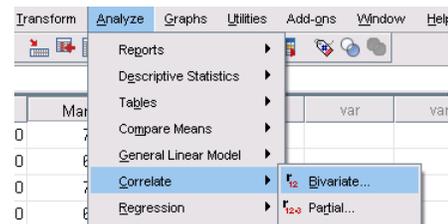
**Example 2.4 Dates' Heights.** The table below gives the heights in inches for a sample of women and the last men whom they dated. (a) Make a scatterplot. (b) Compute the correlation coefficient  $r$  between the heights of these men and women. (c) How would  $r$  change if all the men were six inches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

<b>Women (<math>x</math>)</b>	66	64	66	65	70	65
<b>Men (<math>y</math>)</b>	72	68	70	68	71	65

*Solution.* (a) We define two variables called Woman and Man and enter the heights.

	Woman	Man
1	66.00	72.00
2	64.00	68.00
3	66.00	70.00
4	65.00	68.00
5	70.00	71.00
6	65.00	65.00

(b) To simply compute the correlation, we use **Analyze, Correlate, Bivariate**. Click in the dialog box to add the variables, then **OK**.



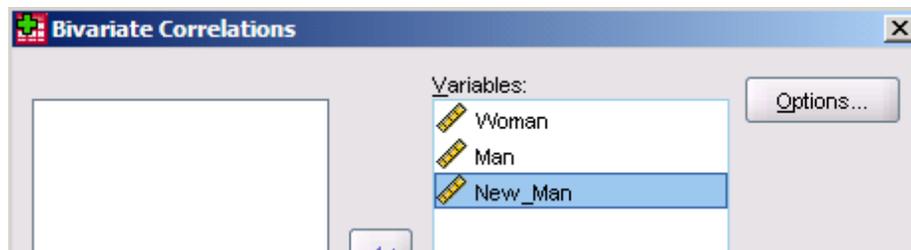
We actually obtain a correlation matrix. We see that the correlation between the men's and women's heights is 0.565. SPSS also gives a  $p$ -value for a test of whether or not this correlation is statistically significant. We'll talk about tests of hypotheses later.

		Woman	Man
Woman	Pearson Correlation	1	.565
	Sig. (2-tailed)		.242
	N	6	6
Man	Pearson Correlation	.565	1
	Sig. (2-tailed)	.242	
	N	6	6

(c) In order to get heights for all the males that are six inches shorter than they originally were, we could manually subtract 6 from each entry in **Man** and store the result into a new variable, or we could define a new variable and do the computations using **Transform, Compute Variable**. Below, I define a new variable called **New\_Man** which is to be calculated as **Man-6**.



Click **OK** to perform the calculation. Return to **Analyze, Correlations, Bivariate** and add **New\_Man** into the box, then click OK to recalculate the correlations.



**Correlations**

		Woman	Man	New_Man
Woman	Pearson Correlation	1	.565	.565
	Sig. (2-tailed)		.242	.242
	N	6	6	6
Man	Pearson Correlation	.565	1	1.000**
	Sig. (2-tailed)	.242		.000
	N	6	6	6
New_Man	Pearson Correlation	.565	1.000**	1
	Sig. (2-tailed)	.242	.000	
	N	6	6	6

\*\* . Correlation is significant at the 0.01 level (2-tailed).

We see that subtracting six inches from each man's height does not change the correlation with the woman's height; in fact, the men's heights have a correlation of 1 with the new height. Since each new height is exactly six inches less than the original, we have a perfect straight line relationship.

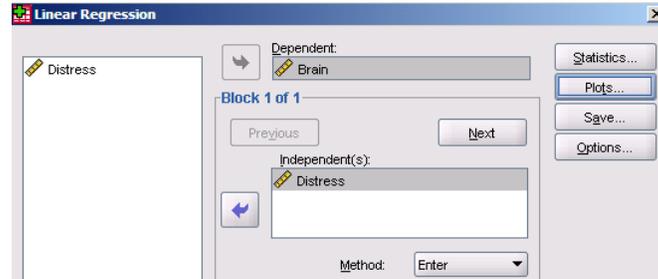
### 2.3 Least-Squares Regression

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

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

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

*Solution.* (a) Click **Analyze, Regression, Linear**. All we need to do here is click to place Brain into the Dependent box, and Distress in the Independent box, then click **OK**.



There are actually four parts to the output. The first (that tells us that Distress was entered into the model) can be ignored. The third can also be ignored for now. To find the equation of the regression line, we are interested in the last table.

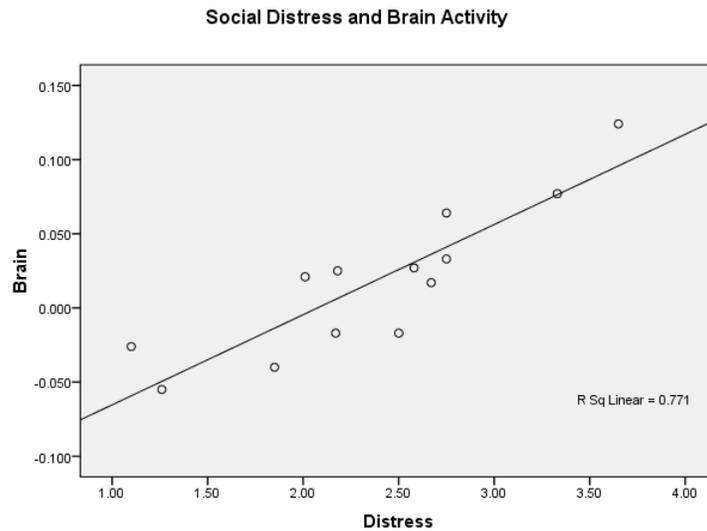
**Coefficients<sup>a</sup>**

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-.126	.025		-5.116	.000
	Distress	.061	.010	.878	6.091	.000

a. Dependent Variable: Brain

The coefficients we want are in the first numeric column, labeled B. Reading this column, we see we have the equation Brain Activity =  $-0.126 + 0.061 \cdot \text{Social-Distress}$ . Notice at the bottom of the table we're even reminded that our dependent (y) variable is Brain.

We could have added the regression line to the plot when we first created it. Recreate a **Simple Scatterplot** of the data, then click in the plot and select **Edit Contents in Separate Window**. In the Chart Editor, select **Elements, Fit Line at Total**. A Properties dialog box will pop up, and the linear regression line (the default) will be added into your graph. Click Close and then close the Chart Editor. The coefficient of determination ( $r^2$ ) is also now included in the graph.



(b) Visually, from our plot above, it appears that a distress rating of 2 corresponds to brain activity of about -0.01. Using the equation, we have  
 Brain Activity =  $-.126 + .061 * 2 = -.004$ .

(c) Adding the fit line to the graph has already told us that the amount of brain activity explained by social distress is  $r^2 = 77.1\%$ . This is also found in the second table of the regression output, along with the correlation coefficient,  $r$ , which is 0.878, indicating strong relationship. Ignore the Adjusted R Square column; it only pertains in multiple regression (where there are several predictor variables).

**Model Summary**

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.878 <sup>a</sup>	.771	.751	.025090

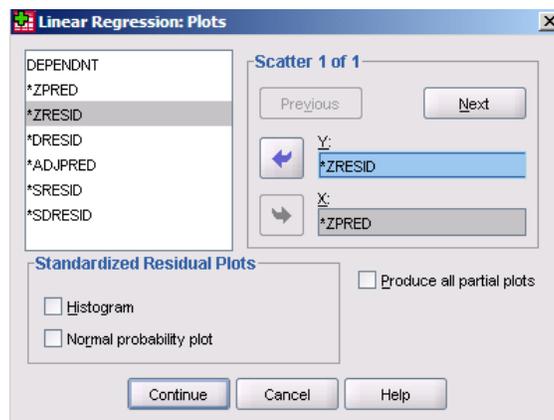
a. Predictors: (Constant), Distress

## 2.4 Cautions about Correlation and Regression

We now complete an exercise to demonstrate how to work with the residuals of a least-squares regression line. Ideally, these will be randomly distributed around the  $x$  axis ( $y = 0$  line). Any indications of pattern (curvature, widening [or narrowing] as  $x$ -values increase) indicates a violation of assumptions for the regression. They can also be used to help identify outliers and potentially influential points in a regression. Residuals should also have a Normal distribution. SPSS can easily create the necessary plots to examine these properties.

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

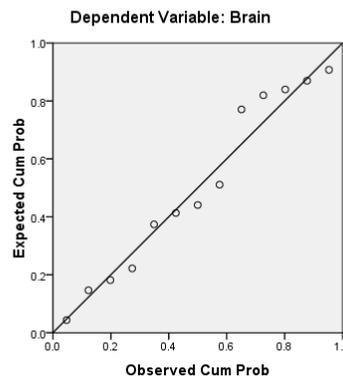
*Solution.* We could have done this when finding our original regression line. Return to **Analyze, Regression, Linear**. Click the **Plots** button at the upper right of the dialog box. You can define up to two scatterplots that involve residuals. In addition, you can ask for a histogram of residuals or a Normal plot. Here I ask for a plot of standardized residuals (their  $z$ -scores) against standardized predicted values (the fitted values from the regression equation for each  $x$  value). (You can alternately save actual residuals and predicted values using **Save**.)



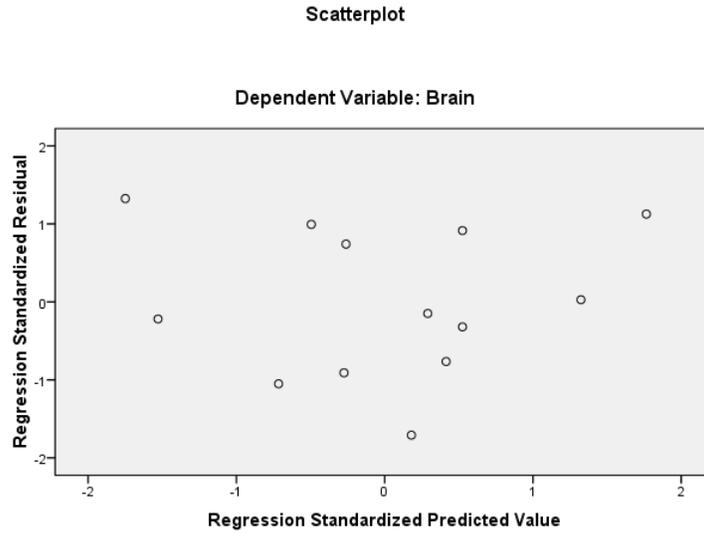
Click **Continue** then **OK** to perform the regression and generate the plots.

First we see the Normal plot. Recall from Chapter 1 that if the data have an approximately Normal Distribution, this plot should resemble a straight line. SPSS adds a line in for reference. Since all the points follow the line, we can say the residuals are approximately Normally distributed.

Normal P-P Plot of Regression Standardized Residual



The next plot is the (standardized) residuals against fitted values. Ideally, this plot shows random scatter in a relatively even band around  $y = 0$  (since the mean of the residuals must be 0). The plot below is pretty much ideal. There is no overt pattern, and the residuals are relatively evenly spaced around  $y = 0$ .



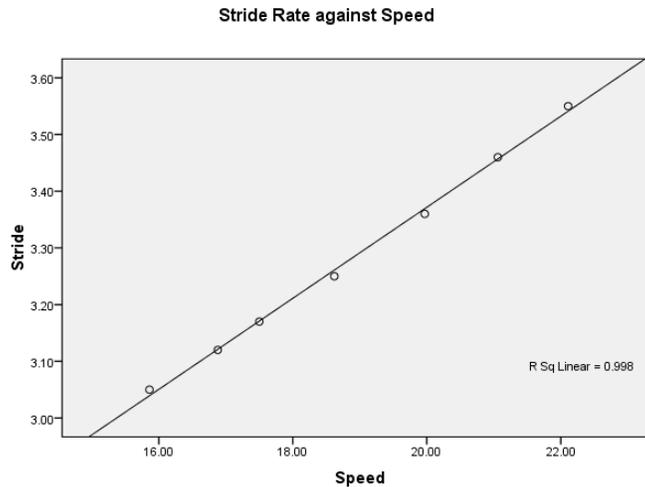
**Example 2.7 Runners’ Stride Rates.** The following table gives the speeds (in feet per second) and the mean stride rates for some of the best female American runners. We’ll perform a complete analysis of these data, attempting to predict stride rate ( $y$ ) using the speed ( $x$ ) of the runner.

<b>Speed</b>	15.86	16.88	17.50	18.62	19.97	21.06	22.11
<b>Stride Rate</b>	3.05	3.12	3.17	3.25	3.36	3.46	3.55

*Solution:* First, enter the data into two variables columns. I have named my variables **Speed** and **Stride**.

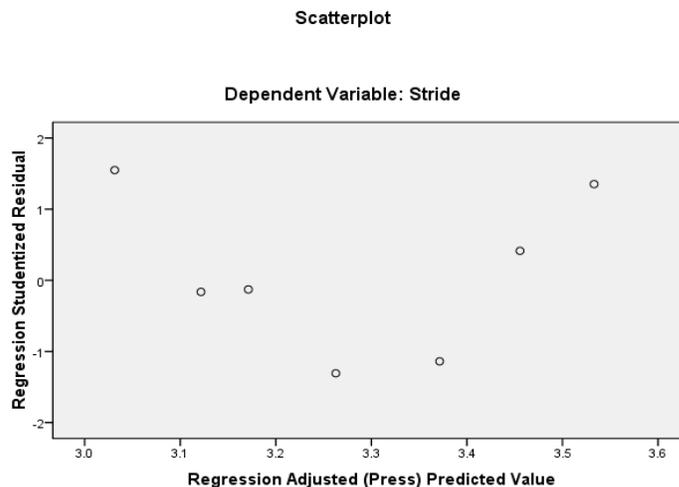
	Speed	Stride
1	15.86	3.05
2	16.88	3.12
3	17.50	3.17
4	18.62	3.25
5	19.97	3.36
6	21.06	3.46
7	22.11	3.55

Next, use **Graphs, Legacy Dialogs, Scatter/Dot** and define a **Simple Scatter** plot using **Stride** on the y axis and **Speed** on the x. Click **Titles** to give your graph a descriptive title. Click **OK** to generate the initial graph. The graph looks very linear. Now click in the graph, and select and select **Edit Contents in Separate Window**. In the Chart Editor, select **Elements, Fit Line at Total**. Click Close in the Properties box since a Linear model is the default and then close the Chart Editor. The coefficient of determination ( $r^2$ ) is also now included in the graph. My finished graph is below.



So far, it appears that a linear model should be a good one. We already know  $r^2 = 0.998$ ; practically perfect!

Perform the regression and ask for both the Normal plot of residuals and the plot of standardized residuals (**SRESID**) as y against adjusted predicted values (**ADJPRED**) as x. Click **Continue** then **OK** to perform the regression and generate the graphs. Let's look at the plot of standardized residuals against predicted values first—is a linear model a “correct” one here? Based on our original data plot, we expect it to be.



This plot shows a clear curve! Since this indicates a linear model is not appropriate for these data, we'll ignore the rest. There's a moral to the story—even if a plot (or  $r^2$ ) indicate a linear relationship exists, we are not through until residuals have been examined!

## 2.5 Relations in Categorical Variables

Categorical data are most typically summarized with a two-way table of counts. Each “cell” in the table represents the number of individuals possessing a particular characteristic of each of the two variables. Here, we examine ways of computing different frequencies from the data, and how to display these data with a stacked bar chart. SPSS is most happy with actual data on individuals; however, we can work with summarized data as shown below.

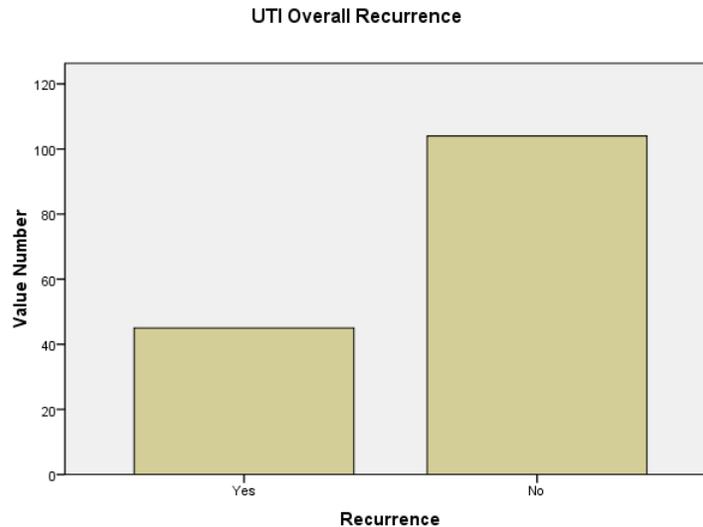
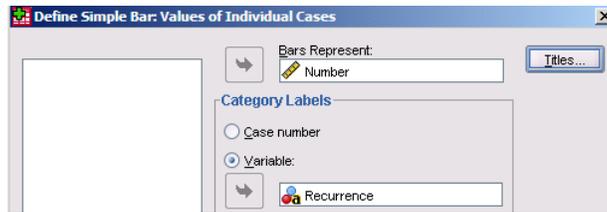
**Example 2.8 Benefits of Cranberry Juice.** The table below presents the results of a study of the recurrence of urinary tract infections (UTIs) in women just cured of a UTI with antibiotics and then taking one of three preventive treatments over a six-month period (cranberry juice daily, a lactobacillus drink daily, or neither drink). Find and graph the marginal distribution of recurrence.

Treatment	Recurrence	No Recurrence	Total
Cranberry juice	8	42	50
Lactobacillus drink	19	30	49
Neither drink	18	32	50
Total	45	104	149

*Solution:* Based on this table, we compute *marginal* frequencies (the percent in one category of one of the variables, without considering the second). This is the row (or column) total divided by the grand total. The overall percent of individuals in the study who had a recurrence is  $45/149 = 30.2\%$ . Similarly, the frequency of no recurrence is  $104/149 = 69.2\%$ . To show this marginal distribution in a bar graph, we enter the summarized data as below.

	Recurrence	Number
1	Yes	45
2	No	104

We next define a simple bar graph from **Graphs**, **Legacy Dialogs**, **Bar** and select Simple and Data in Chart are **Values of individual cases**. Click the **Define** button to continue. Here, the bar heights represent the Number who had each results, and Yes/No for recurrence will label the bars. Give the plot a Title and click **OK** to generate the graph.



We can clearly see the majority of women in the study had no recurrence of the UTI. However, this is not the relationship of most interest to us.

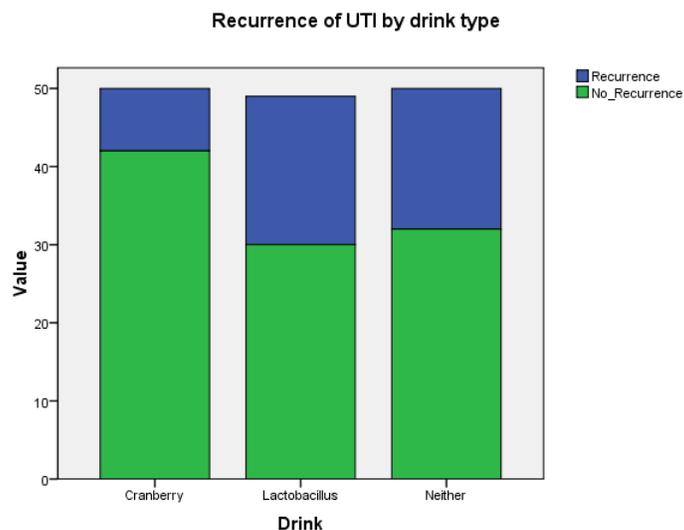
**Example 2.9 Benefits of Cranberry Juice, Continued.** Find and display the conditional distribution of recurrence given the type of drink.

*Solution:* This is what may answer the question of whether or not cranberry juice is effective in preventing UTIs. We can manually find the conditional distributions by dividing each count in the body of the table by the total number of individuals who had that type of drink. For example, we see that for those who drank cranberry juice,  $8/50 = 16\%$  had a recurrence and  $42/50 = 84\%$  had no recurrence. So that for each type of drink we have a complete set of conditional relative frequencies. One way to display these is in a *stacked bar* graph. Each bar represents the 100% of the women who had each type of drink.

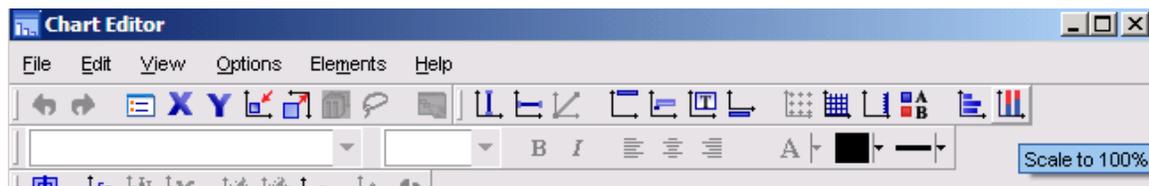
We begin by entering the summarized data as shown on the following page. Notice that since these are count data, the variables have been defined with 0 decimal places, and **Drink** is a string variable (length 13) to accommodate the word “Lactobacillus.”

	Drink	Recurrence	No_Recurrence
1	Cranberry	8	42
2	Lactobacillus	19	30
3	Neither	18	32

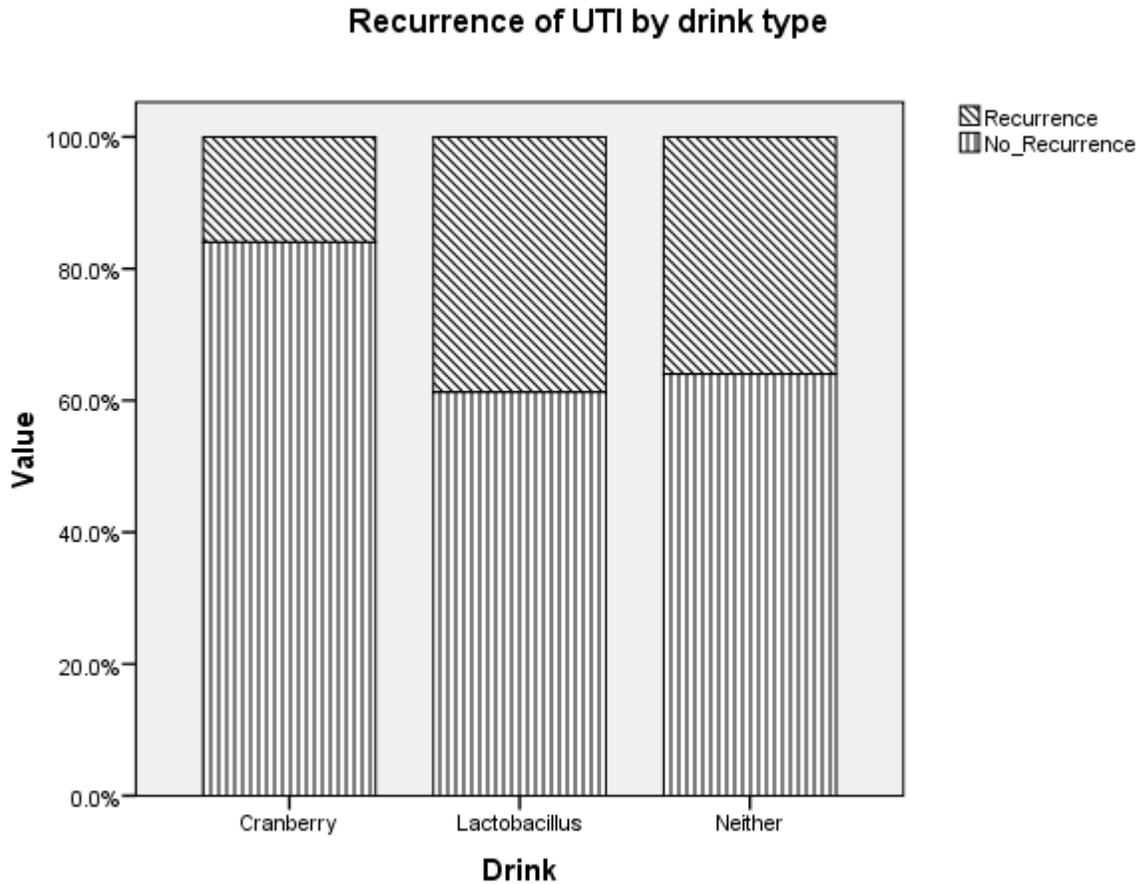
To define the plot, select **Graphs, Legacy Dialogs, Bar** but select **Stacked** and that Data in chart are **Values of individual cases**. Click **Define** to continue. We want one bar for each type of **Drink**, so it is the **Category Labels** variable. Our two variables of counts are entered in the **Bars Represent** box. Give the graph a **Title** and click **OK** to generate the plot. Below is the initial plot.



Since the number of individuals who took each type of drink was relatively even here, we can easily visually compare the bars and conclude that cranberry juice may reduce the recurrence of UTIs; certainly that group had the fewest recurrences in our sample. However, not all data sets have this even a split between groups. We really should adjust the bars to represent the 100% who had each drink type. (You also may want to change from colored bars to fill patterns for a black-and-white printer.) Right-click in the graph and select **Edit Contents in Separate Window**. In the Chart Editor, the right-most button on the menu bar is **Scale to 100%**. Click there to rescale the bars to all reach 100% instead of the actual count in each group.



If you want to change fill colors to a pattern, right-click on a bar, and select the Properties Window. This was discussed in Chapter 1. When finished, close the Chart Editor window. My black-and-white finished stacked bar graph is below.



Since all the bars are normalized to 100% of the individuals in the group, this type of graph is much easier to comprehend (especially when group sizes differ greatly).

## CHAPTER

# 3

	IQobs	SimIQ
1	1	104.50
2	.	64.66
3	.	93.69
4	.	126.20
5	.	92.24
6	.	128.37

## Producing Data

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

### Introduction

In this chapter, we use SPSS to simulate the collection of random samples. Good data collection practice involves randomly selecting individuals from the population, or randomly assigning treatments in a controlled experiment. The randomization can be done with a random digits table, a calculator, or a computer. When your text says “start on line xxx of Table B” the sample drawn in that manner will *not* be random—this is merely a mechanism to be able to write an answer for the back of the book.

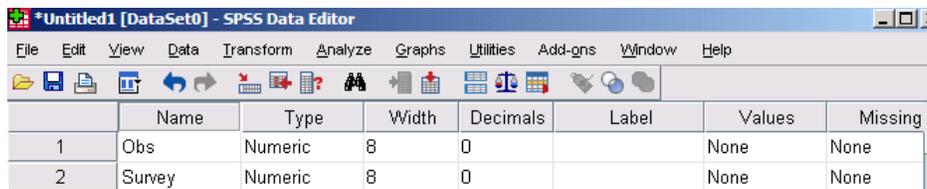
SPSS is primarily for data analysis, so if some of the set-up necessary to generate “random samples” seems contrived, it is.

### 3.1 First Steps

In this section, we demonstrate how to generate count data, or *Bernoulli trials*, for a specified proportion  $p$ . The data simulates observational “Yes/No” outcomes obtained from a random survey. To generate the data, we will use **Transform, Compute Variable**.

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

*Solution.* We want 200 observations (0 = “No,” 1 = “Yes”) where the population proportion of “Yes” answers should be 62%. First, define a variable named **Obs** and one named **Survey**. Each should have 0 decimal places.

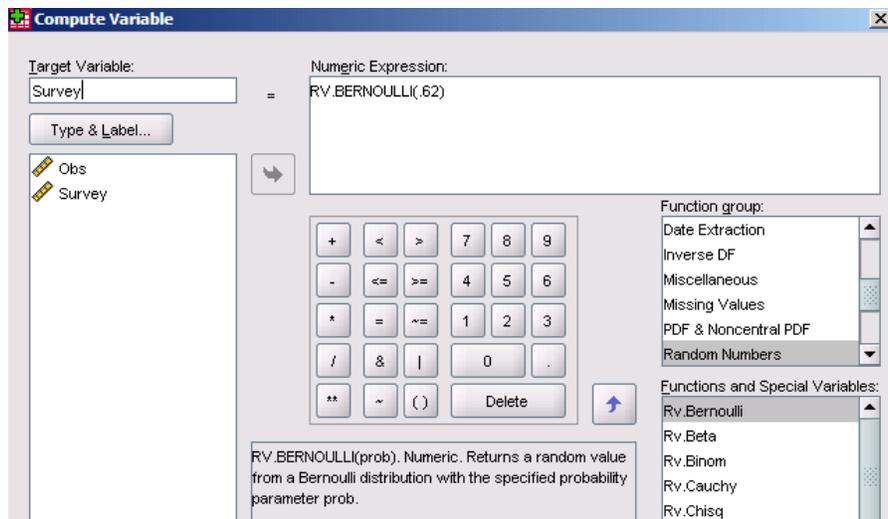


	Name	Type	Width	Decimals	Label	Values	Missing
1	Obs	Numeric	8	0		None	None
2	Survey	Numeric	8	0		None	None

We now need to set the “population” size to be 200 (the number of desired survey “results.” In the Data View, scroll down as far as possible, entering a 1 followed by **Enter** until you can get to row 200. Enter a 1 in row 200. This sets the number of random observations that will be generated.

	Obs	Survey
199	.	.
200	1	.

Now, click **Transform, Compute Variable**. On the right-hand side of the box, scroll down to find the function group **Random Numbers**, then in the lower box select **Rv.Bernoulli**. Set the **Target Variable** to **Survey** and enter the desired proportion of “successes,” here .62. Click **OK** to generate the data.



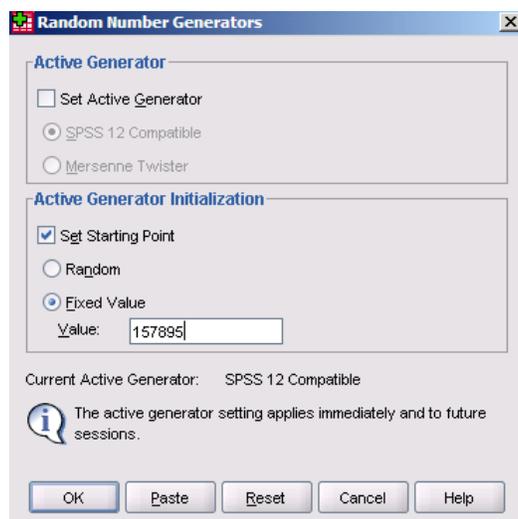
You may see a box asking whether to change the existing variable. Click **OK**. We now have 200 observations of 0 (“No”) and 1 (“Yes”). To see our sample proportion of “Yes” answers, use **Analyze, Descriptive Statistics, Descriptives** for the variable **Survey**. In our particular set of generated “answers,” there were exactly 62% “Yes” answers as seen below.

	N	Minimum	Maximum	Mean	Std. Deviation
Survey	200	0	1	.62	.487
Valid N (listwise)	200				

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

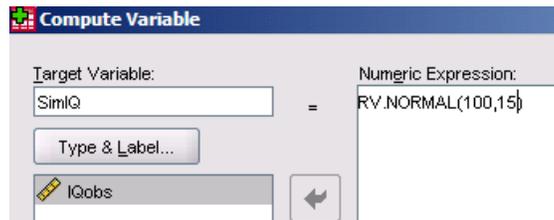
**Example 3.2 Changing the Seed.** We really want random numbers each time. However, just as your text will instruct you to “start in line xx” when using the table of random digits, your instructor may ask that you use a particular seed so that each student will get the same random numbers. How do we do that?

*Solution.* Click **Transform, Random Number Generators**. The default selection in this dialog box is to have a **Random** Starting Point. For a particular sequence, change the button to **Fixed Value** and enter the desired seed. Click **OK**. If you want to return to “truly” random numbers, set this option back to **Random**.



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

*Solution.* Click **File, New, Data** for a clean worksheet. Define variable **IQobs** to have 0 decimal places. As before, we'll set our "population size" to 150 by entering a 1 in row 150 of a new variable called **IQobs**. Use **Transform, Compute Variable** and select **Random Numbers** and **Rv.Normal**. Click the up arrow to transfer the shell to the Expression box and replace the question marks with the mean (100) and standard deviation (15) for our desired data. Click **OK** to generate the values.



To compute summary statistics for our generated data, use **Analyze, Descriptive Statistics, Descriptives** for variable **SimIQ**. Our sample mean of 99.0092 is very close to the population mean of 100; the standard deviation is close to 15, but just a little low at 14.431.

**Descriptive Statistics**

	N	Minimum	Maximum	Mean	Std. Deviation
SimIQ	150	64.66	149.66	99.0092	14.43076
Valid N (listwise)	150				

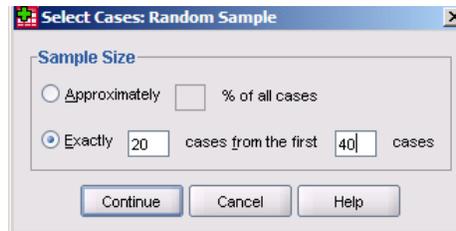
## 3.2 Design of Experiments

Assigning treatments for a randomized experiment can easily be done in SPSS after first setting the "population" of available subjects/experimental units.

**Example 3.3 Treatment Assignments.** For an experiment with two treatments, we will randomly choose 20 subjects from a group of 40 to receive Treatment A. The remaining subjects will receive Treatment B.

*Solution.* Define a variable called **Patientnum**. As described before, set the "target" population to have size 40 by scrolling down to row 40 and entering a 1 followed by

**Enter.** Click **Data, Select Cases**. Click to change the selection method to **Random sample of cases**, then click the **Sample** button. We want to randomly select exactly 20 of the 40 cases as shown below to receive Treatment A.



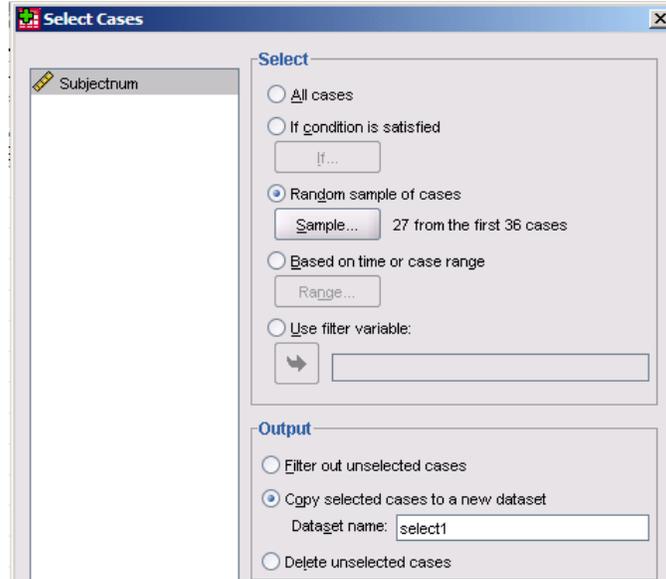
Click **Continue** and **OK**. Upon returning to the Data Editor, a new variable called filter\_\$ has been created. It is set to 1 for selected cases and 0 for unselected cases. Unselected cases also have had a slash imposed on their row number. Among the first seven patients in our list, only patients 3 and 6 were not selected for Treatment A, they will receive Treatment B.

	Patientnum	filter_\$
1	.	1
2	.	1
<del>3</del>	.	0
4	.	1
5	.	1
<del>6</del>	.	0
7	.	1

**Example 3.4 More Treatment Assignments.** We have thirty-six subjects numbered 1 through 36 available for a small clinical trial. We want to randomly assign them to four treatment groups, each of size nine.

*Solution.* Since we do not want any individual assigned more than once, we will successively sample smaller groups out of our original 36 subjects. First we will sample 27 of the 36 into a new data set. Individuals not included at this point will receive Treatment A. From the remaining 27, we will then sample 18 into yet another new data set. Individuals not included here will get Treatment B. Lastly, we will sample 9 of the remaining 18 into yet another new dataset. Those will receive Treatment D. Those not remaining in the last set will receive Treatment C.

If necessary, click **File, New, Data** to obtain a clean data window. Define a variable called **Subjectnum**. Enter numbers 1 through 36 in this column to correspond with our 36 subjects. As in Example 3.3 above, we use **Data, Select Cases** to select a random sample of exactly 27 of the first 36 cases. However, we will copy the selected cases to a new data set called **Select1**.



Click **OK** to perform the selection. A part of the selection we obtained is shown. In this selection, subjects 2, 6, 9, and 12 do not appear. They will be assigned (with any others not selected at this step) to Treatment A. Continue with the process, narrowing down the “selected” individuals as described above.

Subjectnum	
1	1.00
2	3.00
3	4.00
4	5.00
5	7.00
6	8.00
7	10.00
8	11.00
9	13.00
10	14.00
11	15.00

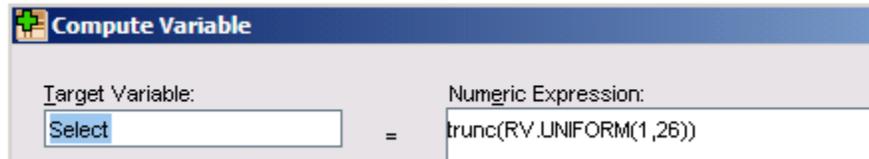
### 3.3 Sampling Design

We can easily use Select Cases as described above to select a simple random sample from a “population” of available objects. There are other types of sampling to consider.

**Example 3.5 Systematic Sampling.** Choose a systematic random sample of four addresses from a list of 100.

*Solution.* Because the list of 100 divides evenly into four groups of 25, we will choose one address from each of the groups 1–25, 26–50, 51–75, and 76–100. However, because we will use **Rv.Uniform(a,b)** from the **Transform, Compute Variable** option to get these in SPSS we will need to adjust our parameters a little. This Uniform random

number generator gives results as a continuous variable between  $a$  and  $b$ . Since our first selection should be between 1 and 25, we'll use a low end of 1 and a high end of 26 in our command specification, then truncate the result to eliminate the decimal places. (If we use 25 as the upper end, we couldn't possibly select address number 25.) Since we only want one number, all we need to do is put the cursor (highlight) in a blank cell in the Data View (however, there must be some data—SPSS doesn't like a blank worksheet!).



For the second, we repeat the process, but use low end 26 and high end 51 (so that a large value could truncate to 50). Then repeat again using a low end of 50 and high end of 76. Lastly, we use low end 76 and high end 101. VAR00001 was the dummy (to have data in the spreadsheet). The selected addresses are 4, 36, 66, and 83.

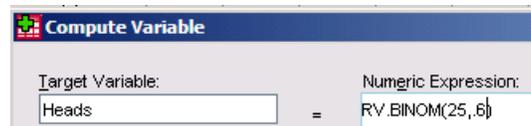
	VAR00001	Select	Select2	Select3	Select4
1	1.00	4.00	36.00	66.00	83.00

### 3.4 Toward Statistical Inference

In Example 3.1 (page 53), we generated a random sample of “Yes/No” responses. In this section, we will demonstrate how to simulate the collection of multiple samples of this type. In particular, we are concerned with the total number of “Yes” responses in several repetitions of the survey (or experiment), the sample proportion for each sample, and the resulting average of all sample proportions.

**Example 3.7 Simulating Coin Flips** (a) We have a coin for which the probability of heads is 0.60. We toss the coin 25 times and count the number of heads in this sample. Then we repeat the process for a total of 50 samples of size 25. Simulate the counts of heads for these 50 samples of size 25, compute the sample proportion for each sample, and make a histogram of the sample proportions.

*Solution.* As we have done before, we'll first set up a column to define the number of repetitions we want (50). We defined a variable named Repeat and scrolled down (and entered 1's) until we could enter a 1 in the 50<sup>th</sup> row. This defines the number of repetitions of our random number generator command. Use **Transform, Compute Variable** and use random Number type **Rv.Binom(25,.6)** to generate our simulated numbers of heads into a new variable called Heads.



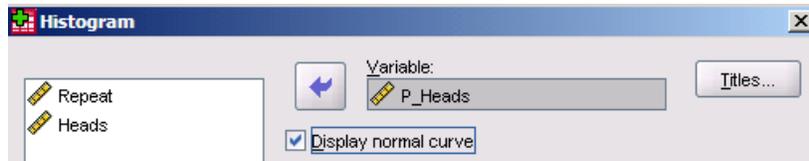
Click **OK** to generate the results, a portion of which can be seen below.

	Repeat	Heads
1	1.00	17.00
2	.	15.00
3	.	16.00
4	.	15.00
5	.	11.00
6	.	15.00
7	.	17.00

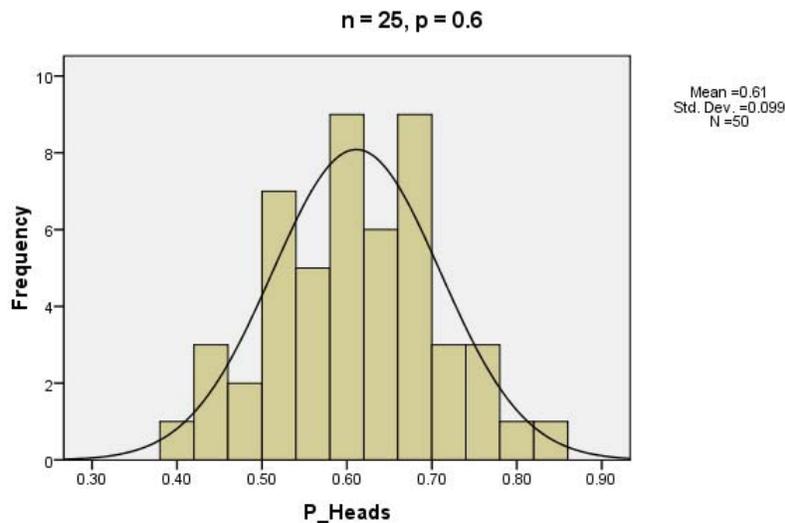
We'll use Transform, Compute Variable to compute a new variable,  $P\_Heads = Heads/25$  to find the sample proportions.



Lastly, we define a **Histogram** (from the **Graphs, Legacy Dialogs** options) to use variable  $P\_Heads$ . (Be sure to give the graph a title!) If you want to see how well this sampling distribution is modeled by a Normal curve, check the box to have the curve added to the histogram. Click **OK** to generate the plot. Our plot is relatively symmetric, but somewhat shorter than Normal in the middle. The mean of the sample proportions (0.61) is very close to 0.6.

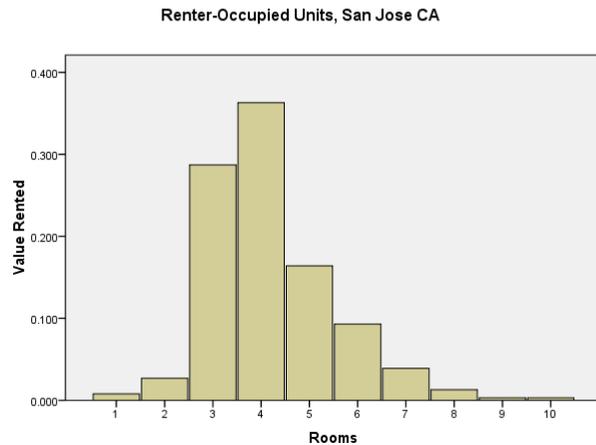


Sampling Distribution of Heads



## CHAPTER

# 4



# Probability: The Study of Randomness

- |     |                           |
|-----|---------------------------|
| 4.1 | Randomness                |
| 4.2 | Probability Models        |
| 4.3 | Random Variables          |
| 4.4 | Means of Random Variables |
| 4.5 | General Probability       |

## Introduction

In this chapter, we show how to use SPSS to generate some random sequences. We then see how to make a probability histogram for a discrete random variable and how to compute its mean. Since SPSS is primarily a data analysis computer package, it can't help much with some of the typical probability calculations; a calculator is much more useful with those.

Beware—since we are simulating pseudorandom values, your results most likely will not agree with those shown here; they should, however, be fairly similar.

## 4.1 Randomness

In this section, we work some examples that use SPSS to generate various random sequences.

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

*Solution.* This is similar to Example 3.1 (page 53). As we did there, scroll down and enter 1's until you can enter a 1 in row 100 of a variable. This allows us to generate 100 random Bernoulli observations where 1 = a made shot, and 0 = a missed shot. Again, we use **Transform, Compute Variable**, selecting **Rv.Bernoulli** from the set of Random Number functions. Replace the question mark with the parameter value .5. Click **OK** to generate the observations.



In the Data View we can examine the sequence of hits and misses. In this small segment, we see the player made both of the first shots, missed the third, made the fourth, missed the fifth, etc.

	VAR00001	Shot
1	.	1.00
2	.	1.00
3	.	0.00
4	.	1.00
5	.	0.00
6	.	1.00
7	.	0.00

What percentage of shots were made? Using **Analyze, Descriptive Statistics, Descriptives** for the variable **Shot**, we see the mean of the variable Shot (the sample proportion in this case) is 56%.

	N	Minimum	Maximum	Mean	Std. Deviation
Shot	100	.00	1.00	.5600	.49889
Valid N (listwise)	100				

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

*Solution.* We’ll use 100 rolls as an example. If you just did the last example in this manual, you already have a 1 entered in row 100 of VAR0001 as a place-holder to tell SPSS that we want 100 observations generated. If not, scroll down entering 1 until you can enter a 1 in row 100. We will truncate a Uniform random variable to get rid of decimal places. Since we want values 1 through 6, we will generate our Uniform observations as being between 1 and 7 (so that 6’s are possible).



Clicking **OK** generates data for the first roll. Repeat to generate rolls for **Die2**, **Die3**, and **Die4** (all you need to do is change the Target Variable name).

To find the percentage of the time that there was at least one “6” in the set, we could manually scan down the data sheet and count, but that’s inefficient and we might miss some. To answer this question, we’ll create a new variable called **Sixes**. It’s created as the maximum value of the four rolls.



At this point, the worksheet looks like this. Again, we could count the number of times that **Sixes** has the value 6, but we could still miss some.

		Die1	Die2	Die3	Die4	Sixes
1	00	2.00	4.00	5.00	1.00	5.00
2	00	3.00	1.00	3.00	6.00	6.00
3	00	4.00	3.00	2.00	2.00	4.00
4	00	4.00	2.00	4.00	2.00	4.00
5	00	1.00	5.00	1.00	3.00	5.00
6	00	2.00	1.00	3.00	6.00	6.00
7	00	5.00	6.00	4.00	4.00	6.00

We’ll use **Analyze**, **Descriptive Statistics**, **Frequencies** on the variable **Sixes** to get a table of the values in this variable.

Sixes					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	3	4	4.0	4.0	4.0
	4	20	20.0	20.0	24.0
	5	25	25.0	25.0	49.0
	6	51	51.0	51.0	100.0
	Total	100	100.0	100.0	

Here, we can see that in our 100 repetitions of this experiment, 6 was included at least once (so it was the largest value) 51 times out of 100, or 51% of the time.

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

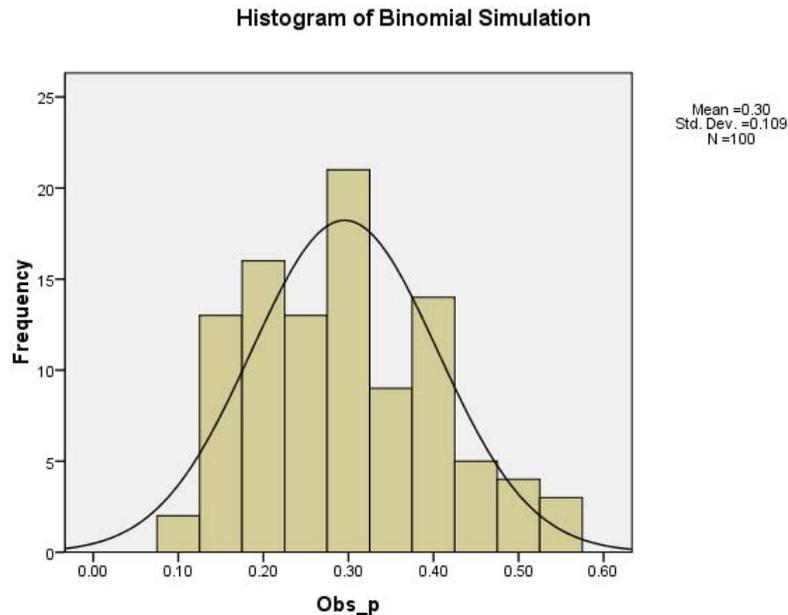
*Solution.* If we still are using the same data worksheet as before, all we need to do to start is to use **Transform, Compute Variable** to generate our binomial observations as below.



To convert the observed number of successes to percents, we compute a new variable we'll call **Obs\_p** by dividing **Binom** by 20.



Lastly, we use **Graphs, Legacy Dialogs, Histogram** to create the histogram for **Obs\_p**. Notice that the histogram of observed proportions is centered at the parameter  $p = 0.3$ .



## 4.2 Probability Models

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

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

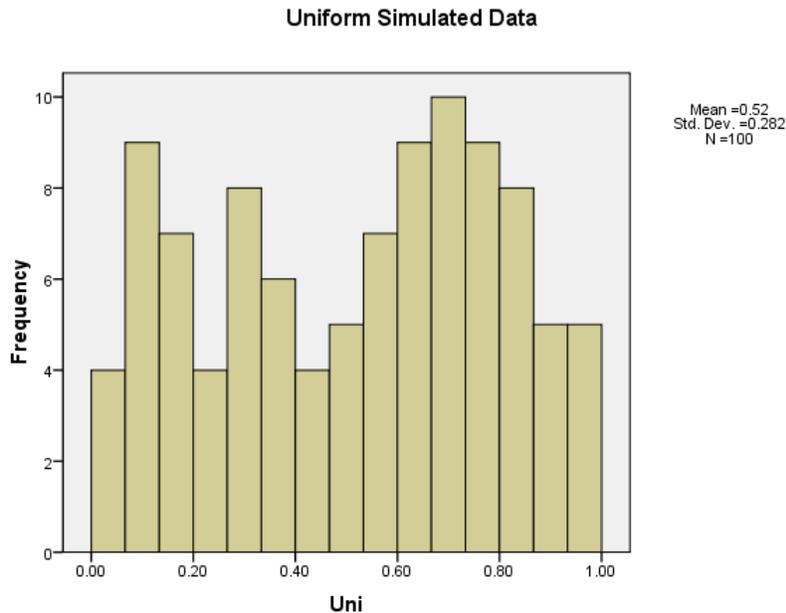
*Solution.* (a) Assuming we're still using the same worksheet of data as in the two previous examples, we will again use **Transform, Compute Variable** to create variable **Uni** as below.



We can observe the first few values in the Data View.

Uni
0.82
0.51
0.58
0.62
0.66
0.45
0.54
0.87

Next, define a histogram for **Uni** using **Graphs, Legacy Dialogs, Histogram**.



While not perfectly flat, this simulated data is quite different from Normally distributed data. There is no tapering towards the ends of the distribution, nor is there an obvious central peak.

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

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

*Solution.* (a) Define two variables to contain the first digit and its probability. The leading digit cannot have a decimal portion, and our probabilities are given to three places, so adjust decimal places accordingly. Enter the data. There is no need to type the leading 0; SPSS will supply it for you.

	Name	Type	Width	Decimals	Label	Values	Missing
1	Digit	Numeric	8	0		None	None
2	Freq	Numeric	8	3		None	None

To verify that the table defines a legitimate probability distribution, the sum of the variable Freq must be 1. Use **Analyze, Descriptive Statistics, Frequencies** for **Freq**. Click the Statistics button and select **Sum**. Click **Continue** and **OK**. Our values do indeed sum to 1.

**Statistics**

Freq		
N	Valid	9
	Missing	0
Sum		1.000

(b) Unfortunately, SPSS can't help with this part of the question. We will manually add the probabilities that the first digit is 6 or higher.  $0.067 + 0.058 + 0.051 + 0.046 = 0.222$ . There is a 22.2% chance the leading digit is at least a 6.

### 4.3 Random Variables

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

**Example 4.6 How Many Rooms?** The table below gives the distributions of rooms for owner-occupied units and for renter-occupied units in San Jose, California. Make probability histograms of these two distributions.

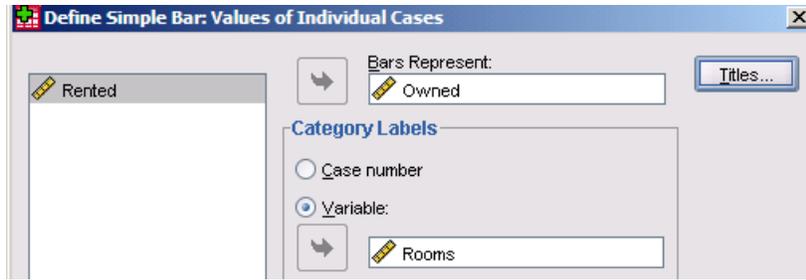
Rooms	1	2	3	4	5	6	7	8	9	10
Owned	0.003	0.002	0.023	0.104	0.210	0.224	0.197	0.149	0.053	0.035
Rented	0.008	0.027	0.287	0.363	0.164	0.093	0.039	0.013	0.003	0.003

*Solution.* To create a histogram for data like these, we start with a bar graph which will be modified using the Chart Editor to look like a histogram. We first define three variables as below.

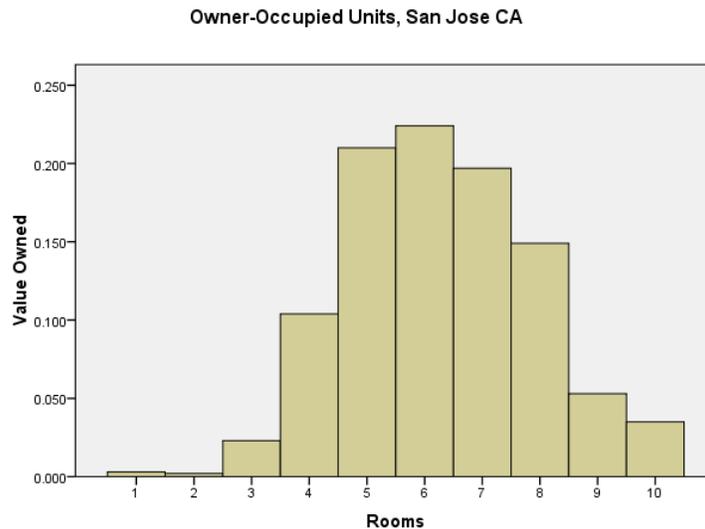
	Name	Type	Width	Decimals	Label	Values
1	Rooms	Numeric	8	0		None
2	Owned	Numeric	8	3		None
3	Rented	Numeric	8	3		None

Next, enter the information given in the table above. To verify that the frequencies have been entered correctly, you can use **Analyze, Descriptive Statistics, Frequencies** as in Example 4.5 earlier. To get statistics for both variables at once, hold **Ctrl** while clicking to select the variable names.

Select **Graphs, Legacy Dialogs, Bar**. We want a **Simple** chart where Data in Graph are **Values of individual cases**. Click the **Define** button to continue.



Above is the basic definition for the owner-occupied units. (Renters would be similar.) Give the plot a title and click OK to generate the graph. The initial graph will have spaces between the bars. To eliminate these, right-click in the graph and select **Edit Contents in Separate Window**. Right-click on a bar and select **Properties Window**. Then select **Bar Options** and set the bar width to 100%. Clearly from the graph below, the majority of owner-occupied units have between 5 and 7 rooms.

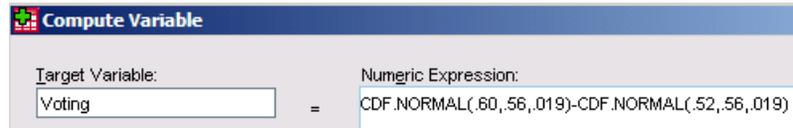


We conclude this section by working an exercise using the normal density curve that reviews the normal distribution calculations from Section 1.3 of this manual.

**Example 4.7 Voting in Oregon.** After an election in Oregon, voter records showed that 56% of registered voters actually voted. A survey of 663 registered voters is conducted and the sample proportion  $\hat{p}$  of those who claim to have voted is obtained. For all random samples of size 663, these values of the sample proportions  $\hat{p}$  will follow an approximate normal distribution with mean  $\mu = 0.56$  and standard deviation  $\sigma = 0.019$ . Use this distribution to compute  $P(0.52 < \hat{p} < 0.60)$  and  $P(\hat{p} \geq 0.72)$ .

*Solution.* We'll again use **Transform, Compute Variable** to answer this question. Our Function group is **CDF and Noncentral CDF**. The function is **CDF.Normal**.

Remember, this finds the probability that random variable  $X$  is less than or equal to the specified value. Here, we find the area below  $\hat{p} = .60$  and subtract the area to the left of  $\hat{p} = .52$  to find the desired probability. To two decimal places, we find this is 96%. If you want more decimal places in your answer, go to the **Variable View** and increase them.



The probability that at least 72% of the sample would have claimed to have voted (if they told the truth) is  $1 - P(\hat{p} \leq .72)$ . Simply change the first CDF.NORMAL to a 1 and the .52 to .72.



We find this probability is 0.

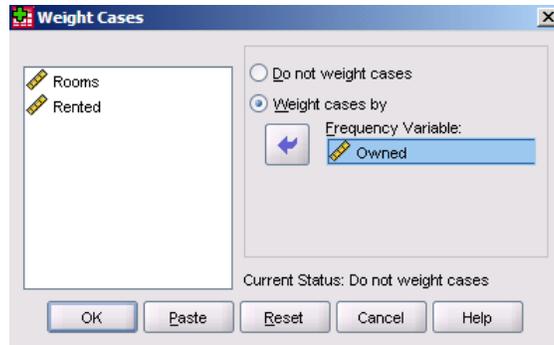
#### 4.4 Means of Random Variables

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

**Example 4.8 More About Rooms.** The table below again gives the distributions of the number of rooms for owner-occupied units and renter-occupied units in San Jose, California. We want to calculate the mean and the standard deviation of the number of rooms for each type.

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

*Solution.* If we simply asked SPSS to compute the mean (average) number of rooms for this distribution with **Analyze, Descriptive Statistics, Descriptives**, every room number would have an equal weighting. We want to use the frequencies as weights. We do this using **Data, Weight Cases**. To find the mean and standard deviation for owner-occupied units, we set the weight variable to **Owned**. Click **OK** to set the weights.



We then use **Analyze, Descriptive Statistics, Descriptives** to compute the mean. SPSS does not give the standard deviation in this case. We find the average (mean) number of rooms for owner-occupied units in San Jose is 6.28. Since this is based on a probability distribution, we would report  $\mu = 6.28$ . (Notice that SPSS does not give the standard deviation here.)

**Descriptive Statistics**

	N	Minimum	Maximum	Mean	Std. Deviation
Rooms	1	1	10	6.28	
Valid N (listwise)	1				

## 4.5 General Probability

We conclude this chapter with a small example that illustrates how we might do some general calculations; as stated in the introduction to this chapter, SPSS as a data analysis package is not really suited to answer most general probability questions.

**Example 4.9 World War II Bomber Survival.** During World War II, the British found that Allied bombers had a 95% probability of surviving any one mission. A tour of duty was 30 missions. Since it is reasonable to assume missions are independent, what is the probability that an airman would survive 30 such missions and be eligible to return home?

*Solution.* Define two variables in a new data sheet as below. We use several decimal places for **SurvivalProb** because the answer will involve raising the individual mission survival probability to a large power.

	Name	Type	Width	Decimals	Label
1	Probability	Numeric	8	2	
2	SurvivalProb	Numeric	8	6	

Enter 0.95 for the individual mission survival probability under **Probability**. Using **Transform**, **Compute Variable**, we will compute **SurvivalProb** to be Probability raised to the 30<sup>th</sup> power. Exponentiation in SPSS (and many other computer packages) is indicated by \*\*. Click **OK**. If you get the “Change existing variable?” message, click **OK**.

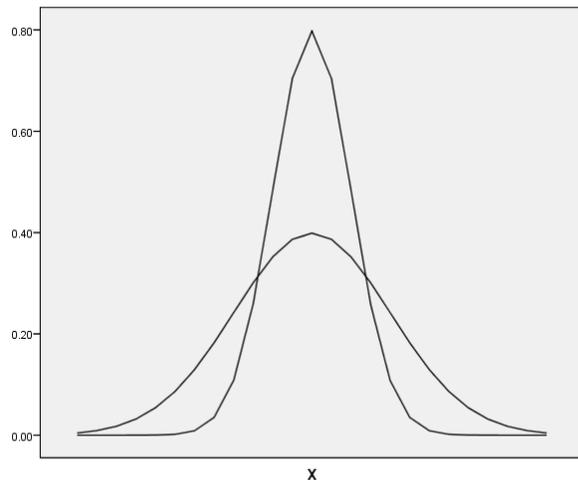


We see below that based on this information, the probability of an airman surviving his tour of duty was only about 21.5%. Not very good odds of going home in one piece.

	Probability	SurvivalProb
1	0.95	0.214639

## CHAPTER

# 5



# Sampling Distributions

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

### Introduction

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

Many of these calculations are really best done using a calculator; however, SPSS can compute them using our old friend **Transform**, **Compute Variable**.

## 5.1 Sampling Distributions for Counts and Proportions

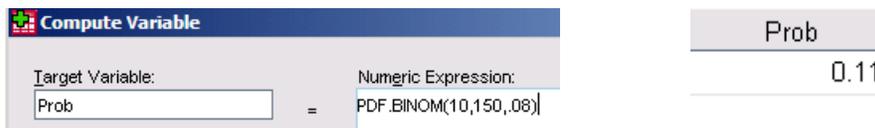
We begin by demonstrating how to compute various probabilities for a given binomial distribution. To do so, we will need the **PDF.Binom** and **CDF.Binom** commands found under the PDF and Noncentral PDF and CDF and Noncentral CDF Function Groups.

### Binomial Probabilities

For a binomial distribution,  $X \sim B(n, p)$ , we compute the probability of exactly  $k$  successes,  $P(X = k)$ , by entering the command **PDF.Binom(k,n,p)**. The probability  $P(X \leq k) = P(0 \leq X \leq k)$  of at most  $k$  successes is computed with the command **CDF.Binom(k,n,p)**. The probability of there being at least  $k$  successes is given by  $P(X \geq k) = 1 - P(X \leq k - 1)$ , and is computed with the command **1-CDF.Binom(k-1,n,p)**. The following three examples demonstrate these calculations.

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

*Solution.* To find the probability of exactly 10 misclassified records, we locate the PDF and Noncentral PDF function group in the **Transform, Compute Variable** dialog box, then locate **PDF.Binom** in the function group at the bottom right. Click the up arrow to transfer the shell to the Numerical Expression box, then complete the command by giving a Target Variable name and the parameters  $k = 10$ ,  $n = 150$ , and  $p = .08$ . Click **OK** to perform the calculation. We find that the probability of exactly 10 misclassified records is 0.11. If you want additional decimal places for the probability, go to the Variable View and increase them.



To find the probability of at most 10 misclassified records, the command is similar, but we use **CDF.Binom** from the CDF and Noncentral CDF function group. We see the probability of at most 10 misclassified records is about 34% (to two decimal places)



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

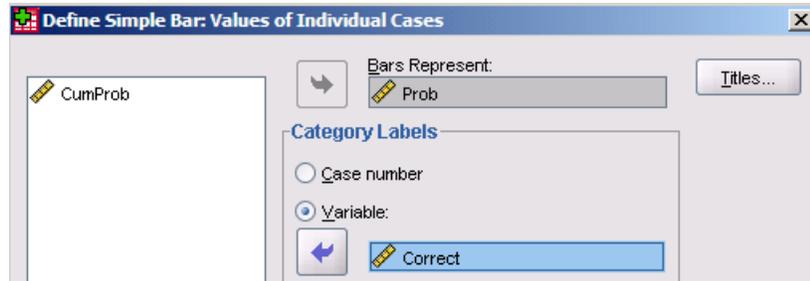
*Solution.* Because there are  $n = 15$  attempts, the possible number of successes range from 0 to 15. So we first enter the integer values 0, 1, . . . , 15 into a variable that has been named **Correct**. Since there cannot be fractions of correct answers, the number of decimal places for this variable has been set to 0. Since the probability distribution is the probability of each possible value of the number of correct answers, we use the variable **Correct** as the parameter for this **PDF.Binom** command instead of specifying a particular value. To find the Cumulative Distribution, the command is similar, but use **CDF.Binom** instead; these results have been stored into a variable named **CumProb**.



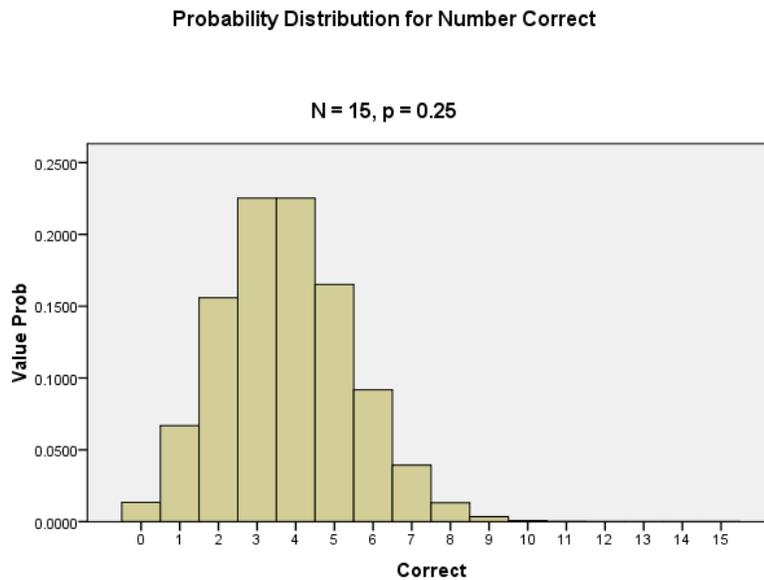
We see our results below. From **CumProb**, we see the probability of at most one correct answer is 0.0802, or about 8%.

	Correct	Prob	CumProb
1	0	0.0134	0.0134
2	1	0.0668	0.0802
3	2	0.1559	0.2361
4	3	0.2252	0.4613
5	4	0.2252	0.6865
6	5	0.1651	0.8516
7	6	0.0917	0.9434
8	7	0.0393	0.9827
9	8	0.0131	0.9958
10	9	0.0034	0.9992
11	10	0.0007	0.9999
12	11	0.0001	1.0000
13	12	0.0000	1.0000

Making a probability histogram of these results is similar to Example 4.8. Define a Simple Bar Graph using Values of Individual Cases. Bars Represent the **Probability** of each particular number of correct answers, and the Category Labels are the number **Correct**.

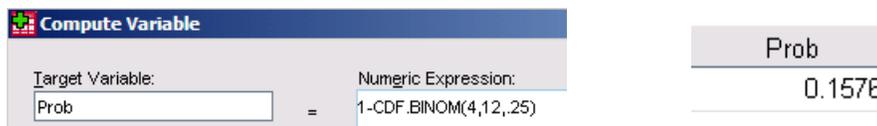


As we did in Chapter 4, edit the content of the generated graph using **Properties**, **Bar Options** to make the bars 100% wide. Below is the finished graph.



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

*Solution.* We use the probability of the complement to obtain  $P(X \geq 5) = 1 - P(X \leq 4)$ , which is computed by **1-CDF.Binom(4,12,.25)**  $\approx 0.1576$ .



### Probabilities for $\hat{p}$

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

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

*Solution.* Because 58% of 2500 is 1450, we must compute  $P(X \geq 1450)$ , where  $X \sim B(2500, 0.6)$ . Instead, we may compute  $1 - P(X \leq 1499)$  using **1-CDF.Binom(1499,2500,.6)**  $\approx 0.5087$ .



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

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

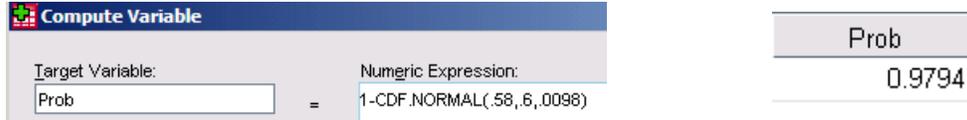


### Normal Approximations

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

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

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



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

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

(b) We now let  $Y \sim N(180, 12.586)$ . Then  $P(165 < X < 195)$  is found using **CDF.NORMAL(195,180,12.586)-CDF.NORMAL(165,180,12.586)**. We see that the desired probability  $\approx 0.7667$ .



### 5.3 Poisson Random Variables

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

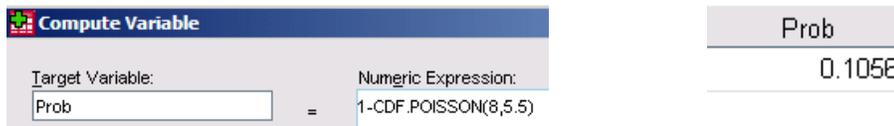
**Example 5.8 Mumps Outbreaks.** Mumps is an acute viral infection that is generally mild, and in about 20% of infected individuals, even asymptomatic (meaning they do not know they have the disease). However, severe complications can arise. Mandatory vaccinations have largely eradicated the disease in the United States. For the state of Iowa, the average monthly number of reported cases is about 0.1 per month. Assuming that cases are independent, what is the probability that in a given month, there will be no more than one case of mumps in Iowa?

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



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

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



### 5.3 The Sampling Distribution of a Sample Mean

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

**Example 5.10 Measuring Blood Glucose.** Sheila’s glucose level one hour after ingesting a sugary drink varies according to the Normal distribution with  $\mu = 125$  mg/dl and  $\sigma = 10$  mg/dl.

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

*Solution.* (a) We compute  $P(X > 140)$  for  $X \sim N(125, 10)$  with the command **1-CDF.Normalcdf(140,125,10)**. Then,  $P(X > 140) \approx 0.0668$ .

Compute Variable		Prob
Target Variable:	Numeric Expression:	
Prob	= 1-CDF.NORMAL(140,125,10)	0.0668

(b) For an SRS of size  $n = 4$ ,  $\bar{x}$  has a mean of  $\mu = 125$  and a standard deviation of  $\sigma/\sqrt{n} = 10/\sqrt{4} = 5$ . So now we compute  $P(\bar{x} > 140)$  for  $\bar{x} \sim N(125, 5)$  and obtain a value of about 0.00135.

Compute Variable		Prob
Target Variable:	Numeric Expression:	
Prob	= 1-CDF.NORMAL(140,125,5)	0.0013

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

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

Compute Variable		Level
Target Variable:	Numeric Expression:	
Level	= IDF.NORMAL(.95,125,5)	133.22

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

*Solution.* If the total weight of 12 eggs falls between 750 g and 825 g, then the sample mean  $\bar{x}$  falls between  $750/12 = 62.5$  g and  $825/12 = 68.75$  g. So, we compute  $P(62.5 < \bar{x} < 68.75)$  for  $\bar{x} \sim N(65, 5/\sqrt{12})$  using the command **CDF.Normal(68.75,65,5/Sqrt(12))-CDF.Normal(62.5,65,5/sqrt(12))**.

Compute Variable		Prob
Target Variable:	Numeric Expression:	
Prob	= CDF.NORMAL(68.75,65,5/sqrt(12))-CDF.NORMAL(62.5,65,5/sqrt(12))	0.9584

### Sum of Independent Normal Measurements

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

**Example 5.13 A Golf Competition.** Tom and George are playing in the club golf tournament. Their scores vary as they play the course repeatedly. Tom's score  $X$  has the  $N(110, 10)$  distribution and George's score  $Y$  has a  $N(100, 8)$  distribution. If the scores are independent, then what is the probability that Tom's score is less than George's?

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

Compute Variable		Prob
Target Variable:	Numeric Expression:	
Prob	CDF.NORMAL(0,10,12.806)	0.2174

### Sum and Difference of Sample Means

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

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

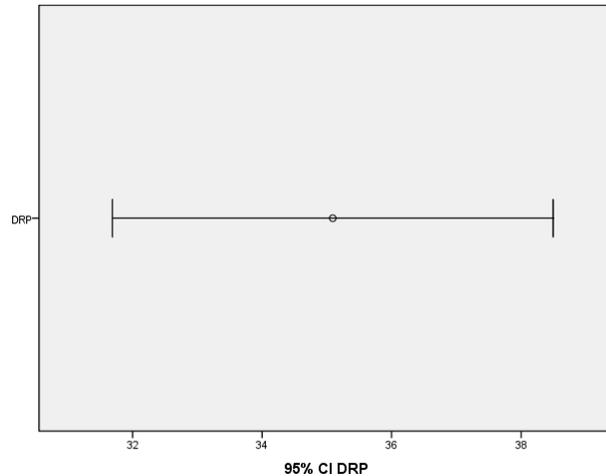
*Solution.* First,  $\bar{y} - \bar{x} \sim N\left(4.8 - 2.4, \sqrt{\frac{1.5^2}{30} + \frac{1.6^2}{30}}\right) = N(2.4, 0.4004)$ . We find

$P(\bar{y} - \bar{x}) \geq 1$  using **1-CDF.normal(1,2.4,0.4004)** and obtain a value of 0.9998. It's virtually certain the *Journal* will have a higher mean score at least one point higher than the *Enquirer*.

Compute Variable		Prob
Target Variable:	Numeric Expression:	
Prob	1-CDF.NORMAL(1,2.4,0.4004)	0.9998

## CHAPTER

# 6



# Introduction to Inference

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

## Introduction

In this chapter, we show how to use SPSS (primarily as a calculator) to compute confidence intervals and conduct hypothesis tests for the mean  $\mu$  of a normally distributed population.

Since SPSS is really a data analysis package, it has no built-in capability to compute confidence intervals or hypothesis tests for summarized data. Also, since in practice the population standard deviation is almost never known, the built-in inference procedures in SPSS are not based on standard Normal distributions, but  $t$ -distributions which will be discussed in Chapter 7.

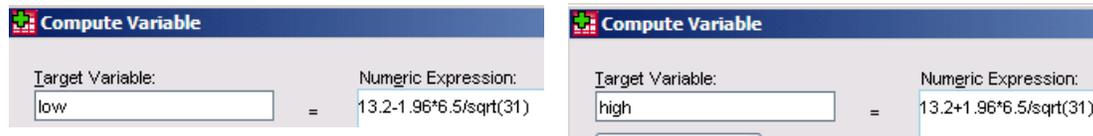
**Note:** Because of rounding, some differences may occur with different technologies. These are usually not a major cause for concern.

## 6.1 Confidence Intervals with $\sigma$ Known

In this section, we show how to compute a confidence interval for the mean of a normal population with known standard deviation  $\sigma$ . Here, as mentioned in the introduction to this chapter, we can really only use the **Transform, Compute Variable** function of SPSS to mimic a calculator.

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

*Solution.* For a 95% confidence interval with  $\sigma$  “known,” the value of  $z^*$  is 1.96. The confidence interval is of the form  $\bar{x} \pm z^* \frac{\sigma}{\sqrt{n}}$ . We compute the lower and upper ends of the interval as



Our results are

low	high
10.91	15.49

We normally report one more significant digit than was in the information provided, but check with your instructor for his or her rounding rules. I would say that based on this information, I am 95% confident the mean TRAP level for young women represented by this sample is between 10.92 and 15.49 Units/liter.

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

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

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

*Solution.* Here, we'll use SPSS to compute the mean speed, then calculate the confidence interval as in Example 6.1. Define a variable called **Speed** and enter the data.

Now use **Analyze, Descriptive Statistics, Descriptives** to obtain the mean of this set of data.

	N	Minimum	Maximum	Mean	Std. Deviation
Speed	20	16.0	49.0	30.800	11.2041
Valid N (listwise)	20				

Notice the sample standard deviation (11.2041) is somewhat different from the assumed value (10.3). We proceed to calculate the lower and upper ends of the interval.

Compute Variable	
Target Variable:	Numeric Expression:
low	$30.8 - 1.96 * 10.3 / \sqrt{20}$

Compute Variable	
Target Variable:	Numeric Expression:
high	$30.8 + 1.96 * 10.3 / \sqrt{20}$

The computed values are

low	high
10.91	35.31

We can say, based on this sample with 95% confidence, the mean speed for vehicles undergoing this fuel efficiency test is between 10.91 and 35.31 miles per hour.

### Choosing the Sample Size

Suppose we want to find the minimum sample size  $n$  that will produce a desired margin of error  $m$  with a specific level of confidence. To do so, we can use the formula

$$n \geq \left( \frac{z^* \sigma}{m} \right)^2, \text{ where } z^* \text{ is the appropriate critical value.}$$

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

*Solution.* The critical value for 95% confidence is  $z^* = 1.96$ . Using this value in the formula  $n \geq \left( \frac{1.96 * 49000}{2000} \right)^2$ , with  $\sigma = 49,000$  and  $m = 2000$ , we obtain a necessary

sample size of  $n = 2306$ . Working part (b) similarly with  $m = 1500$ , we obtain a required sample size of  $n = 4100$ .

## 6.2 Tests of Significance

We now show how to perform one-sided and two-sided hypothesis tests about the mean  $\mu$  of a normally distributed population for which the standard deviation  $\sigma$  is known. As with our confidence intervals, we really can only use SPSS as a calculator.

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

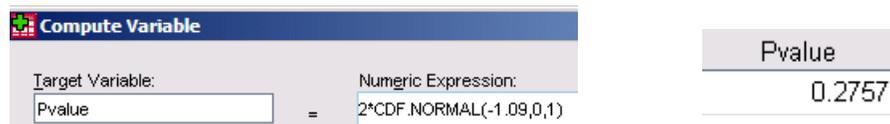
*Solution.* To test if the mean is *different* from 128, we use the null hypothesis  $H_0: \mu = 128$  with a two-sided alternative  $H_A: \mu \neq 128$ . We compute the  $z$  statistic for the test as



The image shows the SPSS 'Compute Variable' dialog box with 'zstat' as the target variable and the numeric expression  $(126.07-128)/(15/\text{sqrt}(72))$ . To the right, a small table shows the result for 'zstat' as -1.09.

zstat
-1.09

We have a  $z$  test statistic of  $-1.09$  and now need to find the  $p$ -value for the test. For this two-sided test, the  $p$ -value comes from the sum of both tail probabilities:  $P(z < -1.09) + P(z > 1.09)$ , so we double the area under the standard normal curve below  $z = -1.09$ .



The image shows the SPSS 'Compute Variable' dialog box with 'Pvalue' as the target variable and the numeric expression  $2*\text{CDF.NORMAL}(-1.09,0,1)$ . To the right, a small table shows the result for 'Pvalue' as 0.2757.

Pvalue
0.2757

If the true mean for all the company's executives in this age group were equal to 128, then there would be a 27.57% chance of obtaining an  $\bar{x}$  as far away as 126.07 with a sample of size 72. This rather high  $p$ -value does not give us good evidence to reject the null hypothesis. These executives *might* have a mean systolic blood pressure of 128; we have not shown their mean systolic blood pressure is different.

**Example 6.5 DRP Scores.** The following table gives the DRP scores for a sample of 44 third-grade students in a certain district. It is known that  $\sigma = 11$  for all such scores in the district. A researcher believes that the mean score of all third graders in this district is higher than the national mean of 32. State the appropriate  $H_0$  and  $H_A$ , then conduct the test and give the  $p$ -value.

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

*Solution.* Here we test  $H_0: \mu = 32$  with a one-sided alternative  $H_A: \mu > 32$ . Enter the data and use **Analyze, Descriptive Statistics, Descriptives** to find the sample mean.

	N	Minimum	Maximum	Mean	Std. Deviation
DRP	44	14	54	35.09	11.189
Valid N (listwise)	44				

As before, we'll compute the  $z$  statistic for the test and its one-sided  $p$ -value “manually.” Remember, since we have the one-sided “greater than” alternate hypothesis, we want the area above our calculated test statistic, so we need to use **1-CDF.Normal**.

We obtain a  $p$ -value of 0.0314. If the average of the district were equal to 32, then there would be only a 3.14% chance of a sample group of 44 averaging as high as  $\bar{x} = 35.09$ . There is strong evidence to reject  $H_0$  and conclude that the district's average DRP score is higher than 32.

### 6.3 Use and Abuse of Tests

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

**Example 6.6 SAT Coaching.** Suppose that SATM scores vary normally with  $\sigma = 100$ . Calculate the  $p$ -value for the test of  $H_0: \mu = 480$ ,  $H_A: \mu > 480$  in each of the following situations:

- A sample of 100 coached students yielded an average of  $\bar{x} = 483$ .
- A sample of 1000 coached students yielded an average of  $\bar{x} = 483$ .
- A sample of 10,000 coached students yielded an average of  $\bar{x} = 483$ .

*Solution.* We again use **Transform, Compute Variable** to find both the  $z$ -scores and  $p$ -values with the different sample sizes.

(a)

Compute Variable		zstata
Target Variable:	Numeric Expression:	0.30
zstata	$(483-480)/(100/\sqrt{100})$	

Compute Variable		Pvala
Target Variable:	Numeric Expression:	0.38
Pvala	$1-\text{CDF.NORMAL}(.3,0,1)$	

(b)

Compute Variable		zstatb
Target Variable:	Numeric Expression:	0.95
zstatb	$(483-480)/(100/\sqrt{1000})$	

Compute Variable		Pvalb
Target Variable:	Numeric Expression:	0.17
Pvalb	$1-\text{CDF.NORMAL}(.95,0,1)$	

(c)

Compute Variable		zstatac
Target Variable:	Numeric Expression:	3.00
zstatac	$(483-480)/(100/\sqrt{10000})$	

Compute Variable		Pvalc
Target Variable:	Numeric Expression:	0.00
Pvalc	$1-\text{CDF.NORMAL}(3,0,1)$	

We can clearly see the impact of the increasing sample size on the  $z$  statistics and their decreasing  $p$ -values. If the sample size is large enough, any (practically meaningless) difference between an observed value and  $H_0$  can be “statistically significant.”

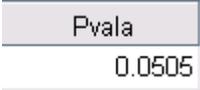
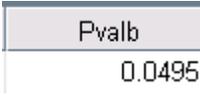
**Example 6.7 More SAT Coaching.** For the same hypothesis test as in Example 6.6 above, consider the sample mean of 100 coached students. (a) Is  $\bar{x} = 496.4$  significant at the 5% level? (b) Is  $\bar{x} = 496.5$  significant at the 5% level?

*Solution:* As above, we compute the  $z$ -statistic and its associated  $p$ -value for these sample means.

(a)

Compute Variable		zstata
Target Variable:	Numeric Expression:	1.64
zstata	$(496.4-480)/(100/\sqrt{100})$	

(b)

 <p>Target Variable: Pvala = Numeric Expression: 1-CDF.NORMAL(1.64,0,1)</p>	 <p>Pvala 0.0505</p>
 <p>Target Variable: zstatb = Numeric Expression: (496.5-480)/(100/sqrt(100))</p>	 <p>zstatb 1.65</p>
 <p>Target Variable: Pvalb = Numeric Expression: 1-CDF.NORMAL(1.65,0,1)</p>	 <p>Pvalb 0.0495</p>

In the first case,  $p = 0.0505 > 0.05$ ; so the value of  $\bar{x} = 496.4$  is not significant at the 5% level. However, in the second case,  $p = 0.0495 < 0.05$ ; so the value of  $\bar{x} = 496.5$  is significant at the 5% level. However, for SATM scores, there is no real “significant” difference between means of 496.4 and 496.5. There is nothing magical about the 0.05 significance level; the smaller the  $p$ -value the more evidence we have against  $H_0$ .

## 6.4 Power and Inference as a Decision

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

**Example 6.8 More California SATs.** In Example 6.6 we considered the hypotheses  $H_0: \mu = 450$ ,  $H_A: \mu > 450$ . Suppose a sample of size  $n = 500$  is taken from a normal population having  $\sigma = 100$ . If we performed our test at the 1% level of significance, find the power of this test against the alternative  $\mu = 462$ .

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

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

or equivalently if  $\bar{x} > 450 + 2.326 \cdot 100/\sqrt{500} = 460.4022$ . Now we must find the probability that  $\bar{x}$  is greater than 460.4022, given that the alternative  $\mu = 462$  is true.

Given that  $\mu = 462$ , then  $\bar{x} \sim N(462, 100/\sqrt{500} = 4.472)$ , and we must compute  $P(\bar{x} > 460.4022)$ . To do so, we use **1-CDF.Normal(460.4022,462,4.472)** and find that the power against the alternative  $\mu = 462$  is about 0.64.

Compute Variable		Power
Target Variable:	Numeric Expression:	
Power	= 1-CDF.NORMAL(460.4022,462,4.472)	0.6396

**Example 6.9** (a) An SRS of size 584 is taken from a population having  $\sigma = 58$  to test the hypothesis  $H_0: \mu = 100$  versus a two-sided alternative at the 5% level of significance. Find the power against the alternative  $\mu = 99$ .

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

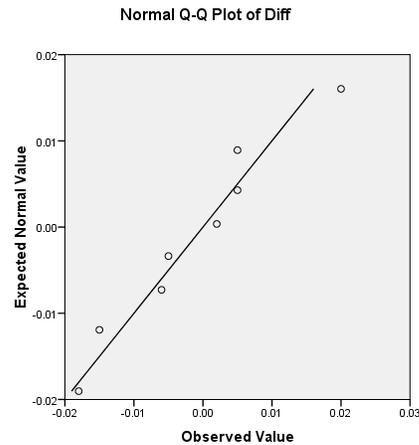
$$\frac{\bar{x} - 100}{58/\sqrt{584}} < -1.96 \quad \text{or} \quad \frac{\bar{x} - 100}{58/\sqrt{584}} > 1.96$$

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

Compute Variable		Power
Target Variable:	Numeric Expression:	
Power	= CDF.Normal(95.2959,99,2.400)+1-CDF.Normal(104.7041,99,2.400)	0.0701

# CHAPTER

# 7



## Inference for Distributions

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

### Introduction

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

Just as with computing confidence intervals or hypothesis tests when  $\sigma$  is “known,” SPSS can really only be used as a calculator when we have only summary statistics. Its power comes into play when there are actual data.

## 7.1 Inference for the Mean of a Population

We begin with a short exercise that allows us to find a critical value  $t^*$  upon specifying the degrees of freedom and confidence level. This makes use of our old friend **Transform, Compute Variable**.

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

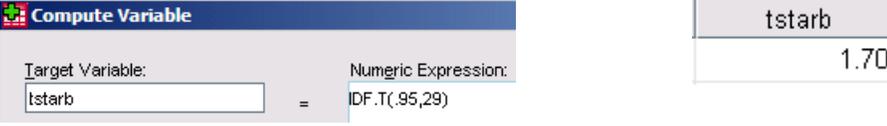
- (a) A 95% confidence interval based on  $n = 20$  observations
- (b) A 90% confidence interval from an SRS of 30 observations
- (c) An 80% confidence interval from a sample of size 50

*Solution.* The confidence intervals are based on  $t$  distributions with  $n - 1$  degrees of freedom, since we have single samples. So we need 19 degrees of freedom for part (a), 29 degrees of freedom for part (b), and 49 degrees of freedom for part (c). Since our confidence region is in the middle of the curve, we add half of the leftover area to the desired amount of confidence to find the area to the left of  $t^*$ .

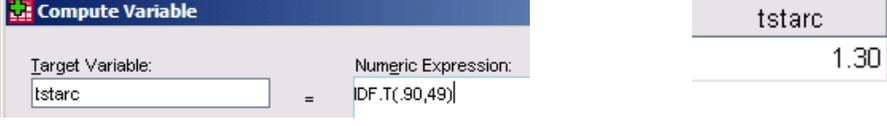
We will use **Idf.T** from the Inverse DF function group.

(a) 

tstara
2.09

(b) 

tstarb
1.70

(c) 

tstarc
1.30

### One-sample $t$ Confidence Interval

We now examine confidence intervals for one mean. SPSS is happiest if we have data in the worksheet; if we have only summary statistics, we'll work like we did in Chapter 6, but use CDF.T to find the  $p$ -value.

**Example 7.2 Vitamin C in Corn.** The amount of vitamin C in a factory’s production of corn soy blend (CSB) is measured from 8 samples giving  $\bar{x} = 22.50$  (mg/100 g) and  $s = 7.19$ . Find a 95% confidence interval for the mean vitamin C content of the CSB produced during this run.

*Solution.* Here we have only summary statistics, so we will “manually” find the upper and lower ends of the confidence interval. We first find the correct value of  $t^*$  as we did in Example 7.1. We then compute the lower and upper ends of the interval.

The image shows three sequential screenshots of the SPSS 'Compute Variable' dialog box. Each dialog box has a 'Target Variable' field and a 'Numeric Expression' field. To the right of each dialog box is a small window showing the result of the calculation.

- First dialog:** Target Variable: `tstar`; Numeric Expression: `IDF.T(.975,7)`. Result: `2.36`.
- Second dialog:** Target Variable: `lower`; Numeric Expression: `22.5-2.36*7.18/sqrt(8)`. Result: `16.51`.
- Third dialog:** Target Variable: `upper`; Numeric Expression: `22.5+2.36*7.18/sqrt(8)`. Result: `28.49`.

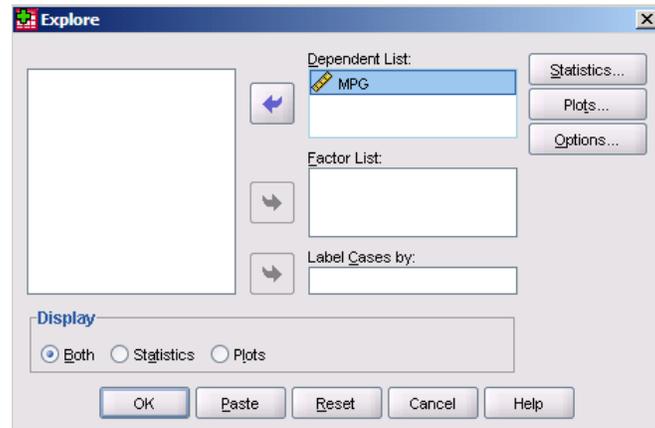
So we can say that based on this sample, with 95% confidence, the mean vitamin content in this run is between 16.51 and 28.49 mg/100g.

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

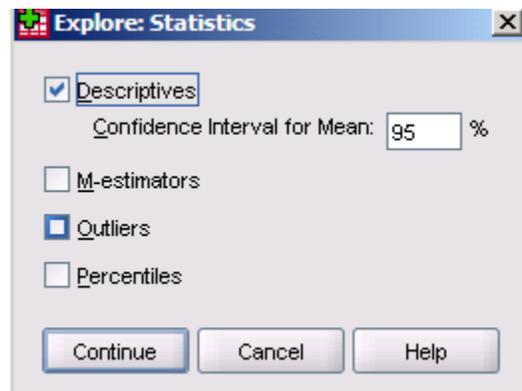
15.8	13.6	15.6	19.1	22.4	15.6	22.5	17.2	19.4	22.6
19.4	18.0	14.6	18.7	21.0	14.8	22.6	21.5	14.3	20.9

*Solution.* First, define variable **MPG** with one decimal place, and input the data. SPSS can give us all we need—the summary statistics, the confidence interval, and a stem-and-leaf plot (to check that our data are approximately Normally distributed) using **Analyze**, **Descriptive Statistics**, **Explore**.

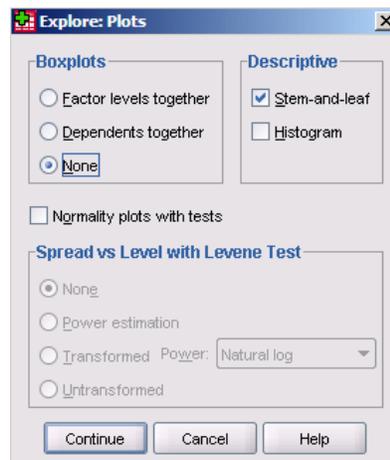
First, click to add variable **MPG** in the Dependent List. Notice at the bottom left we are asking for both Statistics and Plots. Click Statistics to see what will be generated.



We are asking for **Descriptives** and a 95% confidence interval for the mean. If you want another confidence level, here's where you would indicate that. Click **Continue** to return to the main dialog box.



We can get boxplots (including side-by-side boxplots if we have independent samples, histograms, Normal quantile plots, and stem-and-leaf plots). For this simple one-sample data set, just a stem-and-leaf plot will be enough. Click Continue to return to the main dialog box, and click OK to process our request.



As indicated below, we have a sample mean of 18.48 miles per gallon. The 95% confidence interval indicates we are 95% confident mean gas mileage for this vehicle is between 17.02 and 19.94 miles per gallon (rounding to one more place than the original data).

		Statistic	Std. Error
MPG	Mean	18.480	.6967
	95% Confidence Interval for Mean		
	Lower Bound	17.022	
	Upper Bound	19.938	
	5% Trimmed Mean	18.522	
	Median	18.900	
	Variance	9.708	
	Std. Deviation	3.1158	
	Minimum	13.6	
	Maximum	22.6	
	Range	9.0	
	Interquartile Range	5.8	
	Skewness	-.081	.512
	Kurtosis	-1.464	.992

Finally, we see the stem-and-leaf plot. Since this is roughly symmetric with no outliers, our inference is appropriate.

MPG Stem-and-Leaf Plot

Frequency	Stem &	Leaf
4.00	1 .	3444
9.00	1 .	555788999
7.00	2 .	0112222
Stem width:	10.0	
Each leaf:	1 case(s)	

### One-Sample $t$ test

We now perform some significance tests about the mean using **One-Sample T Test** from the **Analyze, Compare Means** menu. We begin with an example using already summarized data.

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

*Solution.* We compute the  $t$ -statistic for the test using our old friend, **Transform, Compute Variable**. Since there were eight data values, our  $t$ -statistic has 7 degrees of freedom when we compute the  $p$ -value.

tstat
-6.88

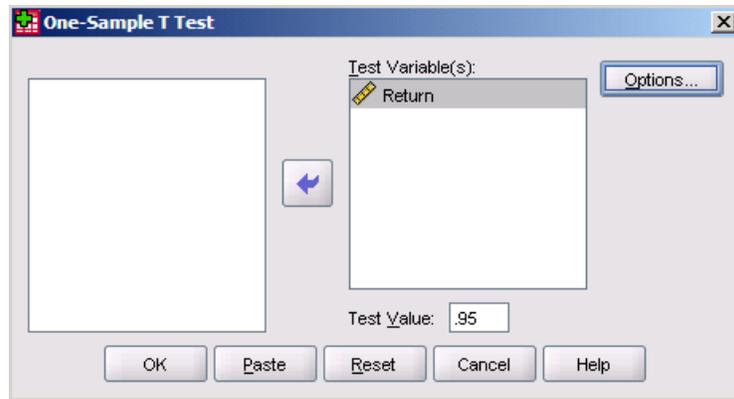
Pvalue
0.0001

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

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

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

*Solution.* Define a variable (we have called it **Return**) and enter the data. Then select **Analyze, Compare Means, One-Sample T Test**. If you click the Options button at the right, you can change the confidence level for the interval that is also calculated from the default  $95\%$ .



We are first given basic summary statistics for the variable **Return**.

	N	Mean	Std. Deviation	Std. Error Mean
Return	39	-1.0997	5.99089	.95931

Next is the inference information.

	Test Value = .95					
	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
					Lower	Upper
Return	-2.137	38	.039	-2.04974	-3.9918	-.1077

We obtain a  $p$ -value of 0.039. If the average return were equal to 0.95%, then there would be only a 3.9% chance of a sample of 39 months averaging as far away as  $-1.1\%$ . There is sufficient evidence to reject  $H_0$  and conclude that the mean monthly return differs from (was worse than) 0.95%. Further, the 95% confidence interval ( $-3.992$  to  $-0.108$ ) is clearly well below 0.95. This confirms our idea that the broker was probably mismanaging the funds; at least, his performance was significantly worse than the S&P 500 during this time.

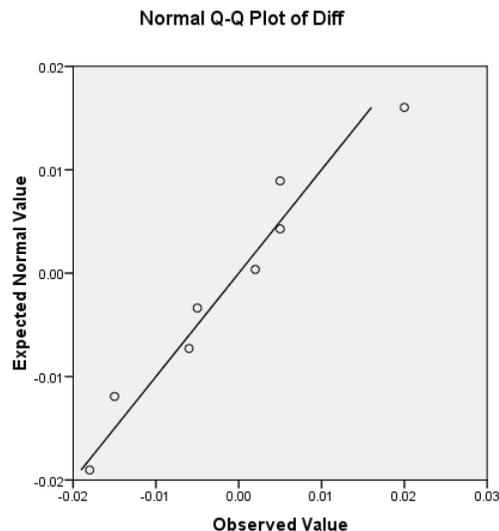
### Matched Pair $t$ Procedure

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

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

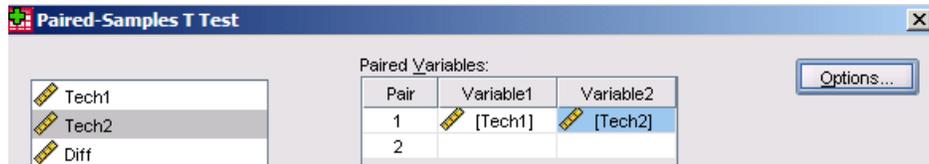
Use a significance test to examine the null hypothesis that the two operators have the same mean. Use a 95% confidence interval to provide a range of differences that are compatible with these data.

*Solution.* These are paired data because each subject was measured by both technicians. Two variables (called **Tech1** and **Tech2**) have been defined and the data entered. Paired data work with the differences. Since this is a small sample, we should check that the differences are approximately Normally distributed (that there is no evidence of skewness or outliers). We can make a Normal quantile plot of the differences (after having computed them using **Transform, Compute Variable**). We use **Analyze, Descriptive Statistics, Q-Q Plots** for the variable **Diff**.



Since all the differences fall roughly around the line we can believe the differences are approximately Normally distributed. We may proceed with our test. This can be done in either of two ways (since we have already computed the differences). From **Analyze, Compare Means**, we could use a **One-Sample T Test** using the differences, or a **Paired Samples T Test** using the original data. Both will give the same results. Using the Paired

Samples test, the completed dialog box should look like the one below. The **Options** button is again where one can change the confidence level for the interval which is also computed.



There are three tables generated for output. The first gives sample statistics for each technician.

**Paired Samples Statistics**

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	Tech1	1.17250	8	.151810	.053673
	Tech2	1.17400	8	.149481	.052850

The second gives the correlation between the paired samples. Here, we see the two technicians are almost perfectly correlated.

**Paired Samples Correlations**

		N	Correlation	Sig.
Pair 1	Tech1 & Tech2	8	.997	.000

The third gives the inference results. Since this table is too wide to fit on a page of this manual, I show below the important parts for our purposes.

		95% Confidence Interval of the Difference		t	df	Sig (2-tailed)
		Lower	Upper			
Pair 1	Tech1 - Tech2	-.011720	.008720	-.347	7	.739

First, the difference has been calculated as **Tech1 – Tech2**. It is important to keep in mind the direction of the subtraction. The confidence interval for the mean difference spans 0; this indicates that there is not a significant difference in measurements by the two technicians, on average. This is further confirmed by a *t*-statistic of  $-.347$  with a *p*-value of  $0.739$ . It appears that results from the two technicians *are* consistent.

## The Sign Test

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

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

*Solution.* Because so many patients had a change in behavior, we shall test the hypothesis  $H_0: p = 0.50$  with the alternative  $H_A: p > 0.50$ . We must compute the probability of there being as many as 14 changes with a  $B(15,0.50)$  distribution. Equivalently, we can compute the probability of there being as few as one non-change. Thus, we could compute either  $P(B \geq 14) = 1 - P(B \leq 13)$  or  $P(B \leq 1)$ . To do so, we use **CDF.Binomcdf** from **Transform, Compute Variable**.



We obtain the very low  $p$ -value of 0.0005. If there were no moon effect, then there would be almost no chance of having as many as 14 out of 15 showing a change. Therefore, we can reject  $H_0$  in favor of the alternative that the moon generally causes more aggressive behavior.

## 7.2 Comparing Two Means

We next consider confidence intervals and significance tests for the difference of means  $\mu_1 - \mu_2$  given two normal populations that have unknown standard deviations. The results are based on independent random samples of sizes  $n_1$  and  $n_2$ . SPSS requires the data for **Analyze, Compare Means, Independent Samples T Test** be entered so that one variable indicates group membership and a second contains the actual data. If the data are already summarized, you can use SPSS simply as a calculator as had already been shown with the one-sample  $t$  test and intervals.

**Example 7.8 College Study Habits.** The Survey of Study Habits and Attitudes was given to first-year students at a private college. The tables on the next page are a random sample of the scores.

**Women's scores**

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

**Men's scores**

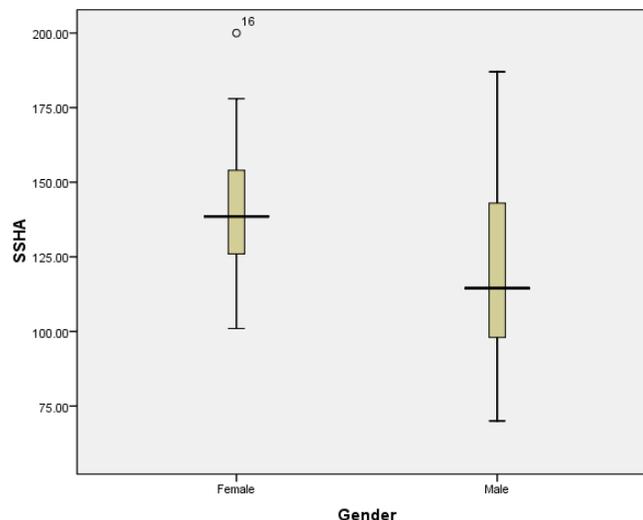
108	140	114	91	180	115	126	92	169	146
109	132	75	88	113	151	70	115	187	104

- (a) Examine each sample graphically to determine if the use of a  $t$  procedure is acceptable.
- (b) Test the supposition that the mean score for all men is lower than the mean score for all women among first-year students at this college.
- (c) Give a 90% confidence interval for the mean difference between the SSHA scores of male and female first-year students at this college.

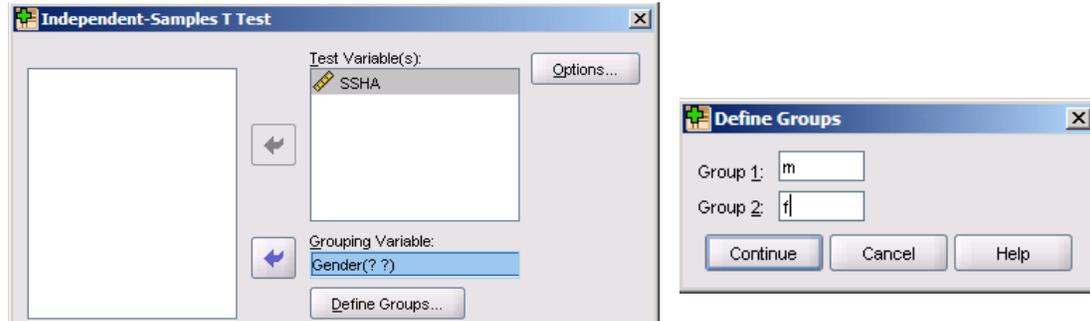
*Solution.* (a) First define a variable for **Gender** and one for the **SSHA** score and enter the data. Define a **Graphs**, **Legacy Dialogs**, **Boxplot** with **Summaries for groups of cases**.



The boxplot indicates one outlier in the female distribution; however, it is mild (and a normal plot is reasonably straight) so we will rely on the robustness of  $t$  procedures in continuing our analysis.



(b) Next, let  $\mu_1$  be the mean SSHA score among all first-year women and let  $\mu_2$  be the mean score among all first-year men. We shall test the hypothesis  $H_0: \mu_1 = \mu_2$  versus the alternative  $H_A: \mu_1 > \mu_2$ . We use **Analyze, Compare Means, Independent Samples T Test** for our inference about these population means. The Test Variable is **SSHA** and the Grouping Variable is **Gender**. Notice we have question marks for the actual groups. Click **Define Groups**.



Once groups have been defined, click **Continue** to return to the main dialog box. Since we want a 90% confidence interval for the difference in means, click **Options** to change the confidence level from the default 95%. Click **OK** to start the procedure.

The first part of the output is sample statistics for each group. Females had a mean of 141.06 (notice the E2 in the output) while males had a mean of 121.25. The females had a higher average score, but is it different enough to be statistically significant, or is it due merely to chance?

	Gender	N	Mean	Std. Deviation	Std. Error Mean
SSHA	Male	20	1.2125E2	32.85194	7.34592
	Female	18	1.4106E2	26.43632	6.23110

The next is the  $t$  test and confidence results. To fit the page, the output table has been split.

		Levene's Test for Equality of Variances	
		F	Sig.
SSHA	Equal variances assumed	.862	.359
	Equal variances not assumed		

This first part gives the result of a test of  $H_0: \sigma_1^2 = \sigma_2^2$  against  $H_A: \sigma_1^2 \neq \sigma_2^2$ . In this case, with a large  $p$ -value, it is possible males and females have the same variance in their SSHA scores. The second portion of the table gives the results both (top line) assuming equal variances and (bottom line) not assuming equal variances. In this case, we have a

slight difference in the computed  $t$  statistics, as well as the  $p$ -values. Be careful—SPSS gives two-sided  $p$ -values. Since our alternate hypothesis of interest was  $\mu_1 > \mu_2$  we need to divide these  $p$ -values by two to get the correct one-sided value. Either way, we have a  $p$ -value of 0.025 or 0.0235.

t-test for Equality of Means						
t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	90% Confidence Interval of the Difference	
					Lower	Upper
-2.032	36	.050	-19.80556	9.74479	-36.25767	-3.35344
-2.056	35.587	.047	-19.80556	9.63271	-36.07342	-3.53769

If the true means were equal, then there would be only a 2.35% chance of  $\bar{x}_1$  being so much larger than  $\bar{x}_2$  with samples of these sizes. The relatively low  $p$ -value gives us evidence to reject  $H_0$  and conclude that  $\mu_1 > \mu_2$ . That is, the mean score for all men is lower than the mean score for all women among first-year students at this college.

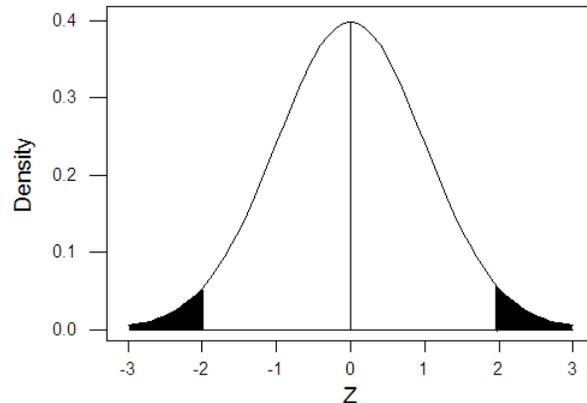
(c) Reading the confidence interval (from the bottom row, where equal variances are not assumed), we are 90% confident that males at this college have SSHA scores between 3.54 and 36.07 point lower on average than the females. (Remember, we have males as group 1 in our group definition.)

### 7.3 Optional Topics in Comparing Distributions

We have already seen SPSS compute a Levene's test for equality of variances. This is related to, but different from, the  $F$  test mentioned in your text. The  $F$  test is extremely sensitive to any departures from normality and is generally considered to be very dangerous outside the hands of a statistician; Levene's test is not as sensitive. SPSS, in fact, does not do the  $F$  test described in your text (although you could perform the calculations yourself using **Transform, Compute Variable**).

## CHAPTER

# 8



# Inference for Proportions

8.1	Inference for a Single Proportion
8.2	Comparing Two Proportions

## Introduction

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

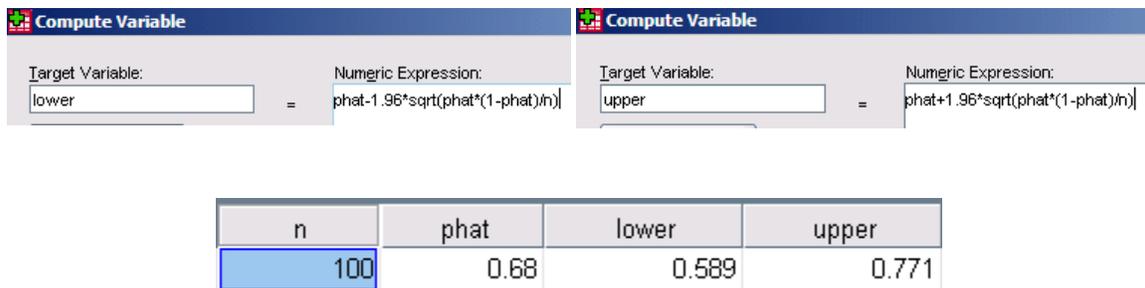
Once again, SPSS does not have built-in functions for these procedures; we'll use **Transform**, **Compute Variable** to do our calculations. If you had a variable that consisted of 0's and 1's, you could use the One-Sample T procedures (for large samples) to approximate the interval and test, since  $t$  distributions become closer to standard Normal as degrees of freedom (sample sizes) become large.

## 8.1 Inference for a Single Proportion

### A Large-Sample Confidence Interval

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

*Solution.* Define variables **n** and **phat** in the Variable View. Enter 100 for **n** and .68 (68/100) for **phat**. Remember that for 95% confidence,  $z^* = 1.96$ . Now we will use **Transform**, **Compute Variable** to find the lower and upper ends of the confidence interval.



n	phat	lower	upper
100	0.68	0.589	0.771

Having set decimal places to three for both **lower** and **upper**, we can say with 95% confidence that based on this information, between 58.9% and 77.1% of restaurant workers would agree that work stress has had a negative impact on their personal lives.

### A Plus-Four Confidence Interval

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

Because we have a small number of studies that discussed blinding (fewer than 10), we use the “plus four” method: we simply add four to the number of studies (it becomes 101) and two to the number of “successes” (we’ll consider that nine discussed blinding). After that, we compute the confidence interval just as we did above.

*Solution.* We find **phat** (really  $\tilde{p}$ ) is  $9/101 = .0891$ . We enter that and **n** (the adjusted 101) and proceed to compute the lower and upper ends of the interval. Increase the number of decimal places on **phat**, if needed.

n	phat	lower	upper
101	0.0891	0.034	0.145

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

### Choosing a Sample Size

As with confidence intervals for the mean, we often would like to know in advance what sample size would provide a certain maximum margin of error  $m$  with a certain level of confidence. The required sample size  $n$  satisfies

$$n \geq \left( \frac{z^*}{m} \right)^2 p^*(1-p^*)$$

where  $z^*$  is the appropriate critical value depending on the level of confidence and  $p^*$  is a guessed value of the true proportion  $p$ . If  $p^* = 0.50$ , then the resulting sample size insures that the margin of error is no more than  $m$ , regardless of the true value of  $p$ .

**Example 8.3 Alcohol Awareness—Find a Sample Size.** Among students who completed an alcohol awareness program, you want to estimate the proportion who state that their behavior towards alcohol has changed since the program. Using the guessed value of  $p^* = 0.30$  from previous surveys, find the sample size required to obtain a 95% confidence interval with a maximum margin of error of  $m = 0.10$ .

*Solution.* Again we use **Transform, Compute Variable** to find our sample size. We'll define variables **pstar** and **m** and use these to find **n**. Enter .3 for **pstar** and .10 for **m**, then compute the sample size as shown below.

Compute Variable		n	pstar	m
Target Variable:	Numeric Expression:	80.67	0.30	0.10
n	(1.96/m)**2*pstar*(1-pstar)			

Remember, this is the smallest sample size that will give the desired margin of error. For any fraction or decimal part in the computed **n** we must go to the next larger sample size, and not just round. In this case, we'll need a sample of at least 81 to get the desired margin of error.

### Significance Tests

We now show how to conduct hypothesis tests for a single population proportion  $p$ .

**Example 8.4 More Stress in Restaurant Workers.** In the restaurant worker survey, 68 of a sample of 100 employees agreed that work stress had a negative impact on their

personal lives. Let  $p$  be the true proportion of restaurant employees who agree. Test the hypothesis  $H_0: p = 0.75$  versus  $H_A: p \neq 0.75$ .

*Solution.* Define variables **n**, **phat**, and **p0** as in Example 8.1 and enter 100 for **n**, .68 for **phat**, and .75 for **p0**. Use **Transform, Compute Variable** to find the  $z$  statistic and  $p$ -value as shown below. Since our alternate hypothesis is of the two-tailed form, we multiply the area below our calculated  $z$  statistic by 2 to find the  $p$ -value. (Look at the  $z$  statistic before computing the  $p$ -value to be safe here.)

**Compute Variable**

Target Variable: zstat = Numeric Expression: (phat-p0)/sqrt(p0\*(1-p0)/n)

**Compute Variable**

Target Variable: Pvalue = Numeric Expression: 2\*CDF.NORMAL(zstat,0,1)

n	phat	p0	zstat	Pvalue
100.00	0.68	0.75	-1.62	0.1060

With a  $p$ -value of 0.1060, we fail to reject the null hypothesis; the proportion of restaurant workers whose work stress has negatively impacted their personal life *may* be 75%; we have failed to show this is wrong. Compare these results with the confidence interval found in Example 8.1—notice that 75% *is* included in that interval.

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

*Solution.* We calculate this as we did in example 8.4, but notice that the alternate hypothesis is a “less than” one-tailed one, so we simply use **CDF.Normal** to find the  $p$ -value. We obtain a test statistic of  $-2.53$  and a  $p$ -value of 0.0057. If  $p$  were 0.50, then there would be only a 0.0057 probability of  $\hat{p}$  being as low as 0.3 with 40 subjects. There is strong evidence to reject  $H_0$ , and conclude that those who prefer instant coffee are a minority of the population.

**Compute Variable**

Target Variable: zstat = Numeric Expression: (phat-p0)/sqrt(p0\*(1-p0)/n)

**Compute Variable**

Target Variable: Pvalue = Numeric Expression: CDF.NORMAL(zstat,0,1)

n	phat	p0	zstat	Pvalue
40.00	0.30	0.50	-2.53	0.0057

## 8.2 Comparing Two Proportions

We now demonstrate confidence intervals and significance tests for the difference of two population proportions  $p_1$  and  $p_2$ . These calculations are again done using **Transform**, **Compute Variable**.

### A Large-Sample Confidence Interval for a Difference of Proportions

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

Population	$n$	$X$
Men	7180	1630
Women	9916	1684

*Solution.* Call the men group 1 and the women group 2. Define variables **n1**, **n2**, **x1**, and **x2**. We’ll first find the observed proportions **Phat1** and **Phat2**. We do not show the computation for **Phat2**, but only its result. We have also increased the displayed decimal places on these sample proportions from the default 2.

Compute Variable dialog box showing Target Variable: Phat1 and Numeric Expression: x1/n1.

n1	n2	x1	x2	Phat1	Phat2
7180.00	9916.00	1630.00	1684.00	0.2270	0.1698

We now find the lower end of the confidence interval (to find the upper, change the minus before 1.96 to a plus).

Compute Variable dialog box showing Target Variable: lower and Numeric Expression: (Phat1-Phat2)-1.96\*SQRT(Phat1\*(1-Phat1)/n1+Phat2\*(1-Phat2)/n2).

lower	upper
0.0450	0.0694

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

### A Plus–Four Confidence Interval for a Difference of Proportions

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

*Solution.* We proceed as before to find a plus-four confidence interval for  $p_1 - p_2$ . But for  $\mathbf{x1}$  and  $\mathbf{x2}$ , enter 1 more than the actual number of positive responses (we split the two successes added in the one sample case). For  $\mathbf{n1}$  and  $\mathbf{n2}$ , enter 2 more than the actual sample sizes (splitting the added four trials). Here, enter **5** for  $\mathbf{x1}$ , **4** for  $\mathbf{x2}$ , and enter **14** for both  $\mathbf{n1}$  and  $\mathbf{n2}$ .

n1	n2	x1	x2	Phat1	Phat2
14.00	14.00	5.00	4.00	0.3571	0.2857

lower	upper
-0.2735	0.4164

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

### Significance Tests for a Difference of Proportions

We now show how to conduct hypothesis tests about  $p_1 - p_2$ .

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

Population	$n$	$X$
Men	7180	1630
Women	9916	1684

*Solution.* Let  $p_1$  be the true proportion of male students who are frequent binge drinkers, and let  $p_2$  be the true proportion of female students who are frequent binge drinkers. We shall test the hypothesis  $H_0: p_1 = p_2$  versus the alternative  $H_A: p_1 \neq p_2$ . Since under the

null hypothesis, the proportion should be the same in both groups, we “pool” the “successes” (as if binge drinking is a success) and sample sizes to find a blended, overall sample proportion, **Phat**. Since the alternate hypothesis is two-tailed, we double the area above our calculated statistic to find the  $p$ -value for the test.

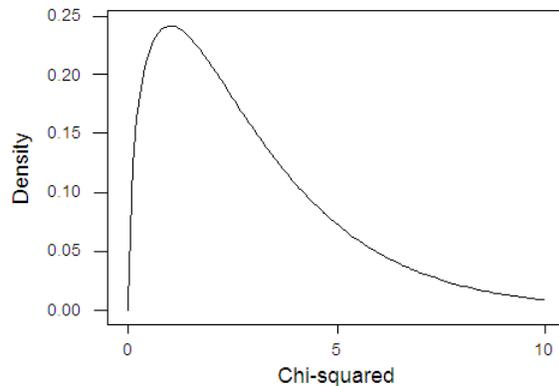
<b>Compute Variable</b> Target Variable: Phat = Numeric Expression: (x1+x2)/(n1+n2)	Phat 0.1938
<b>Compute Variable</b> Target Variable: zstat = Numeric Expression: (p1-p2)/sqrt(Phat*(1-Phat)*(1/n1+1/n2))	zstat 9.34
<b>Compute Variable</b> Target Variable: Pvalue = Numeric Expression: 2*(1-CDF.NORMAL(9.34,0,1))	Pvalue 0.0000

We obtain a  $p$ -value (to four decimal places) of 0 from a test statistic of  $z = 9.34$ . If  $p_1 = p_2$  were true, then there should be no chance of obtaining sample proportions as far apart as  $\hat{p}_1 = 0.227$  and  $\hat{p}_2 = 0.1698$  with samples of these sizes. So, we can reject  $H_0$  and conclude that not only are the genders different in (admitting to) binge drinking, but that males are more likely to engage in this (or at least to admit to it). This reinforces our results obtained in Example 8.6; the confidence interval was wholly positive, so there would be no belief that the two proportions should be the same.

## CHAPTER

# 9

A Chi-squared Distribution



# Inference for Two-Way Tables

9.1	Data Analysis for Two-Way Tables
9.2	Inference for Two-Way Tables
9.3	Goodness of Fit

## Introduction

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

As always, SPSS is happiest with raw data in the data sheet; it will then construct the table for you and conduct the  $\chi^2$  test of association. When data have already been summarized, we need to enter our information differently from the usual matrix form.

## 9.1 Data Analysis for Two-Way Tables

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

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

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

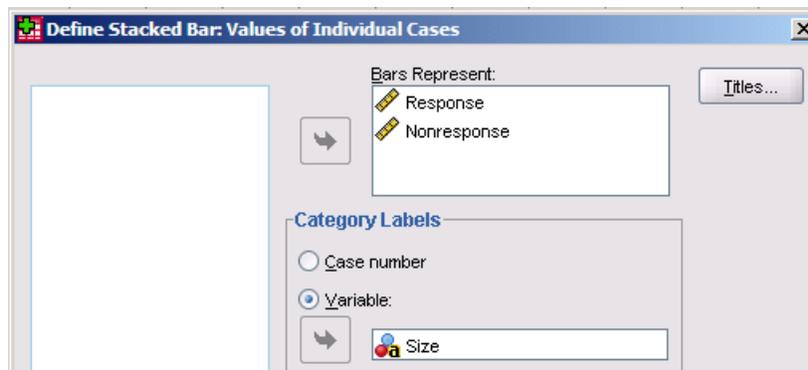
*Solution.* (a) To answer this question, we divide the total number of responses ( $125 + 155 + 180 = 460$ ) by the total number of sent questionnaires (900).  $460/900 = .5111$ . Overall, 51.1% of the businesses did not respond.

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

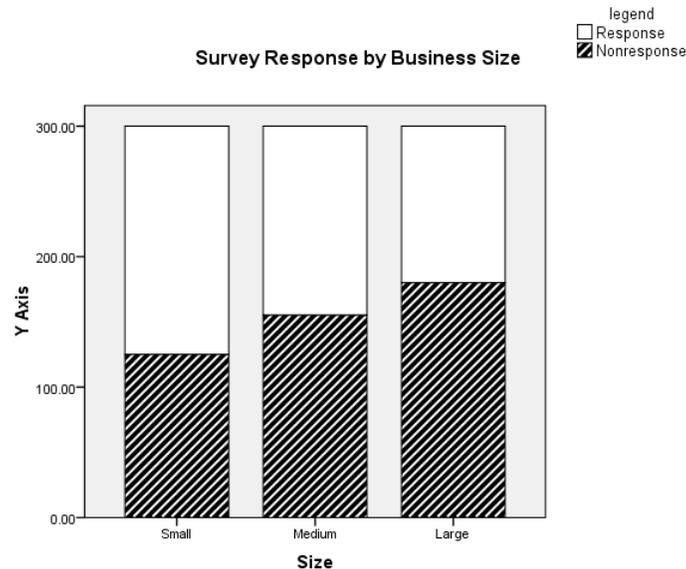
(c) To make a segmented bar graph that will dramatically illustrate the differences, first define variables and enter the data as shown below.

Size	Response	Nonresponse
Small	175.00	125.00
Medium	145.00	155.00
Large	120.00	180.00

Now use **Graphs**, **Legacy Dialogs**, **Bar** and define a **Stacked** Bar Chart where Data in Chart are **Values of Individual Cases**. Be sure to give the graph a title.



Click **OK** to generate the graph. The default graph will use colors to responding and nonresponding businesses; right-click in the graph and select **Edit Content in Separate Window** to change the fill to a pattern (for a black-and-white printer). In the **Properties** dialog box, click **Variables** and change style from Color to **Pattern**. Also, these samples were of equal size; in case they are not, click Options, Scale to 100% to make all bars show 100% of their type. My plot is below.



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

## 9.2 Inference for Two-Way Tables

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

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

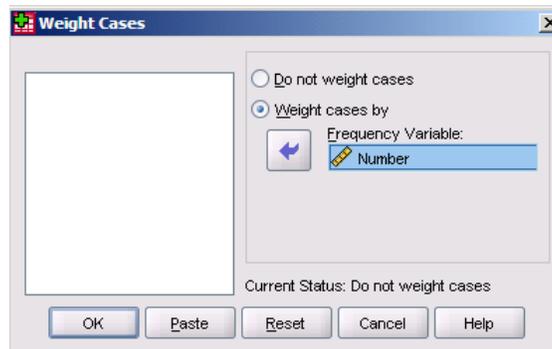
	Observed number of firms		Total
	Exclusive territory		
Success	Yes	No	Total
Yes	108	15	123
No	34	13	47
Total	142	28	170

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

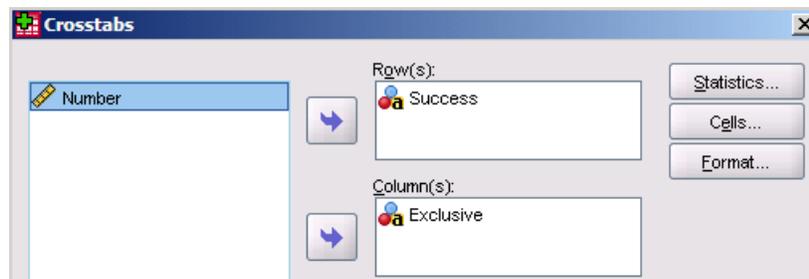
*Solution.* Here, we need to enter the data somewhat differently than we did in the previous example. We'll enter string variables to represent success (or not) and exclusive territory (or not). The observed counts for each type will then be entered as seen below.

Exclusive	Success	Number
Yes	Yes	108
Yes	No	34
No	Yes	15
No	No	13

We use the observed number in each category to weight the outcomes. Click **Data**, **Weight Cases**.



Now we're ready to compute the  $\chi^2$  test. Click **Analyze**, **Descriptive Statistics**, **Crosstabs**. Since our original table had Success as the row variable, we will use the same orientation here.



To obtain expected counts (or percents) click **Cells** and check the appropriate boxes. Click **Statistics** to obtain the  $\chi^2$  statistic and its  $p$ -value (otherwise, you'll just get the table).

The first part of the output gives the case processing summary that indicates how many valid and missing entries there are; this is not shown here. The second part gives the contingency table and (because it was asked for) the expected count for each cell.

**Success \* Exclusive Crosstabulation**

			Exclusive		Total
			No	Yes	
Success	No	Count	13	34	47
		Expected Count	7.7	39.3	47.0
	Yes	Count	15	108	123
		Expected Count	20.3	102.7	123.0
Total		Count	28	142	170
		Expected Count	28.0	142.0	170.0

The third gives the test results. The statistic of interest to us is given in the first row of the table.

**Chi-Square Tests**

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	5.911 <sup>a</sup>	1	.015		
Continuity Correction <sup>b</sup>	4.841	1	.028		
Likelihood Ratio	5.465	1	.019		
Fisher's Exact Test				.021	.016
N of Valid Cases <sup>b</sup>	170				

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 7.74.

b. Computed only for a 2x2 table

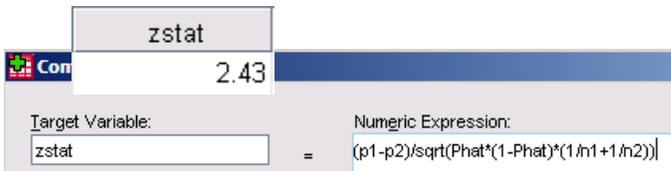
*Conclusions.* Here, the low  $p$ -value of 0.015 gives strong evidence to reject the claim that there is no relationship between exclusive territory rights and franchise success. If there were no relationship between success and exclusive territory rights, then we would expect 102.7 successful exclusive-territory franchises and 20.3 successful non-exclusive-territory franchises, as shown in the second row. These values differ slightly from the observed values of 108 and 15 in the original data. The actual number of successful franchises in exclusive territory was somewhat higher than expected, and the actual number of successful franchises in non-exclusive-territory somewhat lower than expected. It appears that having exclusive territory does help a franchise be successful.

### Comparison with the Two Proportion Z Test

The  $\chi^2$  test for a  $2 \times 2$  table is equivalent to the two-sided  $z$  test for  $H_0: p_1 = p_2$  versus  $H_A: p_1 \neq p_2$ . In Example 9.2, we could let  $p_1$  be the true proportion of all successful exclusive-territory franchises and let  $p_2$  be the true proportion of all successful non-exclusive-territory franchises. Then  $\hat{p}_1 = 108/142$  and  $\hat{p}_2 = 15/28$ . If we use this information and compute a two-sample test of proportions as seen in Chapter 8, we obtain the following:



n1	n2	x1	x2	Phat	p1	p2
142.00	28.00	108.00	15.00	0.72	0.76	0.54



Pvalue
0.0151

The  $p$ -value found here is identical (rounded to three decimal places) to that given by the  $\chi^2$  test.

**Example 9.3 Student Loans.** A recent study of 865 college students found that 42.5% had student loans. The following table classifies the students by field of study and whether or not they have a loan.

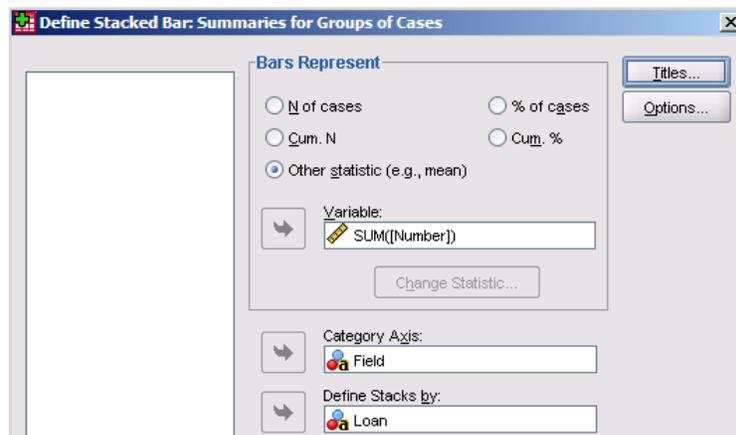
Field of study	Student loan	
	Yes	No
Agriculture	32	35
Child development and family studies	37	50
Engineering	98	137
Liberal arts and education	89	124
Management	24	51
Science	31	29
Technology	57	71

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

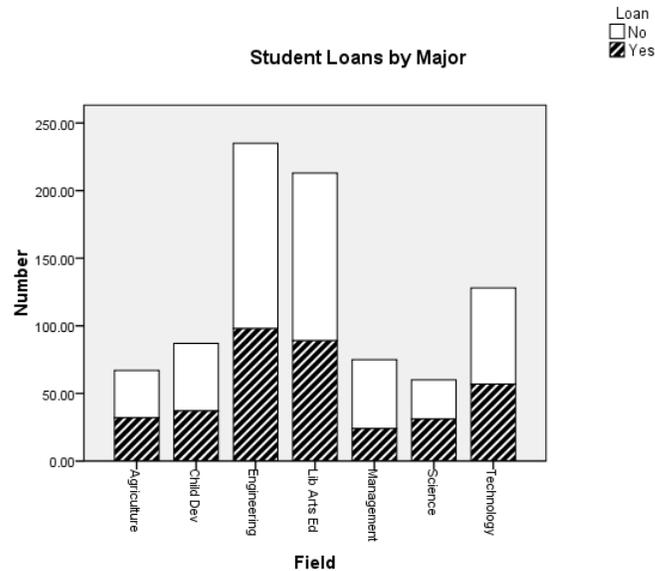
*Solution.* Define variables **Field**, **Loan**, and **Number**. **Field** and **Loan** will be string variables to hold the major and whether or not the student has a loan (Yes/No). Enter the data like the sample shown below.

Field	Loan	Number
Agriculture	Yes	32.00
Agriculture	No	35.00
Child Dev	Yes	37.00
Child Dev	No	50.00
Engineering	Yes	98.00
Engineering	No	137.00
Lib Arts Ed	Yes	89.00
Lib Arts Ed	No	124.00

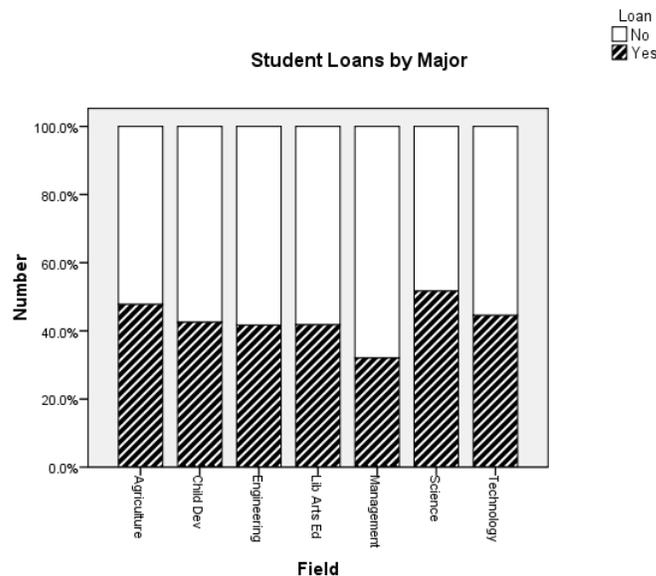
To make a stacked bar graph of these data, select **Graphs, Legacy Dialogs, Bar**. We want a **Stacked** bar graph where values are **Summaries of Groups of Cases**. My plot Definition is below. Notice that bars represent the Sum of **Number** (click to enter **Number** into this box, then click **Change Statistic** as it defaults to Mean). Add a title, and click OK to generate the initial graph.



I've changed this initial graph from Color to Pattern using **Edit Contents In Separate Window** as was detailed in Example 9.1. The bars are all different heights because there were very different numbers of students in the different majors. This is a good example of why we want to normalize all the bars to 100%.



Right-click on the graph (if you haven't already) to **Edit Contents in Separate Window**. Select **Options, Scale to 100%**. Visually, the percent of students having loans looks pretty similar across majors. Are the differences real, or due to chance?



As we've done before, we'll need to weight the field/loan categories by the number of students in each. Click **Data, Weight Cases** and select to use the variable **Number**. Now select **Analyze, Descriptive Statistics, Crosstabs**. Use **Field** for the row variable and **Loan** for the Column Variable. Be sure to click **Statistics** and ask for the chi-squared statistic.

Field \* Loan Crosstabulation

			Loan		Total
			No	Yes	
Field	Agriculture	Count	35	32	67
		Expected Count	38.5	28.5	67.0
	Child Dev	Count	50	37	87
		Expected Count	50.0	37.0	87.0
	Engineering	Count	137	98	235
		Expected Count	135.0	100.0	235.0
	Lib Arts Ed	Count	124	89	213
		Expected Count	122.4	90.6	213.0
	Management	Count	51	24	75
		Expected Count	43.1	31.9	75.0
	Science	Count	29	31	60
		Expected Count	34.5	25.5	60.0
	Technology	Count	71	57	128
		Expected Count	73.5	54.5	128.0
Total		Count	497	368	865
		Expected Count	497.0	368.0	865.0

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	6.525 <sup>a</sup>	6	.367
Likelihood Ratio	6.596	6	.360
N of Valid Cases	865		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 25.53.

If having a loan were independent of field of study (i.e., if all fields had the same proportion of students with loans), then there would be a 0.367 probability of obtaining observed cell counts that differ as much from the expected cell counts. Notice that most of the expected counts are close to the observed counts. Because of the high  $p$ -value, we can say that the observed differences are due to random chance and are not statistically significant. Thus, we will not reject the hypothesis that having a loan is independent of field of study.

### 9.3 Goodness of Fit

There is no built-in function in SPSS to perform a goodness of fit test. However, we can easily do the necessary calculations with **Transform, Compute Variable**.

**Example 9.4 Cell Phones and Accidents.** The following table gives the number of motor vehicle collisions by drivers using a cell phone broken down by days of the week over a 14-month period. Are such accidents equally likely to occur on any day of the week? It appears from the table that weekends have fewer accidents of this type; is this real or just randomness at work?

**Number of collisions by day of the week**

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

*Solution.* If each day were equally likely, then  $1/7$  of all accidents should occur on each day. Define variables for the **Day** (string), **Observed** count, and **Expected**. The **Expected** count for each day of the week is  $1/7 * 699 = 99.86$ . The contribution to the  $\chi^2$  statistic from each cell (day) in the table is  $\frac{(O - E)^2}{E}$ . We compute this into a new variable called **Chisq**.



At this point, our data worksheet looks like this.

Day	Observed	Expected	Chisq
Sunday	20.00	99.86	63.87
Monday	133.00	99.86	11.00
Tuesday	126.00	99.86	6.84
Wednesday	159.00	99.86	35.02
Thursday	136.00	99.86	13.08
Friday	113.00	99.86	1.73
Saturday	12.00	99.86	77.30

Clearly, our  $\chi^2$  statistic (the sum of the values in **Chisq**) is going to be large. To sum these use **Analyze, Descriptive Statistics, Frequencies**. Click **Statistics** and ask for the **Sum**.

**Statistics**

Chisq		
N	Valid	7
	Missing	0
Sum		208.84

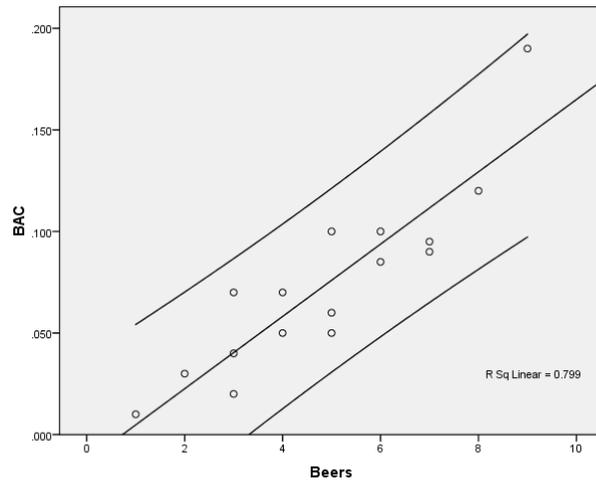
Our test statistic is 208.84. To find the  $p$ -value, use **CDF.Chisq** from **Transform, Compute Variable**. The degrees of freedom are one less than the number of categories, so here we have six.  $\chi^2$  tests are always upper-tailed, so we must subtract the area below our test statistic from 1.

The image shows two screenshots from Minitab. On the left is the 'Compute Variable' dialog box. The 'Target Variable' field contains 'Pvalue' and the 'Numeric Expression' field contains '1-CDF.CHISQ(208.84,6)'. On the right is a small window showing the result for 'Pvalue' as '0.0000'.

*Conclusions.* If these accidents were equally likely to occur on any day of the week, then there would be no chance of obtaining a sample distribution that differs so much from the expected counts of  $1/7 * 699 = 99.86$  for each day. So we can reject the claim that accidents are equally likely on each day. Looking at the components, we clearly see that weekends are much lower than expected; Wednesday is also higher than expected.

## CHAPTER

# 10



## Inference for Regression

10.1	Simple Linear Regression
10.2	More Detail about Simple Linear Regression

### Introduction

In this chapter, we provide details on performing the many difficult calculations for linear regression. In particular, we again find and graph the least-squares regression line and compute the correlation. We then can perform a  $t$  test to check the hypothesis that the correlation (or, equivalently, the regression slope) is equal to 0. We will also show how to compute a test statistic and find a  $p$ -value when the hypothesized slope is something other than 1.

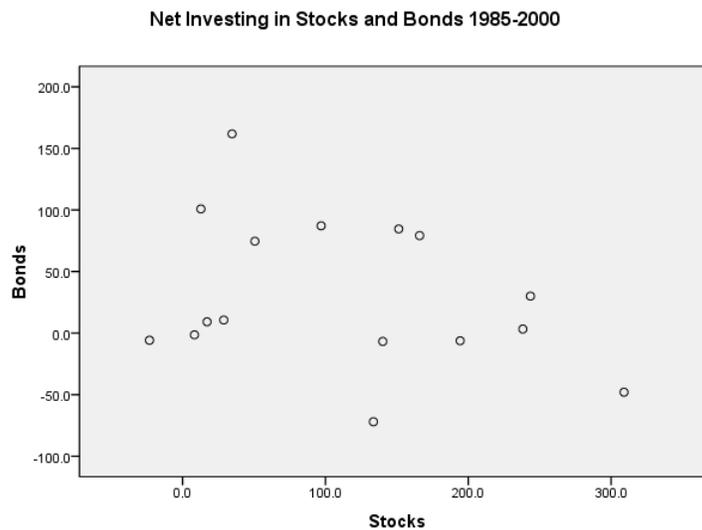
## 10.1 Simple Linear Regression

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

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

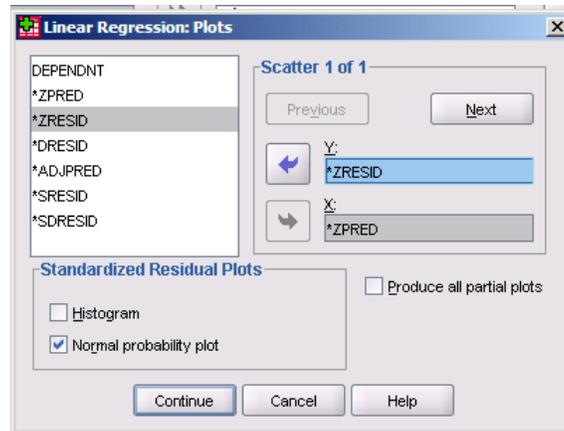
Year	1993	1994	1995	1996	1997	1999	1999	2000
Stocks	151.3	133.6	140.1	238.2	243.5	165.9	194.3	309.0
Bonds	84.6	-72.0	-6.8	3.3	30.0	79.2	-6.2	-48.0

*Solution.* We begin by making a scatterplot as explained in Section 2.1 by defining the variables **Stocks** and **Bonds** and entering the data. We define a scatterplot using **Graphs, Legacy Dialogs, Scatter/Dot, Simple Scatter**. Use **Stocks** as the predictor ( $X$ ) variable and **Bonds** as the response ( $Y$ ) variable. The relationship is fairly linear, but not extremely strong, since there is a good bit of scatter in the plot. The relationship is a decreasing one; if more money went to stocks, less went to bonds.



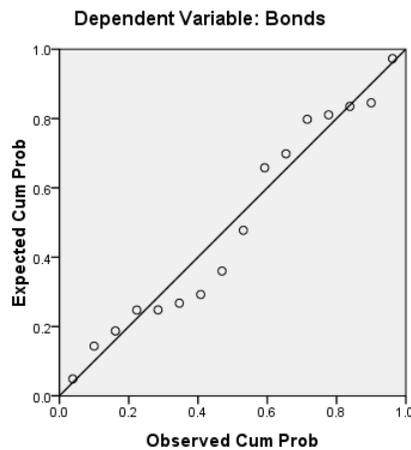
In Chapter 2, we showed how to compute and graph the least-squares line. At that time, we were merely interested in describing a relationship and ignored some of the output from the regression. We'll examine that now. Click **Analyze, Regression, Linear**. Select **Bonds** as the Dependent variable and **Stocks** as the Independent. Click **Plots** and

ask for the **Normal probability plot of residuals** and a plot of standardized residuals (residuals divided by their standard deviation) against standardized predicted values to check for linearity and constant variance. Click **Continue** to return to the main dialog box and **OK** to perform the regression and generate the plots.

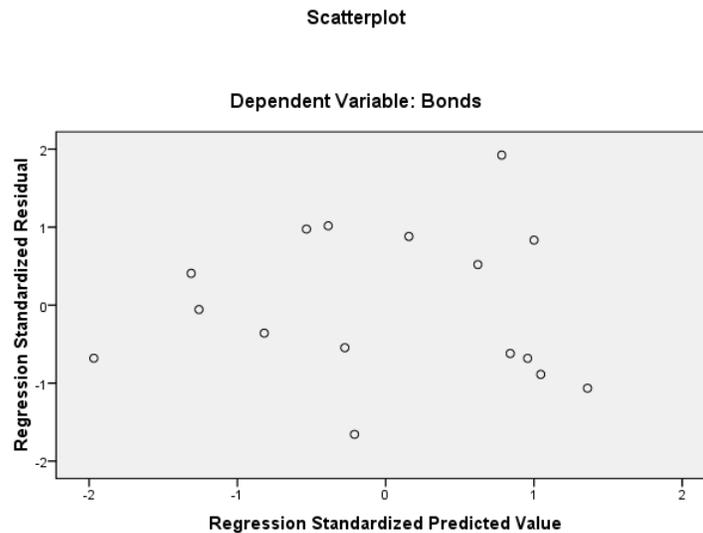


Before looking at the output of the regression, examine the plots for any indications of violations of the basic assumptions: residuals are  $N(0, \sigma)$  and the line is the correct model.

Normal P-P Plot of Regression Standardized Residual



While the points wander somewhat around the line, this Normal plot does not indicate any major skew or outliers; it appears the residuals have an approximately Normal distribution.



This residuals plot shows random scatter so the linear model assumption is reasonable; it also does not show any indications of changing variability with increasing  $x$  values, so the constant variance assumption is also reasonable.

With our plots examined, we turn our attention to the regression output.

**Model Summary<sup>b</sup>**

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.321 <sup>a</sup>	.103	.039	59.8798

a. Predictors: (Constant), Stocks

b. Dependent Variable: Bonds

**Coefficients<sup>a</sup>**

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	53.410	22.993		2.323	.036
	Stocks	-.196	.155	-.321	-1.266	.226

a. Dependent Variable: Bonds

As we have already seen graphically, the relationship is weak. The correlation coefficient,  $R$ , is .321, and **Stocks** explain only 10.3% of the original variation in net **Bonds** investing.

We obtain the regression line  $y = a + bx$  (or  $y = \hat{\beta}_0 + \hat{\beta}_1 x$ ). Here the equation rounds to  $Bonds = 53.410 - 0.196 * Stocks$ .

The  $p$ -value for the  $t$  test of  $H_0: \beta_1 = 0$  against  $H_A: \beta_1 \neq 0$  is given as 0.226 from a  $t$  statistic of  $-1.266$ . If  $\beta_1$  were equal to 0, then there would be a 22.6% chance of obtaining a value for  $b$  as low as  $-0.1962$ , or of obtaining a correlation as low as  $r = -0.32$ , with a sample of this size. This rather high  $p$ -value means that we do *not* have statistically significant evidence that there is some straight-line relationship between the flows of cash into bond funds and stock funds. In other words, we do not have enough evidence to reject that  $\beta_1 = 0$  or to reject that  $\rho = 0$ .

### Hypotheses other than 0

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

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

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

**Example 10.2 Another Hypothesis about the Slope.** In Example 10.1, above, we might be concerned that in any given month there are only so many dollars available to invest. Let's call that  $D$ . If investors were to split their money between stocks and bonds, we would have an equation  $S + B = D$ . This would argue that the regression relationship should be of the form  $B = D - S$ , and we should have slope  $\beta_1 = -1$ . We will test this as  $H_0$  against  $H_A: \beta_1 \neq -1$ . We had  $b_1 = -0.196$ , we need  $s_e$  and  $s_x^2$ . We find  $s_e$  in the Residuals Statistics table as 57.8508.

Residuals Statistics<sup>a</sup>

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	-7.223	57.982	31.312	19.5774	16
Residual	-99.1943	115.1797	.0000	57.8508	16
Std. Predicted Value	-1.968	1.362	.000	1.000	16
Std. Residual	-1.657	1.923	.000	.966	16

a. Dependent Variable: Bonds

To find the standard deviation of **Stocks** use **Analyze, Descriptive Statistics, Frequencies** and ask for just the standard deviation by first clicking **Statistics**.

### Statistics

Stocks		
N	Valid	16
	Missing	0
Std. Deviation		99.7715

Putting this together with the fact that we had  $n = 16$  data points, we find a  $t$  statistic using **Transform, Compute Variable** of 5.1869.

The **Compute Variable** dialog box shows the following configuration:

- Target Variable:** tstat
- Numeric Expression:**  $(-.196+1)*\text{sqrt}(15*99.7715**2)/57.8508$

The result of the calculation is displayed as **tstat = 5.37**.

To find the  $p$ -value for this two-sided test, we will double the area above  $t = 5.37$  under the  $t$  distribution curve with 14 degrees of freedom.

The **Compute Variable** dialog box shows the following configuration:

- Target Variable:** Pvalue
- Numeric Expression:**  $2*(1-\text{CDF.T}(5.37,14))$

The result of the calculation is displayed as **Pvalue = 0.0001**.

Our  $p$ -value is 0.0001, so we have sufficient evidence based on this sample that there is not an inverse relationship between net investments in stocks and bonds (the true slope is not  $-1$ .)

## Confidence Intervals in Regression Inference

**Example 10.3 Stocks and Bonds Continued.** Using the data from Example 10.1, find the 90% confidence intervals for the slope  $\beta_1$ , and intercept  $\beta_0$  of the linear regression model.

**Solution.** SPSS will give 95% intervals explicitly. From the main regression dialog box, click **Statistics**. Click to put a check in the **Confidence intervals** box. Click **Continue** to return to the main dialog box.

The **Linear Regression: Statistics** dialog box shows the following configuration:

- Regression Coefficient:**
  - Estimates
  - Confidence intervals
  - Covariance matrix
- Model fit
- R squared change
- Descriptives
- Part and partial correlations
- Collinearity diagnostics

**Coefficients<sup>a</sup>**

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	53.410	22.993		2.323	.036	4.096	102.724
	Stocks	-.196	.155	-.321	-1.266	.226	-.529	.136

a. Dependent Variable: Bonds

However, our question asked for 90% confidence intervals. These will have to be computed “by hand.” SPSS gives us the standard errors of the slope and intercept in the regression output. These were  $s(b_0) = 22.993$  and  $s(b_1) = .155$ . As always, confidence intervals are of the form

$$\text{estimate} \pm t^* s_{\text{estimate}}$$

We really only need the value of  $t^*$ . This can be found from a table or using **Idf.T** from the Inverse DF function group in **Transform, Compute Variable**. We find  $t^* = 1.76$ . (Remember, the confidence region is in the middle of the distribution, so considering the leftover 10%, the area left of  $t^*$  is 0.95.)



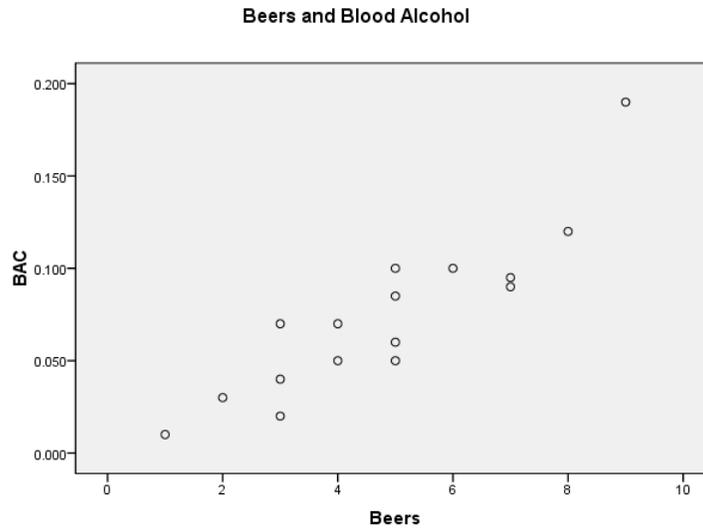
Putting all this together, we find the 90% confidence interval for the intercept is  $53.410 \pm 1.76 * 22.993$  or (12.942, 93.878) and the 90% confidence interval for the slope is  $-.196 \pm 1.76 * .155$  or (-.469, .077).

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

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

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

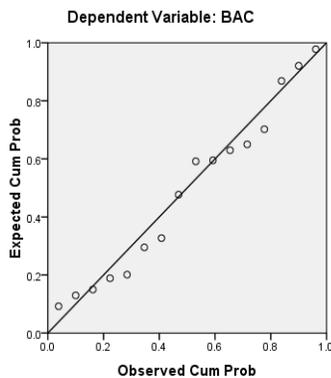
*Solution.* First, plot the data. Is the relationship linear enough for our regression to make sense? Define a scatterplot as detailed above in Example 10.1 (and Chapter 2).



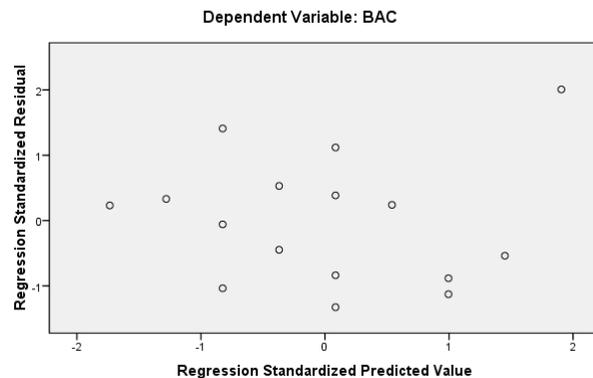
The plot is very linear, but the individual represented at the upper right (9 beers, BAC 0.19) might be influential to this regression.

Performing the regression and asking for the same plots of residuals as was done in Example 10.1 finds these plots.

Normal P-P Plot of Regression Standardized Residual



Scatterplot



The Normal plot is pretty straight, and there are no patterns (except for the possible outlier) in the other plot. Inference for these data is reasonable.

**Model Summary<sup>b</sup>**

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.894 <sup>a</sup>	.800	.786	.020441

a. Predictors: (Constant), Beers

b. Dependent Variable: BAC

**Coefficients<sup>a</sup>**

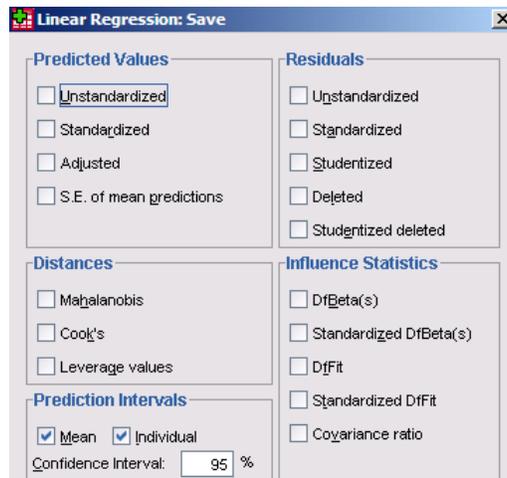
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-.013	.013		-1.005	.332
	Beers	.018	.002	.894	7.480	.000

a. Dependent Variable: BAC

The relationship between the number of beers consumed and BAC is strong;  $R = .894$  and the number of beers explains 80% of the variability in BAC. The regression equation is  $BAC = -0.013 + 0.018 * Beers$ . With a  $t$  statistic of 7.480 and  $p$ -value of 0.000, the linear relationship is statistically significant. Further, we recognize that if no beer has been consumed an individual should have a BAC of 0. We notice that the  $p$ -value for testing a null hypothesis that  $\beta_0 = 0$  has a  $p$ -value of .332; the intercept is not statistically different from 0.

### Confidence Intervals for a Mean or Individual Response

SPSS will give a confidence interval for the mean response and a prediction interval for a new response at each observation in the data set automatically if asked. When performing the regression, click **Save** and click to check **Prediction Intervals Mean** and **Individual**. 95% confidence is the default; change this for a different level.



To compute these manually, the formulas are:

$$\text{Confidence Interval for Mean Response at } x: \hat{y} \pm t^* s_e \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum (x - \bar{x})^2}}$$

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

**Example 10.5. More on Beer.** The legal limit for driving in most states is 0.08. If the average person had four beers, would he be legal to drive? Use a 95% confidence interval estimate of BAC to answer the question. (Of course, it's better to not drink, or use a designated driver.)

*Solution.* We asked SPSS to compute the intervals for mean and individual responses. We look for a row where the individual had four beers (it was Student 11). From this row in the data worksheet, we see the 95% confidence interval for the average BAC level for all individuals who consume four beers is .04632 to .07003. It appears the average person would be legal to drive.

	Beers	BAC	LMCI_1	UMCI_1
11	4	0.070	0.04632	0.07003

If the value for which we are interested in predicting is not one of our observations, we need to compute the interval manually.

**Example 10.6 Steve and Beer.** Steve is planning on going to a party, and he thinks he'll drink about four beers. Will he be legal to drive home?

*Solution.* Steve is one individual, not “the average guy.” In this case, we need a prediction interval for a new observation. This is the individual confidence interval.

Beers	BAC	LMCI_1	UMCI_1	LICI_1	UICI_1
4	0.070	0.04632	0.07003	0.01270	0.10366

For a particular person who consumes four beers, we are 95% confident his BAC will be between .0127 and .10366. This prediction interval extends well beyond the legal 0.08; if Steve drinks four beers, he should find a designated driver.

## 10.2 More Detail about Simple Linear Regression

Analysis of variance (ANOVA) is another method to test the null hypothesis  $H_0: \beta_1 = 0$ , with an alternative  $H_A: \beta_1 \neq 0$ . Although really more useful in a true ANOVA or

multiple regression setting, in the case of simple linear regression it has a use and relation to a  $t$  test similar to that of using a  $\chi^2$  test for two proportions. SPSS produces an ANOVA table as part of its standard regression output.

**Example 10.7 More Stocks and Bonds.** Consider the data from Example 10.1 on the net new money (in billions of dollars) flowing into stock and bond mutual funds from 1985 to 2000.

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

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

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

*Solution.* (a) The ANOVA table created by SPSS is below.

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	5749.114	1	5749.114	1.603	.226 <sup>a</sup>
	Residual	50200.743	14	3585.767		
	Total	55949.857	15			

a. Predictors: (Constant), Stocks

b. Dependent Variable: Bonds

(b) The ANOVA test is about the linear regression slope  $\beta_1$ . We test the null hypothesis  $H_0: \beta_1 = 0$  with the alternative  $H_A: \beta_1 \neq 0$ . With the above  $p$ -value of 0.226, we do not have strong evidence in this case to reject  $H_0$  in favor of the alternative.

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

(d) The  $t$  test was used in Example 10.1. The  $t$  statistic was computed as  $t = -1.266$ . If we square this value, then we obtain 1.602756, which rounds to the value of the  $F$ -statistic in the ANOVA table.

### Sample Correlation and the $t$ test

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

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

**Example 10.8 Avoidance of New Foods.** In a study of 564 children who were two to six years of age, the relationship of food neophobia (avoidance of unfamiliar foods) and the frequency of consumption gave a correlation of  $r = -0.15$  for meat. Perform a significance test about the correlation of meat neophobia and the frequency of meat consumption among children two to six years of age.

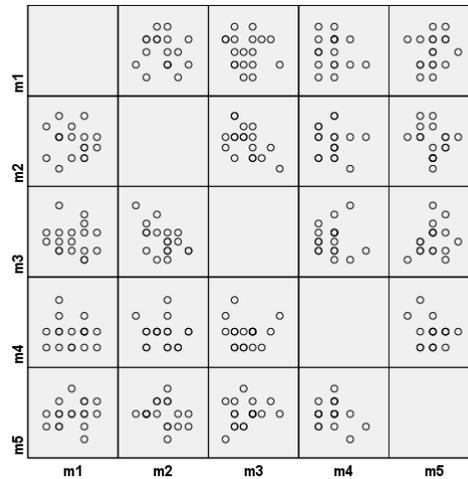
*Solution.* We shall test  $H_0: \rho = 0$  versus  $H_A: \rho < 0$ . The rejection region is a left tail; thus the  $p$ -value will be the left-tail probability of the  $t(n-2) = t(562)$  distribution. We compute the  $t$  statistic and  $p$ -value “manually” by entering  $-.15*\sqrt{562}/\sqrt{1-(-.15)^2}$  to obtain  $t = -3.60$ . Next, we compute the  $p$ -value  $P(t(562) \leq -3.59667)$  with CDF.T to obtain a  $p$ -value of  $2 \times 10^{-4}$ .

 <p>Target Variable: tstat = Numeric Expression: <math>-.15*\sqrt{562}/\sqrt{1-(-.15)^2}</math></p>	<table border="1" data-bbox="966 1081 1159 1207"> <tr><td>tstat</td></tr> <tr><td>-3.60</td></tr> </table>	tstat	-3.60
tstat			
-3.60			
 <p>Target Variable: Pvalue = Numeric Expression: CDF.T(-3.6,562)</p>	<table border="1" data-bbox="966 1291 1159 1383"> <tr><td>Pvalue</td></tr> <tr><td>0.0002</td></tr> </table>	Pvalue	0.0002
Pvalue			
0.0002			

If the true correlation were 0, then there would be only a 0.0002 probability of obtaining a sample correlation as low as  $r = -0.15$  with a random sample of size 564. We have statistical evidence to reject  $H_0$  in favor of the alternative that  $\rho < 0$ . We note that for a two-sided alternative, then our  $p$ -value would be  $2*(2 \times 10^{-4}) = 4 \times 10^{-4}$ , which would still be small enough for us to reject  $H_0$ .

## CHAPTER

# 11



# Multiple Regression

11.1	Inference for Multiple Regression
11.2	A Case Study
11.3	Another Type of Plot

## Introduction

In this chapter, we demonstrate how to use SPSS to calculate a multiple linear regression model. This process uses **Analyze, Regression, Linear** just as simple regression did. We just add more predictors into the box.

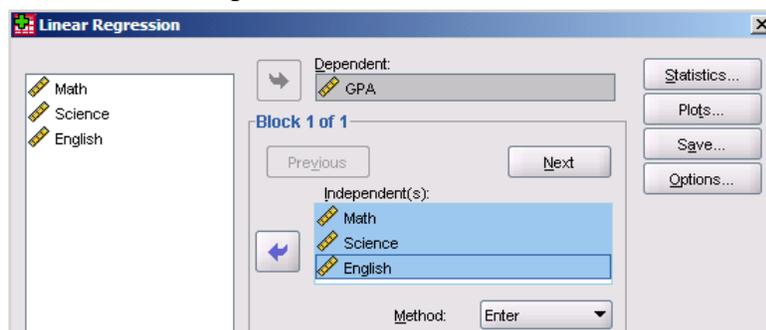
## 11.1 Inference for Multiple Regression

In multiple regression we have several predictor variables for the response variable. Some (or all) may be of use in making predictions.

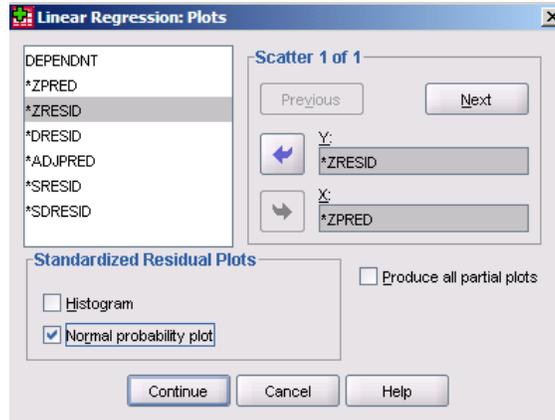
**Example 11.1 Predicting Grades.** Consider the CSDATA of a sample of 24 students at a large university that uses a 4.0 GPA grade scale. Run a multiple regression analysis for predicting the GPA from the three high school grade variables (Math, Science, and English where 10 = ‘A’, 9 = ‘A-’, 8 = ‘B+’,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

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

*Solution.* Define the variables and input the data. Define the linear regression as in the screen below. To select all the predictor variables at once, hold down the **ctrl** key.



Just as with simple regression, good practice says we should examine residual plots for any violations of assumption. Click on Plots and define the scatter of standardized residuals against standardized predicted values and the normal probability plot of residuals.



**Model Summary<sup>b</sup>**

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.573 <sup>a</sup>	.328	.227	.65131

a. Predictors: (Constant), English, Science, Math

b. Dependent Variable: GPA

Overall, the model explains 32.8% of the total variation in GPA. Adjusted R Square penalizes the standard  $r^2$  for adding variables into the model that are not helpful. Since this is so much lower, at least some of the high school grades are not useful in predicting college grades.

**ANOVA<sup>b</sup>**

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	4.139	3	1.380	3.252	.043 <sup>a</sup>
	Residual	8.484	20	.424		
	Total	12.623	23			

a. Predictors: (Constant), English, Science, Math

b. Dependent Variable: GPA

The null hypothesis for the  $F$ -test is that each variable coefficient in the linear regression model is equal to 0 or  $H_0: \beta_1 = \beta_2 = \beta_3 = 0$ . In this case, the  $p$ -value is 0.043, which is a value usually considered low enough to be statistically significant. Thus, even with this small sample of size 24, we have enough evidence to reject  $H_0$ .

Coefficients<sup>a</sup>

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-.228	.992		-.230	.820
	Math	.040	.092	.094	.430	.672
	Science	.242	.121	.429	2.002	.059
	English	.085	.129	.152	.658	.518

a. Dependent Variable: GPA

Our model becomes  $\mu_{GPA} = -0.228 + 0.040 \cdot HSM + 0.242 \cdot HSS + 0.085 \cdot HSE$ .

However, if we look at the  $t$  statistics and the associated  $p$ -values for each of these, we find the Math and English grades are not useful in predicting college GPA.

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

*Solution.* We must evaluate  $-0.228 + 0.040 \cdot HSM + 0.242 \cdot HSS + 0.085 \cdot HSE$  for these values, which is easy enough to do. If you will recall from Chapter 10, SPSS will compute **Predicted** Values (and confidence intervals and prediction intervals for individual observations) from the **Save** dialog box. However, if we want a more accurate (non-rounded) value, then we can insert a dummy case into our data with no value for GPA. This case will be ignored in the regression computation, but will get a predicted value. This regression predicts about a 2.64 GPA for students with these high school grades.

	GPA	Math	Science	English	PRE_1
25	.	8	7	10	2.63786

**Example 11.3 Predicting SAT Math Scores.** With the above 24 sample points from the CSDATA, perform a regression analysis for predicting the SATM from the three high school grade variables. Then find the predicted SATM for a student with  $HSM = 9$ ,  $HSS = 6$ , and  $HSE = 8$ .

*Solution.* We add the **SATM** data into the worksheet and change the dependent variable for the regression. Add a dummy observation with the grades desired.

**Coefficients<sup>a</sup>**

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	330.212	129.195		2.556	.019
	Math	29.738	12.032	.537	2.472	.023
	Science	3.447	15.769	.046	.219	.829
	English	3.766	16.855	.051	.223	.825

a. Dependent Variable: SATM

The new model is  $\mu_{SATM} = 330.212 + 29.738 \cdot HSM + 3.447 \cdot HSS + 3.766 \cdot HSE$ . We also note that Science and English grades are not good predictors of SAT Math scores. Only the coefficient of Math is significantly nonzero.

The displayed *p*-value of 0.0348 gives us statistical evidence to reject that all regression coefficients of HSM, HSS, and HSE are 0. Thus, at least one coefficient is nonzero, and its parameter is correlated to SATM scores. The model predicts students with these high school grades should have a Math SAT score of about 649.

## 11.2 A Case Study

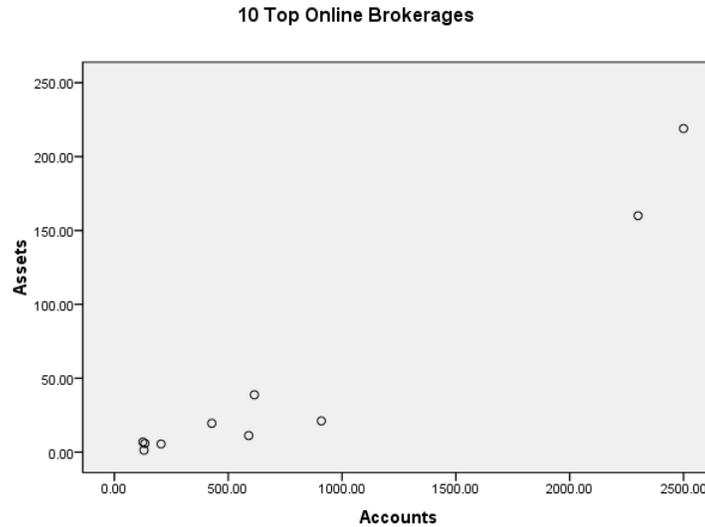
**Example 11.4 Internet Brokerages.** Below are some data for the top ten Internet brokerages. The variables are the Market Share of the firm, the number of Internet accounts in thousands, and the total assets of the firm in billions of dollars.

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

- (a) Use a simple linear regression to predict assets using the number of accounts. Give the regression equation and the results of the significance test for the regression coefficient.
- (b) Do the same using market share to predict assets.

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

*Solution.* (a) Define the variables and input the data. We first examine a scatterplot with number of accounts on the  $x$  axis and Assets on the  $y$  axis.



We recognize from this plot that the two largest brokerages, Schwab and Fidelity, will be influential to this regression.

**Model Summary**

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.968 <sup>a</sup>	.938	.930	20.18767

a. Predictors: (Constant), Accounts

**Coefficients<sup>a</sup>**

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-17.121	8.778		-1.950	.087
	Accounts	.083	.008	.968	10.959	.000

a. Dependent Variable: Assets

The regression equation is  $Assets = -17.121 + 0.083 * Accounts$ . The  $p$ -value for a two-sided significance test is 0 (to three decimal places). If the regression coefficient  $\beta_1$  were equal to 0, then there would be only a very small chance of obtaining a value of  $b$  as large

as 0.083, or a correlation as high as 0.968, even with such a small sample. Thus, we have strong evidence to reject the hypothesis that  $\beta_1 = 0$ .

(b) Now we change our predictor variable to **MShare** to obtain a linear regression equation  $Assets = a + bx = -19.901 + 7.680MarketShare$ .

**Model Summary**

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.780 <sup>a</sup>	.609	.560	50.53590

a. Predictors: (Constant), Mshare

**Coefficients<sup>a</sup>**

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-19.901	25.218		-.789	.453
	Mshare	7.680	2.177	.780	3.527	.008

a. Dependent Variable: Assets

The  $p$ -value for the two-sided significance test is 0.008. If the regression coefficient  $\beta_1$  were equal to 0, then there would be less than a 1% chance of obtaining a value of  $b$  as high as 7.68 or a correlation as high as 0.780. Thus, we again can reject the hypothesis that  $\beta_1 = 0$ .

(c) Return to the Linear Regression and use both predictors.

**Model Summary**

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.971 <sup>a</sup>	.944	.927	20.52160

a. Predictors: (Constant), Accounts, Mshare

**Coefficients<sup>a</sup>**

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-21.453	10.243		-2.094	.074
	Mshare	1.158	1.344	.118	.861	.418
	Accounts	.076	.012	.880	6.443	.000

a. Dependent Variable: Assets

The regression equation is  $\mu_{ASSETS} = -21.4532 + 1.158MarketShare + 0.076Accts$ . The  $p$ -value of (approximately) 0 gives significant evidence to reject the null hypothesis that  $\beta_1 = \beta_2 = 0$ . However, in this multiple regression Market Share is no longer significant.

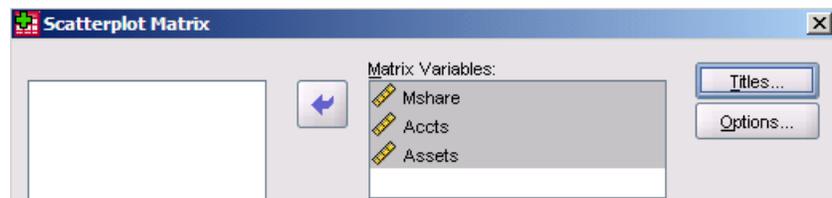
(d) Comparing the  $r^2$  values from each part, we see that the multiple linear regression  $r^2$  of 94.35% is not really different from the 93.8% obtained from the simple model using only the number of accounts. It seems reasonable to believe that market share and number of accounts might be collinear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

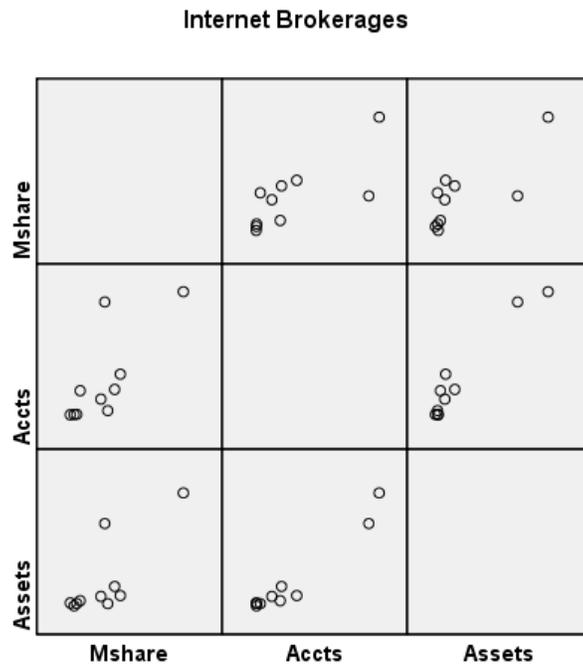
### 11.3 Another Type of Plot

With multiple predictors, it is useful to observe the relationships (if any) among the predictors and the response variable. This can be done with **Graphs**, **Legacy Dialogs**, **Scatter/Dot**, **Matrix Scatter**.

**Example 11.5 More Internet Brokerages.** Examine the relationships among our three variables for the Internet brokerage data of Example 11.4 with a matrix plot.

*Solution:* Click **Graphs**, **Legacy Dialogs**, **Scatter/Dot**, **Matrix Scatter**. Hold down the Ctrl key and click to select all our variables into the matrix. Give the plot a title. Then click **OK** to generate the plot.

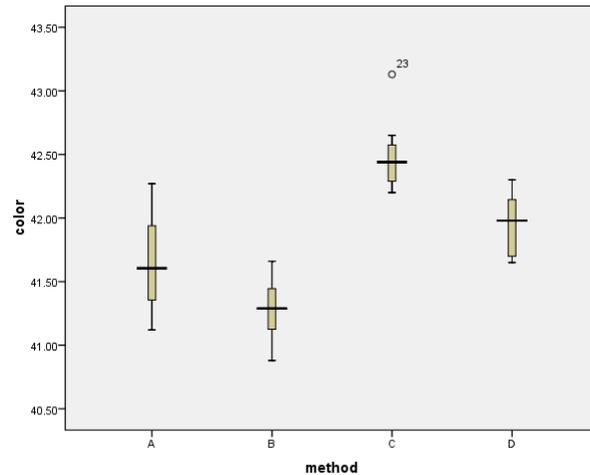




We suspected at the end of Example 11.4 that Accounts and Market share were related; this plot confirms it when we look at the graph in the middle left cell. There is a strong linear relationship between these two variables.

## CHAPTER

# 12



# One-Way Analysis of Variance

12.1	Inference for One-Way Analysis of Variance
------	--

## Introduction

Just as a chi-squared test can be seen as an extension of the two-proportion test, one-way analysis of variance (ANOVA) is an extension of the independent samples  $t$  test to more than two groups.

In this chapter, we perform one-way analysis of variance (ANOVA) to test whether several normal populations, assumed to have the same variance, also have the same mean. As always, SPSS is happiest when it has the actual data; if we have only summary statistics, we will need to use **Transform**, **Compute Variable** to compute the test statistic and  $p$ -value.

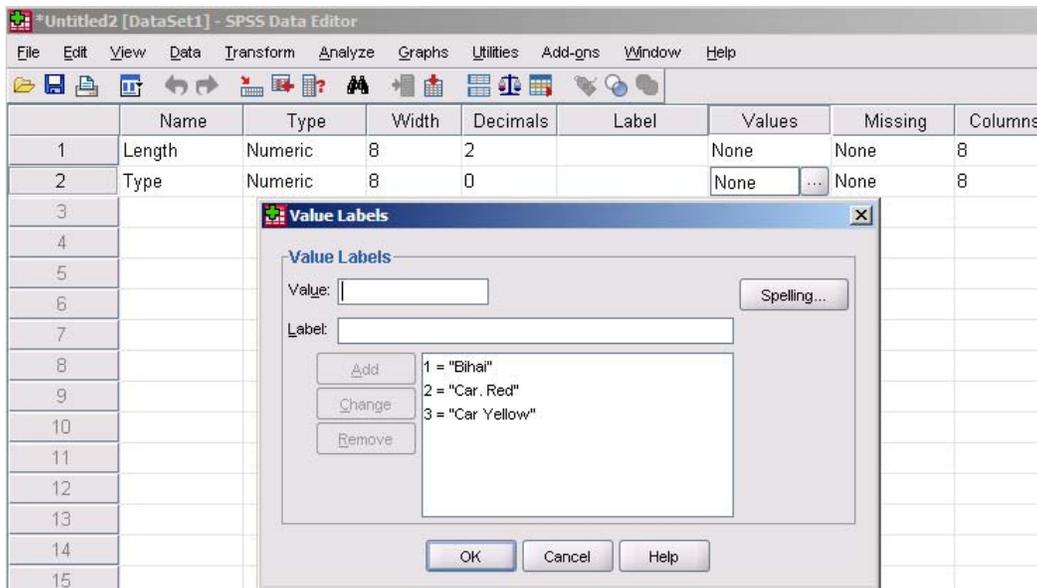
## 12.1 Inference for One-Way Analysis of Variance

We begin with an exercise that demonstrates built-in analysis of variance capabilities using the **One-Way Anova** command from the **Analyze, Compare Means** menu.

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

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

*Solution.* We define two variables: one for the flower type, which must be an integer for the ANOVA procedure, and one for the length, and input the data. We will also define value labels so our output and plots will be labeled with the flower name and not some arbitrary integer.



Our next step might be to compare some descriptive statistics on the three flower types using **Analyze, Compare Means, Means**.

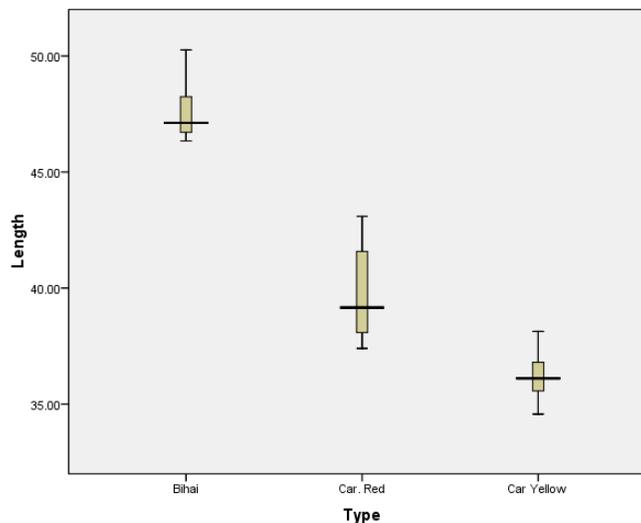
#### Report

Length			
Type	Mean	N	Std. Deviation
Bihai	47.5975	16	1.21288
Car. Red	39.7113	23	1.79876
Car Yellow	36.1800	15	.97532
Total	41.0670	54	4.73732

These sample means range from slightly larger than 36 to more than 47.5. Is the difference real (statistically significant) or just due to randomness? We also see the overall mean in 41.067.

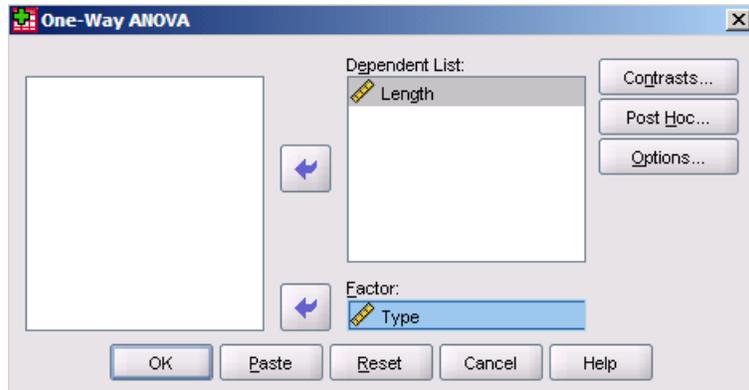
Before proceeding to the ANOVA, we must consider the assumptions of the test. The first is that all groups have the same *population* standard deviation. In practice, the results will be approximately correct if the largest standard deviation is no more than twice as large as the smallest. Since  $2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

The second assumption is that the data come from Normal populations. We could do a Normal plot for each type of flower, but in this case side-by-side boxplots are a good idea. We're looking for indications of skewness or outliers. Use **Graphs, Legacy Dialogs, Boxplot, Simple, Data in Chart are Summaries for groups of cases**. The Variable is **Length** and the Categories are **Type**.



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

Next, we test the hypothesis that the mean lengths of the three species are equal:  $H_0: \mu_1 = \mu_2 = \mu_3$ . **Analyze, Compare Means, One-Way ANOVA.** **Length** is the dependent variable and **Type** is the factor.

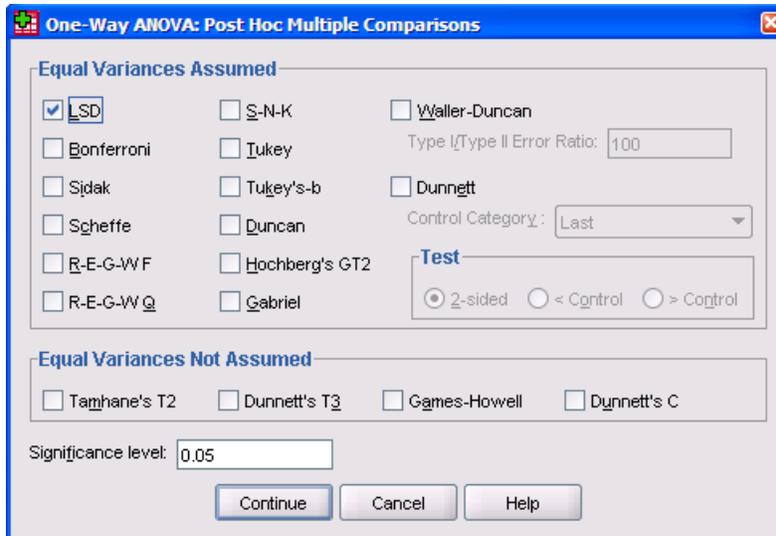


Length	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	1082.872	2	541.436	259.119	.000
Within Groups	106.566	51	2.090		
Total	1189.438	53			

We receive a  $p$ -value of 0 (to three significant digits) from an  $F$ -statistic of 259.119. If the true means were equal, then there would be almost no chance of the sample means varying by as much as they do with samples of these sizes. Thus, we have significant evidence to reject the claim that the mean lengths of these species are equal. We note that the  $r^2$  value is not displayed, but it can be computed from the two displayed SS values. Here we can use  $r^2 = SSF/SSTO = 1082.872/1189.438 = 0.9104$ .

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

If we return to **Analyze, Compare Means, One-Way ANOVA**, we now will click **Post-Hoc**. There are several different mechanisms available to do this follow-up analysis, but perhaps the most common is Tukey's least-significant differences (**LSD**). Check that box, then **Continue** and **OK**.



**Multiple Comparisons**

Length  
LSD

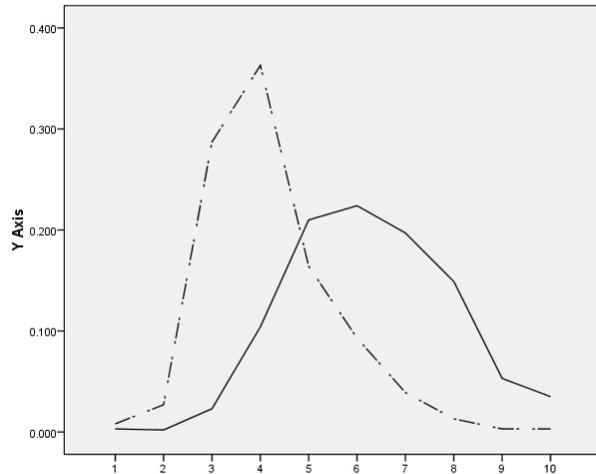
(I) Type	(J) Type	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Bihai	Car. Red	7.88620*	.47058	.000	6.9415	8.8309
	Car Yellow	11.41750*	.51952	.000	10.3745	12.4605
Car. Red	Bihai	-7.88620*	.47058	.000	-8.8309	-6.9415
	Car Yellow	3.53130*	.47974	.000	2.5682	4.4944
Car Yellow	Bihai	-11.41750*	.51952	.000	-12.4605	-10.3745
	Car. Red	-3.53130*	.47974	.000	-4.4944	-2.5682

\*. The mean difference is significant at the 0.05 level.

All the pairwise comparisons have *p*-values of 0, so we conclude that each species has a different mean length.

## CHAPTER

# 13



# Two-Way Analysis of Variance

13.1	Plotting Means
13.2	Inference for Two-Way ANOVA

## Introduction

In this chapter, we demonstrate how to use SPSS to perform a two-way analysis of variance to test for equality of means simultaneously among populations and traits in a two-factor experiment.

### 13.1 Plotting Means

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

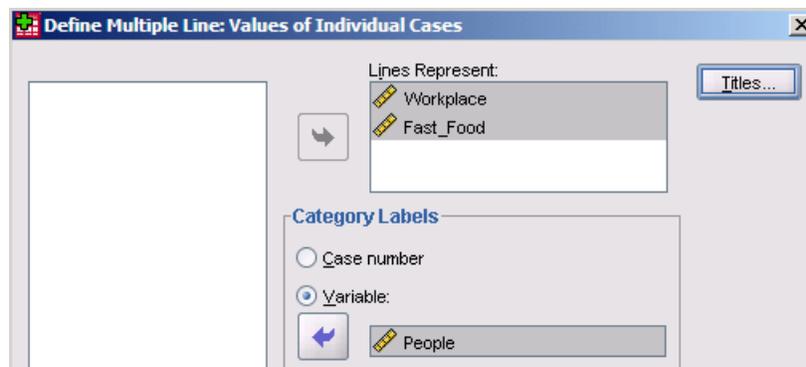
Lunch setting	Number of people eating				
	1	2	3	4	5
Workplace	12.6	23.0	33.0	41.1	44.0
Fast-food restaurant	10.7	18.2	18.4	19.7	21.9

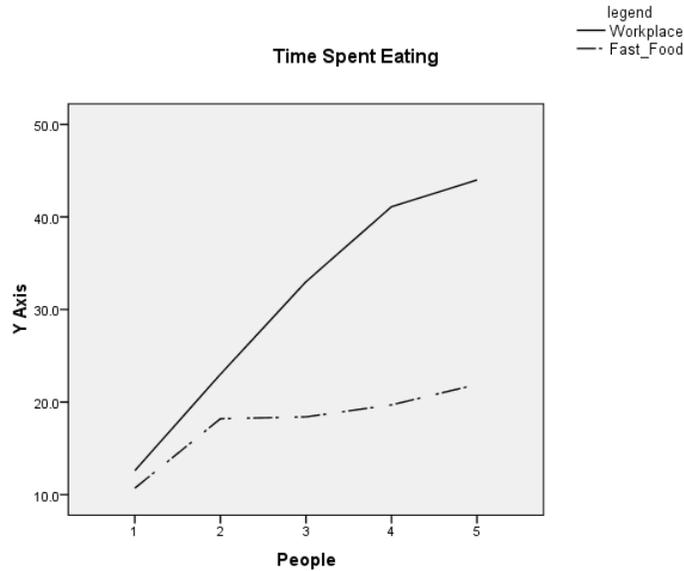
*Solution.* We note that we cannot use ANOVA to analyze this data because each cell contains the *mean* of an unknown number of subjects, and we cannot assume that there are a fixed number of measurements per cell. In order to perform two-way ANOVA, we also would need to know the sample sizes and the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

Define three variables and enter the data.

	People	Workplace	Fast_Food
1	1	12.6	10.7
2	2	23.0	18.2
3	3	33.0	18.4
4	4	41.1	19.7
5	5	44.0	21.9

Click **Graphs**, **Legacy Dialogs**, **Line**, **Multiple**. Data in this chart are **Values of individual cases**. We want a line for each of the places, and our categories are the number of people eating together. Give the plot an appropriate **Title**, and click OK to generate the graph.





As always, the default is to use colors to identify the different sources; if you have only a black-and-white printer, right-click the graph and use **Edit Content In Separate Window**, then right-click to access the **Properties** box. Click the **Variables** tab and change legend style to Dash or thickness.

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

### 13.2 Inference for Two-Way ANOVA

SPSS is fully suited to perform two-way ANOVA. To do this, we'll use **Analyze, General Linear Model, Univariate**.

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

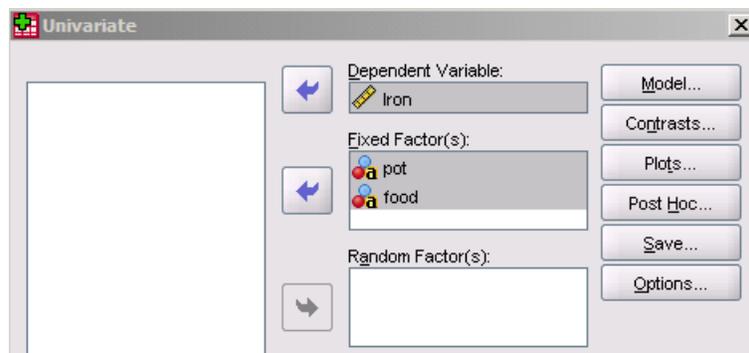
IRON	Food											
	Meat				Legumes				Vegetables			
Aluminum	1.77	2.36	1.96	2.14	2.40	2.17	2.41	2.34	1.03	1.53	1.07	1.30
Clay	2.27	1.28	2.48	2.68	2.41	2.43	2.57	2.48	1.55	0.79	1.68	1.82
Iron	5.27	5.17	4.06	4.22	3.69	3.43	3.84	3.72	2.45	2.99	2.80	2.92

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

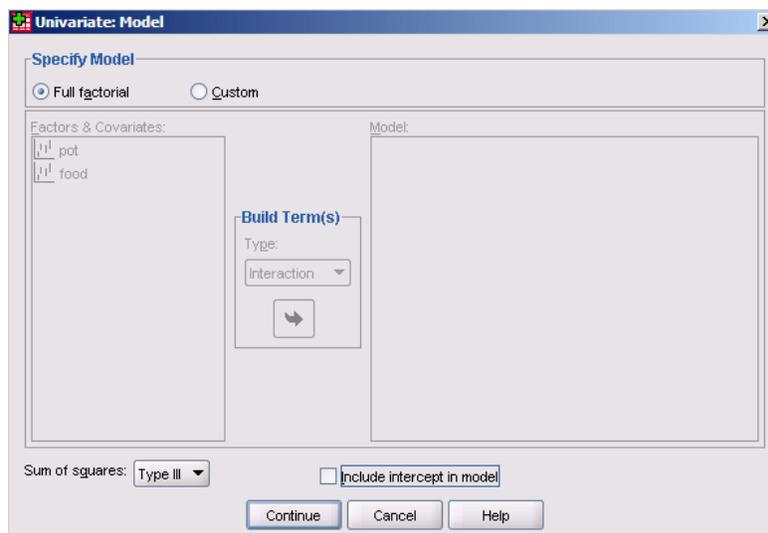
*Solution.* Define variables for the **Iron** content, **pot** type, and **food** type. The factor variables can be either strings or integers; here, we have chosen to use descriptive string names as it makes interpreting the output easier. Enter the data. Our first few rows of data are shown below.

	Iron	pot	food
1	1.77	alum	meat
2	2.36	alum	meat
3	1.96	alum	meat
4	2.14	alum	meat
5	2.27	clay	meat
6	1.28	clay	meat
7	2.48	clay	meat
8	2.68	clay	meat
9	5.27	iron	meat
10	5.17	iron	meat

Click **Analyze**, **General Linear Model**, **Univariate**. Our dependent (response) variable is **Iron** and the fixed factors are the **pot** type and **food** type.



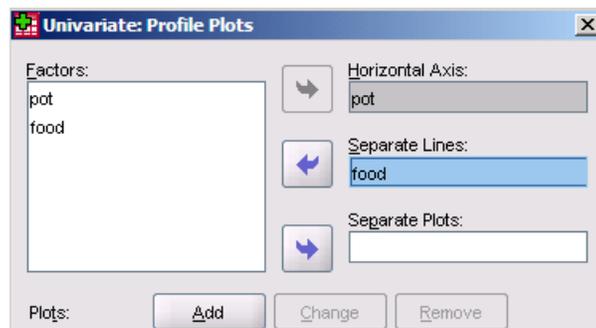
SPSS defaults to fitting an intercept in the general linear model; we do not want that for this ANOVA. Click **Model**, then uncheck the box to fit the intercept. **Continue**.



To see the individual cell means, click **Options**. Display Means for the interaction of pot\*food. **Continue**.



Last, to see the means plot, click **Plots**. Which variable goes where here is judgment or personal preference. We have chosen to put **pot** on the horizontal axis and use separate lines for each value of **food**. Click **Add**, **Continue**. Click **OK** to perform the analysis.



After the case processing summary (a count of valid cases), we find the ANOVA table.

#### Tests of Between-Subjects Effects

Dependent Variable:Iron

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Model	269.292 <sup>a</sup>	9	29.921	221.792	.000
pot	24.894	2	12.447	92.263	.000
food	9.297	2	4.648	34.456	.000
pot * food	2.640	4	.660	4.893	.004
Error	3.643	27	.135		
Total	272.935	36			

a. R Squared = .987 (Adjusted R Squared = .982)

All effects have  $p$ -values of 0.004 or less; so we reject that all combinations of food and pot have the same mean. Further, we have a significant interaction; the pot type does not have the same effect for all foods.

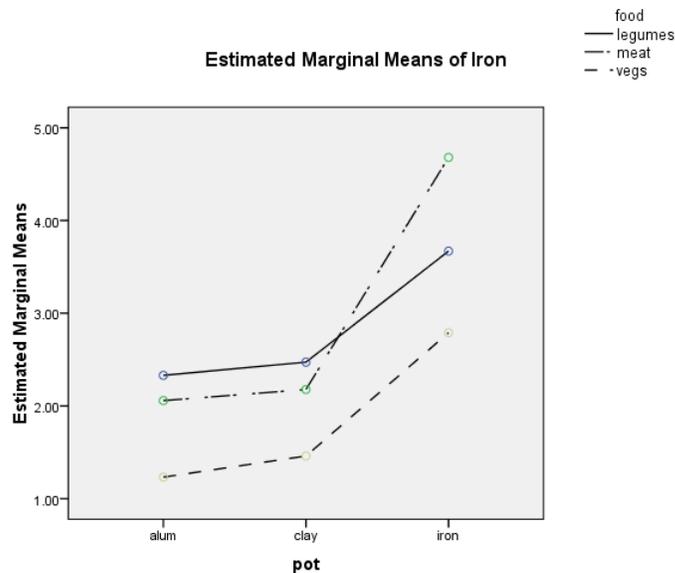
The next part of the output gives the means for each “cell” (combination of food and pot), along with 95% confidence intervals. Notice the standard error for each is the same (due to the assumption on constant variance).

**pot \* food**

Dependent Variable:Iron

pot	food	Mean	Std. Error	95% Confidence Interval	
				Lower Bound	Upper Bound
alum	legumes	2.330	.184	1.953	2.707
	meat	2.058	.184	1.681	2.434
	vegs	1.233	.184	.856	1.609
clay	legumes	2.472	.184	2.096	2.849
	meat	2.177	.184	1.801	2.554
	vegs	1.460	.184	1.083	1.837
iron	legumes	3.670	.184	3.293	4.047
	meat	4.680	.184	4.303	5.057
	vegs	2.790	.184	2.413	3.167

Lastly, we see the means plot. The crossing of the lines for meat and legumes shows the interaction; it appears that meat acts differently in the iron pot.



One could continue with univariate post-hoc tests on the individual factors or contrasts; however, since there is significant interaction this is not recommended.

## CHAPTER

# 14

# Bootstrapping and Permutation Tests

## **Introduction**

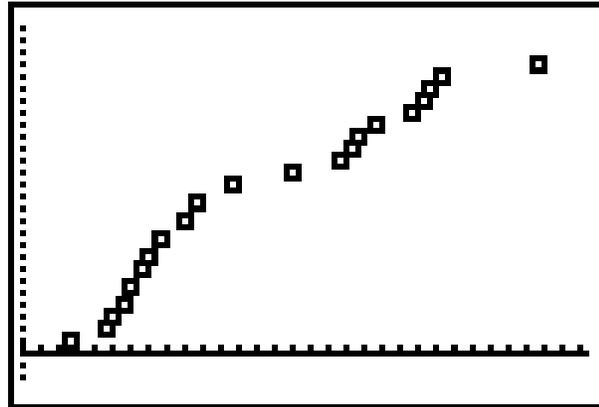
We have already seen that simulating probability experiments is somewhat contrived. We have also seen that performing tests for proportions when the data have been summarized is not something this package is really suited for.

Unfortunately, SPSS does not perform bootstrapping or permutation tests at the level an introductory statistics student could make use of. These require either substantial programming expertise or an add-on module. All statistics packages have strengths and weaknesses; SPSS Base (and the student-oriented Career Starter version) are strictly oriented to analyzing data in the spreadsheet window with standard statistical techniques.

If your instructor wants to study these, this author recommends the S-plus module available from your text's Web site. (S-plus can be obtained by students at no cost.)

## CHAPTER

# 15



# Nonparametric Tests

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

### Introduction

In this chapter, we demonstrate how to perform several nonparametric hypothesis tests. If you have “SPSS Base” these are built-in; however, if you have the student (Career Starter) version SPSS requires an additional “Exact Tests” module to fully automate these; it does, however, allow us to rank cases, which can be manually added. *P*-values can be computed using **Transform, Compute Variable**.

These tests relax the assumption of Normal distribution of the data (or sample mean). They are less powerful than parametric tests since they do not use all the information in the data but only the ranks (sorted order statistics); one has information on the smallest, next smallest, etc., but not the size of the differences.

The Rank Sum test is a stand-in for a two-sample *t* test; the signed rank for a paired samples test, and the Kruskal-Wallis for one-way ANOVA.

## 15.1 The Wilcoxon Rank Sum Test

We first demonstrate the Wilcoxon rank sum test (also called the Mann-Whitney test) on data from two populations. The Wilcoxon test statistic  $W$  is the sum of the ranks from the first sample. Assuming that the two populations have the same continuous distribution (and no ties occur), then  $W$  has a mean and standard deviation given by

$$\mu = \frac{n_1(N+1)}{2} \quad \text{and} \quad \sigma = \sqrt{\frac{n_1 n_2 (N+1)}{12}}$$

where  $n_1$  is the first sample size,  $n_2$  is the second sample size, and  $N = n_1 + n_2$ .

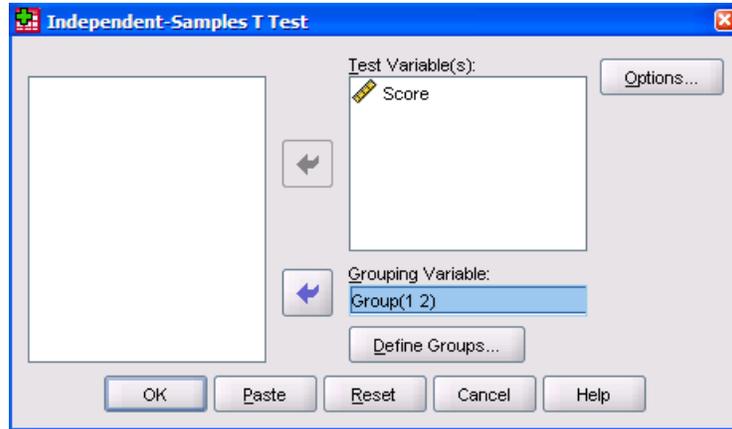
We test the null hypothesis  $H_0$ : there is no difference in distributions. A one-sided alternative is  $H_A$ : the first population yields higher measurements. We use this alternative if we expect or see that  $W$  is a much higher sum than its expected sum of ranks  $\mu$ . In this case, the  $p$ -value is given by a normal approximation. We let  $W \sim N(\mu, \sigma)$  and compute the right-tail  $P(X \geq W)$  (using the continuity correction if  $W$  is an integer).

If we expect or see that  $W$  is the much lower sum than its expected sum of ranks  $\mu$ , then we should use the alternative  $H_A$ : first population yields lower measurements. In this case, the  $p$ -value is given by the left-tail  $P(X \leq W)$ , again using continuity correction if needed. If the two sums of ranks are close, then we could use a two-sided alternative  $H_A$ : there is a difference in distributions. In this case, the  $p$ -value is given by twice the smallest tail value:  $2 * P(X \geq W)$  if  $W > \mu$ , or  $2 * P(X \leq W)$  if  $W < \mu$ .

**Example 15.1 Retelling Stories.** Below are language usage scores of kindergarten students who were classified as high-progress readers or low-progress readers when asked to retell a story that had been read to them. Is there evidence that the scores of high-progress readers are higher than those of low-progress readers? Carry out a two-sample  $t$  test. Then carry out the Wilcoxon rank sum test and compare the conclusions for each test.

Child	Progress	Score
1	high	0.55
2	high	0.57
3	high	0.72
4	high	0.70
5	high	0.84
6	low	0.40
7	low	0.72
8	low	0.00
9	low	0.36
10	low	0.55

*Solution.* After defining variables and entering the data, we use **Analyze, Compare Means, Independent Samples T Test** to test the hypothesis  $H_0: \mu_1 = \mu_2$  versus the alternative  $H_A: \mu_1 < \mu_2$ , where group 1 is the low-progress group and group 2 is the high-progress group.



**Group Statistics**

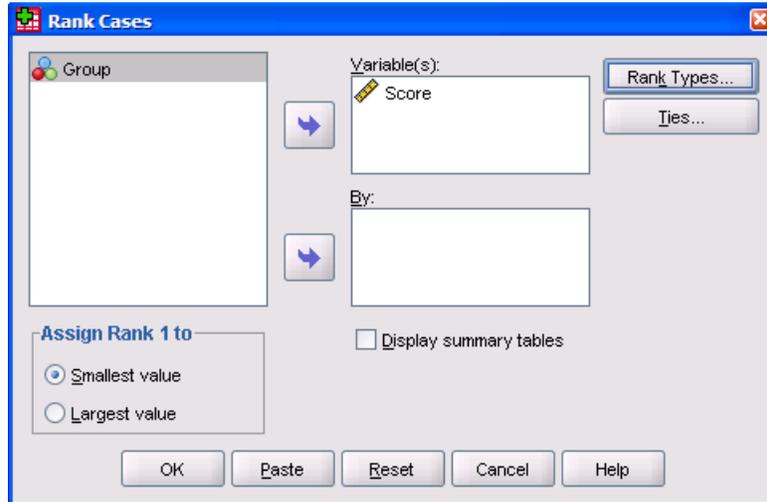
	Group	N	Mean	Std. Deviation	Std. Error Mean
Score	1	5	.4060	.26754	.11965
	2	5	.6760	.11887	.05316

		t-test for Equality of Means						
		t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
							Lower	Upper
Score	Equal variances assumed	-2.062	8	.073	-.27000	.13093	-.57192	.03192
	Equal variances not assumed	-2.062	5.520	.089	-.27000	.13093	-.59724	.05724

We obtain a  $t$  statistic of  $-2.06$  and a one-sided  $p$ -value of  $0.045$  ( $.089/2$ ). With the rather small  $p$ -value, we have significant evidence to reject  $H_0$  and say that the average score of all high-progress readers is higher than the average score of all low-progress readers. For if  $H_0$  were true, then there would be only a  $0.0445$  probability of obtaining a high-progress sample mean that is so much larger than the low-progress sample mean ( $0.676$  compared to  $0.406$ ).

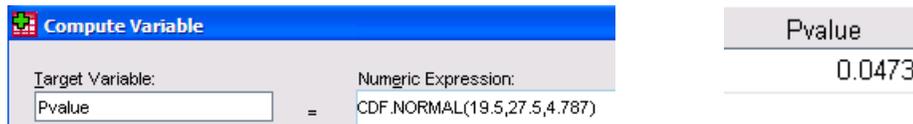
Now for the Wilcoxon (Mann-Whitney) rank sum test, we use  $H_0$ : the distribution is the same for both groups versus  $H_A$ : high-progress readers score higher when retelling the story. The test statistic will be the sum of ranks from the low-progress group.

To calculate these, click **Transform, Rank Cases**. A new variable called RScore will be created.

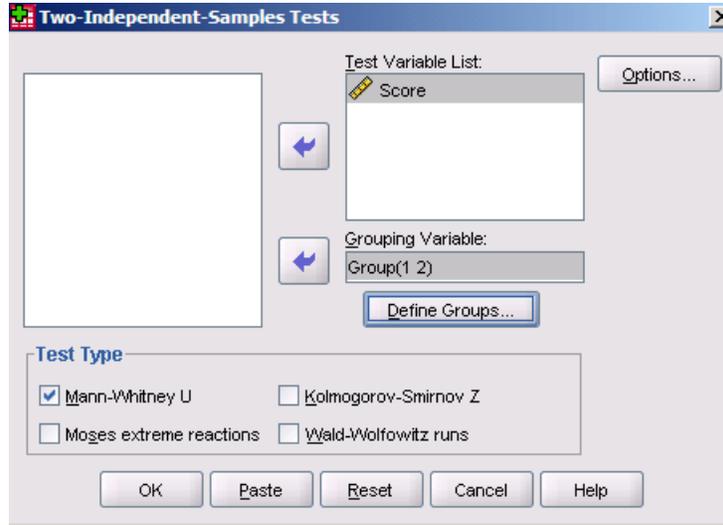


	Score	Group	RScore
1	0.55	2	4.500
2	0.57	2	6.000
3	0.72	2	8.500
4	0.70	2	7.000
5	0.84	2	10.000
6	0.40	1	3.000
7	0.72	1	8.500
8	0.00	1	1.000
9	0.36	1	2.000
10	0.55	1	4.500

The sum of the ranks from the low-progress readers is 19, which is lower than the expected average of  $\mu = (5*11)/2 = 27.5$ . To find the  $p$ -value for the test, we need to calculate the standard deviation  $\sigma = \sqrt{5*5*11/12} = 4.787$ . We now use **Transform, Compute Variable** to find the  $p$ -value for the test using the continuity correction, so instead of an upper bound of  $W = 19$ , we use  $W = 19.5$ .



If you have SPSS Base (where nonparametric tests are implemented), click **Analyze, Nonparametric Tests, Two Independent Samples Tests**. The test definition dialog box is very similar to that for the independent samples  $t$  test. Define the groups and click OK since we want the Mann-Whitney test.



**Ranks**

	Group	N	Mean Rank	Sum of Ranks
Score	1	5	3.80	19.00
	2	5	7.20	36.00
	Total	10		

**Test Statistics<sup>b</sup>**

	Score
Mann-Whitney U	4.000
Wilcoxon W	19.000
Z	-1.786
Asymp. Sig. (2-tailed)	.074
Exact Sig. [2*(1-tailed Sig.)]	.095 <sup>a</sup>

a. Not corrected for ties.

b. Grouping Variable: Group

According to the Wilcoxon test, if the distributions were the same, then there would be only a 0.0473 probability (from the left-tail value) of the low-progress sum of ranks being so much smaller than the expected average of 27.5. Therefore, we should reject  $H_0$  in favor of the alternative; high-progress children are better at retelling stories told to them.

In this case, the Wilcoxon  $p$ -value is slightly higher than the  $t$  test  $p$ -value; however, both are low enough to result in the same conclusion.

**Example 15.2 Logging in the Rainforest.** Below is a comparison of the number of tree species in unlogged plots in the rain forest of Borneo with the number of species in plots logged eight years earlier.

<b>Unlogged</b>	22	18	22	20	15	21	13	13	19	13	19	15
<b>Logged</b>	17	4	18	14	18	15	15	10	12			

Does logging significantly reduce the mean (median) number of species in a plot after eight years? State the hypotheses, do a Wilcoxon rank sum test, and state your conclusion.

*Solution.* We will test the hypothesis  $H_0$ : there is no difference in medians (or distributions) versus the alternative  $H_A$ : the unlogged median is higher. To do so, we first enter the data along with a variable to indicate whether or not the plot was unlogged or logged. Then we use **Transform, Rank Cases** to find the ranks.

	Species	Logged	RSpecies
1	22	No	20.500
2	18	No	14.000
3	22	No	20.500
4	21	No	18.500
5	15	No	9.500
6	21	No	18.500
7	13	No	5.000
8	13	No	5.000
9	19	No	16.500
10	13	No	5.000
11	19	No	16.500
12	15	No	9.500
13	17	Yes	12.000

We note that there are 21 total measurements with 12 unlogged measurements. Adding the unlogged ranks, we find  $W = 159$ . If there were no difference in distributions or medians, then we would expect the sum of ranks from the unlogged plots to be  $\mu = 12 * 22 / 2 = 132$  with  $\sigma = \sqrt{12 * 9 * 22 / 12} = 14.071$ . Again using the continuity correction, we find the  $p$ -value for the test.

The screenshot shows the 'Compute Variable' dialog box in SPSS. The 'Target Variable' is 'Pvalue' and the 'Numeric Expression' is '1-CDF.NORMAL(158.5,132,14.071)'. To the right, a separate box displays the 'Pvalue' as 0.0298.

But if there were no difference in medians, then there would be only a 2.98% chance of the sum of ranks from the unlogged plots being as high as 159. This low  $p$ -value gives significant evidence to reject  $H_0$  in favor of the alternative; there are more species of trees in the unlogged plots.

## 15.2 The Wilcoxon Signed Rank Test

Here we demonstrate how to perform the Wilcoxon signed rank test on data sets of size  $n$  from two populations (this can also be used to test a single median). The Wilcoxon test statistic  $W$  is the sum of the ranks from the positive differences. Assuming that the two populations have the same continuous distribution (and no ties occur), then  $W$  has a mean and standard deviation given by

$$\mu = \frac{n(n+1)}{4} \quad \text{and} \quad \sigma = \sqrt{\frac{n(n+1)(2n+1)}{24}}$$

We test the null hypothesis  $H_0$ : there is no difference in distributions. A one-sided alternative may be  $H_A$ : the second population yields higher measurements. We use this alternative if we expect or see that  $W$  is a much higher sum, which means that there were more positive differences. In this case, the  $p$ -value is again given by a normal approximation. We let  $X \sim N(\mu, \sigma)$  and compute the right-tail  $P(X \geq W)$  (using a continuity correction if  $W$  is an integer).

If we expect or see that  $W$  is the much lower sum, then there were more negative differences. Now we should use the alternative  $H_A$ : the second population yields lower measurements. In this case, the  $p$ -value is given by the left-tail  $P(X \leq W)$ , again using continuity correction if needed. If the two sums of ranks are close, we could use a two-sided alternative  $H_A$ : there is a difference in distributions. In this case, the  $p$ -value is given by twice the smallest tail value.

**Example 15.3 Stepping and Heart Rates.** A student project asked subjects to step up and down for three minutes. There were two treatments: stepping at a low rate (14 steps per minute) and a medium rate (21 steps per minute). Here are data for heart rates for five subjects and the two treatments. For each subject we have the initial resting heart rate and the heart rate at the end of the exercise. Does exercise at the low rate raise the heart rate significantly? State hypotheses in terms of the median increase in heart rate and apply the Wilcoxon signed rank test.

Subject	Low Rate		Medium Rate	
	Resting	Final	Resting	Final
1	60	75	63	84
2	90	99	69	93
3	87	93	81	96
4	78	87	75	90
5	84	84	90	108

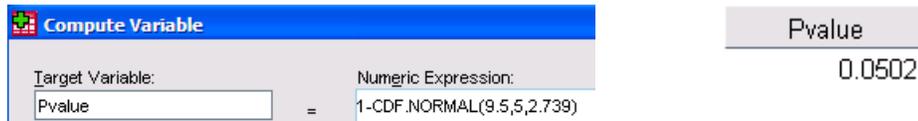
*Solution.* We will test the hypothesis  $H_0$ : For the low rate, resting and final heart rates have the same median versus  $H_A$ : final heart rates are higher. First, we ignore the last individual who had no difference in heart rate. Enter the other four low rate resting heart rates into a variable and the four low rate final heart rates in another. Use **Transform, Compute Variable** to find the differences.



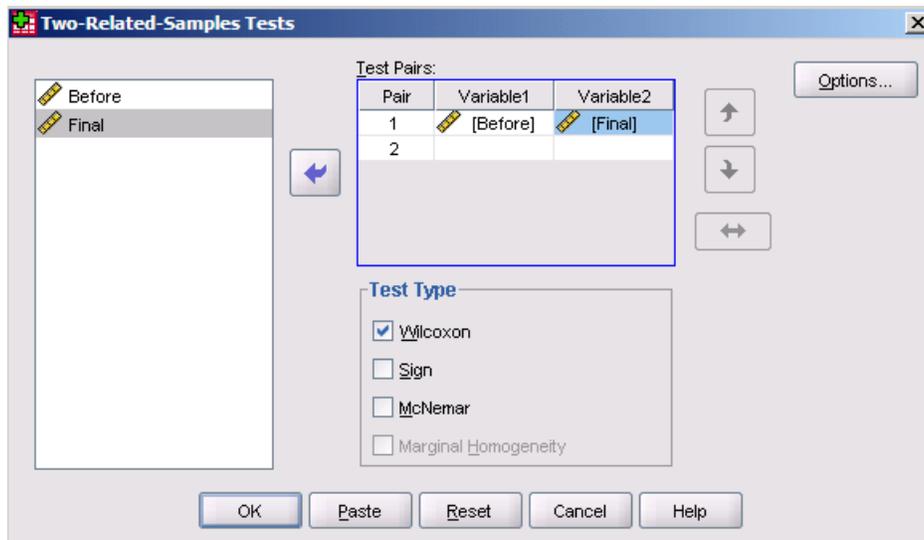
Now, use **Transform, Rank Cases** to rank the differences.

Before	Final	Difference	RDiffere
60	75	15.00	4.000
90	99	9.00	2.500
87	93	6.00	1.000
78	87	9.00	2.500

The alternative means that there should be more positive differences, so that the sum of the positive ranks should be higher. Therefore, the  $p$ -value comes from the right-tail probability created by the test statistic. All the differences are positive, and the sum of these positive ranks is 10. If there were no difference, we would have  $\mu = 4 * 5 / 4 = 5$  and  $\sigma \sqrt{4 * 5 * 9 / 24} = 2.739$ . Again, we use **Transform, Compute Variable** and the continuity correction to find the  $p$ -value for the test.



To do this test with SPSS Base, click **Analyze, Nonparametric Tests, 2 Related Samples**. Click to add in the two variables and **OK** to perform the calculations.



		N	Mean Rank	Sum of Ranks
Final - Before	Negative Ranks	0 <sup>a</sup>	.00	.00
	Positive Ranks	4 <sup>b</sup>	2.50	10.00
	Ties	1 <sup>c</sup>		
	Total	5		

a. Final < Before

b. Final > Before

c. Final = Before

	Final - Before
Z	-1.841 <sup>a</sup>
Asymp. Sig. (2-tailed)	.066

a. Based on negative ranks.

b. Wilcoxon Signed Ranks Test

There is some difference in the final  $p$ -values due to rounding the standard deviation in the “hand” calculations.

We see that the sum of the ranks of the positive differences is much higher than that of the negative differences. If the medians for each rate were the same, then there would be only a 0.0502 (or .033 one-tailed from SPSS) probability of the sum of positive ranks being as high as 10 when expected to be 5 with the four subjects for which there is a difference. The relatively low  $p$ -value provides some evidence to reject  $H_0$  and conclude that the median final heart rate is higher for the low rate test.

**Example 15.4 Radon Detector Accuracy.** Below are the readings from 12 home radon detectors exposed to 105 pCi/l of radon. We want to know if these detectors are accurate. Apply the Wilcoxon signed rank test to determine if the median reading from all such home radon detectors differs significantly from 105.

91.9	97.8	111.4	122.3	105.4	95.0
103.8	99.6	96.6	119.3	104.8	101.7

*Solution.* We will test the null hypothesis  $H_0$ : median = 105 versus  $H_A$ : median  $\neq$  105. First, we enter the given data and then enter 105 twelve times into a second variable. If  $H_0$  were true, then we would expect the sum of ranked positive differences to be  $\mu = (12 \cdot 13) / 4 = 39$  with  $\sigma = \sqrt{12 \cdot 13 \cdot 25 / 24} = 12.7475$ . But  $H_A$  implies that this sum of ranked positive differences will be either much higher than 39 or much lower than 39.



We first find the absolute value of the actual differences and rank that. Then we will add the ranks attached to the positive differences to find the test statistic. Find the ABS function in the Arithmetic function group.



Reading	Exposure	Difference	AbsDiff	RAbsDiff
91.9	105	13.10	13.10	10.000
97.8	105	7.20	7.20	7.000
111.4	105	-6.40	6.40	6.000
122.3	105	-17.30	17.30	12.000
105.4	105	-0.40	0.40	2.000
95.0	105	10.00	10.00	9.000
103.8	105	1.20	1.20	3.000
99.6	105	5.40	5.40	5.000
96.6	105	8.40	8.40	8.000
119.3	105	-14.30	14.30	11.000
104.8	105	0.20	0.20	1.000
101.7	105	3.30	3.30	4.000

We see that there were 12 nonzero differences, and that eight of these were positive differences, meaning that there were eight times in which the home radon detector measured below 105. The sum of the ranks for the positive differences is 47. Again using the continuity correction, we find the *p*-value for the test.



Pvalue
0.2781

The right-tail value is given as 0.2781; thus, the *p*-value for the two-sided alternative is  $2 \times 0.2781 = 0.5562$ . If the median home radon measurement were 105, then there would

be a 0.5562 probability of the sum of positive ranks being as far away (in either direction) from the expected sum of 39 as the resulting sum of 47 is. Thus, we do not have significant evidence to reject  $H_0$  and can conclude that, “on average,” these detectors are accurate.

### 15.3 The Kruskal-Wallis Test

Our next demonstration is for the Kruskal-Wallis test, which simultaneously compares the distribution of more than two populations. We test the null hypothesis  $H_0$ : all populations have the same distribution versus the alternative  $H_A$ : measurements are systematically higher in some populations. To apply the test, we draw independent SRSs of sizes  $n_1, n_2, \dots, n_I$  from  $I$  populations. There are  $N$  observations in all. We rank all  $N$  observations and let  $R_i$  be the sum of the ranks for the  $i$ th sample. The Kruskal-Wallis statistic is

$$H = \frac{12}{N(N+1)} \sum_{i=1}^I \frac{R_i^2}{n_i} - 3(N+1)$$

When the sample sizes are large and all  $I$  populations have the same continuous distribution, then  $H$  has an approximate chi-square distribution with  $I-1$  degrees of freedom. When  $H$  is large, creating a small right-tail  $p$ -value, then we can reject the hypothesis that all populations have the same distribution.

**Example 15.5 Are Insects Colorblind?** An experiment was conducted to determine if insects were equally attracted by different colors. Sticky boards were placed in a field of oats and the number of cereal leaf beetles trapped was counted. Use the Kruskal-Wallis test to see if there are significant differences in the numbers of insects trapped by the different board colors.

Board color	Insects trapped					
Lemon yellow	45	59	48	46	38	47
White	21	12	14	17	13	17
Green	37	32	15	25	39	41
Blue	16	11	20	21	14	7

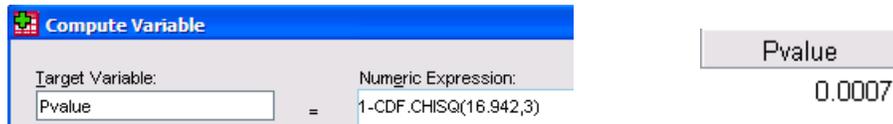
*Solution.* Enter the data and colors, then rank the observations as has been done previously in this chapter.

Insects	Color	RInsects
45	Yellow	20.000
59	Yellow	24.000
48	Yellow	23.000
46	Yellow	21.000
38	Yellow	17.000
47	Yellow	22.000
21	White	12.500
12	White	3.000
13	White	4.500
17	White	9.500
13	White	4.500
17	White	9.500
37	Green	16.000
32	Green	15.000
15	Green	7.000
25	Green	14.000
39	Green	18.000
41	Green	19.000
16	Blue	8.000
11	Blue	2.000
20	Blue	11.000
21	Blue	12.500
14	Blue	6.000
7	Blue	1.000

The sum of the ranks for Yellow is 127; for White the sum is 43.5; for Green the sum is 89; for Blue the sum is 40.5. Our test statistic becomes

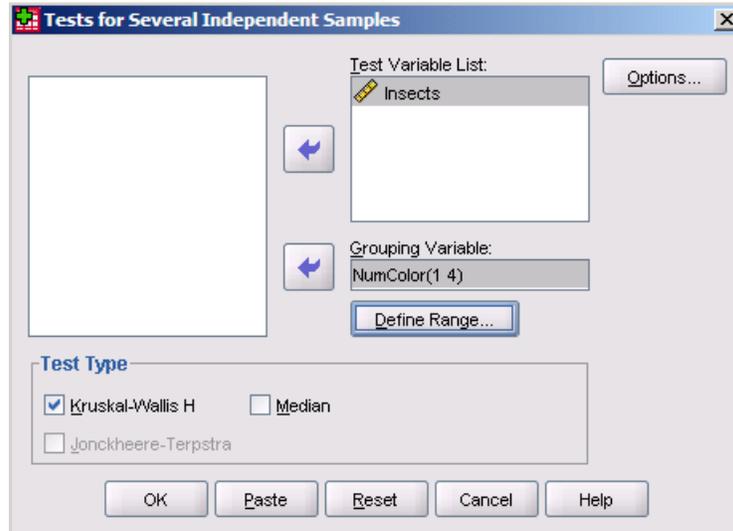
$$H = \frac{12}{24 * 25} \left( \frac{127^2}{6} + \frac{43.5^2}{6} + \frac{89^2}{6} + \frac{40.5^2}{6} \right) - 3 * 25 = 16.942$$

We find the *p*-value for the test using a chi-squared distribution with 4-1 = 3 degrees of freedom.



The low *p*-value of 0.0007 gives good evidence to reject the hypothesis that all colors yield the same distribution of insects trapped.

With SPSS Base, we must have color a numeric variable, instead of a name. Define a new variable called **NumColor**. In it code 1 for Yellow, 2 for White, 3 for Green, and 4 for Blue. Click **Analyze, Nonparametric Tests, K Independent Samples**.



Here, instead of naming the groups, we specify the smallest and largest group number. Click **OK** to perform the test.

**Ranks**

	NumColor	N	Mean Rank
Insects	1	6	21.17
	2	6	7.33
	3	6	14.83
	4	6	6.67
Total		24	

**Test Statistics<sup>a,b</sup>**

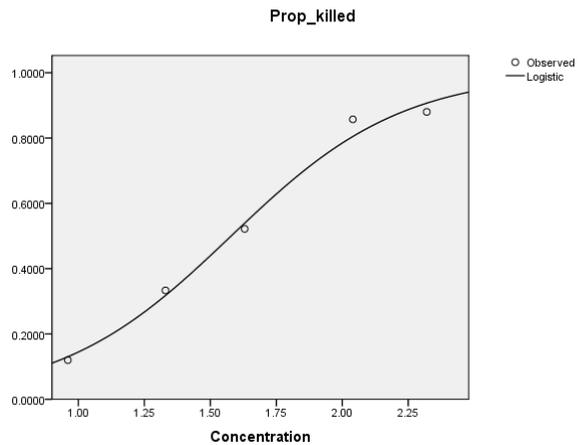
	Insects
Chi-Square	16.975
df	3
Asymp. Sig.	.001

a. Kruskal Wallis Test  
 b. Grouping Variable:  
 NumColor

With SPSS, to three decimal places, we have a  $p$ -value of 0.001. The conclusion is the same as before; all colors do not attract insects (cereal leaf beetles) equally; it seems clear that Yellow is best, then Green. Blue and White are about equal in their attractiveness, based on the mean ranks.

## CHAPTER

# 16



# Logistic Regression

## 16.1 | The Logistic Regression Model

### Introduction

True Logistic Regression in SPSS (and inference for the model) requires an add-on module. However, we can estimate the parameters and odds ratios. In this chapter, we show how to compute odds; we also give a brief discussion of two types of logistic regression fits. The first type is a linear fit for the logarithm of the odds ratio of two population proportions. The second type is the general logistic fit for several population proportions.

## 16.1 The Logistic Regression Model

First, we compute the appropriate mathematical odds for a given probability  $p$  of an event  $A$ . If  $p \leq 0.50$ , then the odds *against*  $A$  are given as the ratio  $(1 - p) : p$ . If  $p > 0.50$ , then the odds *in favor of*  $A$  are given as the ratio  $p : (1 - p)$ . This is done using **Transform, Compute Variable**.

**Example 16.1 Employee Stock Options.** In a study of 91 high-tech companies and 109 non-high-tech companies, 73 of the high-tech companies and 75 of the non-high-tech companies offered incentive stock options to key employees.

- What proportion of high-tech companies offer stock options to their key employees? What are the odds?
- What proportion of non-high-tech companies offer stock options to their key employees? What are the odds?
- Find the odds ratio using the odds for the high-tech companies in the numerator.

*Solution.* (a) and (b): We define two variables and enter the numbers of companies who provide options along with the total number of companies in each category.

	Options	Total
1	73	91
2	75	109

First compute  $\hat{p}$ , the observed proportion of each type of that offers options.

Compute Variable		Phat
Target Variable:	Numeric Expression:	0.8022
Phat	Options/Total	0.6881

We see that 80.22% of the high-tech companies offer stock options; only 68.81% of the non-high-tech offer stock options. To compute the odds (since both proportions are more than half), use the following on **Transform, Compute Variable**.

Compute Variable		Odds
Target Variable:	Numeric Expression:	4.06
Odds	Phat/(1-Phat)	2.21

Notice that SPSS does not give these odds in the familiar ratio form, but rather in a decimal form.

(c) The odds-in-favor ratio can be computed by simple division of the odds  $4.06/2.21=1.84$ . Thus, the odds in favor of a high-tech company offering stock options are about 1.84 times more than the odds for a non-high-tech company.

**Example 16.2 Gender Bias.** In a study on gender bias in textbooks, 48 out of 60 female references were “girl.” Also, 52 out of 132 male references were “boy.” These two types of references were denoted as juvenile references. Compute the odds ratio for comparing the female juvenile references to the male juvenile references.

*Solution.* Following the example above, we simply enter the data, compute **Phat** and **Odds**.

	Juvenile	Total	Phat	Odds
1	48	60	0.8000	4.00
2	52	132	0.3939	0.65

The odds a female reference was juvenile was 4, while the odds for male references was 0.65. To find the odds ratio, we have  $4/.65 = 6.15$ . This means a female reference was more than six times as likely to be juvenile than a male reference.

### Model for Logistic Regression

The logistic regression model is given by the equation

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x$$

where  $\ln$  is the natural (base  $e$ ) logarithm and  $x$  is either 1 or 0 to designate the explanatory variable.

**Example 16.3 Binge Drinking on Campus.** The table below gives data on the numbers of men and women who responded “Yes” to being frequent binge drinkers in a survey of college students. Find the coefficients for the logistic regression model and the odds ratio of men to women.

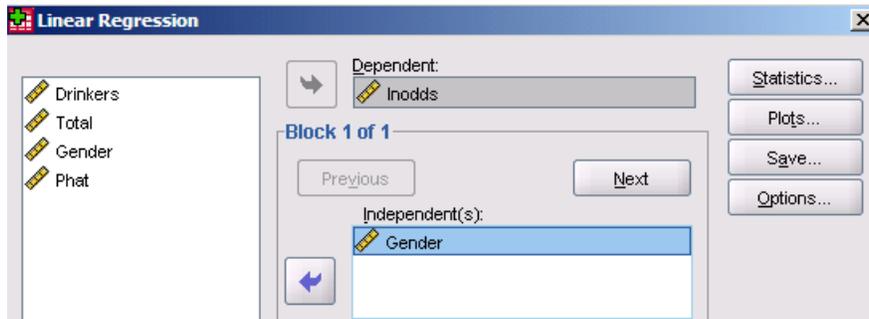
Population	$X$	$n$
Men	1630	7180
Women	1684	9916

*Solution.* As in Examples 16.1 and 16.2, we enter the data and compute **Phat**. We also entered a variable to indicate Gender (1 = Male, 0 = Female). Next, we compute the log of the odds ratio as **Inodds**.



Drinkers	Total	Gender	Phat	Inodds
1630	7180	1	0.23	-1.23
1684	9916	0	0.17	-1.59

Now, we compute the linear regression using **Analyze, Regression, Linear** with **Inodds** as the dependent (y) variable and **Gender** as the independent variable.



**Coefficients<sup>a</sup>**

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-1.587	.000			.
	Gender	.362	.000	1.000		.

a. Dependent Variable: Inodds

We obtain the regression equation  $\log(ODDS) = -1.587 + .362Gender$ . No *t* statistic is shown because with only two data points, the fit has an  $r^2$  of 1. As mentioned in the introduction to this chapter, inference for logistic regression requires an add-on module.

**Example 16.4 Insecticide Effectiveness.** An experiment was designed to examine how well an insecticide kills a certain type of insect. Find the logistic regression for the proportion of insects killed as a function of the insecticide (log) concentration.

(log) Concentration	Number of insects	Number killed
0.96	50	6
1.33	48	16
1.63	46	24
2.04	49	42
2.32	50	44

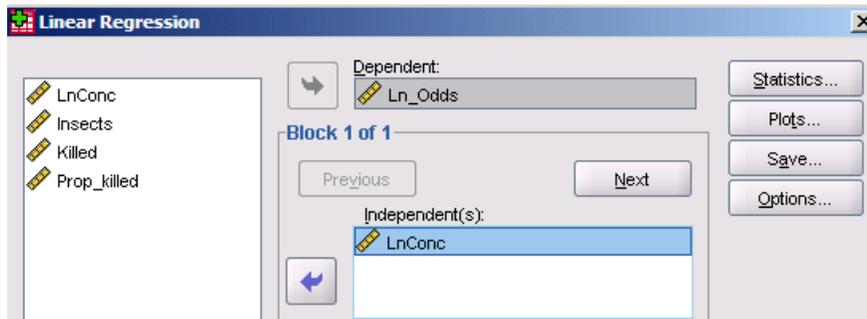
*Solution.* First, enter the Concentrations total number of insects and number killed. So that we do not encounter rounding errors, compute the proportion of insects killed at each concentration.



Next, compute the log odds.



Next, use **Analyze, Regression, Linear** to fit the logistic model.



**Coefficients<sup>a</sup>**

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-4.846	.449		-10.786	.002
	LnConc	3.070	.260	.989	11.790	.001

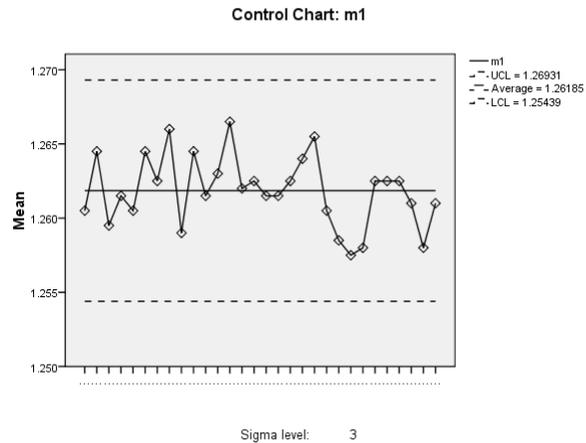
a. Dependent Variable: Ln\_Odds

The logistic regression equation is  $\ln(Odds) = -4.846 + .3070 * \ln(Concentration)$ , or

$\frac{P}{1-p} = e^{-4.846 + .3070 * \ln(Concentration)}$ . Obtaining the Wald Statistic requires the logistic regression add-on module; however, we can see from the linear regression output that with a  $t$  statistic of 11.79 and  $p$ -value .001, the slope of this logistic regression is significantly nonzero.

## CHAPTER

# 17



# Statistics for Quality: Control and Capability

17.1	Statistical Process Control
17.2	Process Capability Indexes
17.3	Control Charts for Sample Proportions

## Introduction

In this chapter, we examine graphing control charts, and computing the capability indices of a process. SPSS will generate the charts for you, based on sampled data. SPSS will *not* create control charts for data in which the samples have already been summarized, or if there is only a single observation per “sample.”

## 17.1 Statistical Process Control

SPSS constructs control charts based solely on past data; however, if you want to specify the limits based on  $\mu \pm 3\sigma/\sqrt{n}$ , this is an option.

**Example 17.1** A manufacturer of computer monitors must control the tension on the mesh of fine vertical wires that lies behind the surface of the viewing screen. Too much tension will tear the mesh, and too little will allow wrinkles. The following mesh tension data gives the data from 20 different samples of size 4.

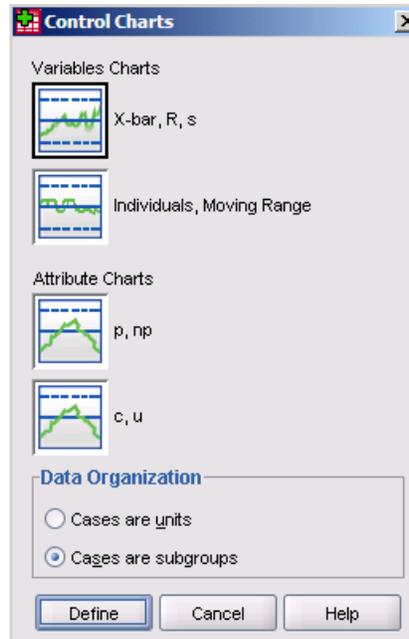
Sample	Measure 1	Measure 2	Measure 3	Measure 4
1	234.5	272.3	234.5	272.3
2	311.1	305.8	238.5	286.2
3	247.1	205.3	252.6	316.1
4	215.4	296.8	274.2	256.8
5	327.9	247.2	283.3	232.6
6	304.3	236.3	201.8	238.5
7	268.9	276.2	275.6	240.2
8	282.1	247.7	259.8	272.8
9	260.8	259.9	247.9	345.3
10	329.3	231.8	307.2	273.4
11	266.4	249.7	231.5	265.2
12	168.8	330.9	333.6	318.3
13	349.9	334.2	292.3	301.5
14	235.2	283.1	245.9	263.1
15	257.3	218.4	296.2	275.2
16	235.1	252.7	300.6	297.6
17	286.3	293.8	236.2	275.3
18	328.1	272.6	329.7	260.1
19	316.4	287.4	373.0	286.0
20	296.8	350.5	280.6	259.8

The target mean tension is  $\mu = 275$  mV with a target standard deviation of 43 mV. Find the center line and control limits for  $\bar{x}$  and for  $s$ . Graph the control charts for each.

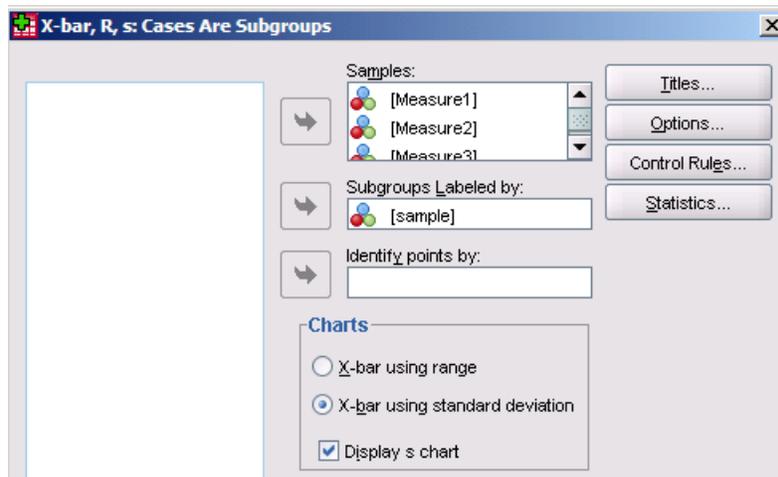
*Solution.* Enter the data into five columns. The first few rows are shown below.

	sample	Measure1	Measure2	Measure3	Measure4
1	1	234.5	272.3	234.5	272.3
2	2	311.1	305.8	238.5	286.2
3	3	247.1	205.3	252.6	316.1
4	4	215.4	296.8	274.2	256.8

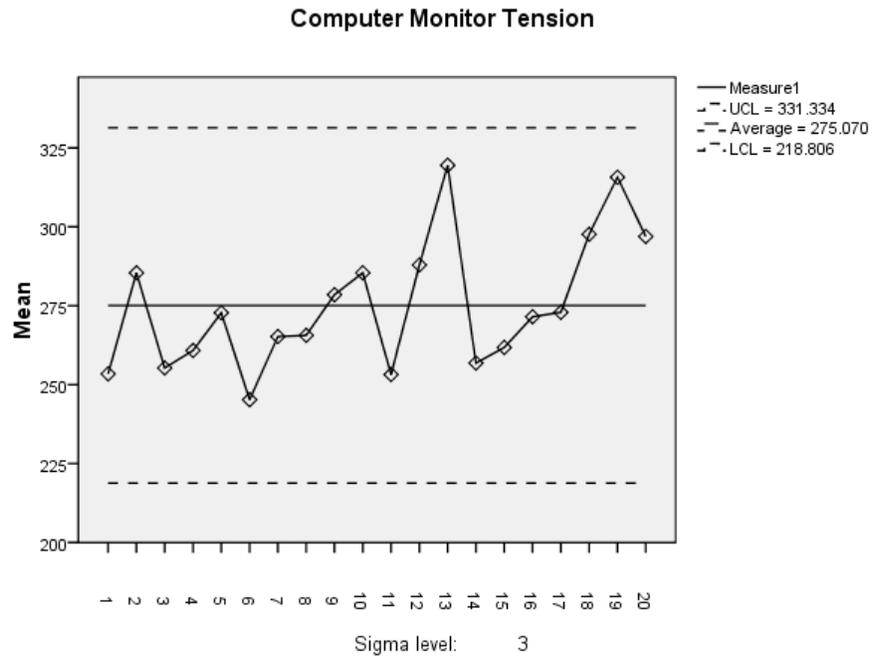
Click **Analyze, Quality Control, Control Charts**. Since our data have one row for each sample, move the button to **Cases are subgroups** on the initial dialog box as shown below. Click **Define** to proceed.



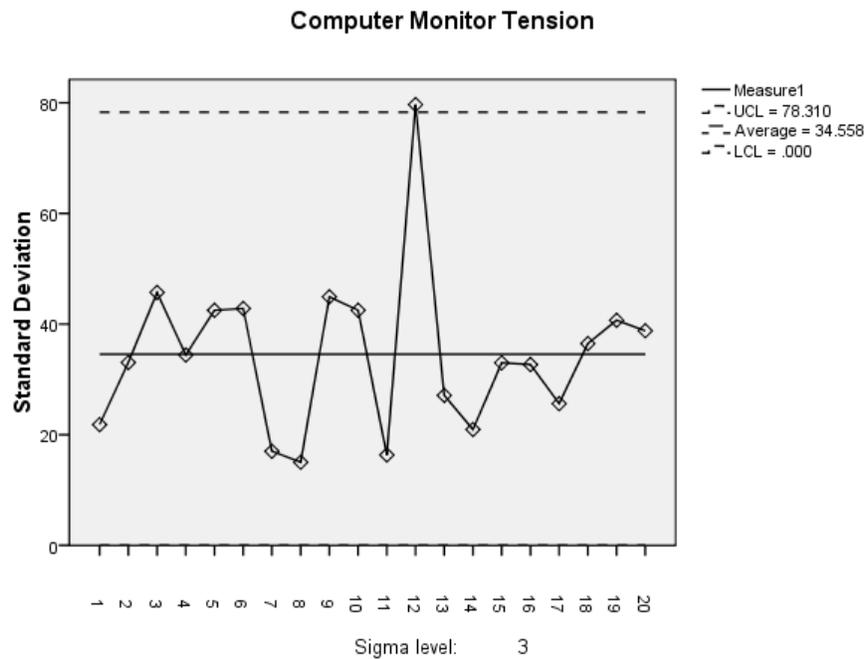
Hold down the **ctrl** key and click to select **Measure1** through **Measure4** as the **Samples** and **sample** as the label for **Subgroups**. Move the button to **X-bar using standard deviation**. Give your chart a title. Leave the Display s chart box checked (since we want that one as well).



The first chart shown is the  $\bar{x}$  chart, constructed at a default  $3\sigma$  level. If you wanted something different, click **Options** in the definition box and change the **Number of Sigmas**.



We see the center line of the chart is 275.07 (very close to the desired 275) and the control limits are 218.806 and 331.334.



The s chart is similar; however, sample 12 has a standard deviation (79.7) above the upper control limit (78.3). This chart is not exactly the one in your text; SPSS uses

sample values while the text uses hypothesized values to construct the chart; the value of  $\bar{s}$  was 34.558, smaller than the specified 43.

Since both charts are basically within the control limits, this process is in control; we see only common cause variation around the desired values of both the mean and standard deviation.

## 17.2 Process Capability Indices

In this section, we show how to compute the capability indexes.

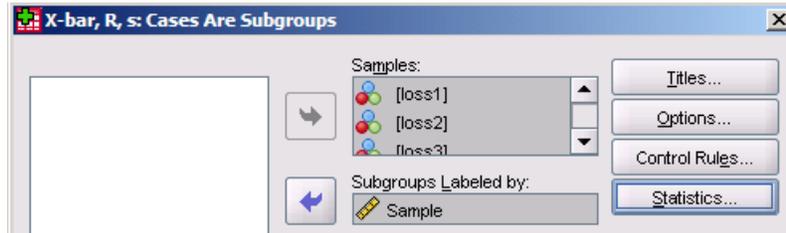
**Example 17.2 Hospital Losses.** Below are data on a hospital’s losses for 120 DRG 209 (major joint replacement) patients collected as 15 monthly samples of eight patients each. The hospital has determined that suitable specification limits for its loss in treating one such patient are LSL = \$4000 and USL = \$8000.

(a) Estimate the percent of losses that meet the specifications.

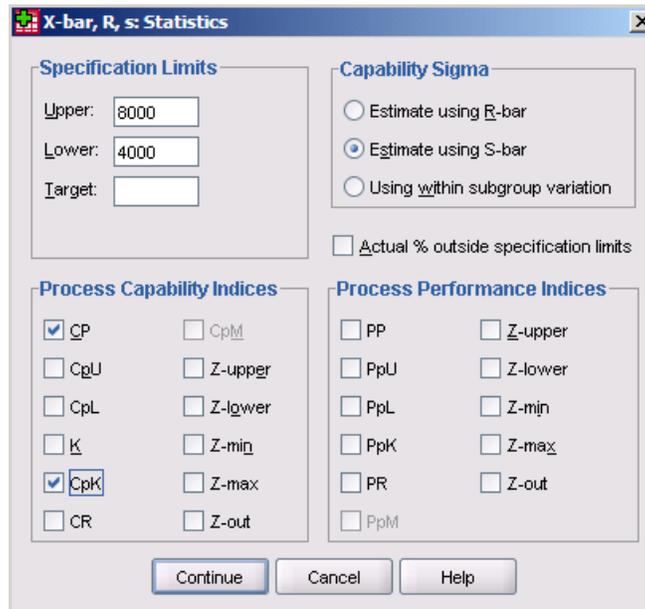
(b) Estimate  $C_p$  and  $C_{pk}$ .

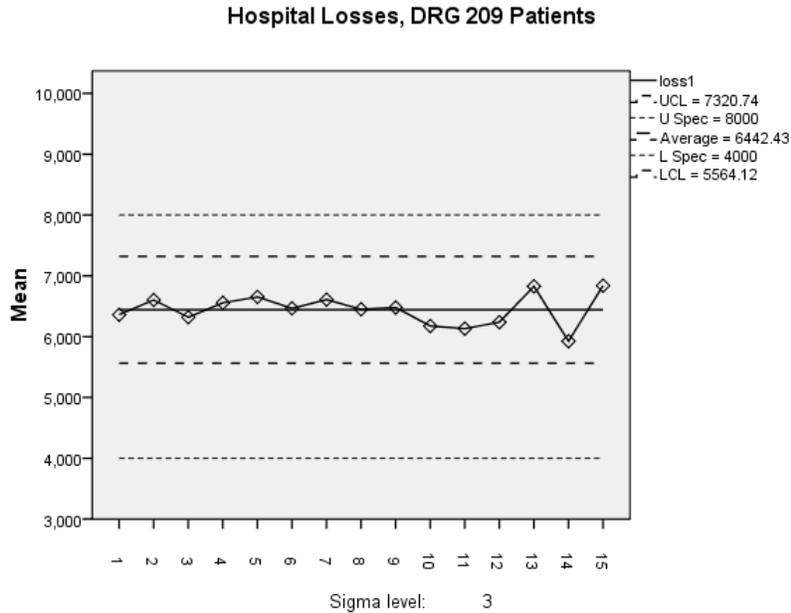
Sample	Loss (dollars)							
1	6835	5843	6019	6731	6362	5696	7193	6206
2	6452	6764	7083	7352	5239	6911	7479	5549
3	7205	6364	6198	6170	6482	4763	7125	6241
4	6021	6347	7210	6384	6807	5711	7952	6023
5	7000	6495	6893	6127	7417	7044	6159	6091
6	7783	6224	5051	7288	6584	7521	6146	5129
7	8794	6279	6877	5807	6076	6392	7429	5220
8	4727	8117	6586	6225	6150	7386	5674	6740
9	5408	7452	6686	6428	6425	7380	5789	6264
10	5598	7489	6186	5837	6769	5471	5658	6393
11	6559	5855	4928	5897	7532	5663	4746	7879
12	6824	7320	5331	6204	6027	5987	6033	6177
13	6503	8213	5417	6360	6711	6907	6625	7888
14	5622	6321	6325	6634	5075	6209	4832	6386
15	6269	6756	7653	6065	5835	7337	6615	8181

*Solution.* First, enter the sample data into eight columns, with a ninth for the sample number. Click **Analyze, Quality Control, Control Charts**. If you have entered the data like my table, we have the case where **Cases are subgroups**.



Next, click **Statistics**. The hospital has determined that an acceptable UCL is \$8000 and the acceptable LCL is \$4000. Enter these in the **Specification Limits**; click to **Estimate using S-bar**, and click to select the indices **CP** and **CpK**. Click **Continue**, then **OK**.





In the above graph, we see the average loss for these patients is well in control and clearly inside the specified limits.

#### Process Statistics

Act. % Outside SL		3.333
Capability Indices	CP <sup>a</sup>	.805
	CpK <sup>a</sup>	.627

The normal distribution is assumed. LSL = 4000 and USL = 8000.

a. The estimated capability sigma is based on the mean of the sample group standard deviations.

SPSS tells us that, assuming the Normal distribution, with our sample data, 3.33% of the actual observations were outside the desired limits. We have a Cp of .085 and CpK of .627. These may not exactly match values calculated with other resources due to rounding.

## 17.4 Control Charts for Sample Proportions

We conclude this chapter with a look at control charts for sample proportions when we have data on each sample's size and number with the desired attribute.

**Example 17.3 School Absenteeism.** Here are data on the total number of absentees among eighth graders with three or more unexcused absences at an urban school district. Because the total number of students varies each month, these totals are also given for each month.

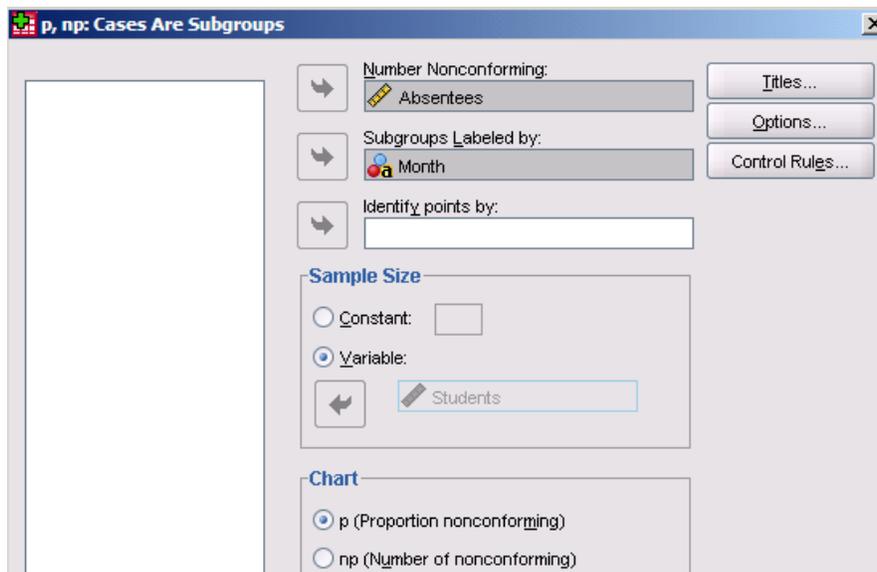
Month	Sep.	Oct.	Nov.	Dec.	Jan.	Feb.	Mar.	Apr.	May	Jun.
Students	911	947	939	942	918	920	931	925	902	883
Absent	291	349	364	335	301	322	344	324	303	344

(a) Find  $\bar{p}$  and  $\bar{n}$ . (b) Make a  $p$  chart using control limits based on  $\bar{n}$  students each month.

*Solution.* Enter the Months, Student numbers, and Absentees.

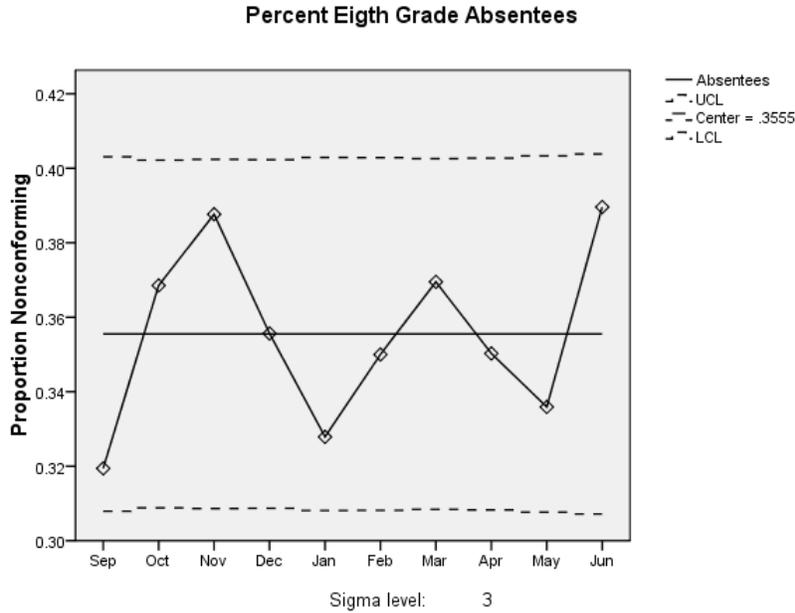
Month	Students	Absentees
Sep	911.00	291.00
Oct	947.00	349.00
Nov	939.00	364.00
Dec	942.00	335.00
Jan	918.00	301.00
Feb	920.00	322.00
Mar	931.00	344.00
Apr	925.00	324.00
May	902.00	303.00
Jun	883.00	344.00

Click **Analyze**, **Quality Control**, **Control Charts**. We want a  $p$ , $np$  chart where data are organized as Cases are subgroups. Click **Define**.



The **Number Nonconforming** are the **Absentees**, the **Subgroups** are the **Months**, and total **Sample Size** is **Students** (this variable does not show well in the screen capture above; it immediately is dimmed by SPSS). Notice that we can select for either the

proportion nonconforming, or the number nonconforming. Give your plot a **Title** and click **OK**.

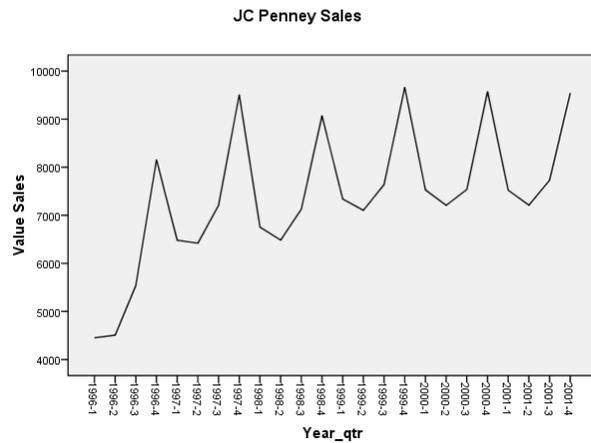


We find  $\bar{p}$  is 0.3555. Since SPSS calculates these proportions based on the total number of students each month, the upper and lower control limits change somewhat. To find  $\bar{n}$ , use **Analyze, Descriptive Statistics, Descriptives**. The average total enrolled students for this school for this year is 921.8.

**Descriptive Statistics**

	N	Minimum	Maximum	Mean	Std. Deviation
Students	10	883.00	947.00	921.8000	19.62312
Valid N (listwise)	10				

# CHAPTER 18



## Time Series Forecasting

18.1	Trends and Seasons
18.2	Time Series Models

### Introduction

A time series is a sequence of observations on a single variable at equally spaced intervals.

In this chapter, we examine some basic ideas in time series forecasting and modeling. The most basic idea is to fit a model to describe a trend (if any), then to add additional terms to describe seasons (or cycles within years). We also examine models where a new observation is related to a prior one (autoregressive models) as well as moving average models.

## 18.1 Trends and Seasons

The most basic component of a time series is assessing whether or not there is a trend (systematic long-term rise or fall) and whether there is some aspect that repeats regularly (a cycle, or seasonal component.)

**Example 18.1 JC Penney Sales.** The table below contains retail sales for JC Penney in millions of dollars beginning with the first quarter of 1996 and ending with the fourth quarter of 2001.

Year-Quarter	Sales	Year-Quarter	Sales
1996-1 <sup>st</sup>	4452	1999-1 <sup>st</sup>	7339
1996-2 <sup>nd</sup>	4507	1999-2 <sup>nd</sup>	7104
1996-3 <sup>rd</sup>	5537	1999-3 <sup>rd</sup>	7639
1996-4 <sup>th</sup>	8157	1999-4 <sup>th</sup>	9661
1997-1 <sup>st</sup>	6481	2000-1 <sup>st</sup>	7528
1997-2 <sup>nd</sup>	6420	2000-2 <sup>nd</sup>	7207
1997-3 <sup>rd</sup>	7208	2000-3 <sup>rd</sup>	7538
1997-4 <sup>th</sup>	9509	2000-4 <sup>th</sup>	9573
1998-1 <sup>st</sup>	6755	2001-1 <sup>st</sup>	7522
1998-2 <sup>nd</sup>	6483	2001-2 <sup>nd</sup>	7211
1998-3 <sup>rd</sup>	7129	2001-3 <sup>rd</sup>	7729
1998-4 <sup>th</sup>	9072	2001-4 <sup>th</sup>	9542

- Make a time plot of the data. Be sure to connect the points in your plot to highlight patterns.
- Is there an obvious trend in JC Penney quarterly sales? If so, is the trend positive or negative?
- Is there an obvious repeating pattern in this data? If so, clearly describe the repeating pattern.

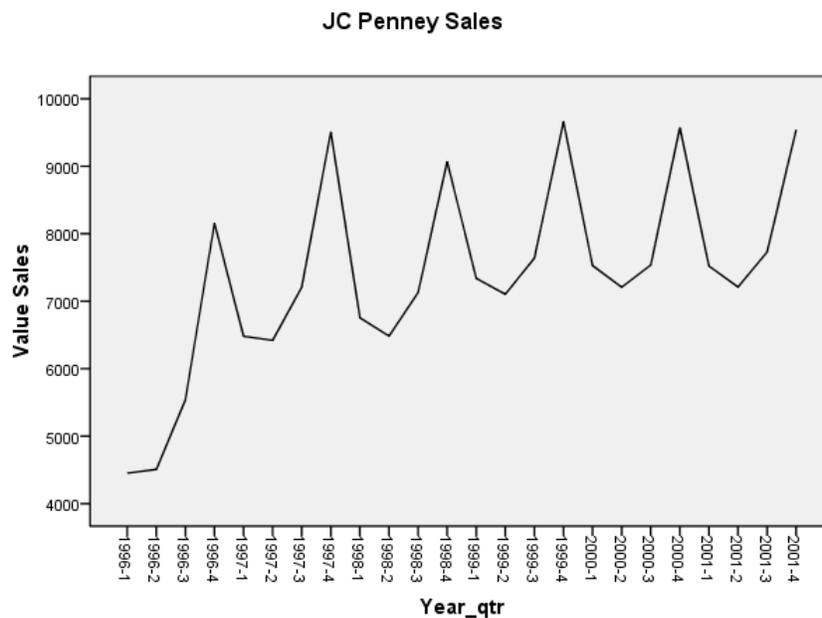
*Solution.* First, enter the data into two variables: one for the year and quarter (this was defined as a string variable to be able to enter the quarter values) and another for the sales amounts. The first few rows are shown below.

	Year_qtr	Sales
1	1996-1	4452
2	1996-2	4507
3	1996-3	5537
4	1996-4	8157
5	1997-1	6481

Click **Graphs, Legacy Dialogs, Line**. We want a **Simple Chart** where data in chart are **Values of individual cases**.



The Line represents **Sales** and **Year\_qtr** will be the  $x$  axis variable. Give your plot a **Title** and click **OK**.



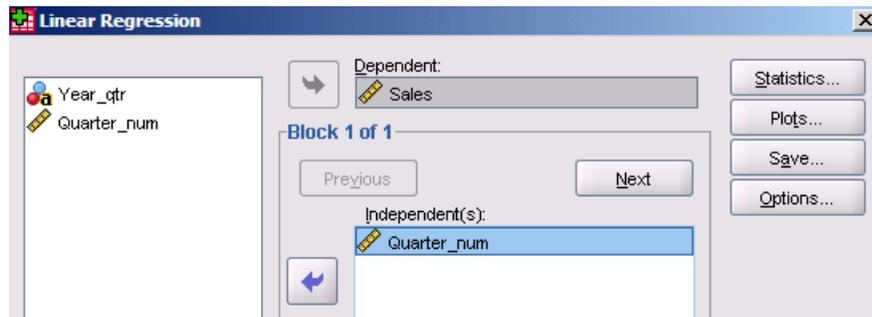
(b) There does appear to be some trend in this plot; however, if we eliminate the first couple of years, it is not as obvious.

(c) There are definitely repeating patterns in this set of data. The fourth quarter (including Christmas) always has the highest sales; the second quarter always has the lowest sales. Third quarter is generally somewhat higher than the first quarter (back-to-school?).

**Example 18.2 JC Penney Trends.** Find any trend in the JC Penney data using  $x = 1$  to correspond with the first quarter of 1996,  $x = 2$  to the second quarter of 1996, etc. Interpret the slope and intercept in the model.

*Solution.* Enter a variable (called **Quarter\_num**) to hold these numeric values. This variable should range from 1 to 24 (for fourth quarter 2001). Compute the regression line

using **Analyze, Regression, Linear** where the Dependent variable is **Sales** and the Independent variable is **Quarter\_num**.

Coefficients<sup>a</sup>

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	5903.217	492.856		11.978	.000
	Quarter_num	118.753	34.493	.592	3.443	.002

a. Dependent Variable: Sales

We find the equation  $Sales = 5903.217 + 118.753 * Quarter\_num$ . With a  $t$  statistic of 3.443 and  $p$ -value of 0.002, this trend is significantly nonzero. The slope indicates that JC Penney sales increase (on average) 118.753 million dollars per quarter. The intercept says estimated sales for quarter 0 (that would be fourth quarter 1995) are 5903.217 million dollars.

**Example 18.3 JC Penney Cycles.** Since sales seem to have an annual cycle, add indicator variables for the quarters to the trend-only model fitted in Example 18.2. Compare the estimated intercept of this model with the intercept found in Example 13.2. Given the patterns of seasonal variation, which appears to be the better estimate?

*Solution.* We create three indicator variables (a fourth would make the model unsolvable) of the form  $x_1 = 1$  if 1<sup>st</sup> quarter, 0 otherwise. Below, we show the first few rows of the new worksheet.

	Year_qtr	Sales	Quarter_num	x1	x2	x3
1	1996-1	4452	1.00	1.00	0.00	0.00
2	1996-2	4507	2.00	0.00	1.00	0.00
3	1996-3	5537	3.00	0.00	0.00	1.00
4	1996-4	8157	4.00	0.00	0.00	0.00
5	1997-1	6481	5.00	1.00	0.00	0.00
6	1997-2	6420	6.00	0.00	1.00	0.00

Return to **Analyze, Regression, Linear** and add the three indicator variables to the model.

**Coefficients<sup>a</sup>**

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	7858.758	331.260		23.724	.000
	Quarter_num	99.541	16.934	.496	5.878	.000
	x1	-2274.210	331.116	-.709	-6.868	.000
	x2	-2564.585	328.944	-.799	-7.796	.000
	x3	-2022.792	327.634	-.630	-6.174	.000

a. Dependent Variable: Sales

Our model is now

$Sales = 7858.758 + 99.541Quarter\_num - 2274.31x1 - 2564.585x2 - 2022.792x3$ . Since the indicator variables only come into play for the selected quarters (as a 1 or 0, which means these values affect the intercept), this really means we have the following set of equations:

$$4thQtrSales = 7858.738 + 99.541Quarter\_num$$

$$1stQtrSales = 5584.448 + 99.541Quarter\_num$$

$$2ndQtrSales = 5294.173 + 99.541Quarter\_num$$

$$3rdQtrSales = 5835.966 + 99.541Quarter\_num$$

Our trend is now that sales increase (on average) 99.541 million dollars per quarter. Since the intercept is still month 0 (4<sup>th</sup> quarter 1995), this model forecasts 7858.738 million for that quarter. This is much more reasonable given that 4<sup>th</sup> quarter sales are always the highest for the year.

## 18.2 Time Series Models

Time series models use past values to predict future values of the series. There are several ways to model these—we have already looked at regression models. There are several other methods whose study can consume entire statistics courses on their own. We will look at two—the autoregressive model and the moving average model.

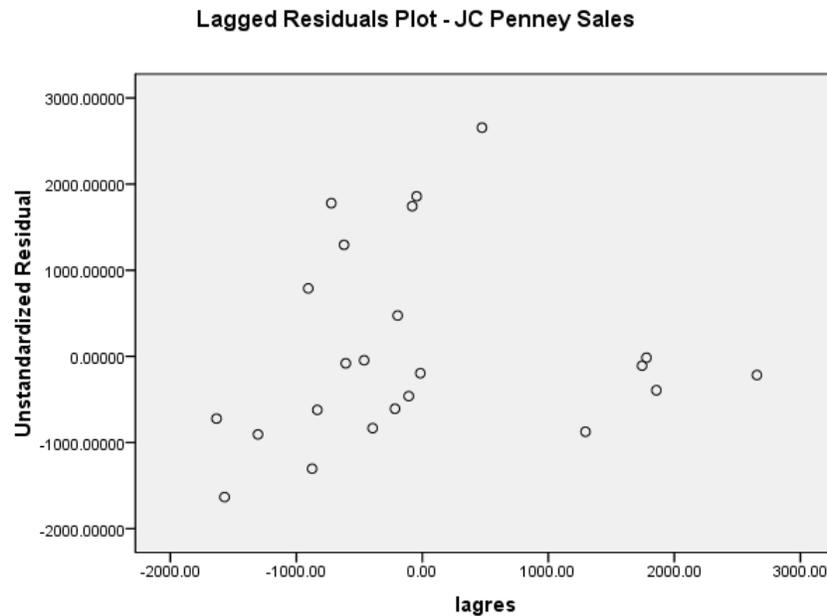
First-order autoregressive models use the equation  $y_t = \beta_0 + \beta_1 y_{t-1} + \varepsilon_t$  as a model. This implies that the value one time period prior is the most useful in forecasting the next value. One way to look for an autoregressive relationship is to plot lagged residuals; that is, plot  $(\varepsilon_1, \varepsilon_2), (\varepsilon_2, \varepsilon_3), \dots, (\varepsilon_{n-1}, \varepsilon_n)$ .

**Example 18.4 JC Penney Trend-Only Residuals.** Return to the linear trend-only model of Example 18.2 and create a lagged residual plot to examine the data for autocorrelation. Calculate the correlation in the lagged residuals.

*Solution.* Recompute the regression; however, in the main dialog box click **Save** and click to check **Unstandardized Residuals**. Click **Transform, Compute Variable**. We want to lag the residuals by 1 time unit. In the Miscellaneous Function Group, there is a Lag command. Our lagged residuals are then formed with the command below.

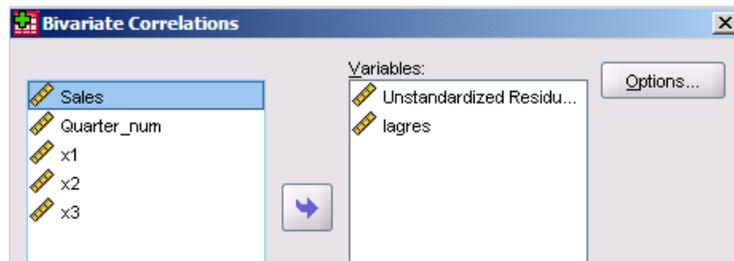


To plot the residuals, click **Graphs, Legacy Dialogs, Scatter/Dot**. Select **Simple Scatter**. The original residuals go on the y axis and the lagged residuals on the x axis. Give the plot a title, then click **OK**.



There appear to be two sets of points here: one with a significant amount of correlation in the left part of this plot, and an interesting group with large positive residuals. These are due to the high 4<sup>th</sup> quarter sales in relationship to sale in the next quarter.

To compute the correlation, use **Analyze, Correlate, Bivariate**. Add the original and lagged residuals in the Variables box and click **OK**.

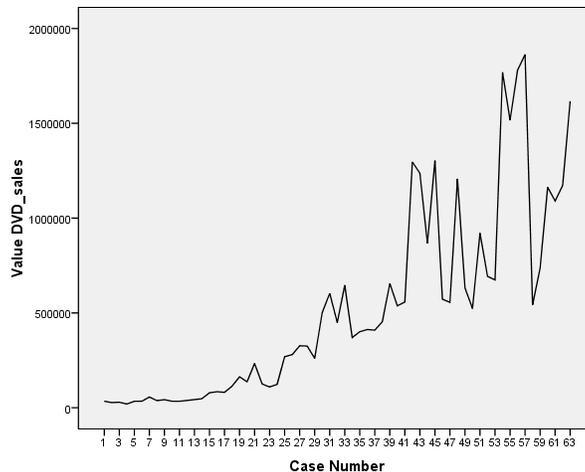


**Correlations**

		Unstandardized Residual	lagres
Unstandardized Residual	Pearson Correlation	1.000	.095
	Sig. (2-tailed)		.667
	N	24.000	23
lagres	Pearson Correlation	.095	1.000
	Sig. (2-tailed)	.667	
	N	23	23.000

The correlation in these residuals is not significant because the  $p$ -value given is 0.667 which is large. This result is most likely due to the clump of large positive residuals; remember, even a single point can make a correlation seem non-significant.

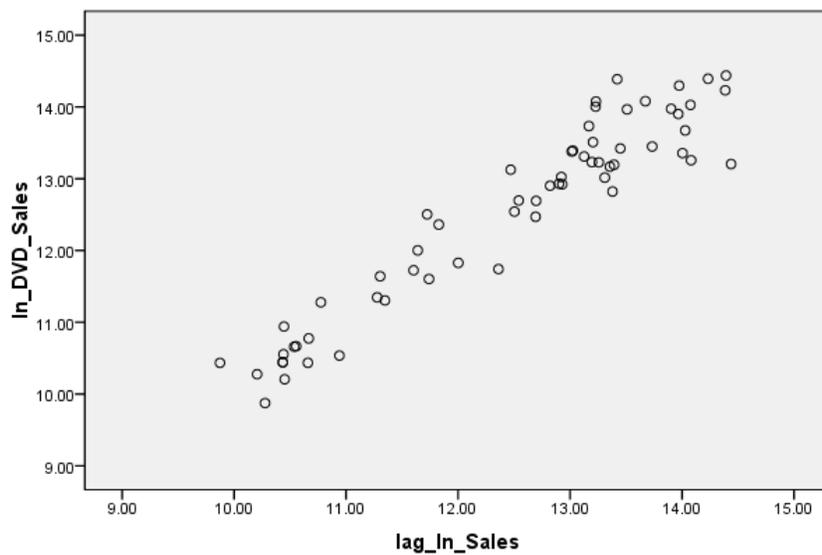
**Example 18.5 Autoregressive DVD Sales.** The popularity of the DVD format has exploded in the relatively short period of time since its introduction in March 1997. The Consumer Electronics Association tracks monthly sales of DVD players. In the table on the next page are data on DVD sales from April 1997 through June 2002. A time plot of the data is below.



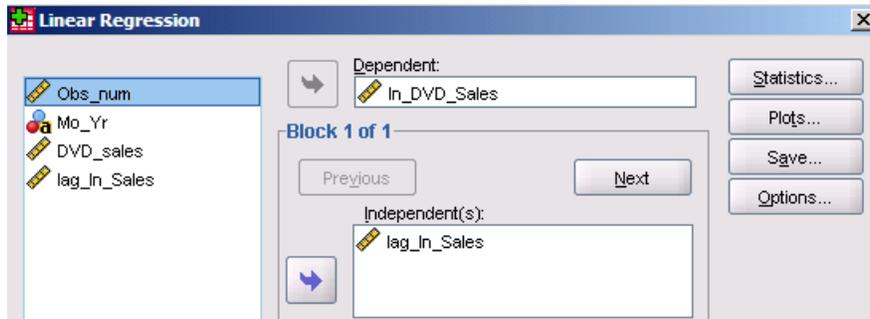
Date	Units Sold	Date	Units Sold	Date	Units Sold
4-97	34601	1-99	125536	10-00	1236658
5-97	27051	2-99	109399	11-00	866507
6-97	29037	3-99	123466	12-00	1303091
7-97	19416	4-99	269107	1-01	572031
8-97	34021	5-99	279756	2-01	555856
9-97	34371	6-99	326668	3-01	1207489
10-97	56407	7-99	325151	4-01	631353
11-97	37657	8-99	260225	5-01	523225
12-97	42575	9-99	501501	6-01	920839
1-98	34027	10-99	603048	7-01	693013
2-98	34236	11-99	449242	8-01	673926
3-98	38336	12-99	646290	9-01	1768821
4-98	42889	1-00	370031	10-01	1516211
5-98	47805	2-00	401035	11-01	1781048
6-98	79044	3-00	412559	12-01	1862772
7-98	84709	4-00	409192	1-02	542698
8-98	81170	5-00	453435	2-02	736118
9-98	113558	6-00	654687	3-02	1162568
10-98	163074	7-00	537453	4-02	1090767
11-98	136908	8-00	557617	5-02	1171984
12-98	233505	9-00	1296280	6-02	1617098

There is very little increase in sales at the beginning; then sales skyrocket and also exhibit considerable variation. A scatterplot of lagged  $\ln(\text{DVD sales})$  versus  $\ln(\text{DVD Sales})$  is very linear, as shown below. (Both these variables were computed using **Transform, Compute Variable.**)

Lag Plot of DVD Sales



Since this plot is so linear, an autoregressive model appears reasonable. We will fit a linear model to the log sales data using the lagged data as the predictor.



**Coefficients<sup>a</sup>**

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	.690	.484		1.427	.159
	lag_In_Sales	.950	.039	.954	24.625	.000

a. Dependent Variable: ln\_DVD\_Sales

We notice the  $t$  statistic for this regression is 24.625 with a  $p$ -value of 0.000. This gives an equation of  $\ln(\text{DVDsales})_t = .690 + .950\ln(\text{DVDsales})_{t-1}$ .

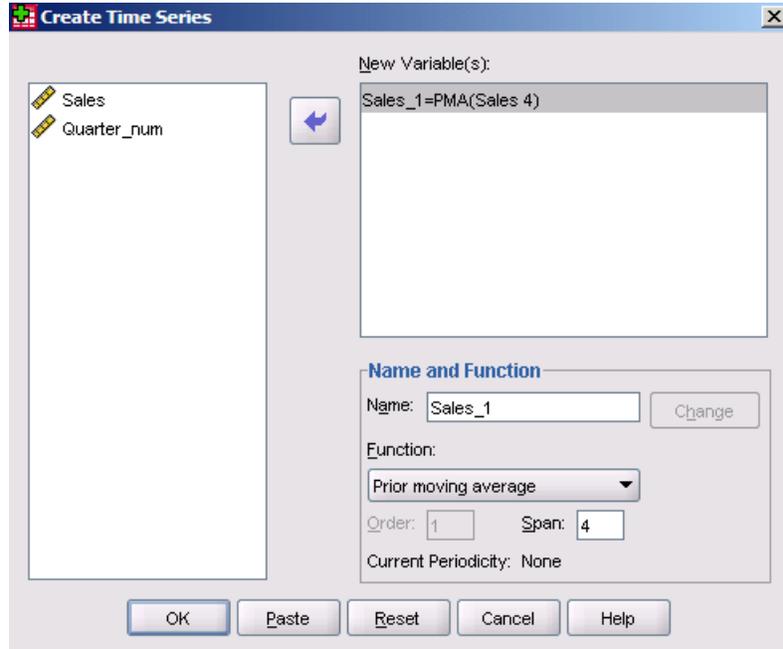
Suppose we want to predict sales for July 2002 based on this model. June 2002 sales were 1617098 units. Putting this into the equation, we have

$$\begin{aligned}\ln(\text{July}) &= .690 + .950\ln(1617098) \\ &= .690 + .950 * 14.2961 \\ &= 14.271 \\ \text{July} &= e^{14.271} = 1577410\end{aligned}$$

Moving Average models forecast  $y_t$  as the average of the preceding  $k$  observations, in other words,  $y_t = \frac{y_{t-1} + y_{t-2} + \dots + y_{t-k}}{k}$ . These models smooth out some of the irregular noise in a typical time series.

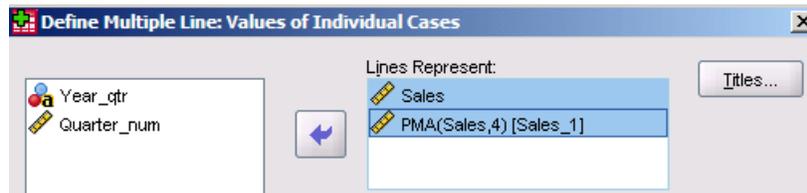
**Example 18.6 JC Penney Moving Averages.** JC Penney sales have an annual periodicity. In every year, the fourth quarter sales are highest, and second quarter sales are lowest. Find the moving average predictor using a span of 4. Graph the moving average model on the original time series.

*Solution.* With the data entered (or a saved file opened), click **Transform, Create Time Series**. Click to place the variable **Sales** into the New Variable box. Click to expand the **Function** box and select **Prior Moving Average**. Change the entry in the **Span** box to 4, then click **Change**. Your dialog box should look like the one below. Click **OK** to create the new time series.

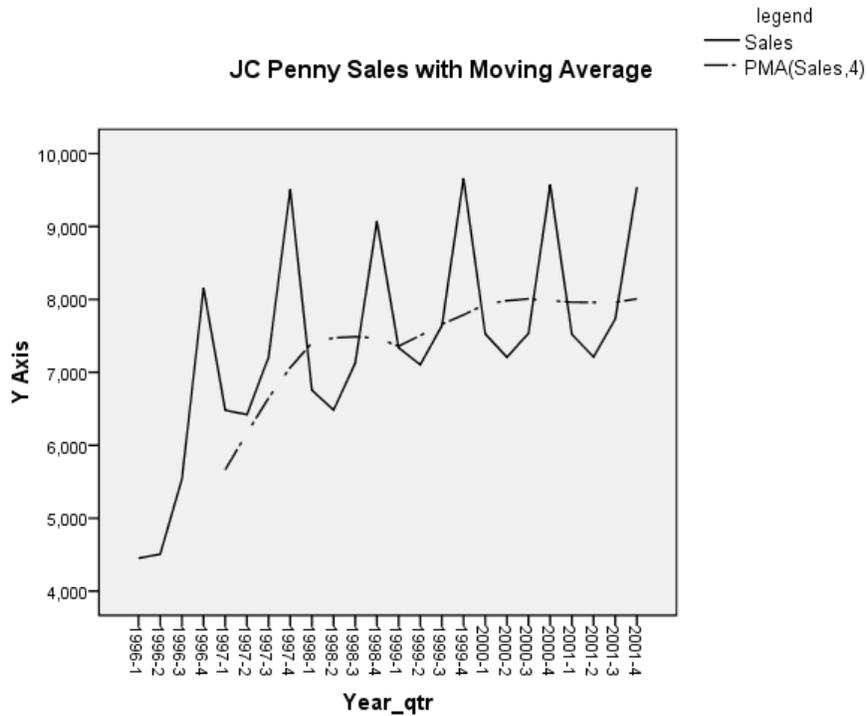


	Year_qtr	Sales	Quarter_num	Sales_1
1	1996-1	4452	1.00	.
2	1996-2	4507	2.00	.
3	1996-3	5537	3.00	.
4	1996-4	8157	4.00	.
5	1997-1	6481	5.00	5663.2
6	1997-2	6420	6.00	6170.5
7	1997-3	7208	7.00	6648.8

Notice that since our new series is the average of the four prior values, there are no values for the first four rows of the worksheet. The average of those values became the entry in the fifth row. To graph both series on the same plot, click **Graphs, Legacy Dialogs, Line, Multiple**. Data in the chart are **Values of individual cases**. The lines represent **Sales** (the original series) and **PMA(Sales,4)**, the moving average series.



You can use either the case number or **Year\_qtr** as the “Category Label.” Give the plot a title, and click **OK** to generate the graph. As always with this type of plot, you may want to right-click the graph and **Edit Contents in Separate Window** to change the default of different colors for each series to different dash types if you have only a black-and-white printer.



Notice that this graph shows the moving average “flattening out” through 2000 and 2001. The country was in a mild recession then, so one would not have expected increasing sales.

## Chapter 1 Problem Statements

**1.3** The rating service Arbitron places U.S. radio stations into more than 50 categories that describe the kind of programs they broadcast. Which formats attract the largest audiences? Here are Arbitron's measurements of the share of the listening audience (aged 12 and over) for the most popular formats

Format	Audience share
Country	12.6%
News/Talk/Information	10.4%
Adult Contemporary	7.1%
Pop Contemporary Hit	5.5%
Classic Rock	4.7%
Rhythmic Contemporary Hit	4.2%
Urban Contemporary	4.1%
Urban Contemporary	3.4%
Oldies	3.3%
Hot Adult Contemporary	3.2%
Mexican Regional	3.1%

- What is the sum of the audience shares for these formats? What percent of the radio audience listens to stations with other formats?
- Make a bar graph to display these data. Be sure to include an "Other format" category.
- Would it be correct to display these data in a pie chart? Why?

**1.5** Births are not, as you might think, evenly distributed across the days of the week. Here are the average numbers of babies born on each day of the week in 2005:

Day	Births
Sunday	7,374
Monday	11,704
Tuesday	13,169
Wednesday	13,038
Thursday	13,013
Friday	12,664
Saturday	8,459

Present these data in a well-labeled bar graph. Would it also be correct to make a pie chart? Suggest some possible reasons why there are fewer births on weekends.

**1.11** Table 1.3 shows the annual spending per person on health care in the world's richer countries. Make a stemplot of the data after rounding to the nearest \$100 (so that stems are thousands of dollars and leaves are hundreds of dollars). Split the stems, placing

leaves 0 to 4 on the first stem and leaves 5 to 9 on the second stem of the same value. Describe the shape, center, and spread of the distribution. Which country is the high outlier?

Country	Dollars	Country	Dollars	Country	Dollars
Argentina	1067	Hungary	1269	Poland	745
Australia	2874	Iceland	3110	Portugal	1791
Austria	2306	Ireland	2496	Saudi Arabia	578
Belgium	2828	Israel	1911	Singapore	1156
Canada	2989	Italy	2266	Slovakia	777
Croatia	838	Japan	2244	Slovenia	1669
Czech Republic	1302	Korea	1074	South Africa	669
Denmark	2762	Kuwait	567	Spain	1853
Estonia	682	Lithuania	754	Sweden	2704
Finland	2108	Netherlands	2987	Switzerland	3776
France	2902	New Zealand	1893	United Kingdom	2389
Germany	3001	Norway	3809	United States	5711
Greece	1997	Oman	419		

**1.25** The most popular colors for cars and light trucks change over time. Silver passed green in 2000 to become the most popular color worldwide, then gave way to shades of white in 2007. Here is the distribution of colors for vehicles sold in North America in 2007:

Color	Popularity
White	19%
Silver	18%
Black	16%
Red	13%
Gray	12%
Blue	12%
Beige, brown	5%
Other	

Fill in the percent of vehicles that are in other colors. Make a graph to display the distribution of color popularity.

**1.27** Among persons aged 15 to 24 years in the United States, the leading causes of death and the number of deaths in 2005 were: accidents, 15,567; homicide, 5359; suicide, 4139; cancer, 1717; heart disease, 1067; congenital defects, 483.

- (a) Make a bar graph to display these data.  
 (b) To make a pie chart, you need one additional piece of information. What is it?

**1.29** Email spam is the curse of the Internet. Here is a compilation of the most common types of spam:

Type of Spam	Percent
Adult	19
Financial	20
Health	7
Internet	7
Leisure	6
Products	25
Scams	9

Make two bar graphs of these percents, one with bars ordered as in the table (alphabetically) and the other with bars in order from tallest to shortest. Comparisons are easier if you order the bars by height.

**1.35** Table 1.5 gives the number of active medical doctors per 100,000 people in each state.

Medical doctors per 100,000 people, by state					
State	Doctors	State	Doctors	State	Doctors
Alabama	213	Louisiana	264	Ohio	261
Alaska	222	Maine	267	Oklahoma	171
Arizona	208	Maryland	411	Oregon	263
Arkansas	203	Massachusetts	450	Pennsylvania	294
California	259	Michigan	240	Rhode Island	351
Colorado	258	Minnesota	281	South Carolina	230
Connecticut	363	Mississippi	181	South Dakota	219
Delaware	248	Missouri	239	Tennessee	261
Florida	245	Montana	221	Texas	212
Georgia	220	Nebraska	239	Utah	209
Hawaii	310	Nevada	186	Vermont	362
Idaho	169	New Hampshire	260	Virginia	270
Illinois	272	New Jersey	306	Washington	265
Indiana	213	New Mexico	240	West Virginia	229
Iowa	187	New York	389	Wisconsin	254
Kansas	220	North Carolina	253	Wyoming	188
Kentucky	230	North Dakota	242	Dist. of Columbia	798

- (a) Why is the number of doctors per 100,000 people a better measure of the availability of health care than a simple count of the number of doctors in a state?
- (b) Make a histogram that displays the distribution of doctors per 100,000 people. Write a brief description of the distribution. Are there any outliers? If so, can you explain them?

**1.37** “Recruitment,” the addition of new members to a fish population, is an important measure of the health of ocean ecosystems. The table gives data on the recruitment of rock sole in the Bering Sea from 1973 to 2000. Make a stemplot to display the distribution of yearly rock sole

recruitment. (Round to the nearest hundred and split the stems.) Describe the shape, center, and spread of the distribution and any striking deviations that you see.

Year	Recruitment (millions)						
1973	173	1980	1411	1987	4700	1994	505
1974	234	1981	1431	1988	1702	1995	304
1975	616	1982	1250	1989	1119	1996	425
1976	344	1983	2246	1990	2407	1997	214
1977	515	1984	1793	1991	1049	1998	385
1978	576	1985	1793	1992	505	1999	445
1979	727	1986	2809	1993	998	2000	767

**1.39** Make a time plot of the rock sole recruitment data in Exercise 1.37. What does the time plot show that your stemplot in Exercise 1.37 did not show? When you have time series data, a time plot is often needed to understand what is happening.

**1.43** The impression that a time plot gives depends on the scales you use on the two axes. If you stretch the vertical axis and compress the time axis, change appears to be more rapid. Compressing the vertical axis and stretching the time axis make change appear slower. Make two more time plots of the college tuition data in Exercise 1.12 (page 24), one that makes tuition appear to increase very rapidly and one that shows only a gentle increase. The moral of this exercise is: pay close attention to the scales when you look at a time plot.

**1.45** Here are data on the number of people bitten by alligators in Florida over a 36-year period:

<b>Year</b>	<b>Number</b>	<b>Year</b>	<b>Number</b>	<b>Year</b>	<b>Number</b>	<b>Year</b>	<b>Number</b>
1972	4	1981	10	1990	17	1999	16
1973	3	1982	7	1991	20	2000	23
1974	4	1983	9	1992	15	2001	25
1975	5	1984	9	1993	19	2002	17
1976	2	1985	7	1994	20	2003	12
1977	14	1986	23	1995	22	2004	13
1978	7	1987	13	1996	13	2005	15
1979	2	1988	18	1997	8	2006	18
1980	5	1989	13	1998	9	2007	18

- (a) Make a histogram of the counts of people bitten by alligators. The distribution has an irregular shape. What is the midpoint of the yearly counts of people bitten?
- (b) Make a time plot. There is great variation from year to year, but also an increasing trend. How many of the 22 years from 1986 to 2007 had more people bitten by alligators than your midpoint from (a)? The trend reflects Florida's growing population, which brings more people close to alligators.

## Chapter 2 Problem Statements

**2.1** Example 1.9 (page 20) gives the breaking strength in pounds of 20 pieces of Douglas fir. Find the mean breaking strength. How many of the pieces of wood have strengths less than the mean? What feature of the stemplot (Figure 1.11, page 21) explains the fact that the mean is smaller than most of the observations?

**2.3** Find the mean of the travel times to work for the 20 New York workers in Example 2.3. Compare the mean and median for these data. What general fact does your comparison illustrate?

**2.5** Table 1.4 (page 33) gives the ratio of two essential fatty acids in 30 food oils. Find the mean and the median for these data. Make a histogram of the data. What feature of the distribution explains why the mean is more than 10 times as large as the median?

**2.11** The mean  $\bar{x}$  and standard deviation  $s$  measure center and spread but are not a complete description of a distribution. Data sets with different shapes can have the same mean and standard deviation. To demonstrate this fact, use your calculator to find  $\bar{x}$  and  $s$  for these two small data sets. Then make a stemplot of each and comment on the shape of each distribution.

<b>Data A</b>	9.14	8.14	8.74	8.77	9.26	8.10	6.13	3.10	9.13	7.26	4.74
<b>Data B</b>	6.58	5.76	7.71	8.84	8.47	7.04	5.25	5.56	7.91	6.89	12.50

**2.13** “Conservationists have despaired over destruction of tropical rain forest by logging, clearing, and burning.” These words begin a report on a statistical study of the effects of logging in Borneo. Charles Cannon of Duke University and his coworkers compared forest plots that had never been logged (Group 1) with similar plots nearby that had been logged 1 year earlier (Group 2) and 8 years earlier (Group 3). All plots were 0.1 hectare in area. Here are the counts of trees for plots in each group:

<b>Group 1</b>	27	22	29	21	19	33	16	20	24	27	28	19
<b>Group 2</b>	12	12	15	9	20	18	17	14	14	1	27	19
<b>Group 3</b>	18	4	22	15	18	19	22	12	12			

To what extent has logging affected the count of trees? Follow the four-step process in reporting your work.

**2.29** An alternative presentation of the flower length data in Table 2.1 reports the five-number summary and uses boxplots to display the distributions. Do this. Do the boxplots

fail to reveal any important information visible in the stemplots in Figure 2.5?

**2.31** Here is the distribution of the weight at birth for all babies born in the United States in 2005:

Weight (grams)	Count	Weight (grams)	Count
Less than 500	6,599	3,000 to 3,499	1,596,944
500 to 999	23,864	3,500 to 3,999	1,114,887
1,000 to 1,499	31,325	4,000 to 4,499	289,098
1,500 to 1,999	66,453	4,500 to 4,999	42,119
2,000 to 2,499	210,324	5,000 to 5,499	4,715
2,500 to 2,999	748,042		

- For comparison with other years and with other countries, we prefer a histogram of the *percents* in each weight class rather than the counts. Explain why.
- How many babies were there? Make a histogram of the distribution, using percents on the vertical scale.
- What are the positions of the median and quartiles in the ordered list of all birth weights? In which weight classes do the median and quartiles fall?

**2.35** Here are the survival times in days of 72 guinea pigs after they were injected with infectious bacteria in a medical experiment. Survival times, whether of machines under stress or cancer patients after treatment, usually have distributions that are skewed to the right.

43	45	53	56	56	57	58	66	67	73	74	79
80	80	81	81	81	82	83	83	84	88	89	91
91	92	92	97	99	99	100	100	101	102	102	102
103	104	107	108	109	113	114	118	1121	123	126	128
137	138	139	144	145	147	156	162	174	178	179	184
191	198	211	214	243	249	329	380	403	511	522	598

- Graph the distribution and describe its main features. Does it show the expected right skew?
- Which numerical summary would you choose for these data? Calculate your chosen summary. How does it reflect the skewness of the distribution?

**2.37** Table 1.1 (page 12) gives the percent of foreign-born residents in each of the states. For the nation as a whole, 12.5% of residents are foreign-born. Find the mean of the 51 entries in Table 1.1. It is *not* 12.5%. Explain carefully why this happens. (*Hint*: The states with the largest populations are California, Texas, New York, and Florida. Look at their entries in Table 1.1.)

**2.43** In 2007, the Boston Red Sox won the World Series for the second time in 4 years. Table 2.2 gives the salaries of the 25 players on the Red Sox World Series roster. Provide the team owner with a full description of the distribution of salaries and a brief summary of its most important features.

Salaries for the 2007 Boston Red Sox World Series team					
Player	Salary	Player	Salary	Player	Salary
Josh Beckett	\$6,666,667	Jon Lester	\$384,000	Jonathan Papelbon	\$425,000
Alex Cora	\$2,000,000	Javier López	\$402,000	Dustin Pedroia	\$380,000
Coco Crisp	\$3,833,333	Mike Lowell	\$9,000,000	Manny Ramirez	\$17,016,381
Manny Delcarmen	\$380,000	Julio Lugo	\$8,250,000	Curt Schilling	\$13,000,000
J. D. Drew	\$14,400,000	D Matsuzaka	\$6,333,333	Kyle Snyder	\$535,000
Jacoby Ellsbury	\$380,000	Doug Mirabelli	\$750,000	Mike Timlin	\$2,800,000
Eric Gagné	\$6,000,000	Hideki Okajimi	\$1,225,000	Jason Varitek	\$11,000,000
Eric Hinske	\$5,725,000	David Ortiz	\$13,250,000	Kevin Youkilis	\$424,000
Bobby Kielty	\$2,100,000				

**2.45** Businesses know that customers often respond to background music. Do they also respond to odors? Nicolas Guéguen and his colleagues studied this question in a small pizza restaurant in France on Saturday evenings in May. On one evening, a relaxing lavender odor was spread through the restaurant; on another evening, a stimulating lemon odor; a third evening served as a control, with no odor. Table 2.3 shows the amounts (in euros) that customers spent on each of these evenings. Compare the three distributions. What effect did the two odors have on customer spending?

Amount spent (euros) by customers in a restaurant when exposed to odors									
No odor									
15.9	18.5	15.9	18.5	18.5	21.9	15.9	15.9	15.9	15.9
15.9	18.5	18.5	18.5	20.5	18.5	18.5	15.9	15.9	15.9
18.5	18.5	15.9	18.5	15.9	18.5	15.9	25.5	12.9	15.9
Lemon Odor									
18.5	15.9	18.5	18.5	18.5	15.9	18.5	15.9	18.5	18.5
15.9	18.5	21.5	15.9	21.9	15.9	18.5	18.5	18.5	18.5
25.9	15.9	15.9	15.9	18.5	18.5	18.5	18.5		
Lavender Odor									
21.9	18.5	22.3	21.9	18.5	24.9	18.5	22.5	21.5	21.9
21.5	18.5	25.5	18.5	18.5	21.9	18.5	18.5	24.9	21.9
25.9	21.9	18.5	18.5	22.8	18.5	21.9	20.7	21.9	22.5

**2.47** Farmers know that driving heavy equipment on wet soil compresses the soil and hinders the growth of crops. Table 2.5 gives data on the “penetrability” of the same soil at

three levels of compression. Penetrability is a measure of the resistance plant roots meet when they grow through the soil. Low penetrability means high resistance. How does increasing compression affect penetrability?

<b>Compressed</b>		<b>Intermediate</b>		<b>Loose</b>	
2.86	3.08	3.14	3.54	3.99	4.11
2.68	2.82	3.38	3.36	4.20	4.30
2.92	2.78	3.10	3.18	3.94	3.96
2.82	2.98	3.40	3.12	4.16	4.03
2.76	3.00	3.38	3.86	4.29	4.89
2.81	2.78	3.14	2.92	4.19	4.12
2.78	2.96	3.18	3.46	4.13	4.00
3.08	2.90	3.26	3.44	4.41	4.34
2.94	3.18	2.96	3.62	3.98	4.27
2.86	3.16	3.02	4.26	4.41	4.91

**2.51** Which members of the Boston Red Sox (Table 2.2) have salaries that are suspected outliers by the  $1.5 \times \text{IQR}$  rule?

**Chapter 3 Problem Statements**

**3.9** The heights of women aged 20 to 29 are approximately Normal with mean 64 inches and standard deviation 2.7 inches. Men the same age have mean height 69.3 inches with standard deviation 2.8 inches. What are the  $z$ -scores for a woman 6 feet tall and a man 6 feet tall? Say in simple language what information the  $z$ -scores give that the actual heights do not.

**3.11** The summer monsoon rains in India follow approximately a Normal distribution with mean 852 millimeters (mm) of rainfall and standard deviation 82 mm.

- (a) In the drought year 1987, 697 mm of rain fell. In what percent of all years will India have 697 mm or less of monsoon rain?
- (b) “Normal rainfall” means within 20% of the long-term average, or between 683 mm and 1022 mm. In what percent of all years is the rainfall normal?

**3.13** Use Table A to find the value  $z$  of a standard Normal variable that satisfies each of the following conditions. (Use the value of  $z$  from Table A that comes closest to satisfying the condition.) In each case, sketch a standard Normal curve with your value of  $z$  marked on the axis.

- (a) The point  $z$  with 20% of the observations falling below it.
- (b) The point  $z$  with 40% of the observations falling above it.

**3.29**

- (a) Find the number  $z$  such that the proportion of observations that are less than  $z$  in a standard Normal distribution is 0.8.
- (b) Find the number  $z$  such that 35% of all observations from a standard Normal distribution are greater than  $z$ .

**3.31** Emissions of sulfur dioxide by industry set off chemical changes in the atmosphere that result in “acid rain.” The acidity of liquids is measured by pH on a scale of 0 to 14. Distilled water has pH 7.0, and lower pH values indicate acidity. Normal rain is somewhat acidic, so acid rain is sometimes defined as rainfall with a pH below 5.0. The pH of rain at one location varies among rainy days according to a Normal distribution with mean 5.4 and standard deviation 0.54. What proportion of rainy days have rainfall with pH below 5.0?

**3.33** Automated manufacturing operations are quite precise but still vary, often with distributions that are close to Normal. The width in inches of slots cut by a milling

machine follows approximately the  $N(0.8750, 0.0012)$  distribution. The specifications allow slot widths between 0.8720 and 0.8780 inch. What proportion of slots meet these specifications?

**3.35** The 2008 Chevrolet Malibu with a four-cylinder engine has combined gas mileage 25 mpg. What percent of all vehicles have worse gas mileage than the Malibu?

**3.37** The quartiles of any distribution are the values with cumulative proportions 0.25 and 0.75. They span the middle half of the distribution. What are the quartiles of the distribution of gas mileage?

**3.39** Reports on a student's ACT or SAT usually give the percentile as well as the actual score. The percentile is just the cumulative proportion stated as a percent: the percent of all scores that were lower than this one. In 2007, composite ACT scores were close to Normal with mean 21.2 and standard deviation 5.0. Jacob scored 16. What was his percentile?

**3.41** The heights of women aged 20 to 29 follow approximately the  $N(64, 2.7)$  distribution. Men the same age have heights distributed as  $N(69.3, 2.8)$ . What percent of young women are taller than the mean height of young men?

**3.43** Changing the mean and standard deviation of a Normal distribution by a moderate amount can greatly change the percent of observations in the tails. Suppose that a college is looking for applicants with SAT math scores 750 and above.

- (a) In 2007, the scores of men on the math SAT followed the  $N(533, 116)$  distribution. What percent of men scored 750 or better?
- (b) Women's SAT math scores that year had the  $N(499, 110)$  distribution. What percent of women scored 750 or better? You see that the percent of men above 750 is almost three times the percent of women with such high scores. Why this is true is controversial. (On the other hand, women score higher than men on the new SAT writing test, though by a smaller amount.)

**3.47** Scores on the ACT test for the 2007 high school graduating class had mean 21.2 and standard deviation 5.0. In all, 1,300,599 students in this class took the test. Of these, 149,164 had scores higher than 27 and another 50,310 had scores exactly 27. ACT scores are always whole numbers. The exactly Normal  $N(21.2, 5.0)$  distribution can include any value, not just whole numbers. What is more, there is *no* area exactly above 27 under the smooth Normal curve. So ACT scores can be only approximately Normal. To illustrate

this fact, find

- (a) the percent of 2007 ACT scores greater than 27.
- (b) the percent of 2007 ACT scores greater than or equal to 27.
- (c) the percent of observations from the  $N(21.2, 5.0)$  distribution that are greater than 27. (The percent greater than or equal to 27 is the same, because there is no area exactly over 27.)

**3.49** Here are the lengths in millimeters of the thorax for 49 male fruit flies:

0.64	0.64	0.64	0.68	0.68	0.68	0.72	0.72	0.72	0.72
0.74	0.76	0.76	0.76	0.76	0.76	0.76	0.76	0.76	0.78
0.80	0.80	0.80	0.80	0.80	0.82	0.82	0.84	0.84	0.84
0.84	0.84	0.84	0.84	0.84	0.84	0.84	0.88	0.88	0.88
0.88	0.88	0.88	0.88	0.88	0.92	0.92	0.92	0.94	

- (a) Make a histogram of the distribution. Although the result depends a bit on your choice of classes, the distribution appears roughly symmetric with no outliers.
- (b) Find the mean, median, standard deviation, and quartiles for these data. Comparing the mean and the median and comparing the distances of the two quartiles from the median suggest that the distribution is quite symmetric. Why?
- (c) If the distribution were exactly Normal with the mean and standard deviation you found in (b), what proportion of observations would lie between the two quartiles you found in (b)? What proportion of the actual observations lie between the quartiles (include observations equal to either quartile value). Despite the discrepancy, this distribution is "close enough to Normal" for statistical work in later chapters.

**3.51** Table 2.5 (page 65) gives data on the penetrability of soil at each of three levels of compression. We might expect the penetrability of specimens of the same soil at the same level of compression to follow a Normal distribution. Make stemplots of the data for loose and for intermediate compression. Does either sample seem roughly Normal? Does either appear distinctly non-Normal? If so, what kind of departure from Normality does your stemplot show?

**3.53** How many standard deviations above and below the mean do the quartiles of any Normal distribution lie? (Use the standard Normal distribution to answer this question.)

## Chapter 4 Problem Statements

**4.5** Airlines have increasingly outsourced the maintenance of their planes to other companies. Critics say that the maintenance may be less carefully done, so that outsourcing creates a safety hazard. As evidence, they point to government data on percent of major maintenance outsourced and percent of flight delays blamed on the airline (often due to maintenance problems):

Airline	Outsource percent	Delay percent	Airline	Outsource percent	Delay percent
AirTran	66	14	Frontier	65	31
Alaska	92	42	Hawaiian	80	70
American	46	26	JetBlue	68	18
America West	76	39	Northwest	76	43
ATA	18	19	Southwest	68	20
Continental	69	20	United	63	27
Delta	48	26	US Airways	77	24

Make a scatterplot that shows how delays depend on outsourcing.

**4.7** Does your plot for Exercise 4.5 show a positive association between maintenance outsourcing and delays caused by the airline? One airline is a high outlier in delay percent. Which airline is this? Aside from the outlier, does the plot show a roughly linear form? Is the relationship very strong?

**4.9** The study of dieting described in Exercise 4.4 collected data on the lean body mass (in kilograms) and metabolic rate (in calories) for both female and male subjects:

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

- Make a scatterplot of metabolic rate versus lean body mass for all 19 subjects. Use separate symbols to distinguish women and men.
- Does the same overall pattern hold for both women and men? What is the most important difference between women and men?

**4.13** The gas mileage of an automobile first increases and then decreases as the speed increases. Suppose that this relationship is very regular, as shown by the following data on speed (miles per hour) and mileage (miles per gallon):

<b>Speed</b>	20	30	40	50	60
<b>Mileage</b>	24	28	30	28	24

Make a scatterplot of mileage versus speed. Show that the correlation between speed and mileage is  $r = 0$ . Explain why the correlation is 0 even though there is a strong relationship between speed and mileage.

**4.27** Coffee is a leading export from several developing countries. When coffee prices are high, farmers often clear forest to plant more coffee trees. Here are five years of data on prices paid to coffee growers in Indonesia and the percent of forest area lost in a national park that lies in a coffee-producing region:

<b>Price (cents per pound)</b>	29	40	54	55	72
<b>Forest Loss (percent)</b>	0.49	1.59	1.69	1.82	3.10

- Make a scatterplot. Which is the explanatory variable? What kind of pattern does your plot show?
- Find the correlation  $r$  between coffee price and forest loss. Do your scatterplot and correlation support the idea that higher coffee prices increase the loss of forest?
- The price of coffee in international trade is given in dollars and cents. If the prices in the data were translated into the equivalent prices in euros, would the correlation between coffee price and percent of forest loss change? Explain your answer.

**4.29** Most people dislike losses more than they like gains. In money terms, people are about as sensitive to a loss of \$10 as to a gain of \$20. To discover what parts of the brain are active in decisions about gain and loss, psychologists presented subjects with a series of gambles with different odds and different amounts of winnings and losses. From a subject's choices, they constructed a measure of "behavioral loss aversion." Higher scores show greater sensitivity to losses. Observing brain activity while subjects made their decisions pointed to specific brain regions. Here are data for 16 subjects on behavioral loss aversion and "neural loss aversion," a measure of activity in one region of the brain:

<b>Neural</b>	-50.0	-39.1	-25.9	-26.7	-28.6	-19.8	-17.6	5.5
<b>Behavioral</b>	0.08	0.81	0.01	0.12	0.68	0.11	0.36	0.34
<b>Neural</b>	2.6	20.7	12.1	15.5	28.8	41.7	55.3	155.2
<b>Behavioral</b>	0.53	0.68	0.99	1.04	0.66	0.86	1.29	1.94

- Make a scatterplot that shows how behavior responds to brain activity.
- Describe the overall pattern of the data. There is one clear outlier.

- (c) Find the correlation  $r$  between neural and behavioral loss aversion both with and without the outlier. Does the outlier have a strong influence on the value of  $r$ ? By looking at your plot, explain why adding the outlier to the other data points causes  $r$  to increase.

**4.31** Japanese researchers measured the growth of icicles in a cold chamber under various conditions of temperature, wind, and water flow. Table 4.2 contains data produced under two sets of conditions. In both cases, there was no wind and the temperature was set at  $-11^{\circ}\text{C}$ . Water flowed over the icicle at a higher rate (29.6 milligrams per second) in Run 8905 and at a slower rate (11.9 mg/s) in Run 8903.

Run 8903				Run 8905			
Time (min)	Length (cm)						
10	0.6	130	18.1	10	0.3	130	10.4
20	1.8	140	19.9	20	0.6	140	11.0
30	2.9	150	21.0	30	1.0	150	11.9
40	4.0	160	23.4	40	1.3	160	12.7
50	5.0	170	24.7	50	3.2	170	13.9
60	6.1	180	27.8	60	4.0	180	14.6
70	7.9			70	5.3	190	15.8
80	10.1			80	6.0	200	16.2
90	10.9			90	6.9	210	17.9
100	12.7			100	7.8	220	18.8
110	14.4			110	8.3	230	19.9
120	16.6			120	9.6	240	21.1

- (a) Make a scatterplot of the length of the icicle in centimeters versus time in minutes, using separate symbols for the two runs.
- (b) What does your plot show about the pattern of growth of icicles? What does it show about the effect of changing the rate of water flow on icicle growth?

**4.33** To detect the presence of harmful insects in farm fields, we can put up boards covered with a sticky material and examine the insects trapped on the boards. Which colors attract insects best? Experimenters placed six boards of each of four colors at random locations in a field of oats and measured the number of cereal leaf beetles trapped. Here are the data:

Board color	Beetles trapped					
Blue	16	11	20	21	14	7
Green	37	32	20	29	37	32
White	21	12	14	17	13	20
Yellow	45	59	48	46	38	47

- (a) Make a plot of beetles trapped against color (space the four colors equally on the horizontal axis). Which color appears best at attracting beetles?
- (b) Does it make sense to speak of a positive or negative association between board color and beetles trapped? Why? Is correlation  $r$  a helpful description of the relationship? Why?

**4.43** We have data from a house in the Midwest that uses natural gas for heating. Will installing solar panels reduce the amount of gas consumed? Gas consumption is higher in cold weather, so the relationship between outside temperature and gas consumption is important. Here are data for 16 consecutive months:

	Nov.	Dec.	Jan.	Feb.	Mar.	Apr.	May	June
<b>Degree-days per day</b>	24	51	43	33	26	13	4	0
<b>Gas used per day</b>	6.3	10.9	8.9	7.5	5.3	4.0	1.7	1.2
	July	Aug.	Sep.	Oct.	Nov.	Dec.	Jan.	Feb.
<b>Degree-days per day</b>	0	1	6	12	30	32	52	30
<b>Gas used per day</b>	1.2	1.2	2.1	3.1	6.4	7.2	11.0	6.9

Outside temperature is recorded in degree-days, a common measure of demand for heating. A day's degree-days are the number of degrees its average temperature falls below  $65^\circ$ . Gas used is recorded in hundreds of cubic feet. Here are data for 23 more months after installing solar panels:

<b>Degree-days</b>	19	3	3	0	0	0	8	11	27	46	38	34
<b>Gas used</b>	3.2	2.0	1.6	1.0	0.7	0.7	1.6	3.1	5.1	7.7	7.0	6.1
<b>Degree-days</b>	16	9	2	1	0	2	3	18	32	34	40	
<b>Gas used</b>	3.0	2.1	1.3	1.0	1.0	1.0	1.2	3.4	6.1	6.5	7.5	

What do the before-and-after data show about the effect of solar panels? (Start by plotting both sets of data on the same plot, using two different plotting symbols.)

**4.49** We often describe our emotional reaction to social rejection as “pain.” Does social rejection cause activity in areas of the brain that are known to be activated by physical pain? If it does, we really do experience social and physical pain in similar ways. Psychologists first included and then deliberately excluded individuals from a social activity while they measured changes in brain activity. After each activity, the subjects filled out questionnaires that assessed how excluded they felt. Here are data for 13 subjects.

The explanatory variable is “social distress” measured by each subject's questionnaire score after exclusion relative to the score after inclusion. (So values greater than 1 show the degree of distress caused by exclusion.) The response variable is change in activity in a region of the brain that is activated by physical pain. Discuss what the data show.

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

## Chapter 5 Problem Statements

**5.3** An outbreak of the deadly Ebola virus in 2002 and 2003 killed 91 of the 95 gorillas in 7 home ranges in the Congo. To study the spread of the virus, measure “distance” by the number of home ranges separating a group of gorillas from the first group infected. Here are data on distance and number of days until deaths began in each later group:

<b>Distance</b>	1	3	4	4	4	5
<b>Days</b>	4	21	33	41	43	46

As you saw in Exercise 4.10 (page 106), there is a linear relationship between distance  $x$  and days  $y$ .

- Use your calculator to find the mean and standard deviation of both  $x$  and  $y$  and the correlation  $r$  between  $x$  and  $y$ . Use these basic measures to find the equation of the least-squares line for predicting  $y$  from  $x$ .
- Enter the data into your software or calculator and use the regression function to find the least-squares line. The result should agree with your work in (a) up to roundoff error.

**5.11** Exercise 4.5 (page 99) gives data for 14 airlines on the percent of major maintenance outsourced and the percent of flight delays blamed on the airline.

- Make a scatterplot with outsourcing percent as  $x$  and delay percent as  $y$ . Hawaiian Airlines is a high outlier in the  $y$  direction. Because several other airlines have similar values of  $x$ , the influence of this outlier is unclear without actual calculation.
- Find the correlation  $r$  with and without Hawaiian Airlines. How influential is the outlier for correlation?
- Find the least-squares line for predicting  $y$  from  $x$  with and without Hawaiian Airlines. Draw both lines on your scatterplot. Use both lines to predict the percent of delays blamed on an airline that has outsourced 76% of its major maintenance. How influential is the outlier for the least-squares line?

**5.35** Keeping water supplies clean requires regular measurement of levels of pollutants. The measurements are indirect—a typical analysis involves forming a dye by a chemical reaction with the dissolved pollutant, then passing light through the solution and measuring its “absorbance.” To calibrate such measurements, the laboratory measures known standard solutions and uses regression to relate absorbance and pollutant concentration. This is usually done every day. Here is one series of data on the absorbance for different levels of nitrates. Nitrates are measured in milligrams per liter of water.

<b>Nitrates</b>	50	50	100	200	400	800	1200	1600	2000	2000
<b>Absorbance</b>	7.0	7.5	12.8	24.0	47.0	93.0	138.0	183.0	230.0	226.0

- (a) Chemical theory says that these data should lie on a straight line. If the correlation is not at least 0.997, something went wrong and the calibration procedure is repeated. Plot the data and find the correlation. Must the calibration be done again?
- (b) The calibration process sets nitrate level and measures absorbance. The linear relationship that results is used to estimate the nitrate level in water from a measurement of absorbance. What is the equation of the line used to estimate nitrate level? What is the estimated nitrate level in a water specimen with absorbance 40?
- (c) Do you expect estimates of nitrate level from absorbance to be quite accurate? Why?

**5.37** Exercise 4.29 (page 116) describes an experiment that showed a linear relationship between how sensitive people are to monetary losses (“behavioral loss aversion”) and activity in one part of their brains (“neural loss aversion”).

- (a) Make a scatterplot with neural loss aversion as  $x$  and behavioral loss aversion as  $y$ . One point is a high outlier in both the  $x$  and  $y$  directions.
- (b) Find the least-squares line for predicting  $y$  from  $x$ , *leaving out the outlier*, and add the line to your plot.
- (c) The outlier lies very close to your regression line. Looking at the plot, you now expect that adding the outlier will increase the correlation but will have little effect on the least-squares line. Explain why.
- (d) Find the correlation and the equation of the least-squares line with and without the outlier. Your results verify the expectations from (c).

**5.39** People with diabetes must manage their blood sugar levels carefully. They measure their fasting plasma glucose (FPG) several times a day with a glucose meter. Another measurement, made at regular medical checkups, is called HbA. This is roughly the percent of red blood cells that have a glucose molecule attached. It measures average exposure to glucose over a period of several months. Table 5.2 gives data on both HbA and FPG for 18 diabetics five months after they had completed a diabetes education class.

<b>Subject</b>	<b>HbA (%)</b>	<b>FPG (mg/ml)</b>	<b>Subject</b>	<b>HbA (%)</b>	<b>FPG (mg/ml)</b>	<b>Subject</b>	<b>HbA (%)</b>	<b>FPG (mg/ml)</b>
1	6.1	141	7	7.5	96	13	10.6	103
2	6.3	158	8	7.7	87	14	10.7	172
3	6.4	112	9	7.9	148	15	10.7	359
4	6.8	153	10	8.7	172	16	11.2	145
5	7.0	143	11	9.4	200	17	13.7	147
6	7.1	95	12	10.5	271	18	19.3	255

- (a) Make a scatterplot with HbA as the explanatory variable. There is a positive linear relationship, but it is surprisingly weak.
- (b) Subject 15 is an outlier in the  $y$  direction. Subject 18 is an outlier in the  $x$  direction. Find the correlation for all 18 subjects, for all except Subject 15, and for all except Subject 18. Are either or both of these subjects influential for the correlation? Explain in simple language why  $r$  changes in opposite directions when we remove each of these points.

**5.41** Add three regression lines for predicting FPG from HbA to your scatterplot from Exercise 5.39: for all 18 subjects, for all except Subject 15, and for all except Subject 18. Is either Subject 15 or Subject 18 strongly influential for the least-squares line? Explain in simple language what features of the scatterplot explain the degree of influence.

**5.51** Do beavers benefit beetles? Researchers laid out 23 circular plots, each 4 meters in diameter, in an area where beavers were cutting down cottonwood trees. In each plot, they counted the number of stumps from trees cut by beavers and the number of clusters of beetle larvae. Ecologists think that the new sprouts from stumps are more tender than other cottonwood growth, so that beetles prefer them. If so, more stumps should produce more beetle larvae. Here are the data:

<b>Stumps</b>	2	2	1	3	3	4	3	1	2	5	1	3
<b>Beetle Larvae</b>	10	30	12	24	36	40	43	11	27	56	18	40
<b>Stumps</b>	2	1	2	2	1	1	4	1	2	1	4	
<b>Beetle Larvae</b>	25	8	21	14	16	6	54	9	13	14	50	

Analyze these data to see if they support the “beavers benefit beetles” idea.

**5.55** Exercise 4.43 (page 121) gives monthly data on outside temperature (in degree-days per day) and natural gas consumed for a house in the Midwest both before and after installing solar panels. A cold winter month in this location may average 45 degree-days per day (temperature  $20^\circ$ ). Use before-and-after regression lines to estimate the savings in gas consumption due to solar panels.

## Chapter 6 Problem Statements

**6.1** Recycling is supposed to save resources. Some people think recycled products are lower in quality than other products, a fact that makes recycling less practical. Here are data on attitudes toward coffee filters made of recycled paper among people who had bought these filters and people who had not:

	Think the quality of the recycled product is		
	Higher	The same	Lower
Buyers	20	7	9
Nonbuyers	29	25	43

- How many people does this table describe? How many of these were buyers of coffee filters made of recycled paper?
- Give the marginal distribution of opinion about the quality of recycled filters. What percent of consumers think the quality of the recycled product is the same or higher than the quality of other filters?

**6.3** Exercise 6.1 gives data on the opinions of people who have and have not bought coffee filters made from recycled paper. To see the relationship between opinion and experience with the product, find the conditional distributions of opinion (the response variable) for buyers and nonbuyers. What do you conclude?

**6.19** Will giving cocaine addicts an antidepressant drug help them break their addiction? An experiment assigned 24 chronic cocaine users to take the antidepressant drug desipramine, another 24 to take lithium, and another 24 to take a placebo. (Lithium is a standard drug to treat cocaine addiction. A placebo is a dummy pill, used so that the effect of being in the study but not taking any drug can be seen.) After three years, 14 of the 24 subjects in the desipramine group had remained free of cocaine, along with 6 of the 24 in the lithium group and 4 of the 24 in the placebo group.

- Make up a two-way table of “Treatment received” by whether or not the subject remained free of cocaine.
- Compare the effectiveness of the three treatments in preventing use of cocaine by former addicts. Use percents and draw a bar graph. What do you conclude?

**6.27** The University of Chicago's General Social Survey asked a representative sample of adults this question: “Which of the following statements best describes how your daily work is organized? 1: I am free to decide how my daily work is organized. 2: I can decide how my daily work is organized, within certain limits. 3: I am not free to decide how my

daily work is organized.” Here is a two-way table of the responses for three levels of education:

Response	Highest Degree Completed		
	Less than high school	High school	Bachelor's
1	31	161	81
2	49	269	85
3	47	112	14

How does freedom to organize you work depend on level of education?

**6.29** “Colleges and universities across the country are grappling with the case of the mysteriously vanishing male.” So said an article in the *Washington Post*. Here are data on the numbers of degrees earned in 2009–2010, as projected by the National Center for Education Statistics. The table entries are counts of degrees in thousands.

	Female	Male
<b>Associate's</b>	447	268
<b>Bachelor's</b>	945	651
<b>Master's</b>	397	251
<b>Professional</b>	49	44
<b>Doctor's</b>	26	25

Briefly contrast the participation of men and women in earning degrees.

## Chapter 7 Problem Statements

**7.3** The Pew Research Center asked a random sample of adults whether they had favorable or unfavorable opinions of a number of major companies. Answers to such questions depend a lot on recent news. Here are the percents with favorable opinions for several of the companies:

Company	Percent Favorable
Apple	71
Ben and Jerry's	59
Coors	53
Exxon/Mobil	44
Google	73
Haliburton	25
McDonald's	71
Microsoft	78
Starbucks	64
Wal-Mart	68

Make a graph that displays these data.

**7.5** Here are the weights (in milligrams) of 58 diamonds from a nodule carried up to the earth's surface in surrounding rock. This represents a single population of diamonds formed in a single event deep in the earth.

13.8 3.7 33.8 11.8 27.0 18.9 19.3 20.8 25.4 23.1 7.8 10.9  
 9.0 9.0 14.4 6.5 7.3 5.6 18.5 1.1 11.2 7.0 7.6 9.0  
 9.5 7.7 7.6 3.2 6.5 5.4 7.2 7.8 3.5 5.4 5.1 5.3  
 3.8 2.1 2.1 4.7 3.7 3.8 4.9 2.4 1.4 0.1 4.7 1.5  
 2.0 0.1 0.1 1.6 3.5 3.7 2.6 4.0 2.3 4.5

Make a graph that shows the distribution of weights of diamonds. Describe the shape of the distribution and any outliers. Use numerical measures appropriate for the shape to describe the center and spread.

**7.9** The Aleppo pine and the Torrey pine are widely planted as ornamental trees in Southern California. Here are the lengths (centimeters) of 15 Aleppo pine needles:

10.2 7.2 7.6 9.3 12.1 10.9 9.4 11.3 8.5 8.5 12.8 8.7 9.0 9.0 9.4

Here are the lengths of 18 needles from Torrey pines:

33.7 21.2 26.8 29.7 21.6 21.7 33.7 32.5 23.1  
 23.7 30.2 29.0 24.2 24.4 25.5 26.6 28.9 29.7

Use five-number summaries and boxplots to compare the two distributions. Given only the length of a needle, do you think you could say which pine species it comes from?

**7.15** The lengths of needles from Aleppo pines follow approximately the Normal distribution with mean 9.6 centimeters (cm) and standard deviation 1.6 cm. According to the 68–95–99.7 rule, what range of lengths covers the center 95% of Aleppo pine needles? What percent of needles are less than 6.4 cm long?

**7.17** Almost all medical schools in the United States require applicants to take the Medical College Admission Test (MCAT). The scores of applicants on the biological sciences part of the MCAT in 2007 were approximately Normal with mean 9.6 and standard deviation 2.2. For applicants who actually entered medical school, the mean score was 10.6 and the standard deviation was 1.7.

- (a) What percent of all applicants had scores higher than 13?  
 (b) What percent of those who entered medical school had scores between 8 and 12?

**7.19** From Rex Boggs in Australia comes an unusual data set: before showering in the morning, he weighed the bar of soap in his shower stall. The weight goes down as the soap is used. The data appear below (weights in grams). Notice that Mr. Boggs forgot to weigh the soap on some days.

Day	Weight	Day	Weight	Day	Weight
1	124	8	84	16	27
2	121	9	78	18	16
5	103	10	71	19	12
6	96	12	58	20	8
7	90	13	50	21	6

Plot the weight of the bar of soap against day. Is the overall pattern roughly linear? Based on your scatterplot, is the correlation between day and weight close to 1, positive but not close to 1, close to 0, negative but not close to  $-1$ , or close to  $-1$ ? Explain your answer. Then find the correlation  $r$  to verify what you concluded from the graph.

**7.23** Animals and people that take in more energy than they expend will get fatter. Here are data on 12 rhesus monkeys: 6 lean monkeys (4% to 9% body fat) and 6 obese monkeys (13% to 44% body fat). The data report the energy expended in 24 hours (kilojoules per minute) and the lean body mass (kilograms, leaving out fat) for each monkey.

Lean		Obese	
Mass	Energy	Mass	Energy
6.6	1.17	7.9	0.93
7.8	1.02	9.4	1.39
8.9	1.46	10.7	1.19
9.8	1.68	12.2	1.49
9.7	1.06	12.1	1.29
9.3	1.16	10.8	1.31

- (a) What is the mean lean body mass of the lean monkeys? Of the obese monkeys? Because animals with higher lean mass usually expend more energy, we can't directly compare energy expended
- (b) Instead, look at how energy expended is related to body mass. Make a scatterplot of energy versus mass, using different plot symbols for lean and obese monkeys. Then add to the plot two regression lines, one for lean monkeys and one for obese monkeys. What do these lines suggest about the monkeys?

**7.25** That animal species produce more offspring when their supply of food goes up isn't surprising. That some animals appear able to anticipate unusual food abundance is more surprising. Red squirrels eat seeds from pine cones, a food source that occasionally has very large crops (called seed masting). Here are data on an index of the abundance of pine cones and average number of offspring per female over 16 years:

<b>Cone Index</b>	0.00	2.02	0.25	3.22	4.68	0.31	3.37	3.09
<b>Offspring</b>	1.49	1.10	1.29	2.71	4.07	1.29	3.36	2.41
<b>Cone Index</b>	2.44	4.81	1.88	0.31	1.61	1.88	0.91	1.04
<b>Offspring</b>	1.97	3.41	1.49	2.02	3.34	2.41	2.15	2.12

Describe the relationship with both a graph and numerical measures, then summarize in words. What is striking is that the offspring are conceived in the spring, *before* the cones mature in the fall to feed the new young squirrels through the winter.

**7.27** The usual way to study the brain's response to sounds is to have subjects listen to "pure tones." The response to recognizable sounds may differ. To compare responses, researchers anesthetized macaque monkeys. They fed pure tones and also monkey calls directly to their brains by inserting electrodes. Response to the stimulus was measured by the firing rate (electrical spikes per second) of neurons in various areas of the brain. Table 7.1 contains the responses for 37 neurons.

Neuron	Tone	Call	Neuron	Tone	Call	Neuron	Tone	Call
1	474	500	14	145	42	26	71	134
2	256	138	15	141	241	27	68	65
3	241	485	16	129	294	28	59	182
4	226	338	17	113	123	29	59	97
5	185	194	18	112	182	30	57	318
6	174	159	19	102	141	31	56	201
7	176	341	20	100	118	32	47	279
8	168	85	21	74	62	33	46	62
9	161	303	22	72	112	34	41	84
10	150	208	23	20	193	35	26	203
11	19	66	24	21	129	36	28	192
12	20	54	25	26	135	37	31	70
13	35	103						

- (a) One important finding is that responses to monkey calls are generally stronger than responses to pure tones. For how many of the 37 neurons is this true?
- (b) We might expect some neurons to have strong responses to any stimulus and others to have consistently weak responses. There would then be a strong relationship between tone response and call response. Make a scatterplot of monkey call response against pure tone response (explanatory variable). Find the correlation  $r$  between tone and call responses. How strong is the linear relationship?

**7.37** We have 91 years of data on the date of ice breakup on the Tanana River. Describe the distribution of the breakup date with both a graph or graphs and appropriate numerical summaries. What is the median date (month and day) for ice breakup?

Year	Day										
1917	11	1933	19	1949	25	1965	18	1981	11	1997	11
1918	22	1934	11	1950	17	1966	19	1982	21	1998	1
1919	14	1935	26	1951	11	1967	15	1983	10	1999	10
1920	22	1936	11	1952	23	1968	19	1984	20	2000	12
1921	22	1937	23	1953	10	1969	9	1985	23	2001	19
1922	23	1938	17	1954	17	1970	15	1986	19	2002	18
1923	20	1939	10	1955	20	1971	19	1987	16	2003	10
1924	22	1940	1	1956	12	1972	21	1988	8	2004	5
1925	16	1941	14	1957	16	1973	15	1989	12	2005	9
1926	7	1942	11	1958	10	1974	17	1990	5	2006	13
1927	23	1943	9	1959	19	1975	21	1991	12	2007	8

1928	17	1944	15	1960	13	1976	13	1992	25	
1929	16	1945	27	1961	16	1977	17	1993	4	
1930	19	1946	16	1962	23	1978	11	1994	10	
1931	21	1947	14	1963	16	1979	11	1995	7	
1932	12	1948	24	1964	31	1980	10	1996	16	

**7.43** Here is one way that nature regulates the size of animal populations: High population density attracts predators, who remove a higher proportion of the population than when the density of the prey is low. One study looked at kelp perch and their common predator, the kelp bass. The researcher set up four large circular pens on the sandy ocean bottom in southern California. He chose young perch at random from a large group and placed 10, 20, 40, and 60 perch in the four pens. Then he dropped the nets protecting the pens, allowing bass to swarm in, and counted the perch left after 2 hours. Here are data on the proportions of perch eaten in four repetitions of this setup:

Perch	Proportion killed			
	10	0.0	0.1	0.3
20	0.2	0.3	0.3	0.6
40	0.075	0.3	0.6	0.725
60	0.517	0.55	0.7	0.817

Do the data support the principle that “more prey attract more predators, who drive down the number of prey”? Follow the four-step process (page 55) in your answer.

**7.49** Table 7.1 (page 188) contains data on the response of 37 monkey neurons to pure tones and to monkey calls. You made a scatterplot of these data in Exercise 7.27.

- Find the least-squares line for predicting a neuron's call response from its pure tone response. Add the line to your scatterplot. Mark on your plot the point (call it A) with the largest residual (either positive or negative) and also the point (call it B) that is an outlier in the  $x$  direction.
- How influential are each of these points for the correlation  $r$ ?
- How influential are each of these points for the regression line?

## Chapter 8 Problem Statements

**8.7** A firm wants to understand the attitudes of its minority managers toward its system for assessing management performance. Below is a list of all the firm's managers who are members of minority groups. Use Table B at line 139 to choose six to be interviewed in detail about the performance appraisal system.

1	Abdulhamid	8	Duncan	15	Huang	22	Puri
2	Agarwal	9	Fernandez	16	Kim	23	Richards
3	Baxter	10	Fleming	17	Lumumba	24	Rodriguez
4	Bonds	11	Gates	18	Mourning	25	Santiago
5	Brown	12	Gomez	19	Nguyen	26	Shen
6	Castro	13	Gupta	20	Peters	27	Vargas
7	Chavez	14	Hernandez	21	Peña	28	Wang

**8.11** Cook County, Illinois, has the second-largest population of any county in the United States (after Los Angeles County, California). Cook County has 30 suburban townships and an additional 8 townships that make up the city of Chicago. The suburban townships are

Barrington	Elk Grove	Maine	Orland	Riverside
Berwyn	Evanston	New Trier	Palatine	Schaumburg
Bloom	Hanover	Niles	Palos	Stickney
Bremen	Lemont	Northfield	Proviso	Thornton
Calumet	Leyden	Norwood	Park Rich	Wheeling
Cicero	Lyons	Oak Park	River Forest	Worth

The Chicago townships are

Hyde Park	Lake	North Chicago	South Chicago
Jefferson	Lake View	Rogers Park	West Chicago

Because city and suburban areas may differ, the first stage of a multistage sample chooses a stratified sample of 6 suburban townships and 4 of the more heavily populated Chicago townships. Use Table B or software to choose this sample. (If you use Table B, assign labels in alphabetical order and start at line 101 for the suburbs and at line 110 for Chicago.)

**8.27** You want to ask a sample of college students the question “How much do you trust information about health that you find on the Internet—a great deal, somewhat, not much, or not at all?” You try out this and other questions on a pilot group of 10 students chosen from your class. The class members are

Anderson	Deng	Glaus	Nguyen	Samuels
Arroyo	De Ramos	Helling	Palmiero	Shen
Batista	Drasin	Husain	Percival	Tse
Bell	Eckstein	Johnson	Prince	Velasco
Burke	Fernandez	Kim	Puri	Wallace
Cabrera	Fullmer	Molina	Richards	Washburn
Calloway	Gandhi	Morgan	Rider	Zabidi
Delluci	Garcia	Murphy	Rodriguez	Zhao

Choose an SRS of 10 students. If you use Table B, start at line 117.

**8.29** To gather data on a 1200-acre pine forest in Louisiana, the U.S. Forest Service laid a grid of 1410 equally spaced circular plots over a map of the forest. A ground survey visited a sample of 10% of these plots.

(a) How would you label the plots?

(b) Choose the first 5 plots in an SRS of 141 plots. (If you use Table B, start at line 105.)

**8.39** At a large block party there are 290 men and 110 women. You want to ask opinions about how to improve the next party. To be sure that women's opinions are adequately represented, you decide to choose a stratified random sample of 20 men and 20 women. Explain how you will assign labels to the names of the people at the party. Give the labels of the first 3 men and the first 3 women in your sample. If you use Table B, start at line 130.

## Chapter 9 Problem Statements

**9.9** The changing climate will probably bring more rain to California, but we don't know whether the additional rain will come during the winter wet season or extend into the long dry season in spring and summer. Kenwyn Suttle of the University of California at Berkeley and his coworkers carried out a randomized controlled experiment to study the effects of more rain in either season. They randomly assigned plots of open grassland to 3 treatments: added water equal to 20% of annual rainfall either during January to March (winter) or during April to June (spring), and no added water (control). Thirty-six circular plots of area 70 square meters were available (see the photo), of which 18 were used for this study. One response variable was total plant biomass, in grams per square meter, produced in a plot over a year.

- (a) Outline the design of the experiment, following the model of Figure 9.4.
- (b) Number all 36 plots and choose 6 at random for each of the 3 treatments. Be sure to explain how you did the random selection.

**9.33** Elementary schools in rural India are usually small, with a single teacher. The teachers often fail to show up for work. Here is an idea for improving attendance: give the teacher a digital camera with a tamper-proof time and date stamp and ask a student to take a photo of the teacher and class at the beginning and end of the day. Offer the teacher better pay for good attendance, verified by the photos. Will this work? A randomized comparative experiment started with 120 rural schools in Rajasthan and assigned 60 to this treatment and 60 to a control group. Random checks for teacher attendance showed that 21% of teachers in the treatment group were absent, as opposed to 42% in the control group

- (a) Outline the design of this experiment.
- (b) Label the schools and choose the first 10 schools for the treatment group. If you use Table B, start at line 108.

**9.35** Some people think that red wine protects moderate drinkers from heart disease better than other alcoholic beverages. This calls for a randomized comparative experiment. The subjects were healthy men aged 35 to 65. They were randomly assigned to drink red wine (9 subjects), drink white wine (9 subjects), drink white wine and also take polyphenols from red wine (6 subjects), take polyphenols alone (9 subjects), or drink vodka and lemonade (6 subjects). Outline the design of the experiment and randomly assign the 39 subjects to the 5 groups. If you use Table B, start at line 107.

**9.37** Doctors identify “chronic tension-type headaches” as headaches that occur almost daily for at least six months. Can antidepressant medications or stress management training reduce the number and severity of these headaches? Are both together more effective than either alone?

- (a) Use a diagram like Figure 9.2 to display the treatments in a design with two factors: medication yes or no and stress management yes or no. Then outline the design of a completely randomized experiment to compare these treatments.
- (b) The headache sufferers named below have agreed to participate in the study. Randomly assign the subjects to the treatments. If you use the *Simple Random Sample* applet or other software, assign all the subjects. If you use Table A, start at line 130 and assign subjects to only the first treatment group.

Abbott	Decker	Herrera	Lucero	Richter
Abdalla	Devlin	Hersch	Masters	Riley
Alawi	Engel	Hurwitz	Morgan	Samuels
Broden	Fuentes	Irwin	Nelson	Smith
Chai	Garrett	Jiang	Nho	Suarez
Chuang	Gill	Kelley	Ortiz	Upasani
Cordoba	Glover	Kim	Ramdas	Wilson
Custer	Hammond	Landers	Reed	Xiang

**9.41** Here's the opening of a Starbucks press release: "Starbucks Corp. on Monday said it would roll out a line of blended coffee drinks intended to tap into the growing popularity of reduced-calorie and reduced-fat menu choices for Americans." You wonder if Starbucks customers like the new "Mocha Frappuccino Light" as well as the regular Mocha Frappuccino coffee.

- (a) Describe a matched pairs design to answer this question. Be sure to include proper blinding of your subjects.
- (b) You have 20 regular Starbucks customers on hand. Use the *Simple Random Sample* applet or Table B at line 141 to do the randomization that your design requires.

**Chapter 10 Problem Statements**

**10.47** A couple plans to have three children. There are 8 possible arrangements of girls and boys. For example, GGB means the first two children are girls and the third child is a boy. All 8 arrangements are (approximately) equally likely.

- (a) Write down all 8 arrangements of the sexes of three children. What is the probability of any one of these arrangements?
- (b) Let  $X$  be the number of girls the couple has. What is the probability that  $X = 2$ ?
- (c) Starting from your work in (a), find the distribution of  $X$ . That is, what values can  $X$  take, and what are the probabilities for each value?

**10.51** A sample survey contacted an SRS of 663 registered voters in Oregon shortly after an election and asked respondents whether they had voted. Voter records show that 56% of registered voters had actually voted. We will see later that in this situation the proportion of the sample who voted (call this proportion  $V$ ) has approximately the Normal distribution with mean  $\mu = 0.56$  and standard deviation  $\sigma = 0.019$ .

- (a) If the respondents answer truthfully, what is  $P(0.52 \leq V \leq 0.60)$ ? This is the probability that the sample proportion  $V$  estimates the population proportion 0.56 within plus or minus 0.04.
- (b) In fact, 72% of the respondents said they had voted ( $V = 0.72$ ). If respondents answer truthfully, what is  $P(V \geq 0.72)$ ? This probability is so small that it is good evidence that some people who did not vote claimed that they did vote.

## Chapter 11 Problem Statements

**11.7** Let's illustrate the idea of a sampling distribution in the case of a very small sample from a very small population. The population is the scores of 10 students on an exam:

<b>Student</b>	0	1	2	3	4	5	6	7	8	9
<b>Score</b>	82	62	80	58	72	73	65	66	74	62

The parameter of interest is the mean score  $\mu$  in this population. The sample is an SRS of size  $n = 4$  drawn from the population. Because the students are labeled 0 to 9, a single random digit from Table B chooses one student for the sample.

- Find the mean of the 10 scores in the population. This is the population mean  $\mu$ .
- Use the first digits in row 116 of Table B to draw an SRS of size 4 from this population. What are the four scores in your sample? What is their mean  $\bar{x}$ ? This statistic is an estimate of  $\mu$ .
- Repeat this process 9 more times, using the first digits in rows 117 to 125 of Table B. Make a histogram of the 10 values of  $\bar{x}$ . You are constructing the sampling distribution of  $\bar{x}$ . Is the center of your histogram close to  $\mu$ ?

**11.9** Suppose that in fact the blood cholesterol level of all men aged 20 to 34 follows the Normal distribution with mean  $\mu = 188$  milligrams per deciliter (mg/dl) and standard deviation  $\sigma = 41$  mg/dl.

- Choose an SRS of 100 men from this population. What is the sampling distribution of  $\bar{x}$ ? What is the probability that  $\bar{x}$  takes a value between 185 and 191 mg/dl? This is the probability that  $\bar{x}$  estimates  $\mu$  within  $\pm 3$  mg/dl.
- Choose an SRS of 1000 men from this population. Now what is the probability that  $\bar{x}$  falls within  $\pm 3$  mg/dl of  $\mu$ ? The larger sample is much more likely to give an accurate estimate of  $\mu$ .

**11.13** An insurance company knows that in the entire population of millions of homeowners, the mean annual loss from fire is  $\mu = 250$  and the standard deviation of the loss is  $\sigma = 1000$ . The distribution of losses is strongly right-skewed: most policies have \$0 loss, but a few have large losses. If the company sells 10,000 policies, can it safely base its rates on the assumption that its average loss will be no greater than \$275? Follow the four-step process as illustrated in Example 11.8.

**11.27** Shelia's doctor is concerned that she may suffer from gestational diabetes (high blood glucose levels during pregnancy). There is variation both in the actual glucose level and in the blood test that measures the level. A patient is classified as having gestational

diabetes if the glucose level is above 140 milligrams per deciliter (mg/dl) one hour after having a sugary drink. Shelia's measured glucose level one hour after the sugary drink varies according to the Normal distribution with  $\mu = 125$  mg/dl and  $\sigma = 10$  mg/dl.

- (a) If a single glucose measurement is made, what is the probability that Shelia is diagnosed as having gestational diabetes?
- (b) If measurements are made on 4 separate days and the mean result is compared with the criterion 140 mg/dl, what is the probability that Shelia is diagnosed as having gestational diabetes?

**11.29** Shelia's measured glucose level one hour after a sugary drink varies according to the Normal distribution with  $\mu = 125$  mg/dl and  $\sigma = 10$  mg/dl. What is the level  $L$  such that there is probability only 0.05 that the mean glucose level of 4 test results falls above  $L$ ? (*Hint*: This requires a backward Normal calculation. See page 83 in Chapter 3 if you need to review.)

**11.31** The number of accidents per week at a hazardous intersection varies with mean 2.2 and standard deviation 1.4. This distribution takes only whole-number values, so it is certainly not Normal.

- (a) Let  $\bar{x}$  be the mean number of accidents per week at the intersection during a year (52 weeks). What is the approximate distribution of  $\bar{x}$  according to the central limit theorem?
- (b) What is the approximate probability that  $\bar{x}$  is less than 2?
- (c) What is the approximate probability that there are fewer than 100 accidents at the intersection in a year? (*Hint*: Restate this event in terms of  $\bar{x}$ .)

**11.33** Andrew plans to retire in 40 years. He plans to invest part of his retirement funds in stocks, so he seeks out information on past returns. He learns that over the entire 20th century, the real (that is, adjusted for inflation) annual returns on U.S. common stocks had mean 8.7% and standard deviation 20.2%. The distribution of annual returns on common stocks is roughly symmetric, so the mean return over even a moderate number of years is close to Normal. What is the probability (assuming that the past pattern of variation continues) that the mean annual return on common stocks over the next 40 years will exceed 10%? What is the probability that the mean return will be less than 5%? Follow the four-step process as illustrated in Example 11.8.

**11.39** Unlike Joe (see the previous exercise) the operators of the numbers racket can rely on the law of large numbers. It is said that the New York City mobster Casper Holstein took as many as 25,000 bets per day in the Prohibition era. That's 150,000 bets in a week if he takes Sunday off. Casper's mean winnings per bet are \$0.40 (he pays out 60 cents of

each dollar bet to people like Joe and keeps the other 40 cents.) His standard deviation for single bets is about \$18.96, the same as Joe's.

- (a) What are the mean and standard deviation of Casper's average winnings  $\bar{x}$  on his 150,000 bets?
- (b) According to the central limit theorem, what is the approximate probability that Casper's average winnings per bet are between \$0.30 and \$0.50? After only a week, Casper can be pretty confident that his winnings will be quite close to \$0.40 per bet.

**Chapter 12 Problem Statements**

**12.9** Suppose that 10% of adults belong to health clubs, and 40% of these health club members go to the club at least twice a week. What percent of all adults go to a health club at least twice a week? Write the information given in terms of probabilities and use the general multiplication rule.

**12.15** Continue your work from Exercise 12.13. What is the conditional probability that exactly 1 of the people will be allergic to peanuts or tree nuts, given that at least 1 of the 5 people suffers from one of these allergies?

**12.27** New York State's "Quick Draw" lottery moves right along. Players choose between one and ten numbers from the range 1 to 80; 20 winning numbers are displayed on a screen every four minutes. If you choose just one number, your probability of winning is  $20/80$ , or 0.25. Lester plays one number 8 times as he sits in a bar. What is the probability that all 8 bets lose?

**12.29** Slot machines are now video games, with outcomes determined by random number generators. In the old days, slot machines were like this: you pull the lever to spin three wheels; each wheel has 20 symbols, all equally likely to show when the wheel stops spinning; the three wheels are independent of each other. Suppose that the middle wheel has 9 cherries among its 20 symbols, and the left and right wheels have 1 cherry each.

- (a) You win the jackpot if all three wheels show cherries. What is the probability of winning the jackpot?
- (b) There are three ways that the three wheels can show two cherries and one symbol other than a cherry. Find the probability of each of these ways.
- (c) What is the probability that the wheels stop with exactly two cherries showing among them?

**12.47** You are tossing a pair of balanced dice in a board game. Tosses are independent. You land in a danger zone that requires you to roll doubles (both faces show the same number of spots) before you are allowed to play again. How long will you wait to play again?

- (a) What is the probability of rolling doubles on a single toss of the dice? (If you need review, the possible outcomes appear in Figure 10.2 (page 267). All 36 outcomes are equally likely.)
- (b) What is the probability that you do not roll doubles on the first toss, but you do on the second toss?

- (c) What is the probability that the first two tosses are not doubles and the third toss is doubles? This is the probability that the first doubles occurs on the third toss.
- (d) Now you see the pattern. What is the probability that the first doubles occurs on the fourth toss? On the fifth toss? Give the general result: what is the probability that the first doubles occurs on the  $k$ th toss?

(*Comment:* The distribution of the number of trials to the first success is called a *geometric distribution*. In this problem you have found geometric distribution probabilities when the probability of a success on each trial is  $1/6$ . The same idea works for any probability of success.)

**Chapter 13 Problem Statements**

**13.5** Typing errors in a text are either nonword errors (as when “the” is typed as “teh”) or word errors that result in a real but incorrect word. Spell-checking software will catch nonword errors but not word errors. Human proofreaders catch 70% of word errors. You ask a fellow student to proofread an essay in which you have deliberately made 10 word errors.

- (a) If the student matches the usual 70% rate, what is the distribution of the number of errors caught? What is the distribution of the number of errors missed?
- (b) Missing 3 or more out of 10 errors seems a poor performance. What is the probability that a proofreader who catches 70% of word errors misses exactly 3 out of 10? If you use software, also find the probability of missing 3 or more out of 10.

**13.11** A small liberal arts college would like to have an entering class of 415 students next year. Past experience shows that about 27% of the students admitted will decide to attend. The college therefore plans to admit 1535 students. Suppose that students make their decisions independently and that the probability is 0.27 that a randomly chosen student will accept the offer of admission.

- (a) What are the mean and standard deviation of the number of students who accept the admissions offer from this college?
- (b) Use the Normal approximation: what is the approximate probability that the college gets more students than they want?
- (c) Use software to compute the exact probability that the college gets more students than they want. How good is the approximation in part (b)?

**13.25** A believer in the random walk theory of stock markets thinks that an index of stock prices has probability 0.65 of increasing in any year. Moreover, the change in the index in any given year is not influenced by whether it rose or fell in earlier years. Let  $X$  be the number of years among the next 5 years in which the index rises.

- (a)  $X$  has a binomial distribution. What are  $n$  and  $p$ ?
- (b) What are the possible values that  $X$  can take?
- (c) Find the probability of each value of  $X$ . Draw a probability histogram for the distribution of  $X$ . (See Figure 13.2 for an example of a probability histogram.)
- (d) What are the mean and standard deviation of this distribution? Mark the location of the mean on your histogram.

**13.27** Many women take oral contraceptives to prevent pregnancy. Under ideal conditions, 1% of women taking the pill become pregnant within one year. In typical use,

however, 5% become pregnant. Choose at random 20 women taking the pill. How many become pregnant in the next year?

- (a) Explain why this is a binomial setting.
- (b) What is the probability that at least one of the women becomes pregnant under ideal conditions? What is the probability in typical use?

**13.29** A study of the effectiveness of oral contraceptives interviews a random sample of 500 women who are taking the pill.

- (a) Based on the information about typical use in Exercise 13.27, what is the probability that at least 25 of these women become pregnant in the next year? (Check that the Normal approximation is permissible and use it to find this probability. If your software allows, find the exact binomial probability and compare the two results.)
- (b) We can't use the Normal approximation to the binomial distributions to find this probability under ideal conditions as described in Exercise 13.27. Why not?

**13.31** According to genetic theory, the blossom color in the second generation of a certain cross of sweet peas should be red or white in a 3:1 ratio. That is, each plant has probability  $3/4$  of having red blossoms, and the blossom colors of separate plants are independent.

- (a) What is the probability that exactly 6 out of 8 of these plants have red blossoms?
- (b) What is the mean number of red-blossomed plants when 80 plants of this type are grown from seeds?
- (c) What is the probability of obtaining at least 60 red-blossomed plants when 80 plants are grown from seeds? Use the Normal approximation. If your software allows, find the exact binomial probability and compare the two results.

**13.33** The Census Bureau says that 21% of Americans aged 18 to 24 do not have a high school diploma. A vocational school wants to attract young people who may enroll in order to achieve high school equivalency. The school mails an advertising flyer to 25,000 persons between the ages of 18 and 24.

- (a) If the mailing list can be considered a random sample of the population, what is the mean number of high school dropouts who will receive the flyer?
- (b) What is the approximate probability that at least 5000 dropouts will receive the flyer?

**13.35** Here is a simple probability model for multiple-choice tests. Suppose that each student has probability  $p$  of correctly answering a question chosen at random from a universe of possible questions. (A strong student has a higher  $p$  than a weak student.) Answers to different questions are independent.

- (a) Jodi is a good student for whom  $p = 0.75$ . Use the Normal approximation to find the probability that Jodi scores between 70% and 80% on a 100-question test.
- (b) If the test contains 250 questions, what is the probability that Jodi will score between 70% and 80%? You see that Jodi's score on the longer test is more likely to be close to her "true score."

**13.39** In 2007, Bob Jones University ended its fall semester a week early because of a whooping cough outbreak; 158 students were isolated and another 1200 given antibiotics as a precaution. Authorities react strongly to whooping cough outbreaks because the disease is so contagious. Because the effect of childhood vaccination often wears off by late adolescence, treat the Bob Jones students as if they were unvaccinated. It appears that about 1400 students were exposed. What is the probability that at least 75% of these students develop infections if not treated? (Fortunately, whooping cough is much less serious after infancy.)

**13.41** We would like to find the probability that exactly 2 of the 20 exposed children in the previous exercise develop whooping cough.

- (a) One way to get 2 infections is to get 1 among the 17 vaccinated children and 1 among the 3 unvaccinated children. Find the probability of exactly 1 infection among the 17 vaccinated children. Find the probability of exactly 1 infection among the 3 unvaccinated children. These events are independent: what is the probability of exactly 1 infection in each group?
- (b) Write down all the ways in which 2 infections can be divided between the two groups of children. Follow the pattern of part (a) to find the probability of each of these possibilities. Add all of your results (including the result of part (a)) to obtain the probability of exactly 2 infections among the 20 children.

### Chapter 14 Problem Statements

**14.3** The critical value  $z^*$  for confidence level 97.5% is not in Table C. Use software or Table A of standard Normal probabilities to find  $z^*$ . Include in your answer a sketch like Figure 14.3 with  $C = 0.975$  and your critical value  $z^*$  marked on the axis.

**14.5** Here are the IQ test scores of 31 seventh-grade girls in a Midwest school district:

114	100	104	89	102	91	114	114	103	105
108	130	120	132	111	128	118	119	86	72
111	103	74	112	107	103	98	96	112	112
									93

- These 31 girls are an SRS of all seventh-grade girls in the school district. Suppose that the standard deviation of IQ scores in this population is known to be  $\sigma = 15$ . We expect the distribution of IQ scores to be close to Normal. Make a stemplot of the distribution of these 31 scores (split the stems) to verify that there are no major departures from Normality. You have now checked the “simple conditions” to the extent possible.
- Estimate the mean IQ score for all seventh-grade girls in the school district, using a 99% confidence interval. Follow the four-step process as illustrated in Example 14.3.

**14.13** The  $P$ -value for the first cola in Example 14.7 is the probability (taking the null hypothesis  $\mu = 0$  to be true) that  $\bar{x}$  takes a value at least as large as 0.3.

- What is the sampling distribution of  $\bar{x}$  when  $\mu = 0$ ? This distribution appears in Figure 14.6.
- Do a Normal probability calculation to find the  $P$ -value. Your result should agree with Example 14.7 up to roundoff error.

**14.17** Exercise 14.7 describes 6 measurements of the electrical conductivity of a liquid. You stated the null and alternative hypotheses in Exercise 14.9.

- One set of measurements has mean conductivity  $\bar{x} = 4.98$ . Enter this  $\bar{x}$ , along with the other required information, into the *P-Value of a Test of Significance* applet. What is the  $P$ -value? Is this outcome statistically significant at the  $\alpha = 0.05$  level? At the  $\alpha = 0.01$  level?
- Another set of measurements has  $\bar{x} = 4.7$ . Use the applet to find the  $P$ -value for this outcome. Is it statistically significant at the  $\alpha = 0.05$  level? At the  $\alpha = 0.01$  level?
- Explain briefly why these  $P$ -values tell us that one outcome is strong evidence against the null hypothesis and that the other outcome is not.

**14.19** Here are 6 measurements of the electrical conductivity of a liquid:

5.32 4.88 5.10 4.73 5.15 4.75

The liquid is supposed to have conductivity 5. Do the measurements give good evidence that the true conductivity is not 5?

The 6 measurements are an SRS from the population of all results we would get if we kept measuring conductivity forever. This population has a Normal distribution with mean equal to the true conductivity of the liquid and standard deviation 0.2. Use this information to carry out a test, following the four-step process as illustrated in Example 14.9.

**14.21** A test of  $H_0: \mu = 1$  against  $H_a: \mu > 1$  has test statistic  $z = 1.776$ . Is this test significant at the 5% level ( $\alpha = 0.05$ )? Is it significant at the 1% level ( $\alpha = 0.01$ )?

**14.23** A random number generator is supposed to produce random numbers that are uniformly distributed on the interval from 0 to 1. If this is true, the numbers generated come from a population with  $\mu = 0.5$  and  $\sigma = 0.2887$ . A command to generate 100 random numbers gives outcomes with mean  $\bar{x} = 0.4365$ . Assume that the population  $\sigma$  remains fixed. We want to test

$$H_0: \mu = 0.5$$

$$H_a: \mu \neq 0.5$$

- Calculate the value of the  $z$  test statistic.
- Use Table C: is  $z$  significant at the 5% level ( $\alpha = 0.05$ )?
- Use Table C: is  $z$  significant at the 1% level ( $\alpha = 0.01$ )?
- Between which two Normal critical values  $z^*$  in the bottom row of Table C does  $z$  lie? Between what two numbers does the  $P$ -value lie? Does the test give good evidence against the null hypothesis?

**14.35** Young men in North America and Europe (but not in Asia) tend to think they need more muscle to be attractive. One study presented 200 young American men with 100 images of men with various levels of muscle. Researchers measure level of muscle in kilograms per square meter ( $\text{kg}/\text{m}^2$ ) of fat-free body mass. Typical young men have about  $20 \text{ kg}/\text{m}^2$ . Each subject chose two images, one that represented his own level of body muscle and one that he thought represented “what women prefer.” The mean gap between self-image and “what women prefer” was  $2.35 \text{ kg}/\text{m}^2$ .

Suppose that the “muscle gap” in the population of all young men has a Normal distribution with standard deviation  $2.5 \text{ kg}/\text{m}^2$ . Give a 90% confidence interval for the

mean amount of muscle young men think they should add to be attractive to women. (They are wrong: women actually prefer a level close to that of typical men.)

**14.41** If young men thought that their own level of muscle was about what women prefer, the mean “muscle gap” in the study described in Exercise 14.35 would be 0. We suspect (before seeing the data) that young men think women prefer more muscle than they themselves have.

- State null and alternative hypotheses for testing this suspicion.
- What is the value of the test statistic  $z$ ?
- You can tell just from the value of  $z$  that the evidence in favor of the alternative is very strong (that is, the  $P$ -value is very small). Explain why this is true.

**14.51** Breast-feeding mothers secrete calcium into their milk. Some of the calcium may come from their bones, so mothers may lose bone mineral. Researchers measured the percent change in mineral content of the spines of 47 mothers during three months of breast-feeding. Here are the data:

-4.7	-2.5	-4.9	-2.7	-0.8	-5.3	-8.3	-2.1	-6.8	-4.3
2.2	-7.8	-3.1	-1.0	-6.5	-1.8	-5.2	-5.7	-7.0	-2.2
-6.5	-1.0	-3.0	-3.6	-5.2	-2.0	-2.1	-5.6	-4.4	-3.3
-4.0	-4.9	-4.7	-3.8	-5.9	-2.5	-0.3	-6.2	-6.8	1.7
0.3	-2.3	0.4	-5.3	0.2	-2.2	-5.1			

- The researchers are willing to consider these 47 women as an SRS from the population of all nursing mothers. Suppose that the percent change in this population has standard deviation  $\sigma = 2.5\%$ . Make a stemplot of the data to see that they appear to follow a Normal distribution quite closely. (Don't forget that you need both a 0 and a  $-0$  stem because there are both positive and negative values.)
- Use a 99 % confidence interval to estimate the mean percent change in the population.

**14.53** Exercise 14.51 gives the percent change in the mineral content of the spine for 47 mothers during three months of nursing a baby. As in that exercise, suppose that the percent change in the population of all nursing mothers has a Normal distribution with standard deviation  $\sigma = 2.5\%$ . Do these data give good evidence that on the average nursing mothers lose bone mineral?

**14.55** Athletes performing in bright sunlight often smear black eye grease under their eyes to reduce glare. Does eye grease work? In one study, 16 student subjects took a test of sensitivity to contrast after 3 hours facing into bright sun, both with and without eye grease. This is a matched pairs design. Here are the differences in sensitivity, with eye grease minus without eye grease:

0.07	0.64	-0.12	-0.05	-0.18	0.14	-0.16	0.03
0.05	0.02	0.43	0.24	-0.11	0.28	0.05	0.29

We want to know whether eye grease increases sensitivity on the average.

- (a) What are the null and alternative hypotheses? Say in words what mean  $\mu$  your hypotheses concern.
- (b) Suppose that the subjects are an SRS of all young people with normal vision, that contrast differences follow a Normal distribution in this population, and that the standard deviation of differences is  $\sigma = 0.22$ . Carry out a test of significance.

## Chapter 15 Problem Statements

**15.5** Example 14.1 (page 360) described NHANES survey data on the body mass index (BMI) of 654 young women. The mean BMI in the sample was  $\bar{x} = 26.8$ . We treated these data as an SRS from a Normally distributed population with standard deviation  $\sigma = 7.5$ .

- Suppose that we had an SRS of just 100 young women. What would be the margin of error for 95% confidence?
- Find the margins of error for 95% confidence based on SRSs of 400 young women and 1600 young women.
- Compare the three margins of error. How does increasing the sample size change the margin of error of a confidence interval when the confidence level and population standard deviation remain the same?

**15.9** Give a 95% confidence interval for the mean pH  $\mu$  for each sample size in the previous exercise. The intervals, unlike the  $P$ -values, give a clear picture of what mean pH values are plausible for each sample.

**15.11** Example 14.1 (page 360) assumed that the body mass index (BMI) of all American young women follows a Normal distribution with standard deviation  $\sigma = 7.5$ . How large a sample would be needed to estimate the mean BMI  $\mu$  in this population to within  $\pm 1$  with 95% confidence?

**15.41** Software can generate samples from (almost) exactly Normal distributions. Here is a random sample of size 5 from the Normal distribution with mean 10 and standard deviation 2:

6.47                  7.51                  10.10                  13.63                  9.91

These data match the conditions for a  $z$  test better than real data will: the population is very close to Normal and has known standard deviation  $\sigma = 2$ , and the population mean is  $\mu = 10$ . Test the hypotheses

$$H_0 : \mu = 8$$

$$H_a : \mu \neq 8$$

- What are the  $z$  statistic and its  $P$ -value? Is the test significant at the 5% level?
- We know that the null hypothesis does not hold, but the test failed to give strong evidence against  $H_0$ . Explain why this is not surprising.

**15.49** The previous exercise shows how to calculate the power of a one-sided  $z$  test. Power calculations for two-sided tests follow the same outline. We will find the power of a test based on 6 measurements of the conductivity of a liquid, reported in Exercise 15.13. The hypotheses are

$$H_0 : \mu = 5$$

$$H_a : \mu \neq 5$$

The population of all measurements is Normal with standard deviation  $\sigma = 0.2$ , and the alternative we hope to be able to detect is  $\mu = 5.1$ . (If you used the *Power of a Test* applet for Exercise 15.15, the two Normal curves for  $n = 6$  illustrate parts (a) and (b) below.)

- (a) Write the  $z$  test statistic in terms of the sample mean  $\bar{x}$ . For what values of  $z$  does this two-sided test reject  $H_0$  at the 5 % significance level?
- (b) Restate your result from part (a): what values of  $\bar{x}$  lead to rejection of  $H_0$ ?
- (c) Now suppose that  $\mu = 5.1$ . What is the probability of observing an  $\bar{x}$  that leads to rejection of  $H_0$ ? This is the power of the test.

**Chapter 16 Problem Statements**

**16.3** A university's financial aid office wants to know how much it can expect students to earn from summer employment. This information will be used to set the level of financial aid. The population contains 3478 students who have completed at least one year of study but have not yet graduated. The university will send a questionnaire to an SRS of 100 of these students, drawn from an alphabetized list.

- (a) Describe how you will label the students in order to select the sample.
- (b) Use Table B, beginning at line 105, to select the first 5 students in the sample.
- (c) What is the response variable in this study?

**16.5** Elephants sometimes damage crops in Africa. It turns out that elephants dislike bees. They recognize beehives in areas where they are common and avoid them. Can this be used to keep elephants away from trees? A group in Kenya placed active beehives in some trees and empty beehives in others. Will elephant damage be less in trees with hives? Will even empty hives keep elephants away?

- (a) Outline the design of an experiment to answer these questions using 72 acacia trees (be sure to include a control group).
- (b) Use software or the *Simple Random Sample* to choose the trees for the active-hive group, or Table B at line 137 to choose the first 4 trees in that group.
- (c) What is the response variable in this experiment?

**16.13** The distribution of blood cholesterol level in the population of young men aged 20 to 34 years is close to Normal with standard deviation  $\sigma = 41$  milligrams per deciliter (mg/dl). You measure the blood cholesterol of 14 cross-country runners. The mean level is  $\bar{x} = 172$  mg/dl. Assuming that  $\sigma$  is the same as in the general population, give a 90% confidence interval for the mean level  $\mu$  among cross-country runners.

**16.15** How large a sample is needed to cut the margin of error in Exercise 16.13 in half? How large a sample is needed to cut the margin of error to  $\pm 5$  mg/dl?

**16.17** The level of pesticides found in the blubber of whales is a measure of pollution of the oceans by runoff from land and can also be used to identify different populations of whales. A sample of 8 male minke whales in the West Greenland area of the North Atlantic found the mean concentration of the insecticide dieldrin to be  $\bar{x} = 357$  nanograms per gram of blubber (ng/g). Suppose that the concentration in all such whales varies Normally with standard deviation  $\sigma = 50$  ng/g. Use a 95% confidence interval to estimate the mean level. Be sure to state your conclusion in plain language.

**16.19** Use the information in Exercise 16.17 to give an 80% confidence interval and a 90% confidence interval for the mean concentration of dieldrin in the whale population. What general fact about confidence intervals do the margins of error of your three intervals illustrate?

**16.21** IQ tests are scaled so that the mean score in a large population should be  $\mu = 100$ . We suspect that the very-low-birth-weight population has mean score less than 100. Does the study described in the previous exercise give good evidence that this is true? State hypotheses, carry out a test assuming that the "simple conditions" (page 360) hold, and give your conclusion in plain language.

**16.27** The time that people require to react to a stimulus usually has a right-skewed distribution, as lack of attention or tiredness causes some lengthy reaction times. Reaction times for children with attention deficit hyperactivity disorder (ADHD) are more skewed, as their condition causes more frequent lack of attention. In one study, children with ADHD were asked to press the spacebar on a computer keyboard when any letter other than X appeared on the screen. With 2 seconds between letters, the mean reaction time was 445 milliseconds (ms) and the standard deviation was 82 ms. Take these values to be the population  $\mu$  and  $\sigma$  for ADHD children.

- What are the mean and standard deviation of the mean reaction time  $\bar{x}$  for a randomly chosen group of 15 ADHD children? For a group of 150 such children?
- The distribution of reaction time is strongly skewed. Explain briefly why we hesitate to regard  $\bar{x}$  as Normally distributed for 15 children but are willing to use a Normal distribution for the mean reaction time of 150 children.
- What is the approximate probability that the mean reaction time in a group of 150 ADHD children is greater than 450 ms?

**16.45** Here are the daily average body temperatures (degrees Fahrenheit) for 20 healthy adults:

98.74 98.83 96.80 98.12 97.89 98.09 97.87 97.42 97.30 97.84  
100.27 97.90 99.64 97.88 98.54 98.33 97.87 97.48 98.92 98.33

- Make stemplot of the data. The distribution is roughly symmetric and single-peaked. There is one mild outlier. We expect the distribution of the sample mean  $\bar{x}$  to be close to Normal.
- Do these data give evidence that the mean body temperature for all healthy adults is not equal to the traditional 98.6 degrees? Follow the four-step process for significance tests (page 378). (Suppose that body temperature varies Normally with standard deviation 0.7 degree.)

**16.47** Use the data in Exercise 16.45 to estimate mean body temperature with 90% confidence. Follow the four-step process for confidence intervals (page 366).

## Chapter 17 Problem Statements

**17.3** Use Table C or software to find

- (a) the critical value for a one-sided test with level  $\alpha = 0.05$  based on the  $t(5)$  distribution.
- (b) the critical value for a 98% confidence interval based on the  $t(21)$  distribution.

**17.7** The composition of the earth's atmosphere may have changed over time. To try to discover the nature of the atmosphere long ago, we can examine the gas in bubbles inside ancient amber. Amber is tree resin that has hardened and been trapped in rocks. The gas in bubbles within amber should be a sample of the atmosphere at the time the amber was formed. Measurements on specimens of amber from the late Cretaceous era (75 to 95 million years ago) give these percents of nitrogen:

63.4 65.0 64.4 63.3 54.8 64.5 60.8 49.1 51.0

Assume (this is not yet agreed on by experts) that these observations are an SRS from the late Cretaceous atmosphere. Use a 90% confidence interval to estimate the mean percent of nitrogen in ancient air. Follow the four-step process as illustrated in Example 17.2.

**17.9** The one-sample  $t$  statistic from a sample of  $n = 25$  observations for the two-sided test of

$$H_0 : \mu = 64$$

$$H_a : \mu \neq 64$$

has the value  $t = 1.12$ .

- (a) What are the degrees of freedom for  $t$ ?
- (b) Locate the two critical values  $t^*$  from Table C that bracket  $t$ . What are the two-sided  $P$ -values for these two entries?
- (c) Is the value  $t = 1.12$  statistically significant at the 10% level? At the 5% level?

**17.11** The usual way to study the brain's response to sounds is to have subjects listen to "pure tones." The response to recognizable sounds may differ. To compare responses, researchers anesthetized macaque monkeys. They fed pure tones and also monkey calls directly to their brains by inserting electrodes. Response to the stimulus was measured by the firing rate (electrical spikes per second) of neurons in various areas of the brain. Table 17.2 contains the responses for 37 neurons. Researchers suspected that the response to monkey calls would be stronger than the response to a pure tone. Do the data support this idea?

Neuron	Tone	Call	Neuron	Tone	Call	Neuron	Tone	Call
1	474	500	14	145	42	26	71	134
2	256	138	15	141	241	27	68	65
3	241	485	16	129	294	28	59	182
4	226	338	17	113	123	29	59	97
5	185	194	18	112	182	30	57	318
6	174	159	19	102	141	31	56	201
7	176	341	20	100	118	32	47	279
8	168	85	21	74	62	33	46	62
9	161	303	22	72	112	34	41	84
10	150	208	23	20	193	35	26	203
11	19	66	24	21	129	36	28	192
12	20	54	25	26	135	37	31	70
13	35	103						

Complete the *Plan*, *Solve*, and *Conclude* steps of the four-step process, following the model of Example 17.4.

**17.13** A group of earth scientists studied the small diamonds found in a nodule of rock carried up to the earth's surface in surrounding rock. This is an opportunity to examine a sample from a single population of diamonds formed in a single event deep in the earth. Table 17.3 (page 460) presents data on the nitrogen content (parts per million) and the abundance of carbon-13 in these diamonds.

Diamond	Nitrogen (ppm)	Carbon-13 Ratio	Diamond	Nitrogen (ppm)	Carbon-13 Ratio
1	487	-2.78	13	273	-2.73
2	1430	-1.39	14	94	-3.57
3	60	-4.26	15	69	-3.83
4	244	-1.19	16	262	-2.04
5	196	-2.12	17	120	-2.82
6	274	-2.87	18	302	-0.84
7	41	-3.68	19	75	-3.57
8	54	-3.29	20	242	-2.42
9	473	-3.79	21	115	-3.89
10	30	-4.06	22	65	-3.87
11	98	-1.83	23	311	-1.58
12	41	-4.03	24	61	-3.97

(Carbon has several isotopes, forms with different numbers of neutrons in the nuclei of their atoms. Carbon-12 makes up almost 99% of natural carbon. The abundance of carbon-13 is measured by the ratio of carbon-13 to carbon-12, in parts per thousand more or less than a standard. The minus signs in the data mean that the ratio is smaller in these diamonds than in standard carbon.)

We would like to estimate the mean abundance of both nitrogen and carbon-13 in the population of diamonds represented by this sample. Examine the data for nitrogen. Can we use a  $t$  confidence interval for mean nitrogen? Explain your answer. Give a 95% confidence interval if you think the result can be trusted.

**17.25** You read in the report of a psychology experiment: “Separate analyses for our two groups of 12 participants revealed no overall placebo effect for our student group (mean = 0.08, SD = 0.37,  $t(11) = 0.49$ ) and a significant effect for our non-student group (mean = 0.35, SD = 0.37,  $t(11) = 3.25$ ,  $p < 0.01$ ).” The null hypothesis is that the mean effect is zero. What are the correct values of the two  $t$  statistics based on the means and standard deviations? Compare each correct  $t$ -value with the critical values in Table C. What can you say about the two-sided  $P$ -value in each case?

**17.27** The Trial Urban District Assessment (TUDA) is a government-sponsored study of student achievement in large urban school districts. TUDA gives a reading test scored from 0 to 500. A score of 243 is a “basic” reading level and a score of 281 is “proficient.” Scores for a random sample of 1470 eighth-graders in Atlanta had  $\bar{x} = 240$  with standard error 1.1.

- We don't have the 1470 individual scores, but use of the  $t$  procedures is surely safe. Why?
- Give a 99% confidence interval for the mean score of all Atlanta eighth-graders. (Be careful: the report gives the standard error of  $\bar{x}$ , not the standard deviation  $s$ .)
- Urban children often perform below the basic level. Is there good evidence that the mean for all Atlanta eighth-graders is less than the basic level?

**17.29** The placebo effect is particularly strong in patients with Parkinson's disease. To understand the workings of the placebo effect, scientists measure activity at a key point in the brain when patients receive a placebo that they think is an active drug and also when no treatment is given. The same six patients are measured both with and without the placebo, at different times.

- Explain why the proper procedure to compare the mean response to placebo with control (no treatment) is a matched pairs  $t$  test.
- The six differences (treatment minus control) had  $\bar{x} = -0.326$  and  $s = 0.181$ . Is there significant evidence of a difference between treatment and control?

**17.31** Blissymbols are pictographs (think of Egyptian hieroglyphics) sometimes used to help learning-disabled children. In a study of computer-assisted learning, 12 normal-ability schoolchildren were assigned at random to each of four computer learning

programs. After they used the program, they attempted to recognize 24 Blissymbols. Here are the counts correct for one of the programs:

12 22 9 14 20 15 9 10 11 11 15 6

- Make a stemplot (split the stems). Are there outliers or strong skewness that would forbid use of the  $t$  procedures?
- Give a 90% confidence interval for the mean count correct among all children of this age who use the program.

**17.33** Our bodies have a natural electrical field that is known to help wounds heal. Does changing the field strength slow healing? A series of experiments with newts investigated this question. In one experiment, the two hind limbs of 12 newts were assigned at random to either experimental or control groups. This is a matched pairs design. The electrical field in the experimental limbs was reduced to zero by applying a voltage. The control limbs were left alone. Here are the rates at which new cells closed a razor cut in each limb, in micrometers per hour:

Newt	1	2	3	4	5	6	7	8	9	10	11	12
Control limb	36	41	39	42	44	39	39	56	33	20	49	30
Experimental limb	28	31	27	33	33	38	45	25	28	33	47	23

- Make a stemplot of the differences between limbs of the same newt (control limb minus experimental limb). There is a high outlier.
- A good way to judge the effect of an outlier is to do your analysis twice, once with the outlier and a second time without it. Carry out two  $t$  tests to see if the mean healing rate is significantly lower in the experimental limbs, one including all 12 newts and another that omits the outlier. What are the test statistics and their  $P$ -values? Does the outlier have a strong influence on your conclusion?

**17.35** Here's a new idea for treating advanced melanoma, the most serious kind of skin cancer. Genetically engineer white blood cells to better recognize and destroy cancer cells, then infuse these cells into patients. The subjects in a small initial study were 11 patients whose melanoma had not responded to existing treatments. One question was how rapidly the new cells would multiply after infusion, as measured by the doubling time in days. Here are the doubling times:

1.4 1.0 1.3 1.0 1.3 2.0 0.6 0.8 0.7 0.9 1.9

- Examine the data. Is it reasonable to use the  $t$  procedures?
- Give a 90% confidence interval for the mean doubling time. Are you willing to use this interval to make an inference about the mean doubling time in a population of similar patients?

**17.37** The concentration of carbon dioxide ( $\text{CO}_2$ ) in the atmosphere is increasing rapidly due to our use of fossil fuels. Because plants use  $\text{CO}_2$  to fuel photosynthesis, more  $\text{CO}_2$  may cause trees and other plants to grow faster. An elaborate apparatus allows researchers to pipe extra  $\text{CO}_2$  to a 30-meter circle of forest. They selected two nearby circles in each of three parts of a pine forest and randomly chose one of each pair to receive extra  $\text{CO}_2$ . The response variable is the mean increase in base area for 30 to 40 trees in a circle during a growing season. We measure this in percent increase per year. The following are one year's data.

Pair	Control plot	Treatment plot
1	9.752	10.587
2	7.263	9.244
3	5.742	8.675

- State the null and alternative hypotheses. Explain clearly why the investigators used a one-sided alternative.
- Carry out a test and report your conclusion in simple language.
- The investigators used the test you just carried out. Any use of the  $t$  procedures with samples this size is risky. Why?

**17.39** Velvetleaf is a particularly annoying weed in cornfields. It produces lots of seeds, and the seeds wait in the soil for years until conditions are right. How many seeds do velvetleaf plants produce? Here are counts from 28 plants that came up in a cornfield when no herbicide was used:

2450 2504 2114 1110 2137 8015 1623 1531 2008 1716  
 721 863 1136 2819 1911 2101 1051 218 1711 164  
 2228 363 5973 1050 1961 1809 130 880

We would like to give a confidence interval for the mean number of seeds produced by velvetleaf plants. Alas, the  $t$  interval can't be safely used for these data. Why not?

**17.43** Give a 90% confidence interval for the difference in healing rates (control minus experimental) in the previous exercise.

**17.45** The design of controls and instruments affects how easily people can use them. Timothy Sturm investigated this effect in a course project, asking 25 right-handed students to turn a knob (with their right hands) that moved an indicator by screw action. There were two identical instruments, one with a right-hand thread (the knob turns clockwise) and the other with a left-hand thread (the knob turns counterclockwise). Table 17.5 gives the times in seconds each subject took to move the indicator a fixed distance.

**Table 17.5 Performance times (seconds) using right-hand and left-hand threads**

Subject	Right Thread	Left Thread	Subject	Right Thread	Left Thread
1	113	137	14	107	87
2	105	105	15	118	166
3	130	133	16	103	146
4	101	108	17	111	123
5	138	115	18	104	135
6	118	170	19	111	112
7	87	103	20	89	93
8	116	145	21	78	76
9	75	78	22	100	116
10	96	107	23	89	78
11	122	84	24	85	101
12	103	148	25	88	123
13	116	147			

- (a) Each of the 25 students used both instruments. Explain briefly how you would use randomization in arranging the experiment.
- (b) The project hoped to show that right-handed people find right-hand threads easier to use. Do an analysis that leads to a conclusion about this issue.

**17.47** Give a 90% confidence interval for the mean time advantage of right-hand over left-hand threads in the setting of Exercise 17.45. Do you think that the time saved would be of practical importance if the task were performed many times—for example, by an assembly-line worker? To help answer this question, find the mean time for right-hand threads as a percent of the mean time for left-hand threads.

## Chapter 18 Problem Statements

**18.5** “Conservationists have despaired over destruction of tropical rain forest by logging, clearing, and burning.” These words begin a report on a statistical study of the effects of logging in Borneo. Here are data on the number of tree species in 12 unlogged forest plots and 9 similar plots logged 8 years earlier:

<b>Unlogged</b>	22	18	22	20	15	21	13	13	19	13	19	15
<b>Logged</b>	17	4	18	14	18	15	15	10	12			

- The study report says, “Loggers were unaware that the effects of logging would be assessed.” Why is this important? The study report also explains why the plots can be considered to be randomly assigned.
- Does logging significantly reduce the mean number of species in a plot after 8 years? Follow the four-step process as illustrated in Examples 18.2 and 18.3.

**18.7** Use the data in Exercise 18.5 to give a 90% confidence interval for the difference in mean number of species between unlogged and logged plots.

**18.9** Businesses know that customers often respond to background music. Do they also respond to odors? One study of this question took place in a small pizza restaurant in France on Saturday evenings in May. On one of these evenings, a relaxing lavender odor was spread through the restaurant. Table 18.2 gives the time (minutes) that two samples of 30 customers spent in the restaurant and the amount they spent (in euros). The two evenings were comparable in many ways (weather, customer count, and so on), so we are willing to regard the data as independent SRSs from spring Saturday evenings at this restaurant. The authors say, “Therefore at this stage it would be impossible to generalize the results to other restaurants.”

- Does a lavender odor encourage customers to stay longer in the restaurant? Examine the time data and explain why they are suitable for two-sample  $t$  procedures. Use the two-sample  $t$  test to answer the question posed.
- Does a lavender odor encourage customers to spend more while in the restaurant? Examine the spending data. In what ways do these data deviate from Normality? With 30 observations, the  $t$  procedures are nonetheless reasonably accurate. Use the two-sample  $t$  test to answer the question posed.

Table 18.2 Time (minutes) and spending (Euros) by restaurant customers			
No odor		Lavender	
Minutes	Euros	Minutes	Euros
103	15.9	92	21.9
68	18.5	126	18.5
79	15.9	114	22.3
106	18.5	106	21.9
72	18.5	89	18.5
121	21.9	137	24.9
92	15.9	93	18.5
84	15.9	76	22.5
72	15.9	98	21.5
92	15.9	108	21.9
85	15.9	124	21.5
69	18.5	105	18.5
73	18.5	129	25.5
87	18.5	103	18.5
109	20.5	107	18.5
115	18.5	109	21.9
91	18.5	94	18.5
84	15.9	105	18.5
76	15.9	102	24.9
96	15.9	108	21.9
107	18.5	95	25.9
98	18.5	121	21.9
92	15.9	109	18.5
107	18.5	104	18.5
93	15.9	116	22.8
118	18.5	88	18.5
87	15.9	109	21.9
101	25.5	97	20.7
75	12.9	101	21.9
86	15.9	106	22.5

**18.25** Equip male and female students with a small device that secretly records sound for a random 30 seconds during each 12.5-minute period over two days. Count the words each subject speaks during each recording period, and from this, estimate how many words per day each subject speaks. The published report includes a table summarizing six such studies. Here are two of the six:

Study	Sample Size		Estimated Average Number (SD) Of Words Spoken per Day	
	Women	Men	Women	Men
1	56	56	16,177 (7520)	16,569 (9108)
2	27	20	16,496 (7914)	12,867 (8343)

Readers are supposed to understand that, for example, the 56 women in the first study had  $\bar{x} = 16,177$  and  $s = 7520$ . It is commonly thought that women talk more than men. Does either of the two samples support this idea? For each study:

- State hypotheses in terms of the population means for men ( $\mu_M$ ) and women ( $\mu_F$ ).
- Find the two-sample  $t$  statistic.
- What degrees of freedom does Option 2 use to get a conservative  $P$ -value?
- Compare your value of  $t$  with the critical values in Table C. What can you say about the  $P$ -value of the test?
- What do you conclude from the results of these two studies?

**18.27** In a study of the presence of whelks along the Pacific coast, investigators put down a frame that covers 0.25 square meter and counted the whelks on the sea bottom inside the frame. They did this at 7 locations in California and 6 locations in Oregon. The report says that whelk densities “were twice as high in Oregon as in California (mean  $\pm$  SEM,  $26.9 \pm 1.56$  versus  $11.9 \pm 2.68$  whelks per 0.25 m<sup>2</sup>, Oregon versus California, respectively; Student's  $t$  test,  $P < 0.001$ ).”

- SEM stands for the standard error of the mean,  $s/\sqrt{n}$ . Fill in the values in this summary table:

Group	Location	$n$	$\bar{x}$	$s$
1	Oregon	?	?	?
2	California	?	?	?

- What degrees of freedom would you use in the conservative two-sample  $t$  procedures to compare Oregon and California?
- What is the two-sample  $t$  test statistic for comparing the mean densities of whelks in Oregon and California?
- Test the null hypothesis of no difference between the two population means against the two-sided alternative. Use your statistic from part (c) with degrees of freedom from part (b). Does your conclusion agree with the published report?

**18.29** The post-lunch dip is the drop in mental alertness after a midday meal. Does an extract of the leaves of the ginkgo tree reduce the post-lunch dip? Assign healthy people aged 18 to 40 to take either ginkgo extract or a placebo pill. After lunch, ask them to read

seven pages of random letters and place an X over every e. Count the number of misses per line read.

- (a) What is a placebo and why was one group given a placebo?  
 (b) What is the double-blind method and why should it be used in this experiment?  
 (c) Here are summaries of performance after 13 weeks of either ginkgo extract or placebo:

<b>Group</b>	<b>Group size</b>	<b>Mean</b>	<b>Std. dev.</b>
<b>Ginkgo</b>	21	0.06383	0.01462
<b>Placebo</b>	18	0.05342	0.01549

Is there a significant difference between the two groups? What do these data show about the effect of ginkgo extract?

**18.31** The SAFE (Social, Attitudinal, Familial, and Environmental Stress) scale measures the stress level of adults adjusting to a different culture. Scores range from 1 (not stressful) to 5 (extremely stressful). In a study of stress among immigrant mothers in a university community, mothers of children between 2 and 10 years of age whose families had come to the United States for professional or academic reasons took the SAFE questionnaire. Here are summaries for mothers from Asia and Europe:

<b>Origin</b>	<b>Sample size</b>	<b>Mean</b>	<b>Std. dev</b>
<b>Asian</b>	12	1.92	0.60
<b>European</b>	9	1.74	0.57

Is there evidence of a difference in mean stress levels between mothers from Asia and Europe?

**18.33** What we really want to know is whether coached students improve more than uncoached students, and whether any advantage is large enough to be worth paying for. Use the information in the previous exercise to answer these questions:

- (a) Is there good evidence that coached students gained more on the average than uncoached students?  
 (b) How much more do coached students gain on the average? Give a 99% confidence interval.  
 (c) Based on your work, what is your opinion: do you think coaching courses are worth paying for?

**18.35** Here are the IQ test scores of 31 seventh-grade girls in a Midwest school district.

114 100 104 89 102 91 114 114 103 105  
 108 130 120 132 111 128 118 119 86 72

111 103 74 112 107 103 98 96 112 112 93

The IQ test scores of 47 seventh-grade boys in the same district are

111 107 100 107 115 111 97 112 104 106 113  
 109 113 128 128 118 113 124 127 136 106 123  
 124 126 116 127 119 97 102 110 120 103 115  
 93 123 79 119 110 110 107 105 105 110 77  
 90 114 106

- (a) Make stemplots or histograms of both sets of data. Because the distributions are reasonably symmetric with no extreme outliers, the  $t$  procedures will work well.  
 (b) Treat these data as SRSs from all seventh-grade students in the district. Is there good evidence that girls and boys differ in their mean IQ scores?

**18.37** Use the data in Exercise 18.35 to give a 95% confidence interval for the difference between the mean IQ scores of all boys and all girls in the district.

**18.39** Of course, the reason for durable press treatment is to reduce wrinkling. "Wrinkle recovery angle" measures how well a fabric recovers from wrinkles. Higher is better. Here are data on the wrinkle recovery angle (in degrees) for the same fabric swatches discussed in the previous exercise:

<b>Permafresh</b>	136	135	132	137	134
<b>Hylite</b>	143	141	146	141	145

Is there a significant difference in wrinkle resistance?

- (a) Do the sample means suggest that one process has better wrinkle resistance?  
 (b) Make stemplots for both samples. There are no obvious deviations from Normality.  
 (c) Test the hypothesis  $H_0: \mu_1 = \mu_2$  against the two-sided alternative. What do you conclude from part (a) and from the result of your test?

**18.41** In Exercise 18.39, you found that the Hylite process results in significantly greater wrinkle resistance than the Permafresh process. How large is the difference in mean wrinkle recovery angle? Give a 90% confidence interval.

**18.45** Kathleen Vohs of the University of Minnesota and her coworkers carried out several randomized comparative experiments on the effects of thinking about money. Here's part of one such experiment. Ask student subjects to unscramble 30 sets of five words to make a meaningful phrase from four of the five words. The control group unscrambled phrases like "cold it desk outside is" into "it is cold outside." The treatment

group unscrambled phrases that lead to thinking about money, turning “high a salary desk paying” into “a high-paying salary.” Then each subject worked a hard puzzle, knowing that he or she could ask for help. Here are the times in seconds until subjects asked for help, for the treatment group,

609 444 242 199 174 55 251 466 443  
531 135 241 476 482 362 69 160

and for the control group,

118 272 413 291 140 104 55 189 126  
400 92 64 88 142 141 373 156

The researchers suspected that money is connected with self-sufficiency, so that the treatment group will ask for help less quickly on the average. Do the data support this idea?

**18.47** A “subliminal” message is below our threshold of awareness but may nonetheless influence us. Can subliminal messages help students learn math? A group of students who had failed the mathematics part of the City University of New York Skills Assessment Test agreed to participate in a study to find out.

All received a daily subliminal message, flashed on a screen too rapidly to be consciously read. The treatment group of 10 students (chosen at random) was exposed to “Each day I am getting better in math.” The control group of 8 students was exposed to a neutral message, “People are walking on the street.” All students participated in a summer program designed to raise their math skills, and all took the assessment test again at the end of the program. Table 18.3 gives data on the subjects' scores before and after the program. Is there good evidence that the treatment brought about a greater improvement in math scores than the neutral message? How large is the mean difference in gains between treatment and control? (Use 90% confidence.)

<b>Treatment group</b>		<b>Control group</b>	
<b>Before</b>	<b>After</b>	<b>Before</b>	<b>After</b>
18	24	18	29
18	25	24	29
21	33	20	24
18	29	18	26
18	33	24	38
20	36	22	27
23	34	15	22
23	36	19	31
21	34		
17	27		

## Chapter 19 Problem Statements

**19.5** Canada has much stronger gun control laws than the United States, and Canadians support gun control more strongly than do Americans. A sample survey asked a random sample of 1505 adult Canadians, “Do you agree or disagree that all firearms should be registered?” Of the 1505 people in the sample, 1288 answered either “Agree strongly” or “agree somewhat”

- The survey dialed residential telephone numbers at random in all ten Canadian provinces (omitting the sparsely populated northern territories). Based on what you know about sample surveys, what is likely to be the biggest weakness in this survey?
- Nonetheless, act as if we have an SRS from adults in the Canadian provinces. Give a 95% confidence interval for the proportion who support registration of all firearms.

**19.7** Sample surveys usually contact large samples, so we can use the large-sample confidence interval if the sample design is close to an SRS. Scientific studies often use small samples that require the plus four method. For example, the small round holes you often see in sea shells were drilled by other sea creatures, who ate the former owners of the shells. Whelks often drill into mussels, but this behavior appears to be more or less common in different locations. Investigators collected whelk eggs from the coast of Oregon, raised the whelks in the laboratory, and then put each whelk in a container with some delicious mussels. Only 9 of 98 whelks drilled into mussels.

- Why can't we use the large-sample confidence interval for the proportion  $p$  of Oregon whelks that will spontaneously drill mussels?
- The plus four method adds four observations, two successes and two failures. What are the sample size and the number of successes after you do this? What is the plus four estimate  $\tilde{p}$  of  $p$ ?
- Give the plus four 90% confidence interval for the proportion of Oregon whelks that will spontaneously drill mussels.

**19.9** The plus four method is particularly useful when there are *no* successes or *no* failures in the data. The study of Spanish currency described in Example 19.5 found that in Seville, all 20 of a sample of 20 euro bills had cocaine traces.

- What is the sample proportion  $\hat{p}$  of contaminated bills? What is the large-sample 95% confidence interval for  $p$ ? It's not plausible that *every* bill in Seville has cocaine traces, as this interval says.
- Find the plus four estimate  $\tilde{p}$  and the plus four 95% confidence interval for  $p$ . These results are more reasonable.

**19.11** PTC is a substance that has a strong bitter taste for some people and is tasteless for others. The ability to taste PTC is inherited. About 75% of Italians can taste PTC, for

example. You want to estimate the proportion of Americans with at least one Italian grandparent who can taste PTC. Starting with the 75% estimate for Italians, how large a sample must you collect in order to estimate the proportion of PTC tasters within  $\pm 0.04$  with 90% confidence?

**19.13** We often judge other people by their faces. It appears that some people judge candidates for elected office by their faces. Psychologists showed head-and-shoulders photos of the two main candidates in 32 races for the U.S. Senate to many subjects (dropping subjects who recognized one of the candidates) to see which candidate was rated “more competent” based on nothing but the photos. On election day, the candidates whose faces looked more competent won 22 of the 32 contests. If faces don't influence voting, half of all races in the long run should be won by the candidate with the better face. Is there evidence that the candidate with the better face wins more than half the time? Follow the four-step process as illustrated in Example 19.7.

**19.25** The Harris Poll asked a sample of smokers, “Do you believe that smoking will probably shorten your life, or not?” Of the 1010 people in the sample, 848 said “Yes.”

- (a) Harris called residential telephone numbers at random in an attempt to contact an SRS of smokers. Based on what you know about national sample surveys, what is likely to be the biggest weakness in the survey?
- (b) We will nonetheless act as if the people interviewed are an SRS of smokers. Give a 95% confidence interval for the percent of smokers who agree that smoking will probably shorten their lives.

**19.29** The Pew Research Center asked a random sample of 1128 adult women, “How satisfied are you with your life overall?” Of these women, 56 said either “Mostly dissatisfied” or “Very dissatisfied.”

- (a) Pew dialed residential telephone numbers at random in the continental United States in an attempt to contact a random sample of adults. Based on what you know about national sample surveys, what is likely to be the biggest weakness in the survey?
- (b) Act as if the sample is an SRS. Give a large-sample 90% confidence interval for the proportion  $p$  of all adult women who are mostly or very dissatisfied with their lives.
- (c) Give the plus four confidence interval for  $p$ . If you express the two confidence intervals in percents and round to the nearest tenth of a percent, how do they differ? (As always, the plus four method pulls results away from 0% or 100%, whichever is closer. Although the condition for the large-sample interval is met, we can place more trust in the plus four interval.)

**19.31** Most soybeans grown in the United States are genetically modified to, for example, resist pests and so reduce use of pesticides. Because some nations do not accept

genetically modified (GM) foods, grain-handling facilities routinely test soybean shipments for the presence of GM beans. In a study of the accuracy of these tests, researchers submitted shipments of soybeans containing 1% of GM beans to 23 randomly selected facilities. Eighteen detected the GM beans.

- (a) Show that the conditions for the large-sample confidence interval are not met. Show that the conditions for the plus four interval are met.
- (b) Use the plus four method to give a 90% confidence interval for the percent of all grain-handling facilities that will correctly detect 1% of GM beans in a shipment.

**19.37** Some shrubs have the useful ability to resprout from their roots after their tops are destroyed. Fire is a particular threat to shrubs in dry climates, as it can injure the roots as well as destroy the aboveground material. One study of resprouting took place in a dry area of Mexico. The investigators clipped the tops of samples of several species of shrubs. In some cases, they also applied a propane torch to the stumps to simulate a fire. Of 12 specimens of the shrub *Krameria cytisoides*, 5 resprouted after fire. Estimate with 90% confidence the proportion of all shrubs of this species that will resprout after fire.

**19.39** A sample survey funded by the National Science Foundation asked a random sample of American adults about biological evolution. One question asked subjects to answer "True," "False," or "Not sure" to the statement "Human beings, as we know them today, developed from earlier species of animals." Of the 1484 respondents, 594 said "True." What can you say with 95% confidence about the percent of all American adults who think that humans developed from earlier species of animals?

**19.41** Does the sample in Exercise 19.39 give good evidence to support the claim "Fewer than half of American adults think that humans developed from earlier species of animals"?

**Chapter 20 Problem Statements**

**20.1** Younger people use online instant messaging (IM) more often than older people. A random sample of IM users found that 73 of the 158 people in the sample aged 18 to 27 said they used IM more often than email. In the 28 to 39 age group, 26 of 143 people used IM more often than email. Give a 95% confidence interval for the difference between the proportions of IM users in these age groups who use IM more often than email. Follow the four-step process as illustrated in Examples 20.1 and 20.2.

**20.3** A government survey randomly selected 6889 female high school students and 7028 male high school students. Of these students, 1915 females and 3078 males met recommended levels of physical activity. (These levels are quite high: at least 60 minutes of activity that makes you breathe hard on at least 5 of the past 7 days.) Give a 99% confidence interval for the difference between the proportions of all female and male high school students who meet the recommended levels of activity.

**20.5** We don't like to find broken crackers when we open the package. How can makers reduce breaking? One idea is to microwave the crackers for 30 seconds right after baking them. Breaks start as hairline cracks called "checking." Assign 65 newly baked crackers to the microwave and another 65 to a control group that is not microwaved. After one day, none of the microwave group and 16 of the control group show checking. Give the 95% plus four confidence interval for the amount by which microwaving reduces the proportion of checking. The plus four method is particularly helpful when, as here, a count of successes is zero. Follow the four-step process as illustrated in Example 20.3.

**20.7** Most alpine skiers and snowboarders do not use helmets. Do helmets reduce the risk of head injuries? A study in Norway compared skiers and snowboarders who suffered head injuries with a control group who were not injured. Of 578 injured subjects, 96 had worn a helmet. Of the 2992 in the control group, 656 wore helmets. Is helmet use less common among skiers and snowboarders who have head injuries? Follow the four-step process as illustrated in Example 20.5. (Note that this is an observational study that compares injured and uninjured subjects. An experiment that assigned subjects to helmet and no-helmet groups would be more convincing.)

**20.17** Many teens have posted profiles on sites such as MySpace. A sample survey asked random samples of teens with online profiles if they included false information in their profiles. Of 170 younger teens (ages 12 to 14), 117 said "Yes." Of 317 older teens (ages 15 to 17), 152 said "Yes."

(a) Do these samples satisfy the guidelines for the large-sample confidence interval?

- (b) Give a 95% confidence interval for the difference between the proportions of younger and older teens who include false information in their online profiles.

**20.19** Genetic influences on cancer can be studied by manipulating the genetic makeup of mice. One of the processes that turn genes on or off (so to speak) in particular locations is called “DNA methylation.” Do low levels of this process help cause tumors? Compare mice altered to have low levels with normal mice. Of 33 mice with lowered levels of DNA methylation, 23 developed tumors. None of the control group of 18 normal mice developed tumors in the same time period.

- (a) Explain why we cannot safely use either the large-sample confidence interval or the test for comparing the proportions of normal and altered mice that develop tumors.
- (b) The plus four method adds two observations, a success and a failure, to each sample. What are the sample sizes and the numbers of mice with tumors after you do this? Give a plus four 99% confidence interval for the difference in the proportions of the two populations that develop tumors.
- (c) Based on your confidence interval, is the difference between normal and altered mice significant at the 1% level?

**20.21** Is there a significant difference in the proportions of papers with and without statistical help that are rejected without review? State hypotheses, find the test statistic, use software or the bottom row of Table C to get a  $P$ -value, and give your conclusion. (This observational study does not establish causation, because studies that include statistical help may also be better in other ways than those that do not.)

**20.23** Give a 95% confidence interval for the difference between the proportions of papers rejected without review when a statistician is and is not involved in the research.

**20.27** The North Carolina State University study in the previous exercise also looked at possible differences in the proportions of female and male students who succeeded in the course. They found that 23 of the 34 women and 60 of the 89 men succeeded. Is there evidence of a difference between the proportions of women and men who succeed?

**20.29** Nicotine patches are often used to help smokers quit. Does giving medicine to fight depression help? A randomized double-blind experiment assigned 244 smokers who wanted to stop to receive nicotine patches and another 245 to receive both a patch and the antidepressant drug bupropion. After a year, 40 subjects in the nicotine patch group and 87 in the patch-plus-drug group had abstained from smoking. Give a 99% confidence interval for the difference (treatment minus control) in the proportion of smokers who quit.

**20.31** Do our emotions influence economic decisions? One way to examine the issue is to have subjects play an “ultimatum game” against other people and against a computer. Your partner (person or computer) gets \$10, on the condition that it be shared with you. The partner makes you an offer. If you refuse, neither of you gets anything. So it's to your advantage to accept even the unfair offer of \$2 out of the \$10. Some people get mad and refuse unfair offers. Here are data on the responses of 76 subjects randomly assigned to receive an offer of \$2 from either a person they were introduced to or a computer:

	<b>Accept</b>	<b>Reject</b>
<b>Human offers</b>	20	18
<b>Computer offers</b>	32	6

We suspect that emotion will lead to offers from another person being rejected more often than offers from an impersonal computer. Do a test to assess the evidence for this conjecture.

**20.35** Are shoppers more or less likely to use credit cards for “impulse purchases” that they decide to make on the spot, as opposed to purchases that they had in mind when they went to the store? Stop every third person leaving a department store with a purchase. (This is in effect a random sample of people who buy at that store.) A few questions allow us to classify the purchase as impulse or not. Here are the data on how the customer paid:

	<b>Credit card?</b>	
	<b>Yes</b>	<b>No</b>
<b>Impulse purchases</b>	13	18
<b>Planned purchases</b>	35	31

Estimate with 95% confidence the percent of all customers at this store who use a credit card. Give numerical summaries to describe the difference in credit card use between impulse and planned purchases. Is this difference statistically significant?

## Chapter 21 Problem Statements

**21.1** A sample survey of 1497 adult Internet users found that 36% consult the online collaborative encyclopedia Wikipedia. Give a 95% confidence interval for the proportion of all adult Internet users who refer to Wikipedia.

**21.3** When the new euro coins were introduced throughout Europe in 2002, curious people tried all sorts of things. Two Polish mathematicians spun a Belgian euro (one side of the coin has a different design for each country) 250 times. They got 140 heads. Newspapers reported this result widely. Is it significant evidence that the coin is not balanced when spun?

**21.5** Ask young men to estimate their own degree of body muscle by choosing from a set of 100 photos. Then ask them to choose what they think women prefer. The researchers know the actual degree of muscle, measured as kilograms per square meter of fat-free mass, for each of the photos. They can therefore measure the difference between what a subject thinks women prefer and the subject's own self-image. Call this difference the "muscle gap." Here are summary statistics for the muscle gap from two samples, one of American and European young men and the other of Chinese young men from Taiwan:

<b>Group</b>	<b><i>n</i></b>	<b><math>\bar{x}</math></b>	<b><i>s</i></b>
<b>American/European</b>	200	2.35	2.5
<b>Chinese</b>	55	1.20	3.2

Give a 95% confidence interval for the mean size of the muscle gap for all American and European young men. On the average, men think they need this much more muscle to match what women prefer.

**21.7** Here's how butterflies mate: a male passes to a female a packet of sperm called a spermatophore. Females may mate several times. Will they remate sooner if the first spermatophore they receive is small? Among 20 females who received a large spermatophore (greater than 25 milligrams), the mean time to the next mating was 5.15 days, with standard deviation 0.18 day. For 21 females who received a small spermatophore (about 7 milligrams), the mean was 4.33 days and the standard deviation was 0.31 day. Is the observed difference in means statistically significant?

**21.9** Give a 90% confidence interval for the difference between the proportions of all Hispanic and all white young people who listen to rap every day.

**21.11** A study of the inheritance of speed and endurance in mice found a trade-off between these two characteristics, both of which help mice survive. To test endurance, mice were made to swim in a bucket with a weight attached to their tails. (The mice were rescued when exhausted.) Here are data on endurance in minutes for female and male mice:

<b>Group</b>	<i>n</i>	Mean	Standard Deviation
<b>Female</b>	162	11.4	26.09
<b>Male</b>	135	6.7	6.69

- (a) Both sets of endurance data are skewed to the right. Why are  $t$  procedures nonetheless reasonably accurate for these data?
- (b) Do the data show that female mice have significantly higher endurance on the average than male mice?

**21.13** Use the information in Exercise 21.11 to give a 95% confidence interval for the mean difference (female minus male) in endurance times.

**21.15** Use the information in the previous exercise to give a 99% confidence interval for the proportion of all students in 2004 who had at least one parent who graduated from college. (The sample excludes 17-year-olds who had dropped out of school, so your estimate is valid for students but is probably too high for all 17-year-olds.)

**21.19** Starting in the 1970s, medical technology allowed babies with very low birth weight (VLBW, less than 1500 grams, about 3.3 pounds) to survive without major handicaps. It was noticed that these children nonetheless had difficulties in school and as adults. A long-term study has followed 242 VLBW babies to age 20 years, along with a control group of 233 babies from the same population who had normal birth weight.

- (a) Is this an experiment or an observational study? Why?
- (b) At age 20, 179 of the VLBW group and 193 of the control group had graduated from high school. Is the graduation rate among the VLBW group significantly lower than for the normal-birth-weight controls?

**21.21** Of the 126 women in the VLBW group, 37 said they had used illegal drugs; 52 of the 124 control group women had done so. The IQ scores for the VLBW women had mean 86.2 (standard deviation 13.4), and the normal-birth-weight controls had mean IQ 89.8 (standard deviation 14.0). Is there a statistically significant difference between the two groups in either proportion using drugs or mean IQ?

**21.23** The Women's Health Initiative is a randomized, controlled clinical trial designed to see if a low-fat diet reduces the incidence of breast cancer. In all, 19,541 women were

assigned at random to a low-fat diet and a control group of 29,294 women were assigned to a normal diet. All the subjects were between ages 50 and 79 and had no prior breast cancer. After 8 years, 655 of the women in the low-fat group and 1072 of the women in the control group had developed breast cancer. Does this clinical trial give evidence that a low-fat diet reduces breast cancer?

**21.25** High levels of cholesterol in the blood are not healthy in either humans or dogs. Because a diet rich in saturated fats raises the cholesterol level, it is plausible that dogs owned as pets have higher cholesterol levels than dogs owned by a veterinary research clinic. “Normal” levels of cholesterol based on the clinic's dogs would then be misleading. A clinic compared healthy dogs it owned with healthy pets brought to the clinic to be neutered. The summary statistics for blood cholesterol levels (milligrams per deciliter of blood) appear below.

<b>Group</b>	<b><math>n</math></b>	<b><math>\bar{x}</math></b>	<b><math>s</math></b>
<b>Pets</b>	26	193	68
<b>Clinic</b>	23	174	44

Is there strong evidence that pets have a higher mean cholesterol level than clinic dogs?

**21.27** Continue your work with the information in Exercise 21.25. Give a 95% confidence interval for the mean cholesterol level in pets.

**21.39** Dogs are big and expensive. Rats are small and cheap. Might rats be trained to replace dogs in sniffing out illegal drugs? A first study of this idea trained rats to rear up on their hind legs when they smelled simulated cocaine. To see how well rats performed after training, they were let loose on a surface with many cups sunk in it, one of which contained simulated cocaine. Four out of six trained rats succeeded in 80 out of 80 trials. How should we estimate the long-term success rate  $p$  of a rat that succeeds in every one of 80 trials?

- What is the rat's sample proportion  $\hat{p}$ ? What is the large-sample 95% confidence interval for  $p$ ? It's not plausible that the rat will *always* be successful, as this interval says.
- Find the plus four estimate  $\tilde{p}$  and the plus four 95 % confidence interval for  $p$ . These results are more reasonable.

**21.41** At what age do infants speak their first word of English? Here are data on 20 children (ages in months):

15 26 10 9 15 20 18 11 8 20  
7 9 10 11 11 10 12 17 11 10

(In fact, the sample contained one more child, who began to speak at 42 months. Child development experts consider this abnormally late, so we dropped the outlier to get a sample of "normal" children. The investigators are willing to treat these data as an SRS.) Is there good evidence that the mean age at first word among all normal children is greater than one year?

**21.45** The color of a fabric depends on the dye used and also on how the dye is applied. This matters to clothing manufacturers, who want the color of the fabric to be just right. The study discussed in the previous exercise went on to dye fabric made of ramie with the same "procion blue" dye applied in two different ways. Here are the lightness scores for 8 pieces of identical fabric dyed in each way:

<b>Method B</b>	40.98	40.88	41.30	41.28	41.66	41.50	41.39	41.27
<b>Method C</b>	42.30	42.20	42.65	42.43	42.50	42.28	43.13	42.45

- (a) This is a randomized comparative experiment. Outline the design.  
 (b) A clothing manufacturer wants to know which method gives the darker color (lower lightness score). Use sample means to answer this question. Is the difference between the two sample means statistically significant? Can you tell from just the  $P$ -value whether the difference is large enough to be important in practice?

**21.47** We wonder what proportion of female students have at least one parent who allows them to drink around him or her. Table 21.1 contains information about a sample of 94 students. Use this sample to give a 95% confidence interval for this proportion.

**21.49** We don't like to find broken crackers when we open the package. How can makers reduce breaking? One idea is to microwave the crackers for 30 seconds right after baking them. Analyze the following results from two experiments intended to examine this idea. Does microwaving significantly improve indicators of future breaking? How large is the improvement? What do you conclude about the idea of microwaving crackers?

- (a) The experimenter randomly assigned 65 newly baked crackers to be microwaved and another 65 to a control group that is not microwaved. Fourteen days after baking, 3 of the 65 microwaved crackers and 57 of the 65 crackers in the control group showed visible checking, which is the starting point for breaks.

(b) The experimenter randomly assigned 20 crackers to be microwaved and another 20 to a control group. After 14 days, he broke the crackers. Here are summaries of the pressure needed to break them, in pounds per square inch:

	<b>Microwave</b>	<b>Control</b>
Mean	139.6	77.0
Standard deviation	33.6	22.6

## Chapter 22 Problem Statements

**22.1** The Pennsylvania State University has its main campus in University Park and more than 20 smaller “commonwealth campuses” around the state. The Penn State Division of Student Affairs polled a random sample of undergraduates about their use of online social networking. (The response rate was only about 20%, which casts some doubt on the usefulness of the data.) Facebook was the most popular site, with more than 80% of students having an account. Here is a comparison of Facebook use by undergraduates at the University Park and commonwealth campuses:

	University Park	Commonwealth
Do not use Facebook	68	248
Several times a month or less	55	76
At least once a week	215	157
At least once a day	640	394

- (a) What percent of University Park students fall in each Facebook category? What percent of commonwealth campus students fall in each category? Each column should add to 100% (up to roundoff error). These are the conditional distributions of Facebook use given campus setting.
- (b) Make a bar graph that compares the two conditional distributions. What are the most important differences in Facebook use between the two campus settings?

**22.3** In the setting of Exercise 22.1, we might do several significance tests to compare University Park with the commonwealth campuses.

- (a) Is there a significant difference between the proportions of students in the two locations who do not use Facebook? Give the  $P$ -value.
- (b) Is there a significant difference between the proportions of students in the two locations who are in the “At least once a week” category? Give the  $P$ -value.
- (c) Explain clearly why  $P$ -values for individual outcomes like these can't tell us whether the two distributions for all four outcomes in the two locations differ significantly.

**22.5** The two-way table in Exercise 22.1 displays data on use of Facebook by two groups of Penn State students. It's clear that nonusers are much more frequent at the commonwealth campuses. Let's look just at students who have Facebook accounts:

Use Facebook	University Park	Commonwealth
Several times a month or less	55	76
At least once a week	215	157
At least once a day	640	394
Total Facebook Users	910	627

The null hypothesis is that there is no relationship between campus and Facebook use.

- (a) If this hypothesis is true, what are the expected counts for Facebook use among commonwealth campus students? This is one column of the two-way table of expected counts. Find the column total and verify that it agrees with the column total for the observed counts.
- (b) Commonwealth campus students as a group are older and more likely to be married and employed than University Park students. What does comparing the observed and expected counts in this column show about Facebook use by these students?

**22.13** Many birds are injured or killed by flying into windows. It appears that birds don't see windows. Can tilting windows down so that they reflect earth rather than sky reduce bird strikes? Place six windows at the edge of a woods: two vertical, two tilted 20 degrees, and two tilted 40 degrees. During the next four months, there were 53 bird strikes, 31 on the vertical windows, 14 on the 20-degree windows, and 8 on the 40-degree windows. If the tilt has no effect, we expect strikes on windows with all three tilts to have equal probability. Test this null hypothesis. What do you conclude?

**22.15** Police may use minor violations such as not wearing a seat belt to stop motorists for other reasons. A large study in Michigan first studied the population of drivers not wearing seat belts during daylight hours by observation at more than 400 locations around the state. Here is the population distribution of seat belt violators by age group:

Age group	16 to 29	30 to 59	60 or older
Proportion	0.328	0.594	0.078

The researchers then looked at court records and called a random sample of 803 drivers who had actually been cited by police for not wearing a seat belt. Here are the counts:

Age group	16 to 29	30 to 59	60 or older
Count	401	382	20

Does the age distribution of people cited differ significantly from the distribution of ages of all seat belt violators? Which age groups have the largest contributions to chi-square? Are these age groups cited more or less frequently than is justified? (The study found that males, blacks, and younger drivers were all over-cited.)

**22.17** For reasons known only to social scientists, the General Social Survey (GSS) regularly asks its subjects their astrological sign. Here are the counts of responses for the most recent GSS:

Sign	Aries	Taurus	Gemini	Cancer	Leo	Virgo
Count	321	360	367	374	383	402

Sign	Libra	Scorpio	Sagittarius	Capricorn	Aquarius	Pisces
Count	392	329	331	354	376	355

If births are spread uniformly across the year, we expect all 12 signs to be equally likely. Are they? Follow the four-step process in your answer.

**22.29** The General Social Survey (GSS) asked this question: “Consider a person who believes that Blacks are genetically inferior. If such a person wanted to make a speech in your community claiming that Blacks are inferior, should he be allowed to speak, or not?” Here are the responses, broken down by the race of the respondent:

	Black	White	Other
Allowed	140	976	121
Not allowed	129	480	131

- Because the GSS is essentially an SRS of all adults, we can combine the races in these data and give a 99% confidence interval for the proportion of all adults who would allow a racist to speak. Do this.
- Find the column percents and use them to compare the attitudes of the three racial groups. How significant are the differences found in the sample?

**22.43** The nonprofit group Public Agenda conducted telephone interviews with a stratified sample of parents of high school children. There were 202 black parents, 202 Hispanic parents, and 201 white parents. One question asked was “Are the high schools in your state doing an excellent, good, fair or poor job, or don't you know enough to say?” Here are the survey results:

	Black Parents	Hispanic Parents	White Parents
Excellent	12	34	22
Good	69	55	81
Fair	75	61	60
Poor	24	24	24
Don't know	22	28	14
Total	202	202	201

Are the differences in the distributions of responses for the three groups of parents statistically significant? What departures from the null hypothesis “no relationship between group and response” contribute most to the value of the chi-square statistic? Write a brief conclusion based on your analysis.

**22.45** Before bringing a new product to market, firms carry out extensive studies to learn how consumers react to the product and how best to advertise its advantages. Here are data from a study of a new laundry detergent. The subjects are people who don't currently use the established brand that the new product will compete with. Give subjects free samples of both detergents. After they have tried both for a while, ask which they prefer. The answers may depend on other facts about how people do laundry.

	Laundry Practices			
	Soft water, warm wash	Soft water, Hot wash	Hard water, Warm wash	Hard water, Hot wash
Prefer standard product	53	27	42	30
Prefer new product	63	29	68	42

How do laundry practices (water hardness and wash temperature) influence the choice of detergent? In which settings does the new detergent do best? Are the differences between the detergents statistically significant?

**22.47** Make a  $2 \times 5$  table by combining the counts in the three rows that mention Democrat and in the three rows that mention Republican and ignoring strict independents and supporters of other parties. We might think of this table as comparing all adults who lean Democrat and all adults who lean Republican. How does support for the two major parties differ among adults with different levels of education?

### Chapter 23 Problem Statements

**23.1** An outbreak of the deadly Ebola virus in 2002 and 2003 killed 91 of the 95 gorillas in 7 home ranges in the Congo. To study the spread of the virus, measure “distance” by the number of home ranges separating a group of gorillas from the first group infected. Here are data on distance and number of days until deaths began in each later group:

<b>Distance <math>x</math></b>	1	3	4	4	4	5
<b>Days <math>y</math></b>	4	21	33	41	43	46

- Examine the data. Make a scatterplot with distance as the explanatory variable and find the correlation. There is a strong linear relationship.
- Explain in words what the slope  $\beta$  of the population regression line would tell us if we knew it. Based on the data, what are the estimates of  $\beta$  and the intercept  $\alpha$  of the population regression line?
- Calculate by hand the residuals for the six data points. Check that their sum is 0 (up to roundoff error). Use the residuals to estimate the standard deviation  $\sigma$  that measures variation in the responses (days) about the means given by the population regression line. You have now estimated all three parameters.

**23.3** One effect of global warming is to increase the flow of water into the Arctic Ocean from rivers. Such an increase may have major effects on the world's climate. Six rivers (Yenisey, Lena, Ob, Pechora, Kolyma, and Severnaya Dvina) drain two-thirds of the Arctic in Europe and Asia. Several of these are among the largest rivers on earth. Table 23.2 presents the total discharge from these rivers each year from 1936 to 1999. Discharge is measured in cubic kilometers of water. Use software to analyze these data.

- Make a scatterplot of river discharge against time. Is there a clear increasing trend? Calculate  $r^2$  and briefly interpret its value. There is considerable year-to-year variation, so we wonder if the trend is statistically significant.
- As a first step, find the least-squares line and draw it on your plot. Then find the regression standard error  $s$ , which measures scatter about this line. We will continue the analysis in later exercises.

**23.5** The most important question we ask of the data in Table 23.2 is this: is the increasing trend visible in your plot (Exercise 23.3) statistically significant? If so, changes in the Arctic may already be affecting the earth's climate. Use software to answer this question. Give a test statistic, its  $P$ -value, and the conclusion you draw from the test.

**23.7** Exercise 23.1 gives data showing that the delay in deaths from an Ebola outbreak in groups of gorillas increases linearly with distance from the origin of the outbreak. There are only 6 observations, so we worry that the apparent relationship may be just chance. Is the correlation significantly greater than 0? Answer this question in two ways.

- Return to your  $t$  statistic from Exercise 23.4. What is the one-sided  $P$ -value for this  $t$ ? Apply your result to test the correlation.
- Find the correlation  $r$  and use Table E to approximate the  $P$ -value of the one-sided test.

**23.9** Exercise 23.1 presents data on distance and days until an Ebola outbreak reached six groups of gorillas. Software tells us that the least-squares slope is  $b = 11.263$  with standard error  $SE_b = 1.591$ . Because there are only 6 observations, the observed slope  $b$  may not be an accurate estimate of the population slope  $\beta$ . Give a 90% confidence interval for  $\beta$ .

**23.11** Use the data in Table 23.2 to give a 90% confidence interval for the slope of the population regression of Arctic river discharge on year. Does this interval convince you that discharge is actually increasing over time? Explain your answer.

**23.29** We know that there is a strong linear relationship. Let's check the other conditions for inference. Figure 23.14 includes a table of the two variables, the predicted values  $\hat{y}$  for each  $x$  in the data, the residuals, and related quantities. (This table is stored as *ex23-29.dat* on the text CD and Web site.)

- Round the residuals to the nearest whole number and make a stemplot. The distribution is single-peaked and symmetric and appears close to Normal.
- Make a residual plot, residuals against boats registered. Use a vertical scale from  $-25$  to  $25$  to show the pattern more clearly. Add the “residual = 0” line. There is no clearly nonlinear pattern. The spread about the line may be a bit greater for larger values of the explanatory variable, but the effect is not large.
- It is reasonable to regard the number of manatees killed by boats in successive years as independent. The number of boats grew over time. Someone says that pollution also grew over time and may explain more manatee deaths. How would you respond to this idea?

**23.33** Exercise 5.53 (page 158) gives data on William Gray's predictions of the number of named tropical storms in Atlantic hurricane seasons from 1984 to 2007. Use these data for regression inference as follows.

- (a) Does Professor Gray do better than random guessing? That is, is there a significantly positive correlation between his forecasts and the actual number of storms? (Report a  $t$  statistic from regression output and give the one-sided  $P$ -value.)
- (b) Give a 95% confidence interval for the mean number of storms in years when Professor Gray forecasts 16 storms.

**23.41** Exercise 7.25 (page 188) gives data on the abundance of the pine cones that red squirrels feed on and the mean number of offspring per female squirrel over 16 years. The strength of the relationship is remarkable because females produce young before the food is available. How significant is the evidence that more cones leads to more offspring? (Use a vertical scale from  $-2$  to  $2$  in your residual plot to show the pattern more clearly.)

**23.43** Exercise 5.51 (page 157) describes a study that found that the number of stumps from trees felled by beavers predicts the abundance of beetle larvae. Is there good evidence that more beetle larvae clusters are present when beavers have left more tree stumps? Estimate how many more clusters accompany each additional stump, with 95% confidence.

## Chapter 24 Problem Statements

**24.9** Bromeliads are tropical flowering plants. Many are epiphytes that attach to trees and obtain moisture and nutrients from air and rain. Their leaf bases form cups that collect water and are home to the larvae of many insects. As a preliminary to a study of changes in the nutrient cycle, Jacqueline Ngai and Diane Srivastava examined the effects of adding nitrogen, phosphorus, or both to the cups. They randomly assigned 8 bromeliads growing in Costa Rica to each of four treatment groups, including an unfertilized control group. A monkey destroyed one of the plants in the control group, leaving 7 bromeliads in that group. Here are the numbers of new leaves on each plant over the 7 months following fertilization:

Nitrogen	Phosphorus	Both	Neither
15	14	14	11
14	14	16	13
15	14	15	16
16	11	14	15
17	13	14	15
18	12	13	11
17	15	17	12
13	15	14	

Analyze these data and discuss the results. Does nitrogen or phosphorus have a greater effect on the growth of bromeliads? Follow the four-step process as illustrated in Example 24.4.

**24.13** What conditions help overweight people exercise regularly? Subjects were randomly assigned to three treatments: a single long exercise period 5 days per week; several 10-minute exercise periods 5 days per week; and several 10-minute periods 5 days per week on a home treadmill that was provided to the subjects. The study report contains the following information about weight loss (in kilograms) after six months of treatment:

Treatment	<i>n</i>	$\bar{x}$	<i>s</i>
Long exercise periods	37	10.2	4.2
Short exercise periods	36	9.3	4.5
Short periods with equipment	42	10.2	5.2

- Do the standard deviations satisfy the rule of thumb for safe use of ANOVA?
- Calculate the overall mean response  $\bar{x}$ , the mean squares MSG and MSE, and the *F* statistic.
- Which *F* distribution would you use to find the *P*-value of the ANOVA *F* test? Software says that  $P = 0.634$ . What do you conclude from this study?

**24.33** Our bodies have a natural electrical field that helps wounds heal. Might higher or lower levels speed healing? An experiment with newts investigated this question. Newts were randomly assigned to five groups. In four of the groups, an electrode applied to one hind limb (chosen at random) changed the natural field, while the other hind limb was not manipulated. Both limbs in the fifth (control) group remained in their natural state.

Table 24.5 gives data from this experiment. The “Group” variable shows the field applied as a multiple of the natural field for each newt. For example, “0.5” is half the natural field, “1” is the natural level (the control group), and “1.5” indicates a field 1.5 times natural. “Diff” is the response variable, the difference in the healing rate (in micrometers per hour) of cuts made in the experimental and control limbs of that newt. Negative values mean that the experimental limb healed more slowly. The investigators conjectured that nature heals best, so that changing the field from the natural state (the “1” group) will slow healing.

Do a complete analysis to see whether the groups differ in the effect of the electrical field level on healing. Follow the four-step process in your work.

Group	Diff								
0	-10	0.5	-1	1	-7	1.25	1	1.5	-13
0	-12	0.5	10	1	15	1.25	8	1.5	-49
0	-9	0.5	3	1	-4	1.25	-15	1.5	-16
0	-11	0.5	-3	1	-16	1.25	14	1.5	-8
0	-1	0.5	-31	1	-2	1.25	-7	1.5	-2
0	6	0.5	4	1	-13	1.25	-1	1.5	-35
0	-31	0.5	-12	1	5	1.25	11	1.5	-11
0	-5	0.5	-3	1	-4	1.25	8	1.5	-46
0	13	0.5	-7	1	-2	1.25	11	1.5	-22
0	-2	0.5	-10	1	-14	1.25	-4	1.5	2
0	-7	0.5	-22	1	5	1.25	7	1.5	10
0	-8	0.5	-4	1	11	1.25	-14	1.5	-4
		0.5	-1	1	10	1.25	0	1.5	-10
		0.5	-3	1	3	1.25	5	1.5	2
				1	6	1.25	-2	1.5	-5
				1	-1				
				1	13				
				1	-8				

**24.35** “Durable press” cotton fabrics are treated to improve their recovery from wrinkles after washing. Unfortunately, the treatment also reduces the strength of the fabric. A study compared the breaking strength of untreated fabric with that of fabrics treated by three commercial durable press processes. Five specimens of the same fabric were assigned at random to each group. Here are the data, in pounds of pull needed to tear the fabric:

Untreated	60.1	56.7	61.5	55.1	59.4
Permafresh 55	29.9	30.7	30.0	29.5	27.6
Permafresh 48	24.8	24.6	27.3	28.1	30.3
Hylite LF	28.8	23.9	27.0	22.1	24.2

The untreated fabric is clearly much stronger than any of the treated fabrics. We want to know if there is a significant difference in breaking strength among the three durable press treatments. Analyze the data for the three processes and write a clear summary of your findings. Which process do you recommend if breaking strength is a main concern? Use the four-step process to guide your discussion. (Although the standard deviations do not quite satisfy our rule of thumb, that rule is conservative and many statisticians would use ANOVA for these data.)

**24.39** Your work in Exercise 24.30 shows that there were significant differences in mean plant biomass among the three treatments in 2003. Do a complete analysis of the data for 2001 and report your conclusions.

## Chapter 25 Problem Statements

**25.1** Our lead example for the two-sample  $t$  procedures in Chapter 18 concerned a study comparing the level of physical activity of lean and mildly obese people who don't exercise. Here are the minutes per day that the subjects spent standing or walking over a 10-day period:

Lean subjects		Obese subjects	
511.100	543.388	260.244	416.531
607.925	677.188	464.756	358.650
319.212	555.656	367.138	267.344
584.644	374.831	413.667	410.631
578.869	504.700	347.375	426.356

The data are a bit irregular but not distinctly non-Normal. Let's use the Wilcoxon test for comparison with the two-sample  $t$  test.

- Find the median minutes spent standing or walking for each group. Which group appears more active?
- Arrange all 20 observations in order and find the ranks.
- Take  $W$  to be the sum of the ranks for the lean group. What is the value of  $W$ ? If the null hypothesis (no difference between the groups) is true, what are the mean and standard deviation of  $W$ ?
- Does comparing  $W$  with the mean and standard deviation suggest that the lean subjects are more active than the obese subjects?

**25.3** In Exercise 25.1, you found the Wilcoxon rank sum  $W$  and its mean and standard deviation. We want to test the null hypothesis that the two groups don't differ in activity against the alternative hypothesis that the lean subjects spend more time standing and walking.

- What is the probability expression for the  $P$ -value of  $W$  if we use the continuity correction?
- Find the  $P$ -value. What do you conclude?

**25.7** Use your software to carry out the one-sided Wilcoxon rank sum test that you did by hand in Exercise 25.3. Use the exact distribution if your software will do it. Compare the software result with your result in Exercise 25.3.

**25.11** Exercise 18.8 (text page 482) compares the breaking strength of polyester strips buried for 16 weeks with that of strips buried for 2 weeks. The breaking strengths in

pounds are

<b>2 weeks</b>	118	126	126	120	129
<b>16 weeks</b>	124	98	110	140	110

- What are the null and alternative hypotheses for the Wilcoxon test? For the two-sample  $t$  test?
- There are two pairs of tied observations. What ranks do you assign to each observation, using average ranks for ties?
- Apply the Wilcoxon rank sum test to these data. Compare your result with the  $P = 0.1857$  obtained from the two-sample  $t$  test in Figure 18.5.

**25.13** The data in Exercise 25.5 for a story told without pictures (Story 1) have tied observations. Is there good evidence that high-progress readers score higher than low-progress readers when they retell a story they have heard without pictures?

- Make a back-to-back stemplot of the 5 responses in each group. Are any major deviations from Normality apparent?
- Carry out a two-sample  $t$  test. State hypotheses and give the two sample means, the  $t$  statistic and its  $P$ -value, and your conclusion.
- Carry out the Wilcoxon rank sum test. State hypotheses and give the rank sum  $W$  for high-progress readers, its  $P$ -value, and your conclusion. Do the  $t$  and Wilcoxon tests lead you to different conclusions?

**25.15** Exercise 7.41 (text page 193) gives data from an experiment in which some bellflower plants in a forest were “fertilized” with dead cicadas and other plants were not disturbed. The data record the mass of seeds produced by 39 cicada plants and 33 undisturbed (control) plants. Do the data show that dead cicadas increase seed mass? Do data analysis to compare the two groups, explain why you would be reluctant to use the two-sample  $t$  test, and apply the Wilcoxon test. Follow the four-step process in your report.

**25.19** Lymphocytes (white blood cells) play an important role in defending our bodies against tumors and infections. Can lymphocytes be genetically modified to recognize and destroy cancer cells? In one study of this idea, modified cells were infused into 11 patients with metastatic melanoma (serious skin cancer) that had not responded to existing treatments. Here are data for an “ELISA” test for the presence of cells that trigger an immune response, in counts per 100,000 cells before and after infusion. High counts suggest that infusion had a beneficial effect.

Patient	1	2	3	4	5	6	7	8	9	10	11
Pre	14	0	1	0	0	0	0	20	1	6	0
Post	41	7	1	215	20	700	13	530	35	92	108

- (a) Examine the differences (post minus pre). Why can't we use the matched pairs  $t$  test to see if infusion raised the ELISA counts?
- (b) We will apply the Wilcoxon signed rank test. What are the ranks for the absolute values of the differences in counts? What is the value of  $W^+$ ?
- (c) What would be the mean and standard deviation of  $W^+$  if the null hypothesis (infusion makes no difference) were true? Compare  $W^+$  with this mean (in standard deviation units) to reach a tentative conclusion about significance.

**25.23** Exercise 17.7 (text page 449) reports the following data on the percent of nitrogen in bubbles of ancient air trapped in amber:

63.4 65.0 64.4 63.3 54.8 64.5 60.8 49.1 51.0

We wonder if ancient air differs significantly from the present atmosphere, which is 78.1% nitrogen.

- (a) Graph the data, and comment on skewness and outliers. A rank test is appropriate.
- (b) We would like to test hypotheses about the median percent of nitrogen in ancient air (the population):

$$H_0 : \text{median} = 78.1$$

$$H_a : \text{median} \neq 78.1$$

To do this, apply the Wilcoxon signed rank statistic to the differences between the observations and 78.1. (This is the one-sample version of the test.) What do you conclude?

**25.25** Cola makers test new recipes for loss of sweetness during storage. Trained tasters rate the sweetness before and after storage. Here are the sweetness losses (sweetness before storage minus sweetness after storage) found by 10 tasters for one new cola recipe:

2.0 0.4 0.7 2.0 -0.4 2.2 -1.5 1.2 1.1 2.3

Are these data good evidence that the cola lost sweetness?

- (a) These data are the differences from a matched pairs design. State hypotheses in terms of the median difference in the population of all tasters, carry out a test, and give your conclusion.
- (b) The output in Figure 17.6 (text page 454) showed that the one-sample  $t$  test had  $P$ -value  $P = 0.0123$  for these data. How does this compare with your result from (a)? What are the hypotheses for the  $t$  test? What conditions must be met for each of the  $t$  and Wilcoxon tests?

**25.27** Exercise 24.30 describes an experiment that examines the effect on plant biomass in plots of California grassland randomly assigned to receive added water in the winter,

added water in the spring, or no added water. The experiment continued for several years. Here are data for 2004 (mass in grams per square meter):

Winter	Spring	Control
254.6453	517.6650	178.9988
233.8155	342.2825	205.5165
253.4506	270.5785	242.6795
228.5882	212.5324	231.7639
158.6675	213.9879	134.9847
212.3232	240.1927	212.4862

The sample sizes are small and the data contain some possible outliers. We will apply a nonparametric test.

- Examine the data. Show that the conditions for ANOVA (text page 644) are not met. What appear to be the effects of extra rain in winter or spring?
- What hypotheses does ANOVA test? What hypotheses does Kruskal-Wallis test?
- What are  $I$ , the  $n_i$ , and  $N$ ? Arrange the counts in order and assign ranks.
- Calculate the Kruskal-Wallis statistic  $H$ . How many degrees of freedom should you use for the chi-square approximation to its null distribution? Use the chi-square table to give an approximate  $P$ -value. What does the test lead you to conclude?

**25.29** Here are the breaking strengths (in pounds) of strips of polyester fabric buried in the ground for several lengths of time:

2 weeks	118	126	126	120	129
4 weeks	130	120	114	126	128
8 weeks	122	136	128	146	140
16 weeks	124	98	110	140	110

Breaking strength is a good measure of the extent to which the fabric has decayed. Do a complete analysis that compares the four groups. Give the Kruskal-Wallis test along with a statement in words of the null and alternative hypotheses.

**25.45** Investigators compared the number of tree species in unlogged plots in the rain forest of Borneo with the number of species in plots logged 8 years earlier. Here are the data:

<b>Unlogged</b>	22	18	22	20	15	21	13	13	19	13	19	15
<b>Logged</b>	17	4	18	14	18	15	15	10	12			

Does logging significantly reduce the number of species in a plot after 8 years?

**25.51** “Second, we found that species richness within tributaries exceeded that within their adjacent upstream mainstem stations.” Again, do a test to confirm significance and report your finding.

### Chapter 26 Problem Statements

**26.13** Exercise 26.10 concerns process control data on the hardness of tablets (measured in kilograms) for a pharmaceutical product. Table 26.4 gives data for 20 new samples of size 4, with the  $\bar{x}$  and  $s$  for each sample. The process has been in control with mean at the target value  $\mu = 11.5$  kg and standard deviation  $\sigma = 0.2$  kg.

Sample	Hardness (kilograms)				Mean	StDev
1	11.432	11.35	11.582	11.184	11.387	0.1660
2	11.791	11.323	11.734	11.512	11.590	0.2149
3	11.373	11.807	11.651	11.651	11.620	0.1806
4	11.787	11.585	11.386	11.245	11.501	0.2364
5	11.633	11.212	11.568	11.469	11.470	0.1851
6	11.648	11.653	11.618	11.314	11.558	0.1636
7	11.456	11.270	11.817	11.402	11.486	0.2339
8	11.394	11.754	11.867	11.003	11.504	0.3905
9	11.349	11.764	11.402	12.085	11.650	0.3437
10	11.478	11.761	11.907	12.091	11.809	0.2588
11	11.657	12.524	11.468	10.946	11.649	0.6564
12	11.820	11.872	11.829	11.344	11.716	0.2492
13	12.187	11.647	11.751	12.026	11.903	0.2479
14	11.478	11.222	11.609	11.271	11.395	0.1807
15	11.750	11.520	11.389	11.803	11.616	0.1947
16	12.137	12.056	11.255	11.497	11.736	0.4288
17	12.055	11.730	11.856	11.357	11.750	0.2939
18	12.107	11.624	11.727	12.207	11.916	0.2841
19	11.933	10.658	11.708	11.278	11.394	0.5610
20	12.512	12.315	11.671	11.296	11.948	0.5641

- Make both  $\bar{x}$  and  $s$  charts for these data based on the information given about the process.
- At some point, the within-sample process variation increased from  $\sigma = 0.2$  kg to  $\sigma = 0.4$  kg. About where in the 20 samples did this happen? What is the effect on the  $s$  chart? On the  $\bar{x}$  chart?
- At that same point, the process mean changed from  $\mu = 11.5$  kg to  $\mu = 11.7$  kg. What is the effect of this change on the  $s$  chart? On the  $\bar{x}$  chart?

**26.15** Figure 26.10 reproduces a data sheet from the floor of a factory that makes electrical meters. The sheet shows measurements on the distance between two mounting holes for 18 samples of size 5. The heading informs us that the measurements are in multiples of 0.0001 inch above 0.6000 inch. That is, the first measurement, 44, stands for

0.6044 inch. All the measurements end in 4. Although we don't know why this is true, it is clear that in effect the measurements were made to the nearest 0.001 inch, not to the nearest 0.0001 inch.

Calculate  $\bar{x}$  and  $s$  for the first two samples. The data file *ex26-15.dat* contains  $\bar{x}$  and  $s$  for all 18 samples. Based on long experience with this process, you are keeping control charts based on  $\mu = 43$  and  $\sigma = 12.74$ . Make  $s$  and  $\bar{x}$  charts for the data in Figure 26.10 and describe the state of the process.

**26.21** Table 26.6 gives data on the losses (in dollars) incurred by a hospital in treating major joint replacement (DRG 209) patients. The hospital has taken from its records a random sample of 8 such patients each month for 15 months.

Sample	Loss (Dollars)								$\bar{x}$	$s$
1	6835	5843	6019	6731	6362	5696	7193	6206	6360.6	521.7
2	6452	6764	7083	7352	5239	6911	7479	5549	6603.6	817.1
3	7205	6374	6198	6170	6482	4763	7125	6241	6319.8	749.1
4	6021	6347	7210	6384	6807	5711	7952	6023	6556.9	736.5
5	7000	6495	6893	6127	7417	7044	6159	6091	6653.2	503.7
6	7783	6224	5051	7288	6584	7521	6146	5129	6465.8	1034.3
7	8794	6279	6877	5807	6076	6392	7429	5220	6609.2	1104
8	4727	8117	6586	6225	6150	7386	5674	6740	6450.6	1033
9	5408	7452	6686	6428	6425	7380	5789	6264	6479.0	704.7
10	5598	7489	6186	5837	6769	5471	5658	6393	6175.1	690.5
11	6559	5855	4928	5897	7532	5663	4746	7879	6132.4	1128.6
12	6824	7320	5331	6204	6027	5987	6033	6177	6237.9	596.6
13	6503	8213	5417	6360	6711	6907	6625	7888	6828.0	879.8
14	5622	6321	6325	6634	5075	6209	4832	6386	5925.5	667.8
15	6269	6756	7653	6065	5835	7337	6615	8181	6838.9	819.5

- (a) Make an  $s$  control chart using center lines and limits calculated from these past data. There are no points out of control.
- (b) Because the  $s$  chart is in control, base the  $\bar{x}$  chart on all 15 samples. Make this chart. Is it also in control?

**26.27** If the mesh tension of individual monitors follows a Normal distribution, we can describe capability by giving the percent of monitors that meet specifications. The old specifications for mesh tension are 100 to 400 mV. The new specifications are 150 to 350 mV. Because the process is in control, we can estimate that tension has mean 275 mV and standard deviation 38.4 mV.

- (a) What percent of monitors meet the old specifications?
- (b) What percent meet the new specifications?

**26.29** Figure 26.10 (page 26-20) displays a record sheet for 18 samples of distances between mounting holes in an electrical meter. The data file *ex26-15.dat* adds  $\bar{x}$  and  $s$  for each sample. In Exercise 26.15, you found that Sample 5 was out of control on the process-monitoring  $s$  chart. The special cause responsible was found and removed. Based on the 17 samples that were in control, what are the natural tolerances for the distance between the holes?

**26.35** Here are data from an urban school district on the number of eighth-grade students with three or more unexcused absences from school during each month of a school year. Because the total number of eighth-graders changes a bit from month to month, these totals are also given for each month.

Month	Sep.	Oct.	Nov.	Dec.	Jan.	Feb.	Mar.	Apr.	May	Jun.
<b>Students</b>	911	947	939	942	918	920	931	925	902	883
<b>Absent</b>	291	349	364	335	301	322	344	324	303	344

- Find  $\bar{p}$ . Because the number of students varies from month to month, also find  $\bar{n}$ , the average per month.
- Make a  $p$  chart using control limits based on  $\bar{n}$  students each month. Comment on control.
- The exact control limits are different each month because the number of students  $n$  is different each month. This situation is common in using  $p$  charts. What are the exact limits for October and June, the months with the largest and smallest  $n$ ? Add these limits to your  $p$  chart, using short lines spanning a single month. Do exact limits affect your conclusions?

**26.43** Painting new auto bodies is a multistep process. There is an “electrocoat” that resists corrosion, a primer, a color coat, and a gloss coat. A quality study for one paint shop produced this breakdown of the primary problem type for those autos whose paint did not meet the manufacturer's standards:

Problem	Percent
Electrocoat uneven---redone	4
Poor adherence of color to primer	5
Lack of clarity in color	2
“Orange peel” texture in color	32
“Orange peel” texture in gloss	1
Ripples in color coat	28
Ripples in gloss coat	4
Uneven color thickness	19
Uneven gloss thickness	5
<b>Total</b>	<b>100</b>

Make a Pareto chart. Which stage of the painting process should we look at first?

**26.51** Calculate control limits for  $s$ , make an  $\bar{x}$ - $s$  chart, and comment on control of short-term process variation.

## Chapter 27 Problem Statements

**27.15** The table below shows the progress of world record times (in seconds) for the 10,000-meter run for both men and women.

Men				Women	
Record year	Time (seconds)	Record year	Time (seconds)	Record year	Time (seconds)
1912	1880.8	1962	1698.2	1967	2286.4
1921	1840.2	1963	1695.6	1970	2130.5
1924	1835.4	1965	1659.3	1975	2100.4
1924	1823.2	1972	1658.4	1975	2041.4
1924	1806.2	1973	1650.8	1977	1995.1
1937	1805.6	1977	1650.5	1979	1972.5
1938	1802.0	1978	1642.4	1981	1950.8
1939	1792.6	1984	1633.8	1981	1937.2
1944	1775.4	1989	1628.2	1982	1895.3
1949	1768.2	1993	1627.9	1983	1895.0
1949	1767.2	1993	1618.4	1983	1887.6
1949	1761.2	1994	1612.2	1984	1873.8
1950	1742.6	1995	1603.5	1985	1859.4
1953	1741.6	1996	1598.1	1986	1813.7
1954	1734.2	1997	1591.3	1993	1771.8
1956	1722.8	1997	1587.8		
1956	1710.4	1998	1582.7		
1960	1698.8	2005	1577.5		

- Make a scatterplot of world record time against year, using separate symbols for men and women. Describe the pattern for each gender. Then compare the progress of men and women.
- Fit the model with two regression lines, one for women and one for men, and identify the estimated regression lines.
- Women began running this long distance later than men, so we might expect their improvement to be more rapid. Moreover, it is often said that men have little advantage over women in distance running as opposed to sprints, where muscular strength plays a greater role. Do the data appear to support these claims?

**27.19** An experiment was conducted using a Geiger-Mueller tube in a physics lab. Geiger-Mueller tubes respond to gamma rays and to beta particles (electrons). A pulse that corresponds to each detection of a decay product is produced, and these pulses were counted using a computer-based nuclear counting board. Elapsed time (in seconds) and

counts of pulses for a short-lived unstable isotope of silver are shown in Table 27.5 (see page 27-36).

- (a) Create a scatterplot of the counts versus time and describe the pattern.
- (b) Since some curvature is apparent in the scatterplot, you might want to consider the quadratic model for predicting counts based on time. Fit the quadratic model and identify the estimated mean response.
- (c) Add the estimated mean response to your scatterplot. Would you recommend the use of the quadratic model for predicting radioactive decay in this situation? Explain.
- (d) Transform the counts using the natural logarithm and create a scatterplot of the transformed variable versus time.
- (e) Fit a simple linear regression model using the natural logarithm of the counts. Provide the estimated regression line, a scatterplot with the estimated regression line, and appropriate residual plots.
- (f) Does the simple linear regression model for the transformed counts fit the data better than the quadratic regression model? Explain.

**27.23** Suppose that the couple shopping for a diamond in Example 27.15 had used a quadratic regression model for the other quantitative variable, *Depth*. Use the data in the file *ta27-04.dat* to answer the following questions.

- (a) What is the estimated quadratic regression model for mean total price based on the explanatory variable *Depth*?
- (b) As you discovered in part (a), it is always possible to fit quadratic models, but we must decide if they are helpful. Is this model as informative to the couple as the model in Example 27.15? What percent of variation in the total price is explained by using the quadratic regression model with *Depth*?

**27.25** Table 27.8 contains data on the size of perch caught in a lake in Finland. Use statistical software to help you analyze these data.

- (a) Use the multiple regression model with two explanatory variables, length and width, to predict the weight of a perch. Provide the estimated multiple regression equation.
- (b) How much of the variation in the weight of perch is explained by the model in part (a)?
- (c) Does the ANOVA table indicate that at least one of the explanatory variables is helpful in predicting the weight of perch? Explain.
- (d) Do the individual  $t$  tests indicate that both  $\beta_1$  and  $\beta_2$  are significantly different from zero? Explain.
- (e) Create a new variable, called interaction, that is the product of length and width. Use the multiple regression model with three explanatory variables, length, width, and interaction, to predict the weight of a perch. Provide the estimated multiple regression equation.

- (f) How much of the variation in the weight of perch is explained by the model in part (e)?
- (g) Does the ANOVA table indicate that at least one of the explanatory variables is helpful in predicting the weight of perch? Explain.
- (h) Describe how the individual  $t$  statistics changed when the interaction term was added.

**27.27** Use explanatory variables length, width, and interaction from Exercise 27.25 (page 27-49) on the 56 perch to provide 95% confidence intervals for the mean and prediction intervals for future observations. Interpret both intervals for the 10th perch in the data set. What  $t$  distribution is used to provide both intervals?

**27.41** A multimedia statistics learning system includes a test of skill in using the computer's mouse. The software displays a circle at a random location on the computer screen. The subject clicks in the circle with the mouse as quickly as possible. A new circle appears as soon as the subject clicks the old one. Table 5.3 (text page 159) gives data for one subject's trials, 20 with each hand. Distance is the distance from the cursor location to the center of the new circle, in units whose actual size depends on the size of the screen. Time is the time required to click in the new circle, in milliseconds.

- (a) Specify the population multiple regression model for predicting time from distance separately for each hand. Make sure you include the interaction term that is necessary to allow for the possibility of having different slopes. Explain in words what each  $\beta$  in your model means.
- (b) Use statistical software to find the estimated multiple regression equation for predicting time from distance separately for each hand. What percent of variation in the distances is explained by this multiple regression model?
- (c) Explain how to use the estimated multiple regression equation in part (b) to obtain the least-squares line for each hand. Draw these lines on a scatterplot of time versus distance.

**27.45** The Sanchez household is about to install solar panels to reduce the cost of heating their house. In order to know how much the solar panels help, they record their consumption of natural gas before the solar panels are installed. Gas consumption is higher in cold weather, so the relationship between outside temperature and gas consumption is important. Here are the data for 16 consecutive months:

	Nov.	Dec.	Jan.	Feb.	Mar.	Apr.	May	June
<b>Degree-days</b>	24	51	43	33	26	13	4	0
<b>Gas used</b>	6.3	10.9	8.9	7.5	5.3	4.0	1.7	1.2
	July	Aug.	Sep.	Oct.	Nov.	Dec.	Jan.	Feb.
<b>Degree-days</b>	0	1	6	12	30	32	52	30
<b>Gas used</b>	1.2	1.2	2.1	3.1	6.4	7.2	11.0	6.9

Outside temperature is recorded in degree-days, a common measure of demand for heating. A day's degree-days are the number of degrees its average temperature falls below 65°F. Gas used is recorded in hundreds of cubic feet.

- (a) Create an indicator variable, say *INDwinter*, which is 1 for the months of November, December, January, and February. Make a plot of all the data using a different symbol for winter months.
- (b) Fit the model with two regression lines, one for winter months and one for other months, and identify the estimated regression lines.
- (c) Do you think that two regression lines were needed to explain the relationship between gas used and degree-days? Explain.

## Chapter 28 Problem Statements

**28.1** The full data for the logging study appear in Table 24.2 (text page 640). The data for counts of individual trees in the plots studied also appear in the data file *ex28-01.dat*. Carry out data analysis and ANOVA to determine whether logging affects the mean count of individual trees in a plot.

**28.3** If you are a dog lover, perhaps having your dog along reduces the effect of stress. To examine the effect of pets in stressful situations, researchers recruited 45 women who said they were dog lovers. The EESEE story “Stress among Pets and Friends” describes the results. Fifteen of the subjects were randomly assigned to each of three groups to do a stressful task alone (the control group), with a good friend present, or with their dog present. The subject's mean heart rate during the task is one measure of the effect of stress. Table 28.2 displays the data. Are there significant differences among the mean heart rates under the three conditions?

**28.7** Using the Minitab output in Figure 28.4, verify the values for the sample contrast  $\hat{L}_2$  and its standard error given in Example 28.7. Give a 95% confidence interval for the population contrast  $L_2$ . Carry out a test of the hypothesis  $H_0 : L_2 = 0$  against the two-sided alternative. Be sure to state your conclusions in the setting of the study.

**28.11** A student project measured the increase in the heart rates of fellow students when they stepped up and down for three minutes to the beat of a metronome. The explanatory variables are step height (Lo = 5.75 inches, Hi = 11.5 inches) and metronome beat (Slow = 14 steps/minute, Med = 21 steps/minute, Fast = 28 steps/minute). The subject's heart rate was measured for 20 seconds before and after stepping. The response variable is the increase in heart rate during exercise. The data appear in Table 28.3.

(a) Display the 6 treatments in a two-way layout.

(b) Find the group means, plot the means, and discuss the interaction and the two main effects.

Lo/Slow	Lo/Med	Lo/Fast	Hi/Slow	Hi/Med	Hi/Fast
15	21	24	39	45	66
9	24	42	33	27	60
6	15	27	15	24	51
9	15	48	15	39	30
0	18	18	16	6	57

**28.13** The researchers who conducted the study in the previous exercise also recorded the number of times each of three types of behavior (object play, locomotor play, and social play) occurred. The file *ex28-13.dat* contains the counts of social play episodes by each rat during the observation period. Use two-way ANOVA to analyze the effects of gender and housing.

**28.29** Exercise 24.33 (text page 661) describes a study on the rate at which the skin of newts heals under the body's natural electrical field (the 1 group in Table 24.5) and under four levels of electric field that differ from the natural level (the 0, 0.5, 1.25, and 1.5 groups). Carry out a one-way ANOVA to compare the mean healing rates. Then perform Tukey multiple comparisons for the 10 pairs of population means. Use the “underline” method illustrated in Example 28.14 to display the complicated results. What do you conclude?

**28.31** The data file *ex28-31.dat* has resting and final heart rates as well as the increase in heart rate for the study described in Exercise 28.11. If the randomization worked well, there should be no significant differences among the 6 groups in mean resting heart rate (variable HRrest in the data file).

- (a) How many pairwise comparisons are there among the means of 6 populations?
- (b) Use Tukey's method to compare these means at the overall 10% significance level.

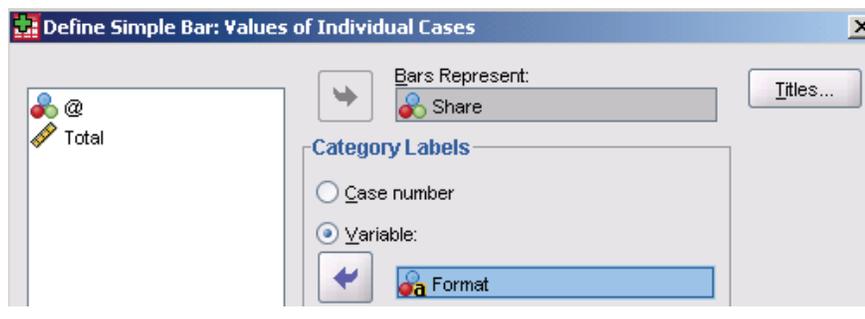
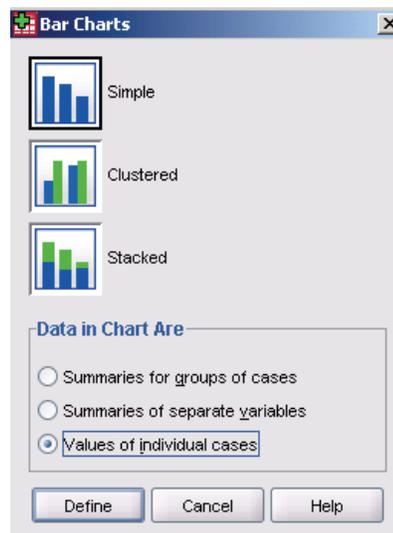
**28.33** The researchers who conducted the study in the previous exercise also recorded the number of times each of three types of behavior (object play, locomotor play, and social play) occurred. The file *ex28-33.dat* contains the counts of object play episodes for each rat during the observation period. Carry out a complete analysis of the effects of gender and housing type.

## Chapter 1 SPSS Solutions

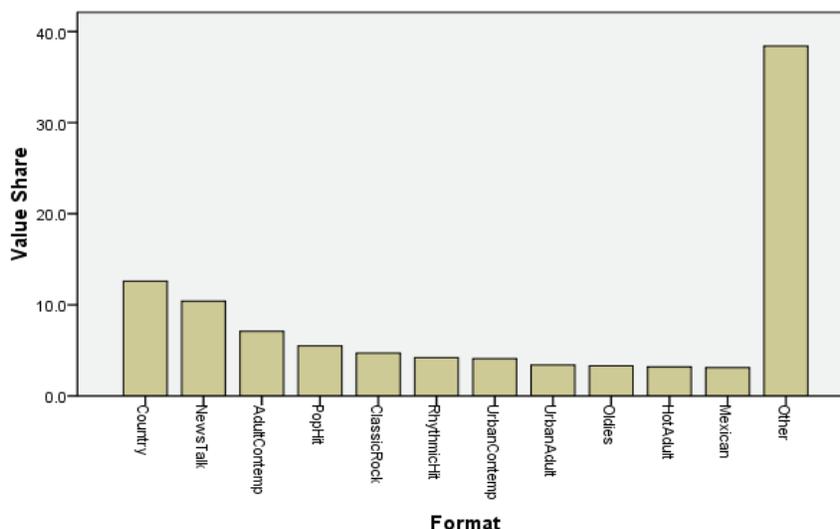
**1.3** Open data file *ex01-03.por*. We add together the given percentages; they add to 61.6%, so 38.4% of radio stations are some other format. Add a category “Other” with 38.4% at the bottom of the spreadsheet.

To create the bar graph, click **Graphs**, **Legacy Dialogs**, **Bar**. The chart type definition box show at right appears. Since our data are already summarized, we want our bars to represent **Values of individual cases**, so click on the button to change this option; we also want a **Simple** bar chart (the default). Click **Define** to proceed.

Click to enter **Share** into the Bars Represent box, then move the Category Labels button to **Variable**. Highlight **Format** and click the arrow to enter it into the box. Click **Titles** and give your graph an appropriate title. Click **Continue** and **OK** to generate the graph.

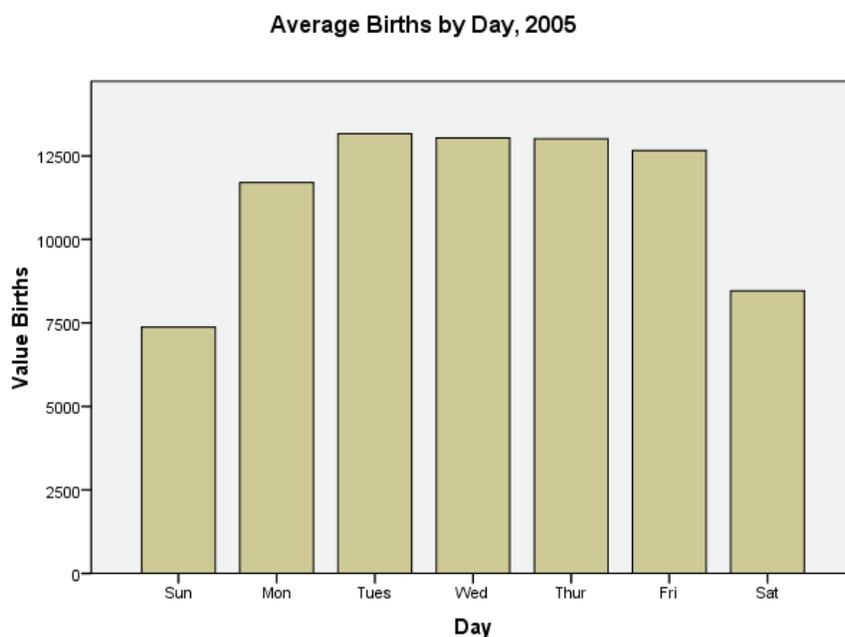


**Radio Station Formats**



Since the addition of the “Other” category makes the percents given add to 100%, we could also make a pie chart; however, given the number of categories (12) this is not a good idea; also, the Other category will dominate the individual categories given.

**1.5** Open data file *ex01-05*. Follow the steps outlined above in the solution to Exercise 1.3 to create the bar graph. Since all days of the week are included in the data given, you don’t need to sum the **Births** variable. Don’t forget to give it a title. Our completed graph is below.



**1.11** Open data file *ta01-03*. To create the stemplot (and compute numerical summaries), click **Analyze, Descriptive Statistics, Explore**. To select **Dollars** as the graph variable, highlight its name on the left, then click the arrow key to move it into the **Dependent List** box. Click **OK**. SPSS automatically does the rounding and splitting of stems. Our graph (produced in the Output window) is below. The shape is skewed right; the United States is the high outlier (indicated by Extremes). The values range from (approximately) \$400 to \$5711; the center (median) is in the interval from 1500 to 1900, at this point we’ll estimate the center is about \$1800.

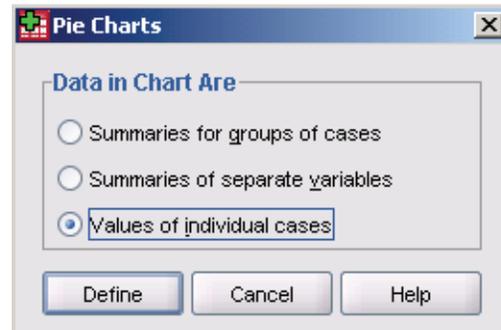
## Dollars Stem-and-Leaf Plot

Frequency	Stem &	Leaf
1.00	0 .	4
8.00	0 .	55667778
5.00	1 .	00123
6.00	1 .	678899
6.00	2 .	122334
7.00	2 .	7788999
2.00	3 .	01
2.00	3 .	78
1.00	Extremes	(>=5711)

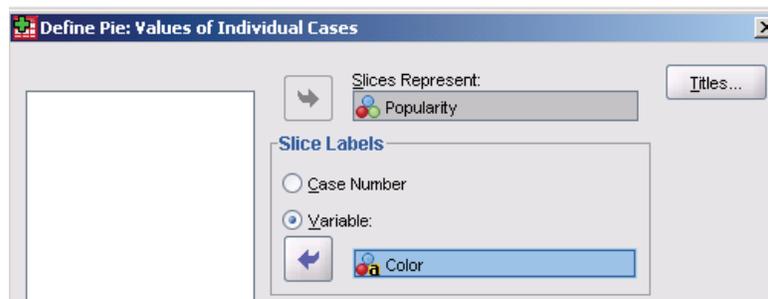
Stem width: 1000  
Each leaf: 1 case(s)

**1.25** Open data file *ex01-25*. To find the percent of other colors we add the percents given and see the sum is 95%. Enter 5 (for 5% into the worksheet for Other.

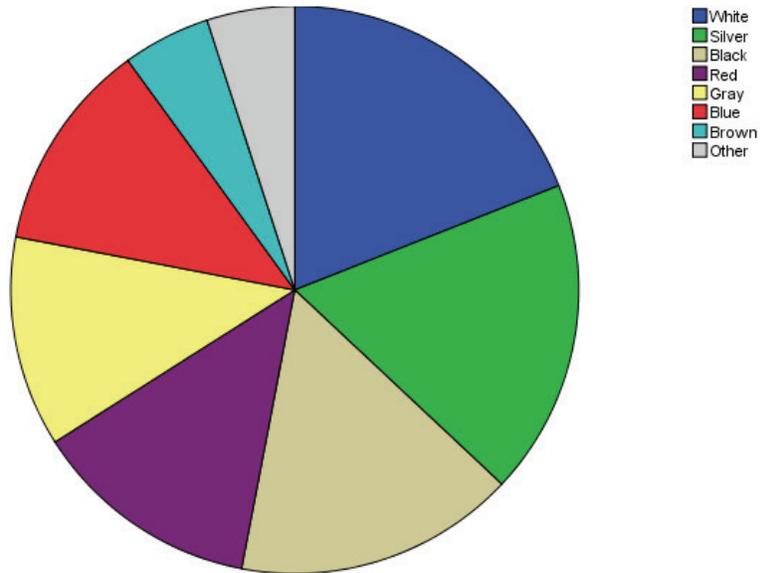
We could make a bar graph of these data as described in the solution of Exercise 1.3; however, since the percents now add to 100 and there are not too many categories, let's make a pie chart instead. Click **Graphs**, **Legacy Dialogs**, **Pie**. Our data are **Values of individual cases**; click **Define** to continue.



Highlight and click the arrow to enter **Popularity** in the Slices represent box; move the button to indicate that Slice Labels are in a **Variable**, then highlight and click the arrow to move **Color** to the box. Click **Titles** to give the graph a descriptive title, then **OK** to generate the graph shown below.

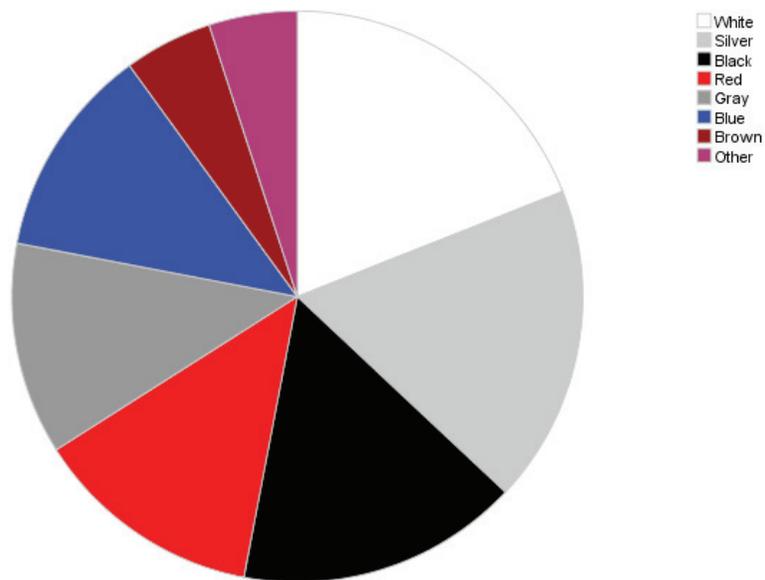


Car Color Preferences, North America, 2007



You can customize each slice's color by double clicking in any slice to bring up the Chart Editor window. Right click on each box of the colors legend, then select **Properties Window**, select the appropriate color to reflect the actual car color and **Apply** and **Close** the window. Our results are below.

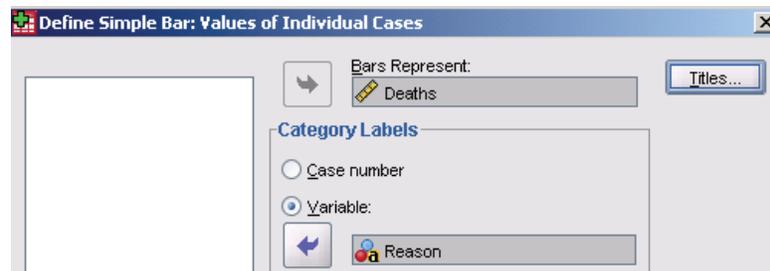
Car Color Preferences, North America, 2007



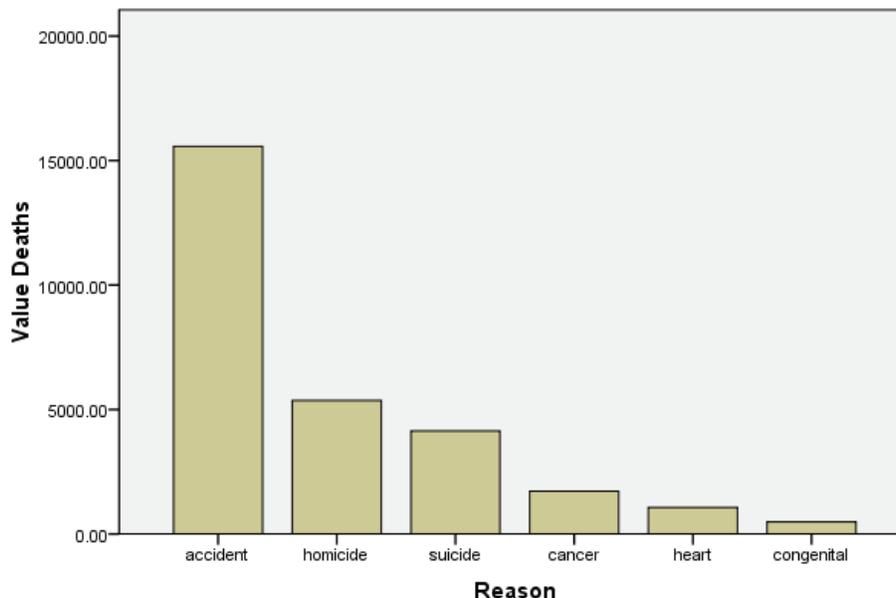
**1.27** These data are not in a worksheet. Define the two variables by clicking on the Variable View tab at the bottom of the worksheet. Reason is a string (alpha) variable that we have allowed 10 characters for. Deaths is numeric (with no decimal places – clicking on the right hand side of the decimals box will allow you to change these from 2 (the default) to 0).

Name	Type	Width	Decimals	Label
Reason	String	10	0	
Deaths	Numeric	8	0	

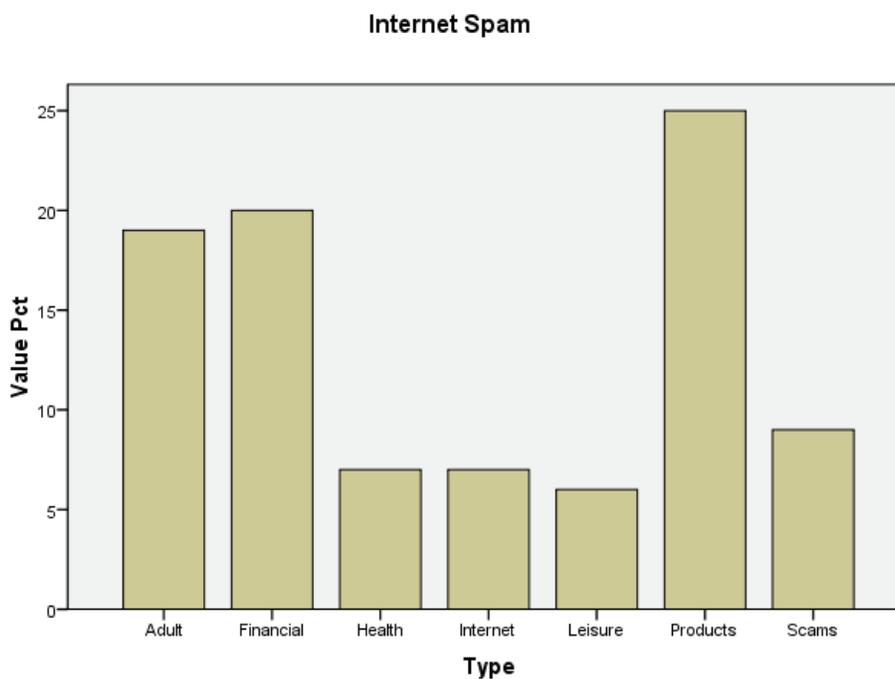
Click on the Data View tab, and enter the death reasons and the corresponding numbers. We can't make a pie chart with these, because we don't know the total number of deaths in this age group. We'll make a bar chart using **Graphs, Legacy Dialogs, Bar for Values of individual cases**. Click to enter that Bars Represent **Deaths**; then move the Category Labels button to **Variable**. Highlight **Reason** and click the arrow to enter it into the box. Click **Titles** and give your graph an appropriate title. Click **Continue** and **OK** to generate the graph.



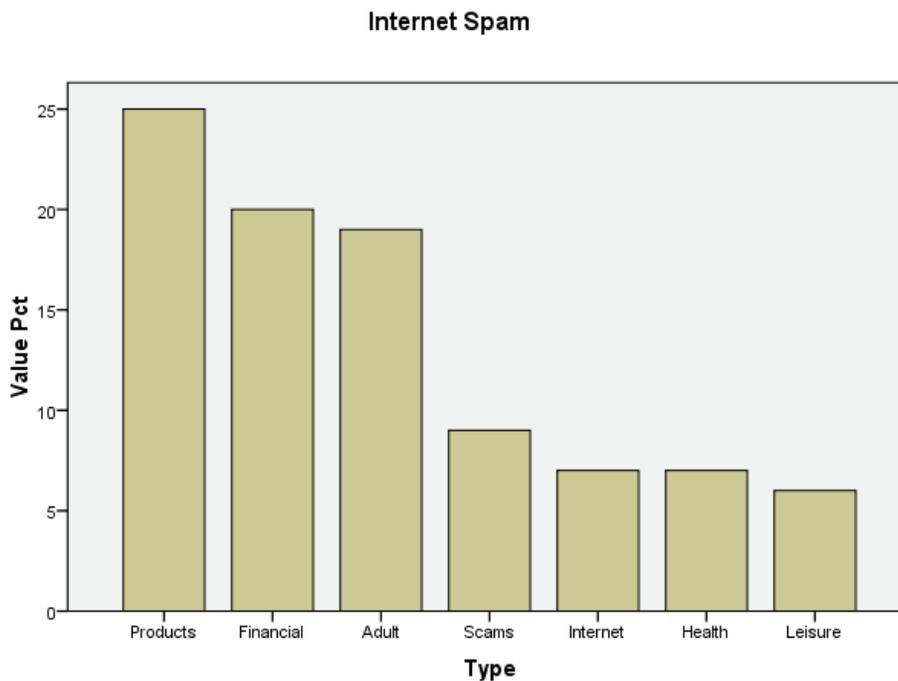
2005 Deaths, Ages 15-24



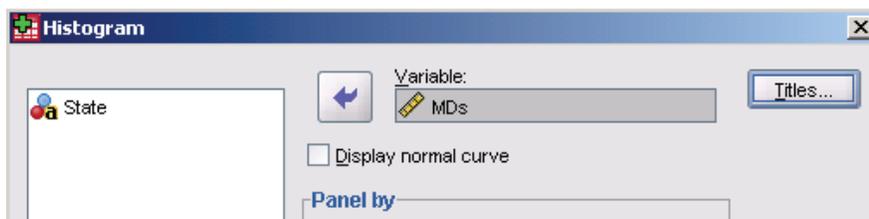
**1.29** Open data file *ex01-29*. We'll make a bar chart using **Graphs, Legacy Dialogs, Bar** for **Values of individual cases**. Click to enter that Bars Represent **Pct**; then move the Category Labels button to **Variable**. Highlight **Type** and click the arrow to enter it into the box. Click **Titles** and give your graph an appropriate title. Click **Continue** and **OK** to generate the graph.

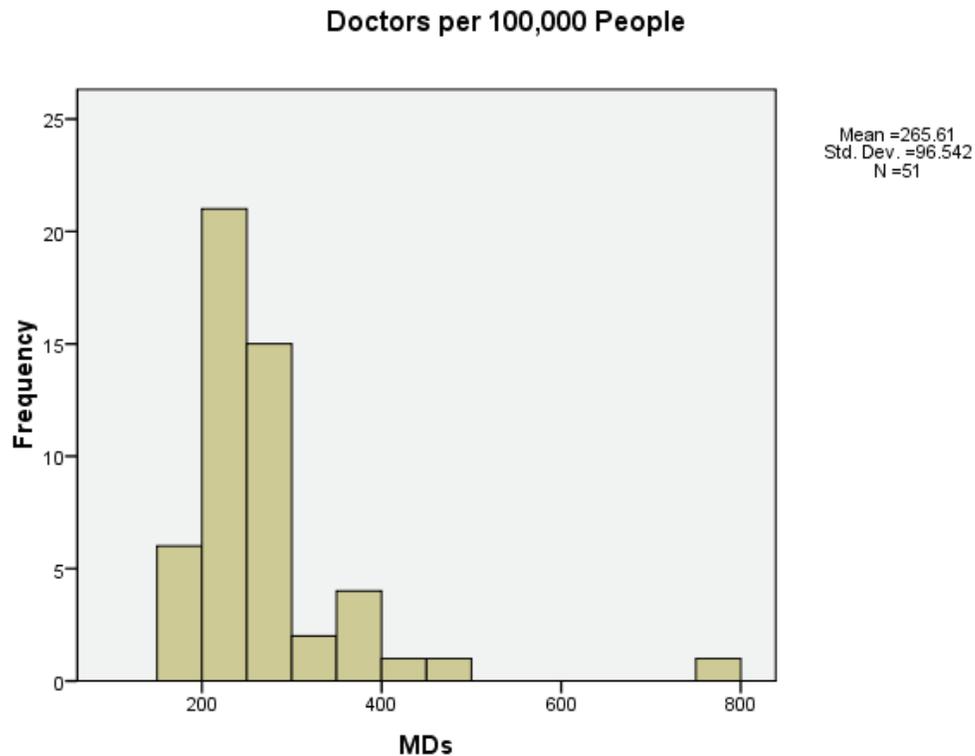


To sort the bars in descending order, double click on the graph to bring up the Chart Editor. Click on the large **X** icon on the menu bar. This brings up the Properties window for the *x* axis. Use the pull down boxes to Sort by **Statistic** in **Descending** direction. **Apply** the change and **Close** the Properties box and Chart Editor.



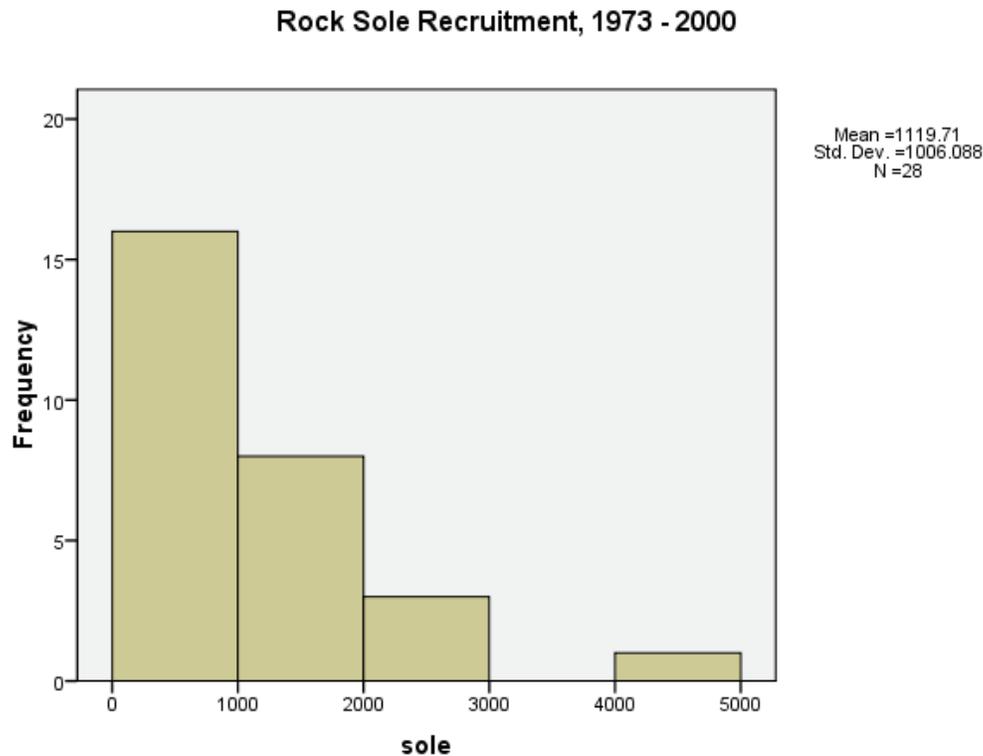
**1.35** We first note that the number of doctors per 100,000 people is a better measure of the availability of health care because states vary greatly in the number of people. Open data file *ta01-05*. To create the histogram of the number of doctors, click **Graphs**, **Legacy Dialogs**, **Histogram**. Click to highlight variable **MDs** and click to enter this into the Variable box. Give your graph an appropriate **Title**, then **Continue** and **OK**.



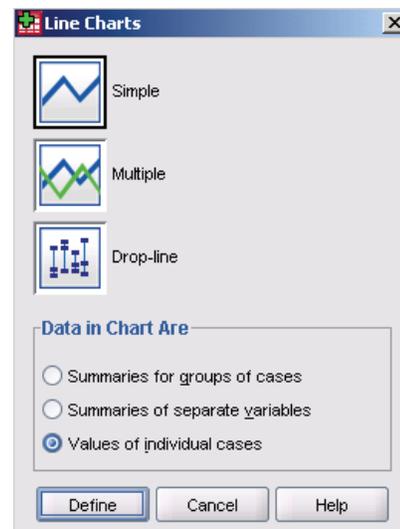


The graph is rather right skewed with an outlier (Washington, D.C.) The range is from about 100 to about 800 with a center about 250. (SPSS actually displays the mean of 265.61 and the standard deviation for you.

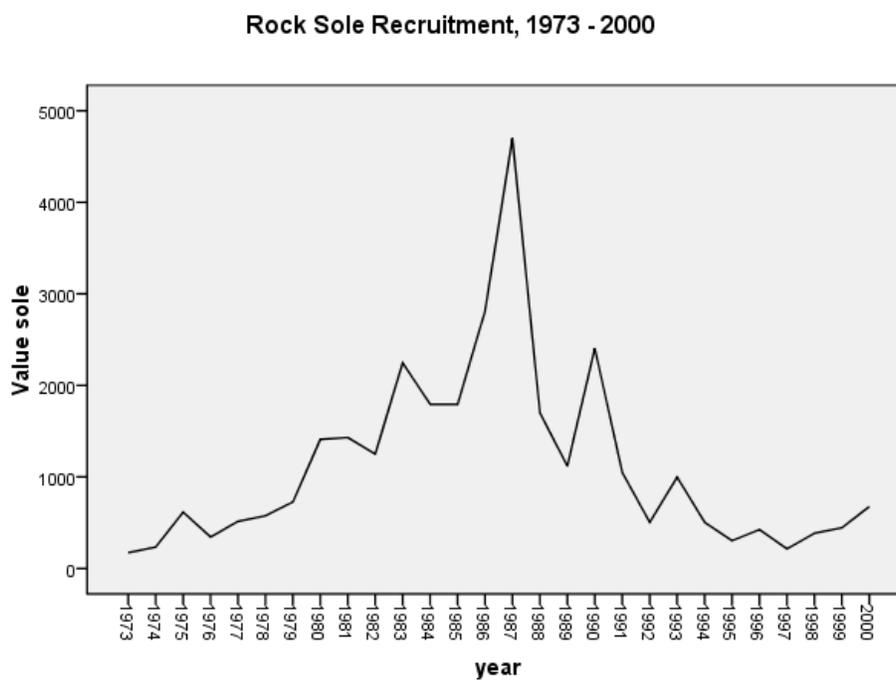
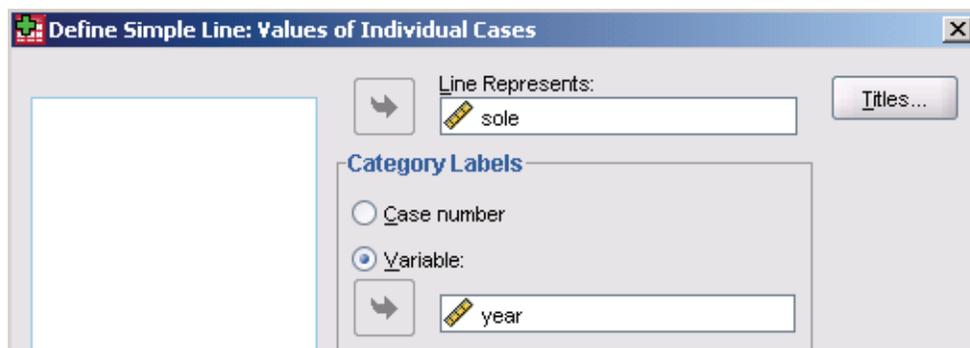
**1.37** Open data file *ex01-37*. Create the histogram as described above in Exercise 1.35. This graph is skewed right with an outlier above 4000 million fish. The center is about 1000 million (SPSS gives the mean as 1119.71 million); the data range from (roughly) 0 to over 4000 million.



**1.39** If you have just finished Exercise 1.37, data file *ex01-37* is still the active worksheet. If not, open this worksheet file. To make the time plot, click **Graphs**, **Legacy Dialogs**, **Line**. We want a Simple chart (the default) with **Values of individual cases**. Click **Define** to continue.

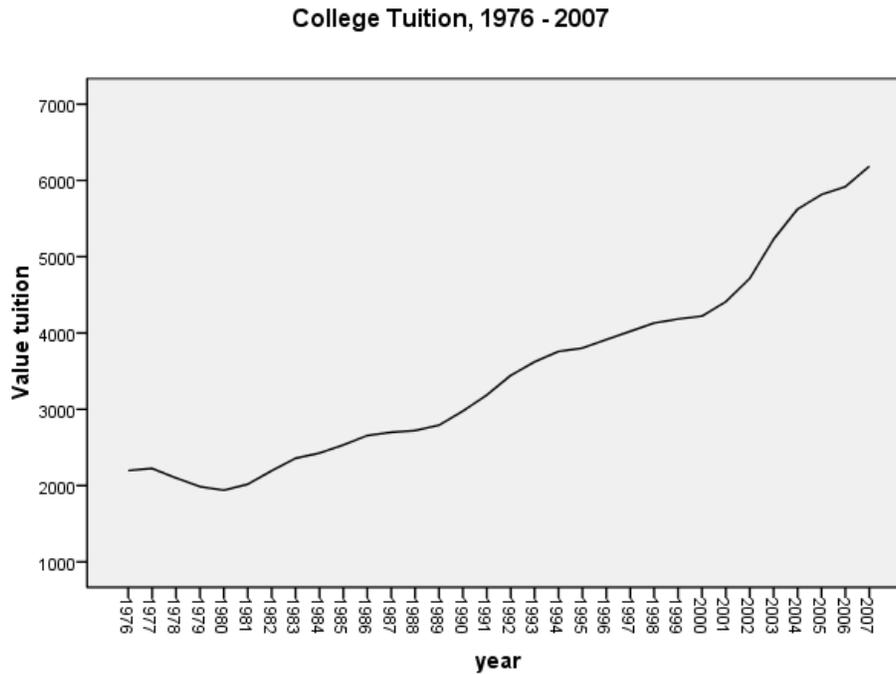


Click to enter **sole** in the Line Represents box ; move the button for Category Labels to **Variable**, and enter **year** as the variable. Give the graph an appropriate **Titles** and **Continue** and **OK** to generate the plot.



The increase in recruitment until the peak in 1987 followed by a sharp decline is clearly evident in this graph.

**1.41** Open data file ex01-12. Follow the instructions from Exercise 1.39 to create an initial time series plot as seen below.



To make the graph appear shallower, we can increase the y axis range. Double-click on the graph in the output window to bring up the Chart Editor. Click on the large **Y** icon on the menu bar. This brings up a Properties box for the scale. Uncheck the Automatic boxes for the minimum and maximum and enter new values (we used 0 and 10000). **OK** applies the change. The new graph is below.

**Properties** ✕

Number Format | Variables | **Scale** | Lines | Labels & Ticks

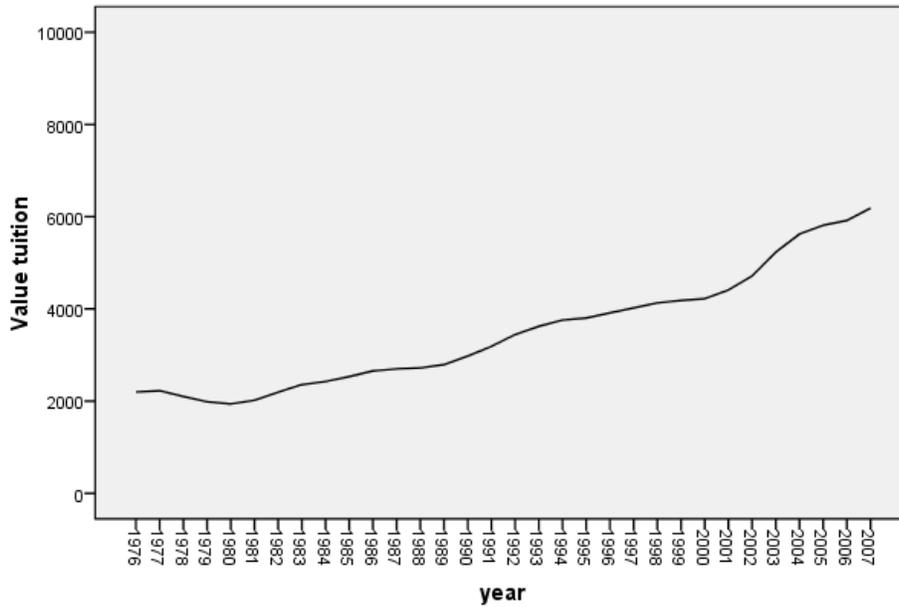
Chart Size | **Scale** | Lines | Labels & Ticks

**Range**

	Auto	Custom	Data
Minimum	<input type="checkbox"/>	<input type="text" value="0"/>	1939
Maximum	<input type="checkbox"/>	<input type="text" value="10000"/>	6185
Major Increment	<input checked="" type="checkbox"/>	<input type="text" value="1000"/>	
Origin	<input checked="" type="checkbox"/>	<input type="text" value="0"/>	

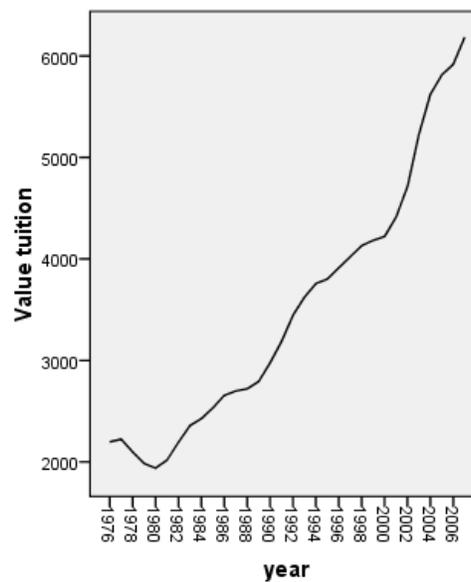
Display line at origin

College Tuition, 1976 - 2007

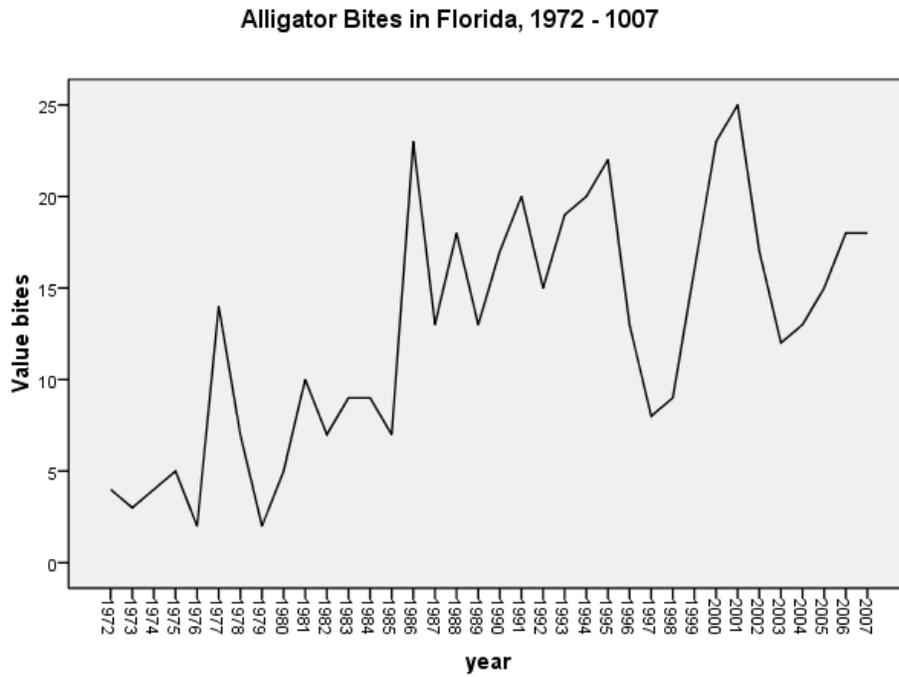


To make the increase seem sharper, repeat the process, but decrease the y axis scale. You can also narrow the width of the graph by clicking on the X icon and then click the Chart Size tab; make the width narrower (be sure to uncheck the Maintain aspect ratio box). (We changed the width to 250 in the graph below).

College Tuition, 1976 - 2007

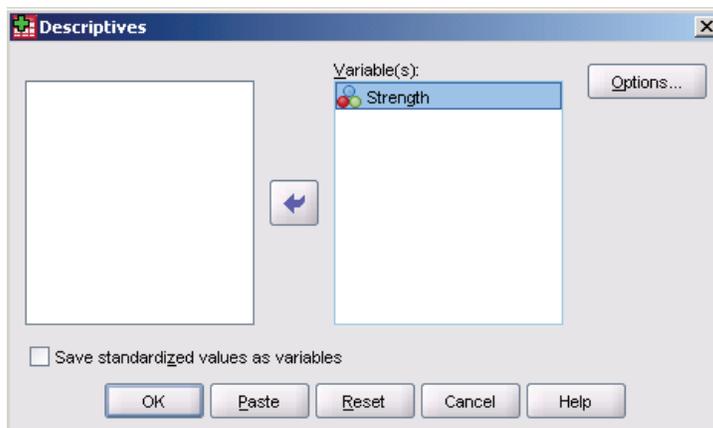


**1.45** Open data file *ex01-45*. Follow the instructions from Exercise 1.39 above to create the time series plot seen below. We can see that alligator bites in Florida are highly variable (there are lots of jagged peaks in the graph) and show a generally increasing trend, perhaps due to encroachment on their natural territory by humans and their continued development of the state.



## Chapter 2 SPSS Solutions

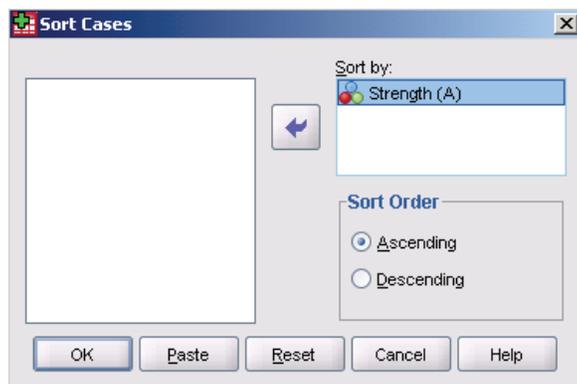
**2.1** Open data file *eg01-09*. To find the mean (and other summary statistics), click **Analyze, Descriptive Statistics, Descriptives**. Click to highlight the variable **Strength**, then click the arrow to select it into the Variables box. **OK** displays the statistics in the output window. Below, we see the mean is 30,841 pounds.



**Descriptive Statistics**

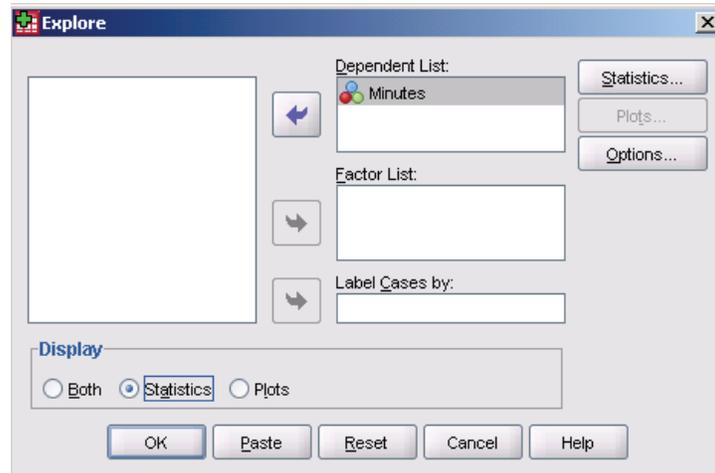
	N	Minimum	Maximum	Mean	Std. Deviation
Strength	20	23040	33650	30841.00	3018.504
Valid N (listwise)	20				

To find out how many observations are less than the mean, we could simply count them, or we can sort the data in ascending order. Click **Data, Sort Cases**. Click to enter the variable **Strength** (Ascending is the default sort order). **OK** sorts the data. We see below that six observations were less than the mean. Referring to the stemplot in Example 1.9, we see that this distribution is skewed left.



	Strength
1	23040
2	24050
3	26520
4	28730
5	30170
6	30460
7	30930

**2.3** Open data file *eg02-03*. To find the mean, median, (and other summary statistics), click **Analyze, Descriptive Statistics, Explore**. Click to highlight the variable **Minutes**, then click the arrow to select it into the Dependent list box. This dialog will compute not only summary statistics but create boxplots and stemplots as well. We have moved the button to ask for only **Statistics**. **OK** displays the statistics in the output window. Below, we see the mean is 31.25 minutes and the median is 22.5 minutes. Since the mean is greater than the median, these data are most likely skewed right (as shown in the stemplot in Example 2.3).

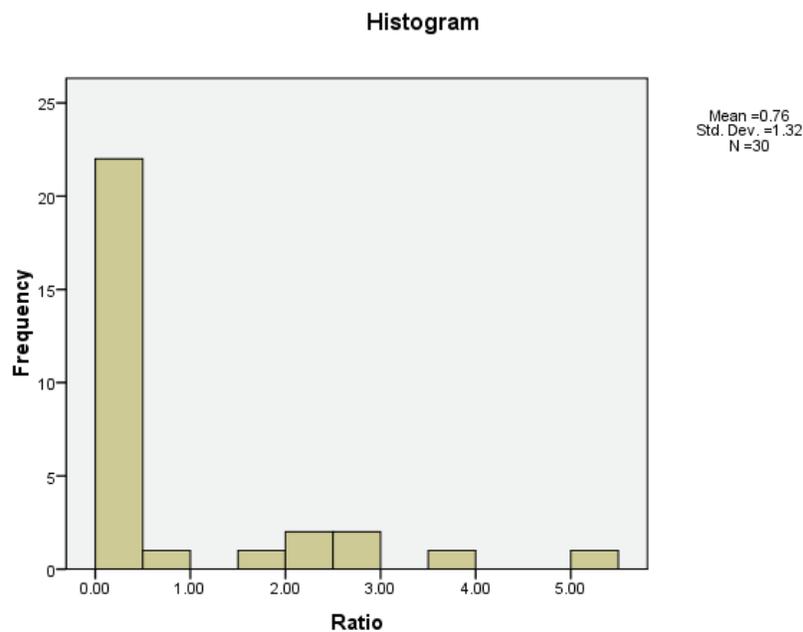


### Descriptives

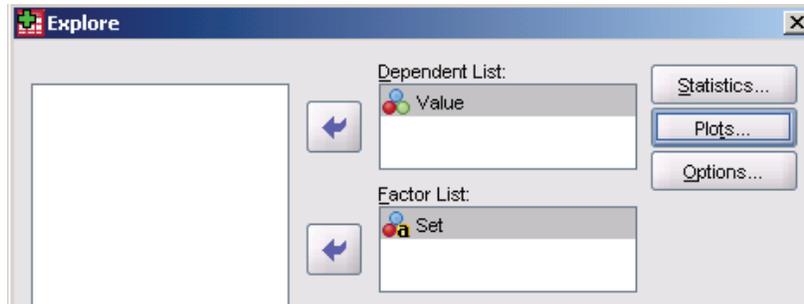
		Statistic	Std. Error	
Minutes	Mean	31.25	4.892	
	95% Confidence Interval for Mean	Lower Bound	21.01	
		Upper Bound	41.49	
	5% Trimmed Mean	29.72		
	Median	22.50		
	Variance	478.618		
	Std. Deviation	21.877		
	Minimum	5		
	Maximum	85		
	Range	80		
	Interquartile Range	29		
	Skewness	1.040	.512	
	Kurtosis	.330	.992	

**2.5** Open data file *ta01-04*. We can compute the summary statistics and create the histogram all at once. Click **Analyze, Descriptive Statistics, Explore**. Click to highlight the variable **Minutes**, then click the arrow to select it into the Dependent list box. Click **Plots**, then put a check in the box to display a histogram of the data. **Continue** and **OK** completes our work. (Notice that this option does not allow us to give the graph our own title). The mean is  $\bar{x} = 0.761$  and the median is 0.075. These two values are not close to one another. We can see from the value of the Max = 5.33 that these data are most likely right skewed. Our histogram confirms this.

Descriptives			Statistic	Std. Error
Ratio	Mean		.7607	.24091
	95% Confidence Interval for Mean	Lower Bound	.2679	
		Upper Bound	1.2534	
	5% Trimmed Mean		.5819	
	Median		.0750	
	Variance		1.741	
	Std. Deviation		1.31954	
	Minimum		.00	
	Maximum		5.33	
	Range		5.33	
	Interquartile Range		.93	
	Skewness		2.071	.427
	Kurtosis		4.122	.833



**2.11** Open data file *ex02-11*. These data are organized differently; we have C1 (**Set**) that indicates which data set the **Value** belongs to. We can find the summary statistics for both sets and create the stemplots using **Analyze, Descriptive Statistics, Explore**. Click to highlight the variable **Value**, use the arrow to select it into the Dependent List box. Next, click on the variable name **Set**, and move it into the Factor List box. **OK** generates the summary statistics seen below (note that we have eliminated some of the statistics that display to focus on the ones of interest to us).



**Descriptives**

Set		Statistic	Std. Error
Value	A	Mean	7.5009
		Median	8.1400
		Std. Deviation	2.03166
		Minimum	3.10
		Maximum	9.26
		Range	6.16
		Interquartile Range	3.00
B	B	Mean	7.5009
		Median	7.0400
		Std. Deviation	2.03058
		Minimum	5.25
		Maximum	12.50
		Range	7.25
		Interquartile Range	2.71

Set A has mean  $\bar{x} = 7.501$  and standard deviation  $s = 2.032$ . Set B also has mean  $\bar{x} = 7.501$  and the standard deviation is very close at  $s = 2.031$ .

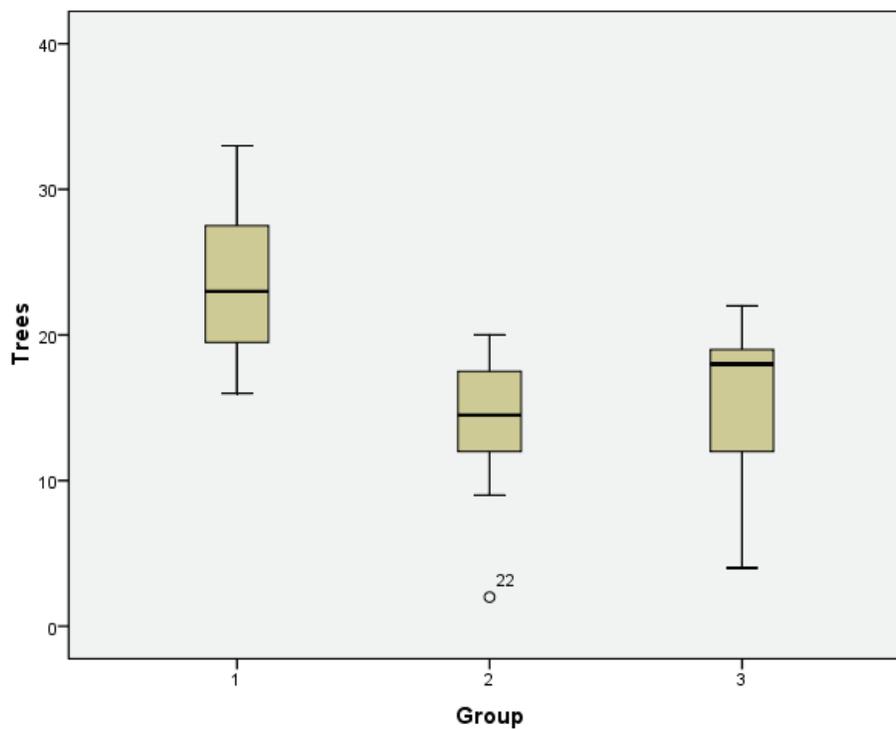
Value Stem-and-Leaf Plot for Set= A			Value Stem-and-Leaf Plot for Set= B		
Frequency	Stem &	Leaf	Frequency	Stem &	Leaf
1.00	Extremes	(=<3.1)	3.00	5 .	257
1.00	4 .	7	2.00	6 .	58
.00	5 .		3.00	7 .	079
1.00	6 .	1	2.00	8 .	48
1.00	7 .	2	1.00	Extremes	(>=12.5)
4.00	8 .	1177	Stem width:	1.00	
3.00	9 .	112	Each leaf:	1 case(s)	
Stem width:	1.00				
Each leaf:	1 case(s)				

The two distributions have dramatically different shapes – the moral is that there is no substitute for failing to plot the data!

**2.13** Open data file *ex02-13*. We'll use **Analyze**, **Descriptive Statistics**, **Explore** to compute summary statistics and create side-by-side boxplots for each group. Click to enter **Trees** in the Dependent List box, then **Group** in the **Factor List** box. Click **Graphs** and check the box for **Boxplot of data**. **Continue** and **OK** gets our results. (We've again deleted some of the statistics computed from our results below).

Descriptives			
Group		Statistic	Std. Error
Trees	1	Mean	23.75
		Median	23.00
		Std. Deviation	5.065
		Minimum	16
		Maximum	33
		Range	17
		Interquartile Range	8
	2	Mean	14.08
		Median	14.50
		Std. Deviation	4.981
		Minimum	2
		Maximum	20
		Range	18

	Interquartile Range	6	
3	Mean	15.78	1.920
	Median	18.00	
	Std. Deviation	5.761	
	Minimum	4	
	Maximum	22	
	Range	18	
	Interquartile Range	8	



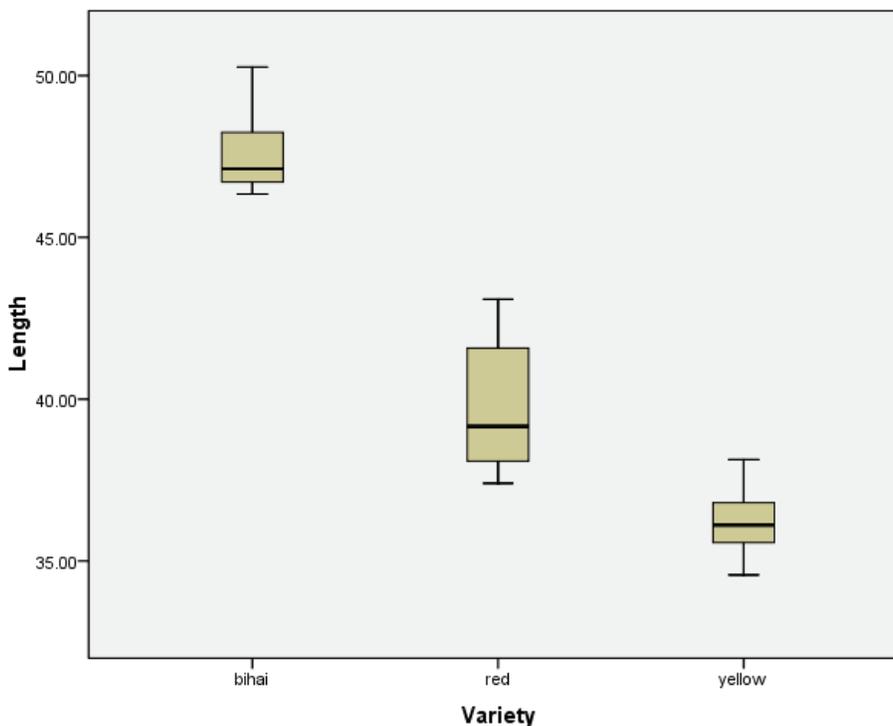
The never logged plots have the largest mean (23.75) and the plots logged 1 year ago the smallest mean (14.08). We see the same pattern in the medians (23 for never logged to 14.5 for the plots logged 1 year ago). This pattern continues when we consider statistics such as the min and max.

We can clearly see the decline in the number of tree species in these plots from the never logged (the top plot) to the plots logged 1 year ago (middle), and those logged 8 years ago (bottom). The minimum value in the plots logged one year ago (2) is a low outlier. Clearly, logging does affect the number of tree species; it also appears that recovery takes a while.

**2.29** Open data file *ta02-01*. We'll use **Analyze**, **Descriptive Statistics**, **Explore** to compute summary statistics and create side-by-side boxplots for each group. Click to enter **Length** in the Dependent List box, then **Variety** in the **Factor List** box. Click **Graphs** and check the box for **Boxplot of data**. **Continue** and **OK** gets our results. (We've again deleted some of the statistics computed from our results below).

Descriptives

Variety			Statistic	Std. Error
Length	bihai	Mean	47.5975	.30322
		Median	47.1200	
		Std. Deviation	1.21288	
		Minimum	46.34	
		Maximum	50.26	
		Range	3.92	
		Interquartile Range	1.60	
red	red	Mean	39.7113	.37507
		Median	39.1600	
		Std. Deviation	1.79876	
		Minimum	37.40	
		Maximum	43.09	
		Range	5.69	
		Interquartile Range	3.62	
yellow	yellow	Mean	36.1800	.25183
		Median	36.1100	
		Std. Deviation	.97532	
		Minimum	34.57	
		Maximum	38.13	
		Range	3.56	
		Interquartile Range	1.37	



We can clearly see the difference in petal lengths for the three types — bihai is clearly the species with the longest petals; yellow has the shortest. All three distributions are somewhat right skewed. There are no outliers in any of the distributions. These plots do not seem to hide any of the features seen in Figure 2.5 (boxplot would tend to hide any gaps in the middle, or indications of bimodality (two peaks in the distribution)).

**2.31** We'd compare different years with percents because the number of births is going to be different; if considerably different, making sense of the actual numbers would be difficult to impossible.

These data are not supplied in a data file. We'll have to enter them. So that we can make a histogram of these data, we enter only the low end of each weight category in a variable we defined as **Weight** and the **Births** in another. Note that each variable has been defined with no decimal places – click on the right side of that box and change from the default of 2.

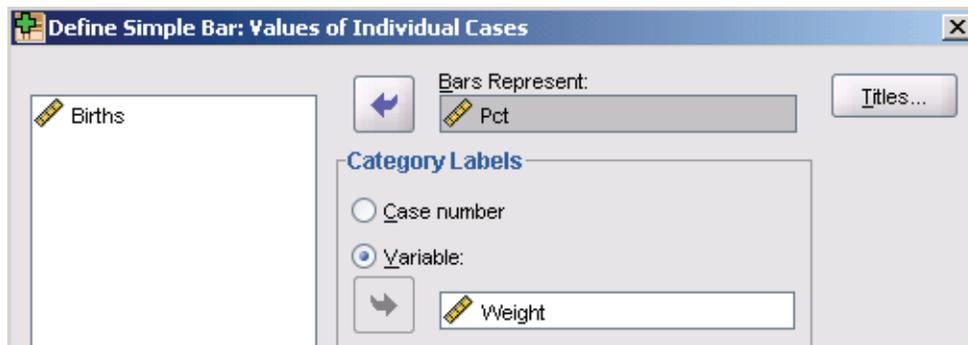
	Name	Type	Width	Decimals
1	Weight	Numeric	8	0
2	Births	Numeric	8	0

	Weight	Births
1	0	6599
2	500	23864
3	1000	31325
4	1500	66453
5	2000	210324
6	1500	748042
7	3000	1596944
8	3500	1114887
9	4000	289098
10	4500	42119
11	5000	4715

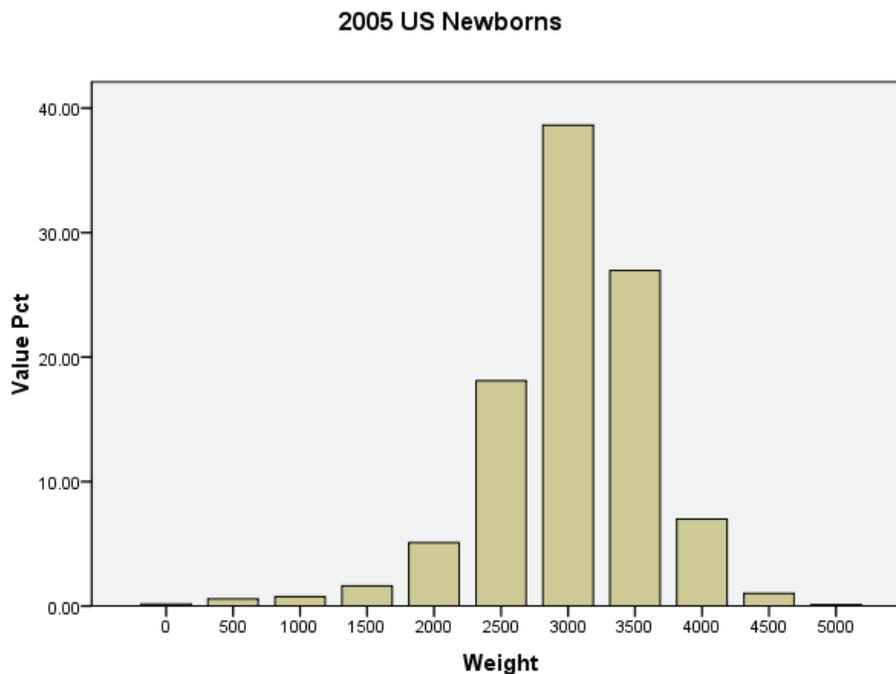
The exercise statement wants us to create the histogram using percents as the y axis values; unfortunately, SPSS won't do this easily. We have (manually) summed all the births – there were 4,143,370. We'll use **Transform, Compute Variable** to create a new variable named **Pct**.



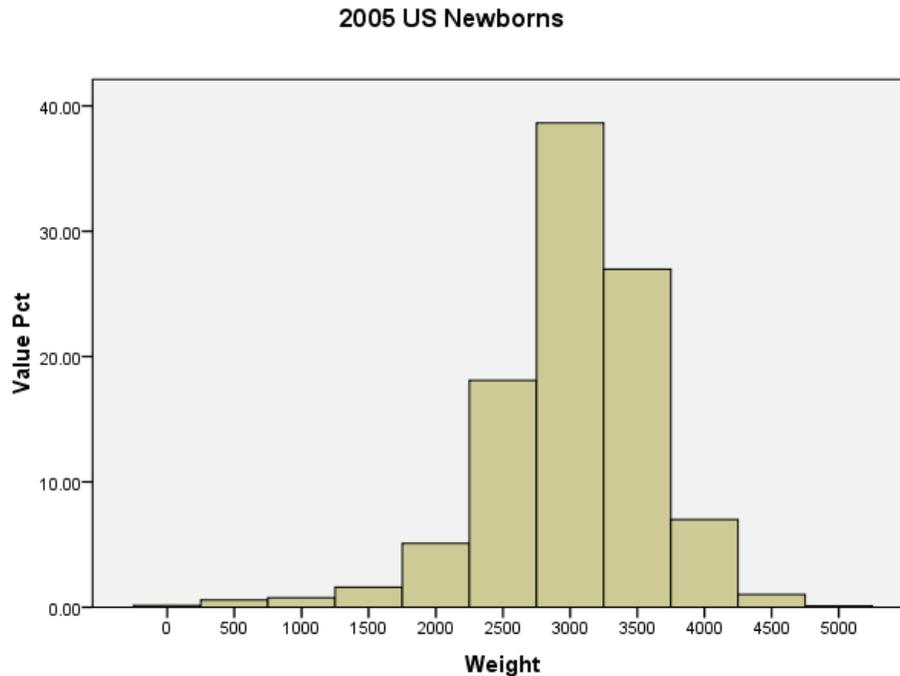
Since the data are already summarized, click **Graphs, Legacy Dialogs, Bar**. We want a **Simple**, where Data in chart are **Values of Individual Cases**. Click **Define** to continue. Click to enter **Pct** for Bars Represent and that Category Labels are the variable **Weight**. Give your graph a descriptive **Title**.



The initial graph is a bar graph (with separated bars) and with the counts on the y axis.



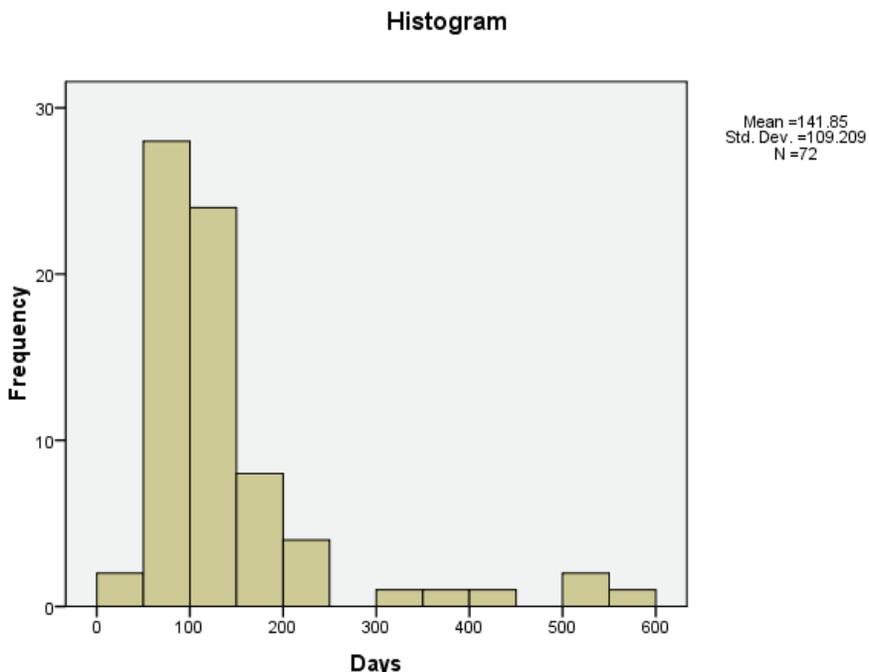
We'll join the bars and change the axis scaling using the Chart Editor. Double click anywhere in the graph to bring up the editor, then double click on any bar in the graph for the Bar properties window. Click on the **Bar options** tab, then move the **Width, Bars** slider to 100% (this connects the bars). **Apply** the change and **Close** the window.



It is clear from this graph that the median falls in the 3,000 to 3,400 gram interval (almost 40% of births are in that interval). It is also clear that  $Q_1$  will be in the 2,500 to 2,999 gram interval;  $Q_3$  will be in the 3,500 to 3,999 gram interval.

**2.35** Open data file *ex02-35*. We'll create a histogram of the survival times and compute summary statistics using **Analyze, Descriptive Statistics, Explore**. Click to enter **Days** as dependent variable. Click **Plots** and ask for a **Histogram**. Click **Statistics** and check the box for **Percentiles** (a cursory look at these data indicates they may be skewed).

		Percentiles						
		5	10	25	50	75	90	95
Weighted Average(Definition 1)	Days	54.95	60.40	82.25	102.50	153.75	247.20	440.80
Tukey's Hinges	Days			82.50	102.50	151.50		



Given this distribution shape, an appropriate numerical summary would be the five-number summary. One advantage of using this procedure is that we have both the graphical summary and both types of numeric summaries available to us.

**2.37** Open data file *ta01-01*. We'll compute summary statistics using **Analyze**, **Descriptive Statistics**, **Descriptives**. Click to enter **PctFor** as the variable, then **OK**.

Descriptive Statistics					
	N	Minimum	Maximum	Mean	Std. Deviation
PctFor	51	1.2	27.2	8.402	6.0539
Valid N (listwise)	51				

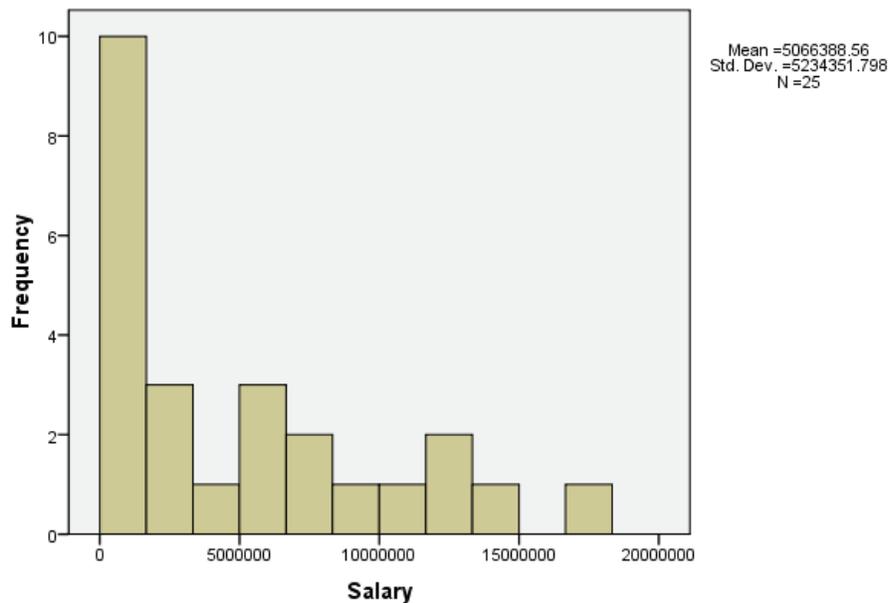
We find the mean (by state) is 8.402%. This is not equal to the nationwide 12.5% because each state is weighted equally here; some states (California, Florida, New Jersey, and New York among them) have many foreign-born residents while others (Alabama, Iowa, Kentucky, Louisiana and others) have very few.

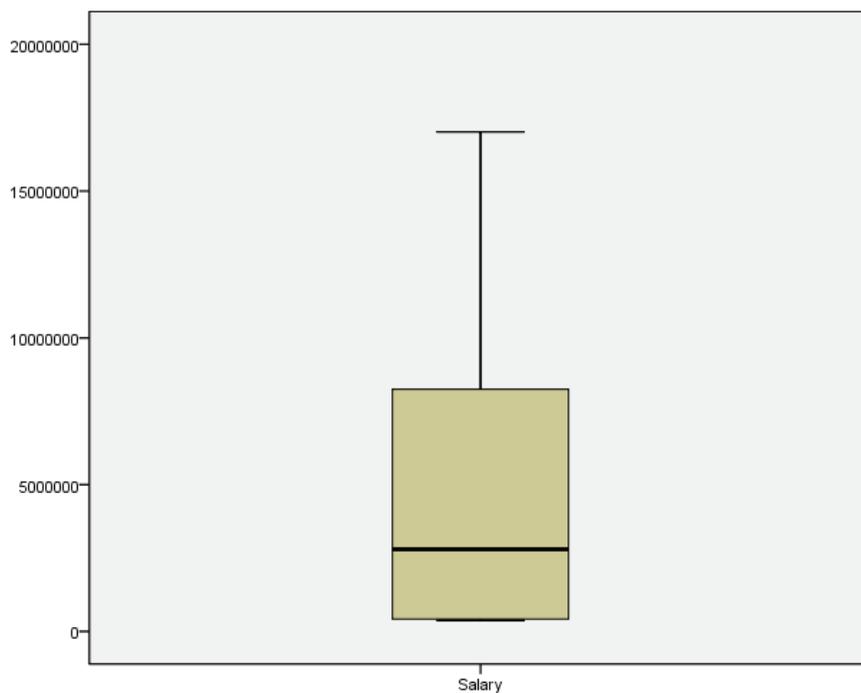
**2.43** Open data file *ta02-02*. We'll create a histogram and boxplot of the salaries and compute summary statistics using **Analyze**, **Descriptive Statistics**, **Explore**. Click to enter **Salary** as the variable; click **Plots** and ask for a **Histogram of the data**. Click **Continue** to return to the main dialog box and **OK** again to perform the actions.

### Descriptives

		Statistic	Std. Error
Salary	Mean	5066388.56	1046870.360
	95% Confidence Interval for Mean		
	Lower Bound	2905754.33	
	Upper Bound	7227022.79	
	5% Trimmed Mean	4691925.91	
	Median	2800000.00	
	Variance	2.740E13	
	Std. Deviation	5234351.798	
	Minimum	380000	
	Maximum	17016381	
	Range	16636381	
	Interquartile Range	8200500	
	Skewness	.905	.464
	Kurtosis	-.382	.902

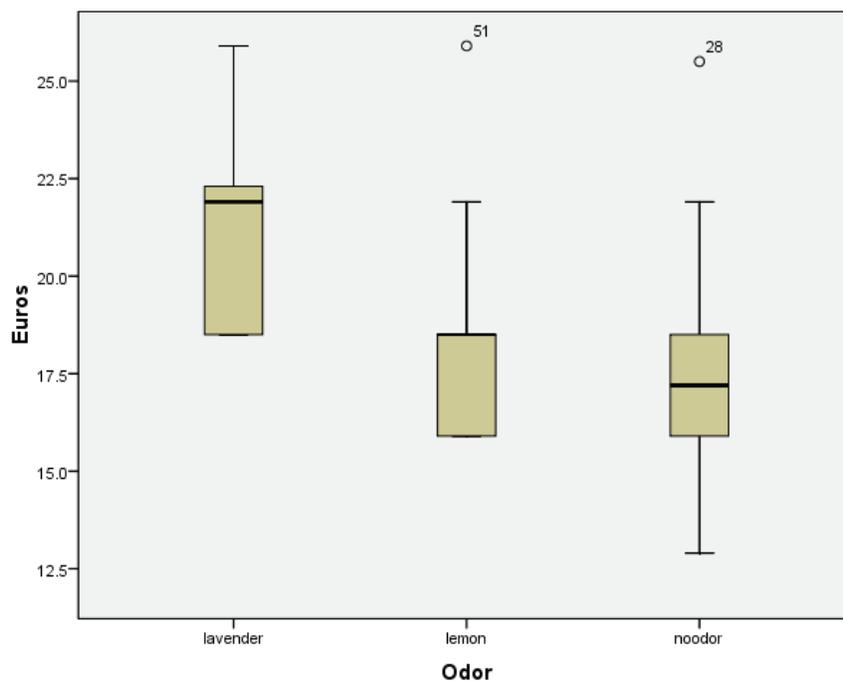
### Histogram





Thirteen of the 25 players earn less than \$4,000,000; the rest are relatively evenly distributed through the range. The boxplot confirms the compression at the low end and also informs us that there are no outliers, at least according to the  $1.5 \times \text{IQR}$  criteria.

**2.45** Open data file *ta02-03*. We'll compute summary statistics for each odor and side-by-side boxplots using **Analyze**, **Descriptive Statistics**, **Explore**. Click to enter **Euros** in the Dependent List and **Odor** as the Factor variable. Click **OK** to perform the actions.



## Descriptives

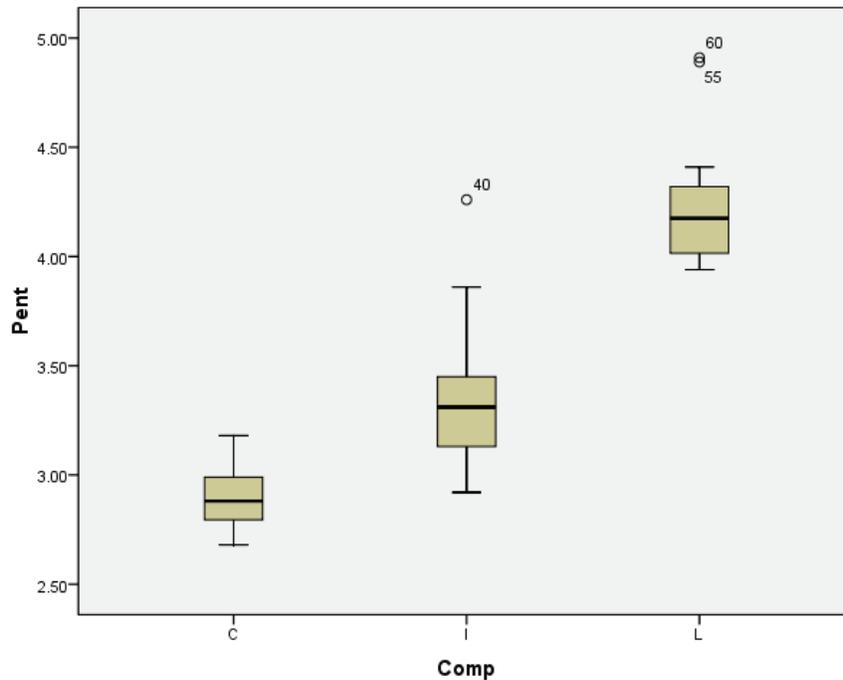
Odor			Statistic	Std. Error
Euros	lavender	Mean	21.123	.4281
		Median	21.900	
		Std. Deviation	2.3450	
		Minimum	18.5	
		Maximum	25.9	
		Range	7.4	
		Interquartile Range	3.9	
	lemon	Mean	18.157	.4192
		Median	18.500	
		Std. Deviation	2.2183	
		Minimum	15.9	
		Maximum	25.9	
		Range	10.0	
		Interquartile Range	2.6	
	noodor	Mean	17.513	.4307
		Median	17.200	
		Std. Deviation	2.3588	
		Minimum	12.9	
		Maximum	25.5	
		Range	12.6	
		Interquartile Range	2.6	

Our boxplots show the differences clearly. We also see that the maximums in both the no odor and lemon distributions are outliers. No line shows in the middle of the box for the lemon scent; checking the summary statistics, we see that the median is the same as  $Q_3$  in this distribution. It appears that if restaurants want their customers to spend more, they should use lavender scents.

**2.47** Open worksheet file *ta02-05*. As in Exercise 2.45 above, we'll compute summary statistics and create side-by-side boxplots to investigate the distributions.

## Descriptives

Comp			Statistic	Std. Error
Pent	C	Mean	2.9075	.03108
		Median	2.8800	
		Std. Deviation	.13898	
		Minimum	2.68	
		Maximum	3.18	
		Range	.50	
		Interquartile Range	.21	
I		Mean	3.3360	.07133
		Median	3.3100	
		Std. Deviation	.31899	
		Minimum	2.92	
		Maximum	4.26	
		Range	1.34	
		Interquartile Range	.33	
L		Mean	4.2315	.06067
		Median	4.1750	
		Std. Deviation	.27134	
		Minimum	3.94	
		Maximum	4.91	
		Range	.97	
		Interquartile Range	.32	



Just as we might suspect, the more compression, the less penetrability there is in the soil. Loose soil has the largest mean (4.232) and median (4.175), while compressed soil has the smallest mean (2.908) and median (2.88). The standard deviation of Intermediate is the largest (0.319), while the standard deviation of compressed soil is the smallest (0.139).

The boxplots of the distributions (Compressed to Loose from top to bottom) also show the change in penetrability with less compression. Except for the outlier in Intermediate soil, there is almost no overlap between that distribution and the one for Loose soil, which also has an outlier.

**2.51** Refer to Exercise 2.43 above. The boxplot of that data indicates no outliers according to the  $1.5 \times \text{IQR}$  criteria. We can confirm this by computing the fences as shown at right. It is clear from the IQR that there are no low end outliers. The upper fence is \$20,925,750 which is larger than the highest salary.

```

8625000-424500
                8200500
8625000+1.5*8200
500                20925750
■

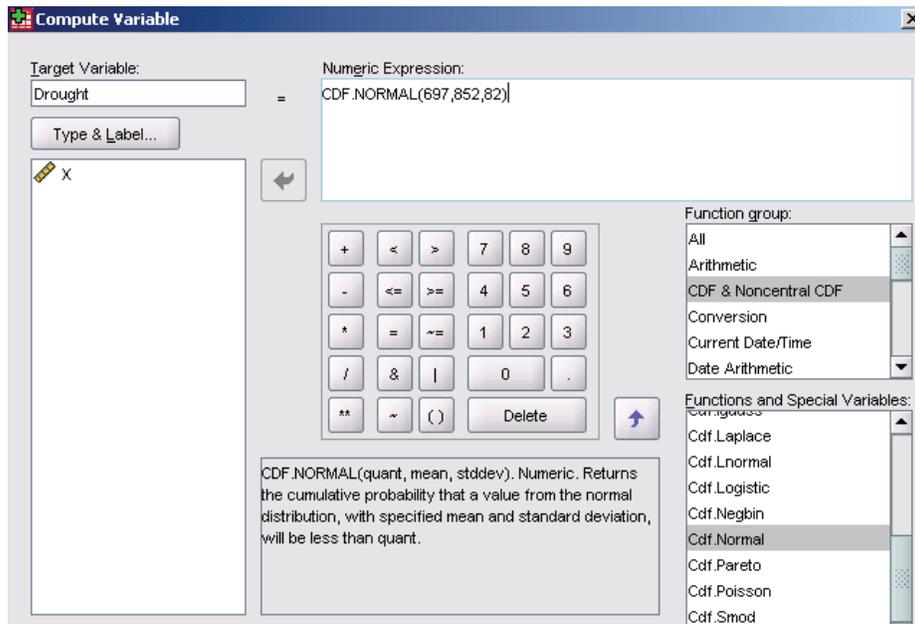
```

### Chapter 3 SPSS Solutions

**3.9** It's inconvenient to use Minitab for a computation such as this. Using a standard calculator, we can easily compute the  $z$ -scores. To compute the  $z$ -scores, we use the formula  $z = (value - \mu) / \sigma$ . Either do the subtraction first, or be sure to use parentheses. A woman six feet (72") tall is 2.96 standard deviations above the mean; the six foot tall man is 0.964 standard deviations above the mean. The woman is *much* taller, relative to other women, than the man is, compared to other men.

```
(72-64)/2.7
2.962962963
(72-69.3)/2.8
.9642857143
```

**3.11** To find the percent of years with less than 697 mm of rain, we use **Transform, Compute Variable**. Locate the **CDF & Noncentral CDF** Function group, then the **CDF.Normal** function in the lower box. Clicking on that will transfer the command shell into the Numeric Expression box. Notice that in the lower center of the box there is a description of the command and its parameters. Enter the parameters as shown, then **OK** computes the probability into the worksheet (as variable **Drought**, here). For more decimal places in your result (remember, the default is two), click on the **Variable view** tab and increase them.



Drought

About 2.9% of all years will have less than 697 mm of rain.

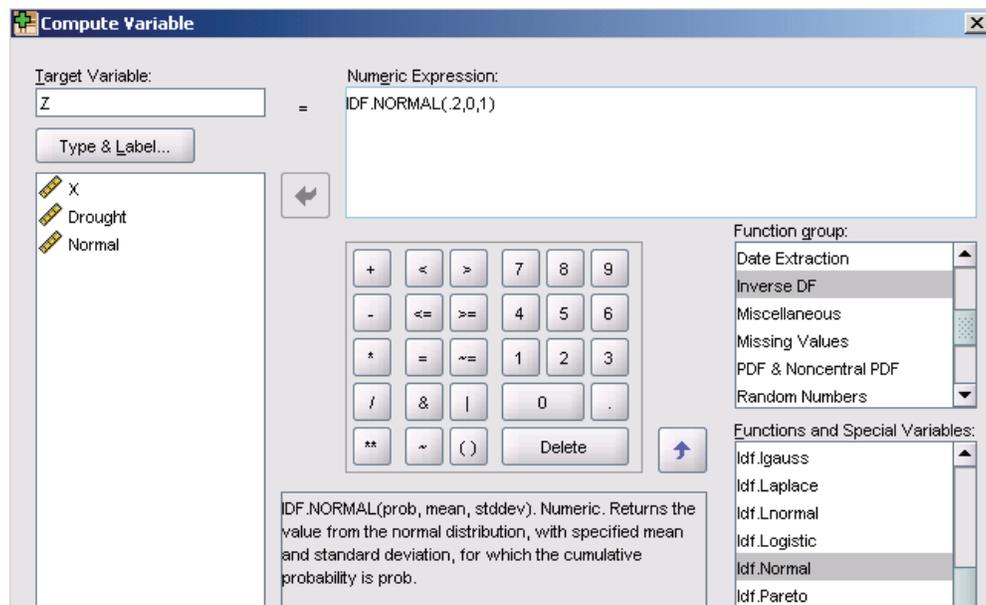
0.0294

To find the percent of “normal rainfall” years (between 683 mm and 1022 mm), we’ll find the cumulative probability for 1022 mm and subtract the cumulative probability of

683. We do this in one combination of CDF.Normal calculations as shown below. About 96.1% of all years will have “normal” rainfall.



**3.13** Here, we are given a relative frequency under the standard Normal curve. We need to find the value of  $z$ . We'll again use **Transform, Compute Variable**. Locate the **Inverse DF** Function group, then the **IDF.Normal** function in the lower box. Clicking on that will transfer the command shell into the Numeric Expression box. Notice that in the lower center of the box there is a description of the command and its parameters. Enter the parameters as shown, then **OK** computes the probability into the worksheet (as variable  $Z$  here).



Z
-0.84

The point  $z$  with 20% of the area below it is  $z = -0.842$ . We repeat for part (b) using 0.6 as the area to the left of the point (since 40% of the observations are above it). This point is  $z = 0.253$ .

**3.29** As with Exercise 3.13 above, use **Transform, Compute Variable**, we want the **Inverse DF** and **IDF.Normal**. As before, enter the area to the left of the desired point on the curve (0.8), the value of the mean (0) and standard deviation (1). This point is  $z = 0.842$ .

Z
0.84

Part (b) asks for the point with 35% of all observations above it; this means that 65% = 0.65 are below it. This point is  $z = 0.39$ .

Z
0.39

**3.31** To find the proportion of rainy days that meet the “acid rain” criteria, we use **Transform, Compute Variable**. Locate the **CDF & Noncentral CDF** Function group, then the **CDF.Normal** function in the lower box. Clicking on that will transfer the command shell into the Numeric Expression box. Notice that in the lower center of the box there is a description of the command and its parameters. Enter the parameters as shown, then **OK** computes the probability into the worksheet (as variable **Acid**, here). For more decimal places in your result (remember, the default is two), click on the **Variable view** tab and increase them. At this location 22.9% of days will qualify as “acid rain” days.

Acid
0.2294

**3.33** To find the proportion of slots that meet specifications, we’ll use **Transform, Compute Variable** and find the cumulative probability for 0.878 inch and subtract the cumulative probability of 0.872 inch. We do this in one combination of CDF.Normal calculations as shown below. About 98.76% of slots will meet the specifications.

Slots
0.9876

**3.35** This problem refers to the information given about 2008 model vehicles. They had mean 18.7 mpg and standard deviation 4.3 mpg. We want to know the area to the left of the Chevy Malibu (with 25 mpg). Use **Transform, Compute Variable** and find the cumulative probability for the Malibu as below. 92.86% of 2008 cars had worse mileage than the Chevy Malibu.

Compute Variable		Malibu
Target Variable:	Numeric Expression:	
Malibu	= CDF.NORMAL(25,18.7,4.3)	0.9286

**3.37** To find the quartiles, we want the points with (respectively) 25% and 75% of the area below them. We can find these values using **Transform, Compute Variable**. We want the **Inverse DF** and **IDF.Normal**. As before, enter the area to the left of the desired point on the curve (0.25, then 0.75), the value of the mean (18.7) and standard deviation (4.3). This point is  $z = 0.842$ . We find that  $Q_1$  (the 25<sup>th</sup> percentile) is 15.80 mpg and  $Q_3$  (the 75<sup>th</sup> percentile) is 21.60 mpg.

Compute Variable		Q1
Target Variable:	Numeric Expression:	
Q1	= IDF.NORMAL(.25,18.7,4.3)	15.80

Compute Variable		Q3
Target Variable:	Numeric Expression:	
Q3	= IDF.NORMAL(.75,18.7,4.3)	21.60

**3.39** The percentile corresponds to the area to the left of the value of interest. We find this using **Transform, Compute Variable** and find the cumulative probability for the Jacob as below. We see that Jacob is not quite at the 15<sup>th</sup> percentile (his is 14.9).

Compute Variable		Percentile
Target Variable:	Numeric Expression:	
Percentile	= CDF.NORMAL(16,21.2,5)	0.149

**3.41** We want to know what proportion of women are taller than the average man (69.3"). We'll use **Transform, Compute Variable** but subtract the percent of women shorter than 69.3" from 1 to find the proportion *taller* than 69.3". Be sure to use the values for the *women's* distribution: mean (64), and the standard deviation (2.7). We see that not quite 2.5% (2.48%) of women should be taller than the average man.

Compute Variable		Taller
Target Variable:	Numeric Expression:	
Taller	= 1-CDF.NORMAL(69.3,64,2.7)	0.0248

**3.43** To find the proportion of students scoring at least 750, we'll use **Transform, Compute Variable** and subtract the proportion scoring less than 750 from 1 as we did in Exercise 3.41.

**Compute Variable**

Target Variable: Over750 = Numeric Expression: 1-CDF.NORMAL(750,533,116)

**Compute Variable**

Target Variable: Over750 = Numeric Expression: 1-CDF.NORMAL(750,499,110)

Over750  
0.0307

Over750  
0.0113

We see that 3.1% of men scored at least 750 while only 1.1% of women did this well.

**3.47** To find the proportion scoring higher than 27, divide the given numbers; to find the proportion scoring 27 or more, add the number that scored 27 to the first. We find that 11.5% scored higher than 27, while 15.3% scored at least 27. To compare this with the Normal computation, use **CDF.Normal** to find the proportion scoring at least than 27 by subtracting the proportion scoring less than 27 from 1.

```
149164/1300599
.1146886934
(149164+50310)/1300599
.1533708699
```

**Compute Variable**

Target Variable: Over27 = Numeric Expression: 1-CDF.NORMAL(27,21.2,5)

Over27  
0.1230

We would expect 12.3% to score at least 27 if the scores were exactly Normal.

**3.49** Open worksheet file *ex03-49*. We'll create a histogram of the lengths and compute summary statistics using **Analyze, Descriptive Statistics, Explore**. Click to enter variable **Length** in the Dependent List. Click **Plots** and be sure the **Histogram** box is checked. To find the quartiles of this distribution, click **Statistics** and ask for **Percentiles**.

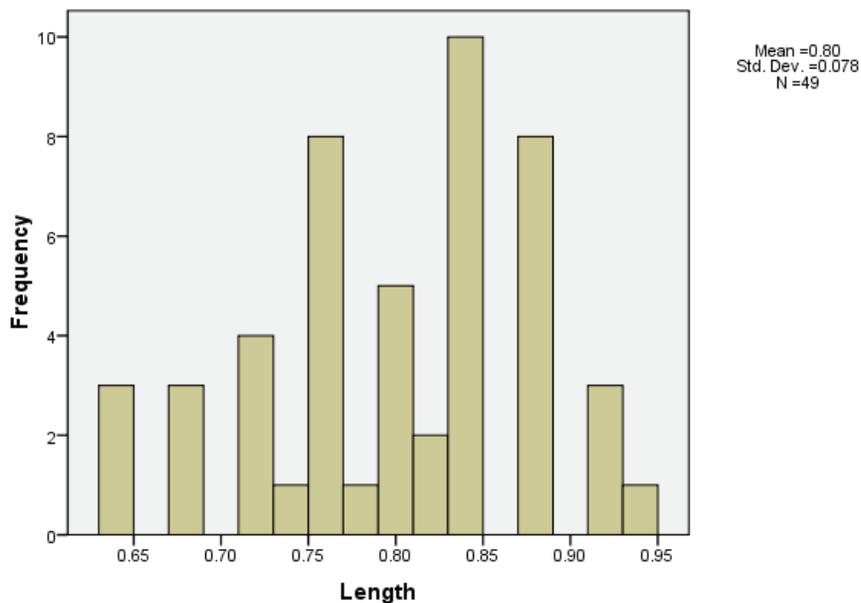
Percentiles

		Percentiles						
		5	10	25	50	75	90	95
Weighted Average(Definition 1)	Length	.6400	.6800	.7600	.8000	.8600	.8800	.9200
Tukey's Hinges	Length			.7600	.8000	.8400		

## Descriptives

		Statistic	Std. Error
Length	Mean	.8004	.01116
	Median	.8000	
	Variance	.006	
	Std. Deviation	.07815	
	Minimum	.64	
	Maximum	.94	
	Range	.30	
	Interquartile Range	.10	
	Skewness	-.361	.340
	Kurtosis	-.566	.668

## Histogram



This distribution actually looks a bit skewed left (other windows also show this same general shape); there are no outliers. The mean ( $\bar{x} = 0.800$ ) is the same (within rounding) as the median (Med = 0.8); the standard deviation is  $s = 0.078$ ; the quartiles are  $Q_1 = 0.76$  and  $Q_3 = 0.86$ . The distances to the quartiles from the median (0.04 and 0.06) are roughly similar. These all suggest the distribution is rather symmetric.

In part (c), we want to find the percent of observations expected to be between the two quartiles (0.76 and 0.86) if the distribution is Normal. We'll use **CDF.Normal** to find the proportion by subtracting the proportion less than 0.76 from the proportion less than 0.86.



About 47.5% of all observations between 0.76 and 0.86. To find what actual proportion lies between these values, sort the list using **Data, Sort Cases**. Enter the variable name **Length** in both the **Sort by** box. Click **OK**.

10	0.72
11	0.74
12	0.76
13	0.76
14	0.76
36	0.84
37	0.84
38	0.88
39	0.88
40	0.88

Examining the worksheet after the sort, we find there are 11 values less than 0.76 and 12 values greater than 0.86; that means  $(49 - 23)/49 = 53.1\%$  of the values are between the quartiles.

**3.51** Open worksheet file *ta02-05*. We want stemplots of the data for both loose and intermediate compression. Use **Analyze, Descriptive Statistics, Explore** and enter **Pent** as the Dependent variable and **Comp** as the Factor.

Pent Stem-and-Leaf Plot for  
Comp= I

Frequency	Stem & Leaf
2.00	2 . 99
14.00	3 . 01111112333444
3.00	3 . 568
1.00	Extremes (>=4.3)
Stem width: 1.00	
Each leaf: 1 case(s)	

Pent Stem-and-Leaf Plot for  
Comp= L

Frequency	Stem & Leaf
4.00	39 . 4689
2.00	40 . 03
5.00	41 . 12369
3.00	42 . 079
2.00	43 . 04
2.00	44 . 11
2.00	Extremes (>=4.89)
Stem width: .10	
Each leaf: 1 case(s)	

We see below that both of these distributions are not Normal; they are skewed right with high outliers (indicated as Extremes).

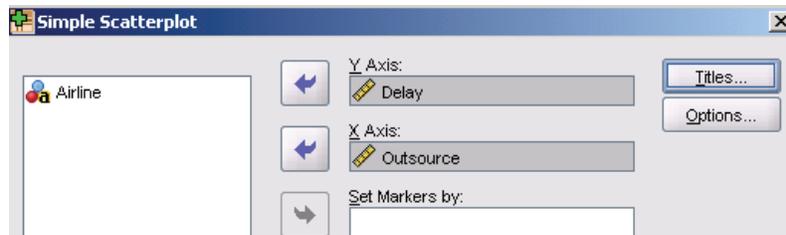
**3.53** We'll find the  $z$ -scores corresponding to the quartiles using **Transform, Compute Variable**, and ask for the **IDF.Normal**. We specify area to the left (0.25) of  $Q_1$ , the mean (0) and standard deviation (1). Since the Normal distribution is symmetric, we'll find only  $Q_1$ . ( $Q_3$  will have the same value, but a positive number).

Compute Variable	
Target Variable:	Numeric Expression:
Q1	IDF.NORMAL(.25,0,1)

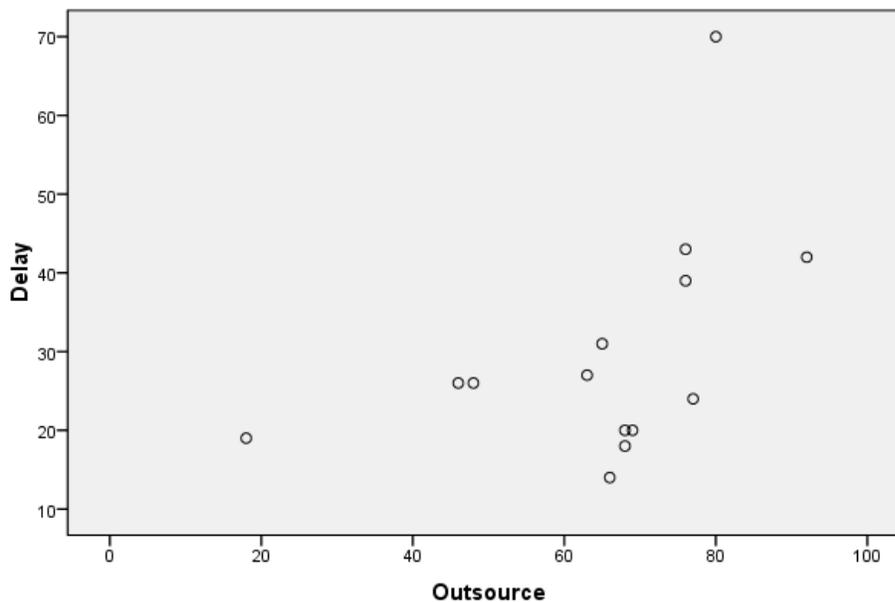
Q1
-0.67

## Chapter 4 SPSS Solutions

4.5 Open data file *ex04-05*. To make a scatterplot of these data, click **Graphs, Legacy Dialogs, Scatter/Dot**. On the first dialog box, select **Simple Scatter** (the default) and **Define**. Click to Enter **Delay** as the Y variable and **Outsource** as the X variable. Click **Titles** to give your graph a title, then **Continue** and **OK** to generate the graph.



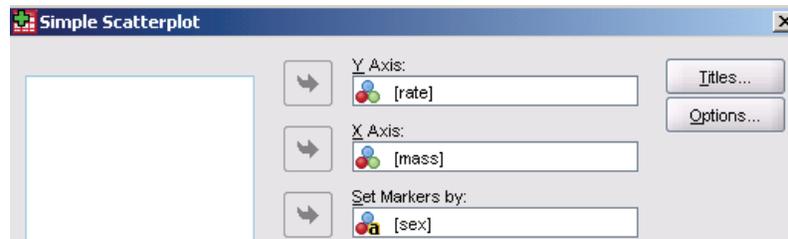
**Airline Outsourcing and Delays**



4.7 The plot shown above shows a positive association because it generally rises from left to right. The high outlier is Hawaiian at 70%. There is also a low outlier (ATA) at (18, 19). Ignoring these, the plot is very roughly linear. The relationship is not particularly strong.

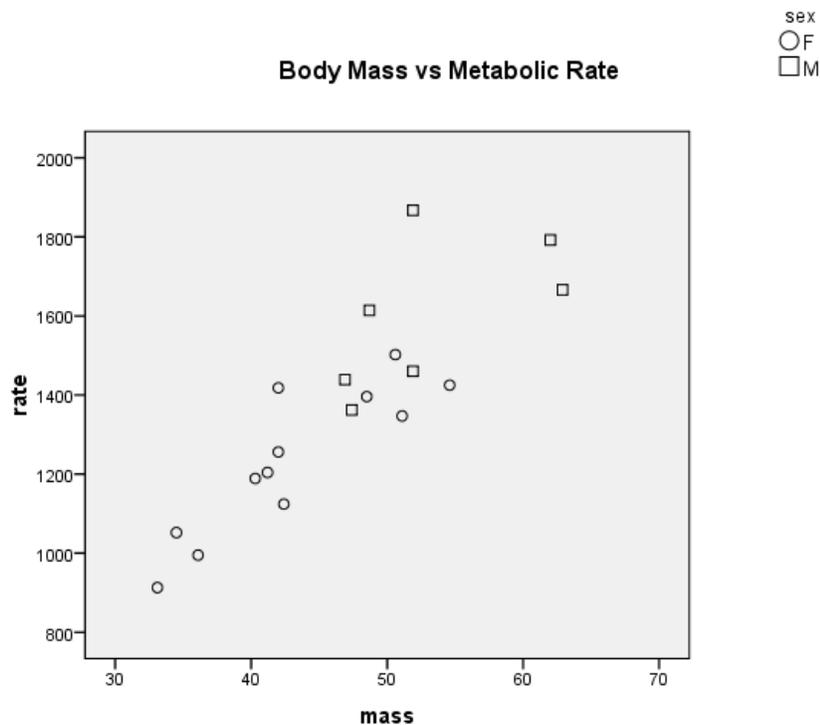
4.9 To create a scatterplot of these data, first open data file *ex04-09.por*. Click **Graphs, Legacy Dialogs, Scatter/Dot**. Select Simple Scatter and click **Define**. Click to move

**rate** to the Y axis and **mass** to the X axis. Set markers by **sex** will give each sex a different plotting symbol. Click to add **Titles**; then **Continue** and **OK**.

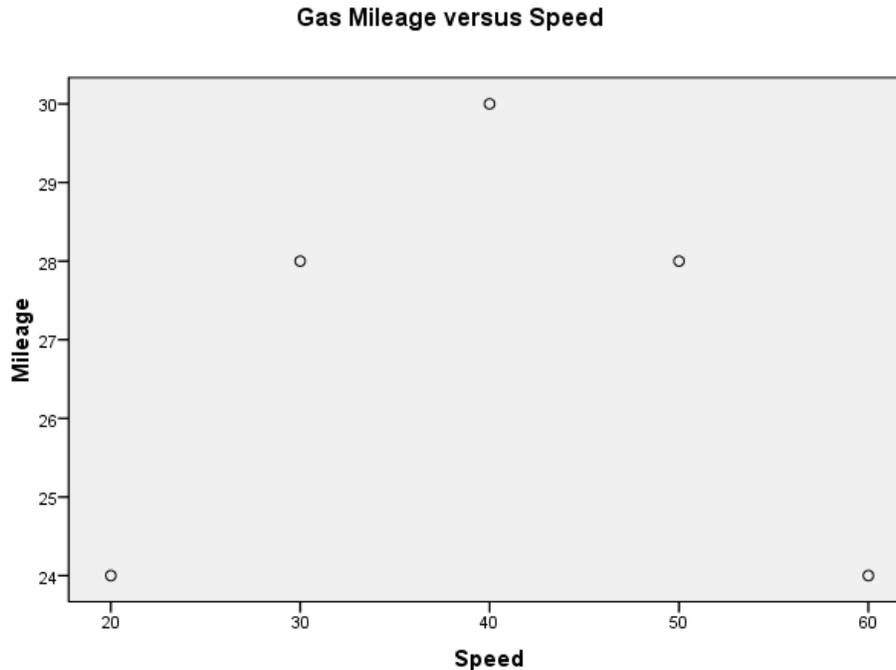
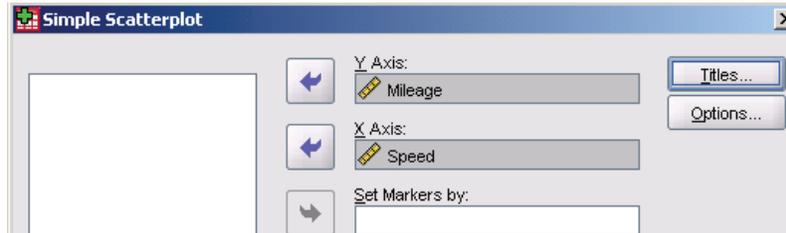


The initial plot marks females with blue circles and males with green circles. To change that (for a black-and-white printer, for example), right click in the graph in the output window and select Edit Contents in Separate Window. Click in the graph again to bring up the Properties box. If necessary, click on the **Variables** tab. Click the Styles button for **sex** and change the style from **Color** to **Shape** (or size). Click **Apply** to actually make the change. In the graph below, we now have Females represented by circles and Males by squares.

The relationship for both genders is positive; but stronger for the females (less scatter) than for the males. Notice the females cluster in the lower left portion of the graph while the males (with generally larger body masses) are in the upper right.



**4.13** We entered the data into two columns called **Speed** and **Mileage**. To make a scatterplot of these data, click **Graphs, Legacy Dialogs, Scatter/Dot**. We want a **Simple Scatterplot** (the default), so click **Define**. Click to Enter **Mileage** as the Y variable and **Speed** as the X variable. Click **Titles** to give your graph a title, then **Continue** and **OK** to generate the graph.

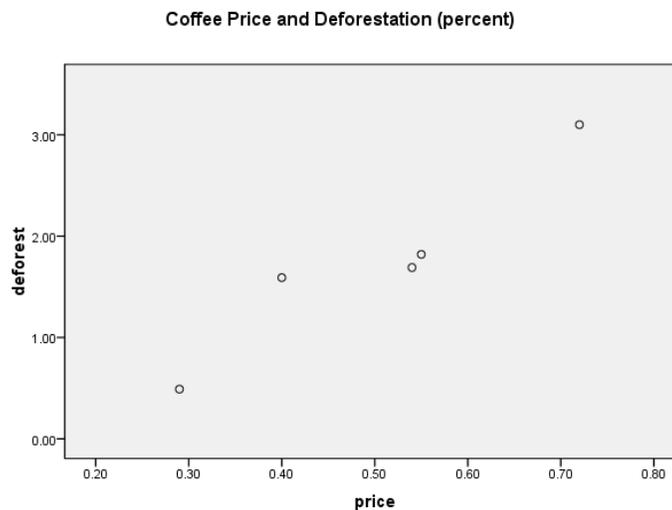


The plot is not linear – it is curved. To calculate the correlation, click **Analyze, Correlate, Bivariate**. Click to enter both variables (you can hold down the Shift key and highlight both names to only **Select** once), then **OK**.

		Speed	Mileage
Speed	Pearson Correlation	1.000	.000
	Sig. (2-tailed)		1.000
	N	5.000	5

The correlation is 0 because a linear model does not make sense here (the best “model” has a slope of 0).

**4.27** Since the question is whether higher coffee prices lead to more forest clearing, coffee price is the explanatory variable. There is no data set for this Exercise, so define variable names and enter the data. Make a scatterplot using **price** as the explanatory variable by clicking **Graphs, Legacy Dialogs, Scatter/Dot**. Select Simple Scatter and click **Define**. Click to move **price** to the Y axis and **deforest** to the X axis. Click to add **Titles**; then **Continue** and **OK**. The form is roughly linear, but there is a rather horizontal region in the center.

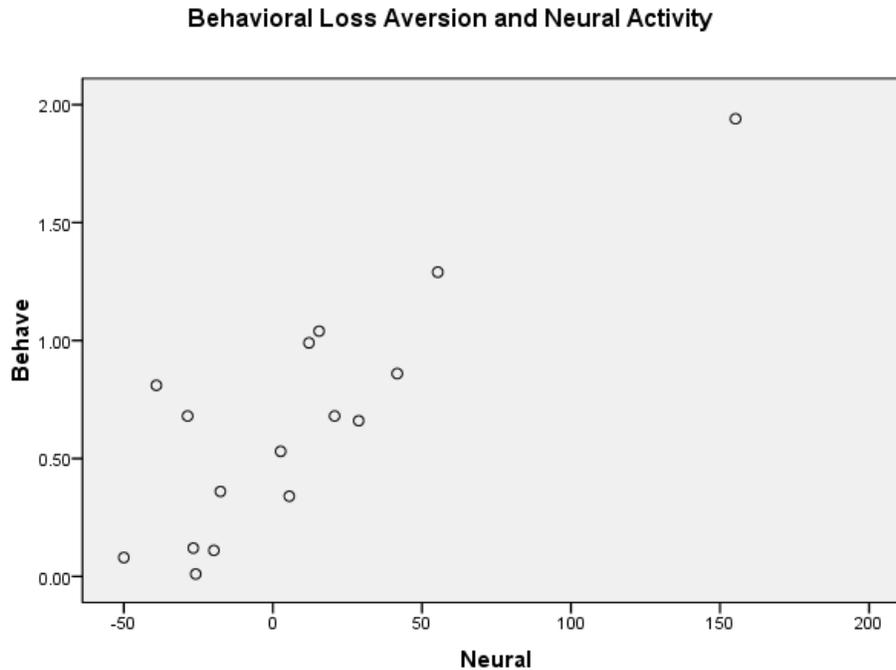


To find the correlation, click **Analyze, Correlate, Bivariate**. Move the two variable names into the box, and then click OK to find the correlation between coffee price and deforestation is a very strong 0.955. The correlation is  $r = 0.955$ . The data do support the idea that higher coffee prices lead to loss of forest. If the units were changed,  $r$  will not change, because it is based on  $z$ -scores.

**Correlations**

		price	deforest
price	Pearson Correlation	1.000	.955
	Sig. (2-tailed)		.011
	N	5.000	5

**4.29** Open data file *ex04-29*. To make a scatterplot of these data, click **Graph, Legacy Dialogs, Scatter/Dot**. The default is **Simple**. Click **Define** to continue on to the plot definition dialog box. Click to Enter **Behave** as the Y axis variable and **Neural** as the X axis variable. Click **Titles** to give your graph a title, then **Continue** and **OK** to generate the graph.



We clearly see the outlier at the upper right (the last data point with neural 155.2). The plot is fairly linear, although it will be less so if the outlier is removed – visually, this point makes the linear relationship look stronger. To find the correlation coefficient, use **Analyze, Correlate, Bivariate**. Click to enter both variables (you can hold down the Shift key and highlight both names to only use the arrow once), then **OK**.

**Correlations**

		Neural	Behave
Neural	Pearson Correlation	1.000	.849**
	Sig. (2-tailed)		.000
	N	16.000	16
Behave	Pearson Correlation	.849**	1.000
	Sig. (2-tailed)	.000	
	N	16	16.000

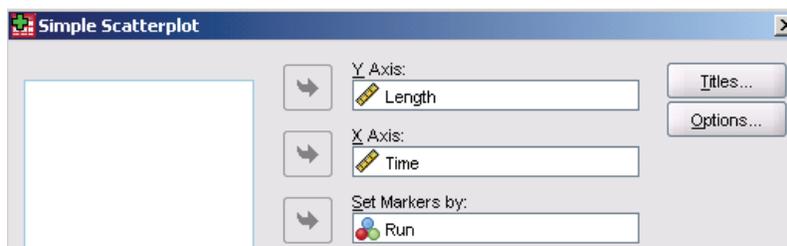
\*\* . Correlation is significant at the 0.01 level (2-tailed).

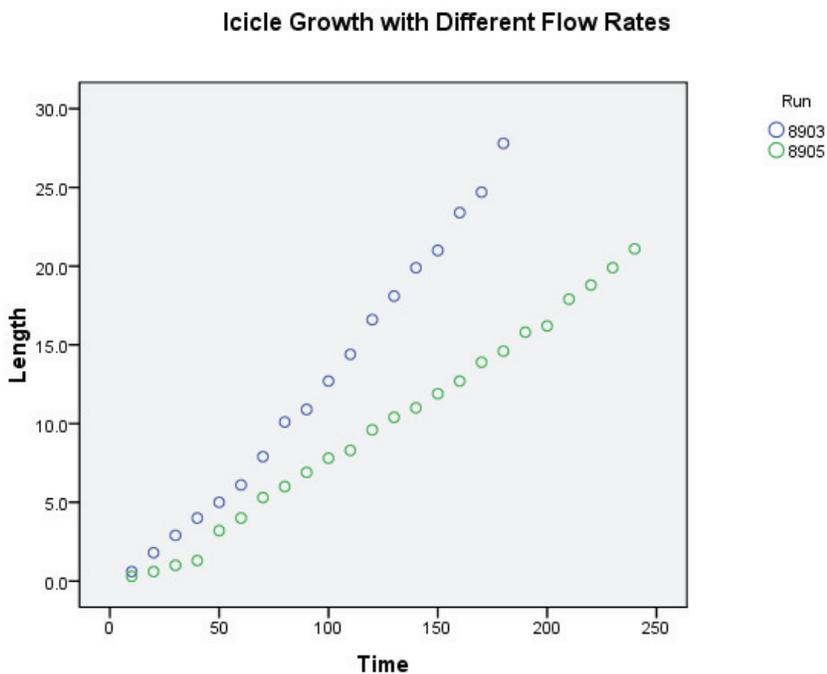
With the outlier included, the correlation is  $r = 0.849$ . Now, delete the last data value from *each* list and recompute the regression and correlation. With the outlier deleted, the correlation has weakened—to  $r = 0.701$ .

		Neural	Behave
Neural	Pearson Correlation	1.000	.701**
	Sig. (2-tailed)		.004
	N	15.000	15
Behave	Pearson Correlation	.701**	1.000
	Sig. (2-tailed)	.004	
	N	15	15.000

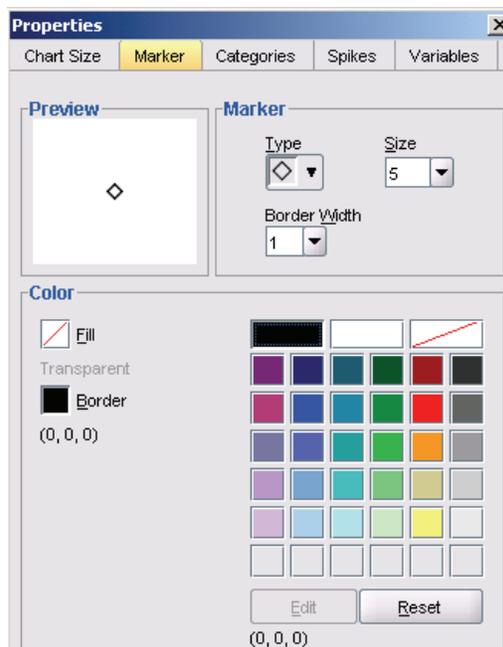
\*\* . Correlation is significant at the 0.01 level (2-tailed).

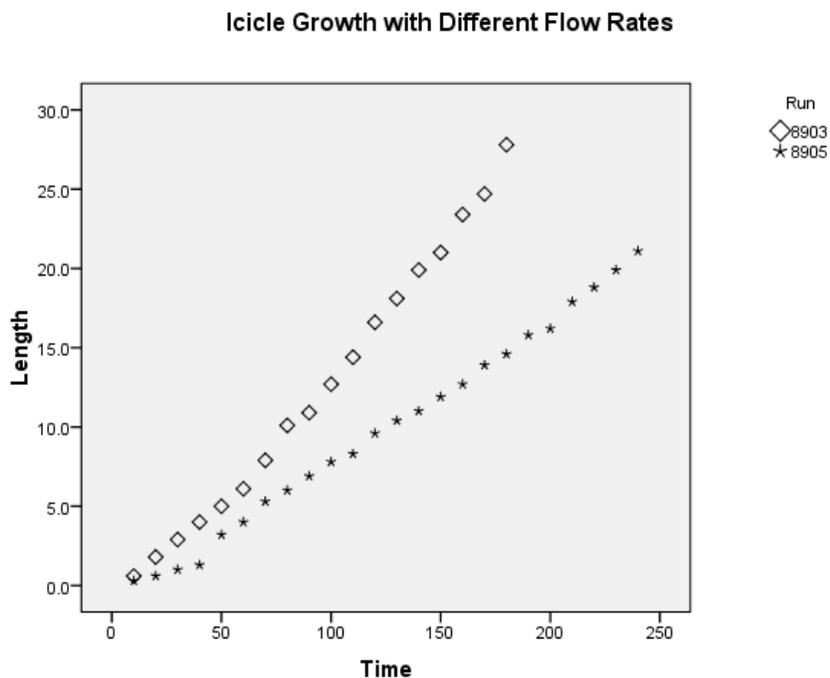
**4.31** Open file *ta04-02* To make a scatterplot with different symbols, click **Graphs**, **Legacy Dialogs**, **Scatter/Dot**. The default is Simple Scatter, so click **Define** to continue. Click to enter **Length** as the Y variable and **Time** as the X variable. Enter **Run** as the Set Markers By variable, then click **Titles** to give your graph a descriptive title. **Continue** and **OK** generates the graph.





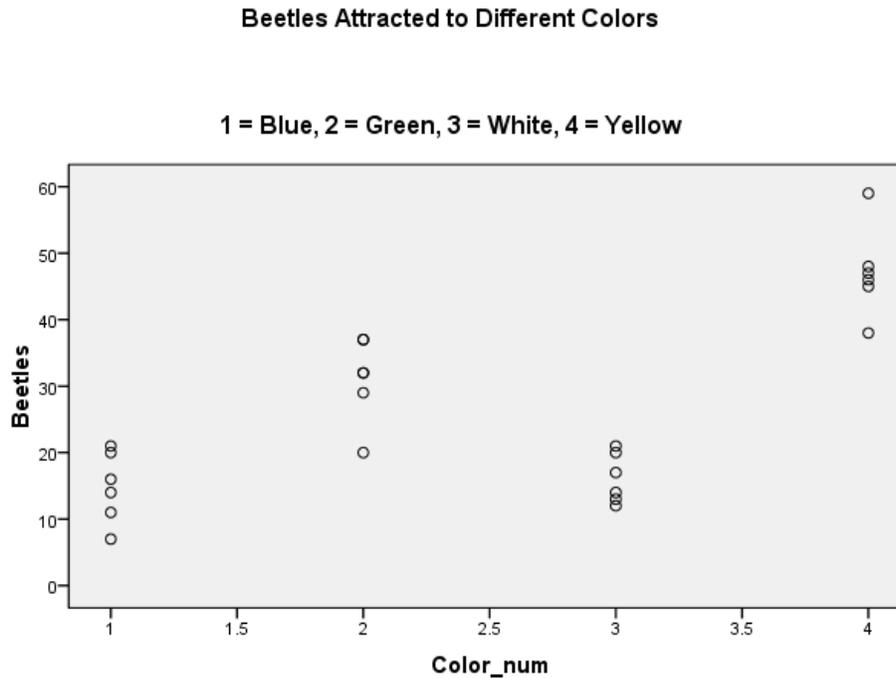
SPSS by default gives both circular marks with different colors (blue for 8903 and green for 8905) to distinguish them. If you have only a black-and-white printer, this is undesirable. To change the graphing symbols, double click in the output graph to bring up the Chart Editor. Double-click on one of the legend symbols for the Properties dialog box. Use the Marker Type drop-down to select a new marker shape; you can also change the color and fill for the symbol. Click **Apply** and **Close** the dialog box. When you are finished changing the markers, **Close** the Chart Editor. Our finished graph is below.





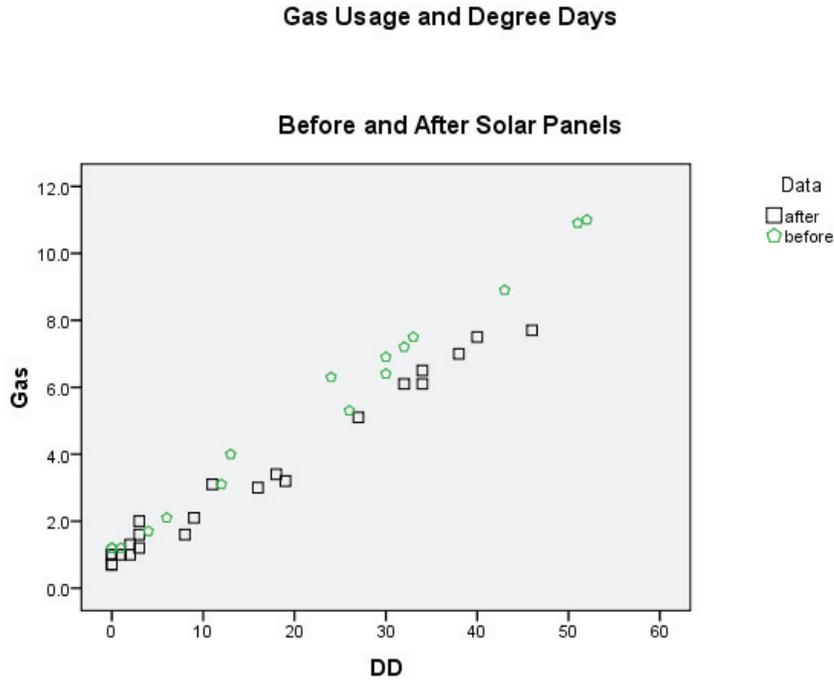
The 8903 icicles (the slower flow rate) grew at a faster rate than the 8905 icicles. Both growth patterns are pretty linear.

**4.33** Open data file *ex04-33*. SPSS will not let you use categorical variables (the colors) in creating a scatterplot. Define a new variable to be a “stand-in” for the color name. We have called our variable `Color_num` and used values 1 = Blue through 4 = Yellow. To make a scatterplot of these data, click **Graph**, **Legacy Dialogs**, **Scatter/Dot**. The default is **Simple**. Click **Define** to continue on to the plot definition dialog box. Click to Enter **Beetles** as the Y axis variable and **Color\_num** as the X axis variable. Click **Titles** to give your graph a title, then **Continue** and **OK** to generate the graph.



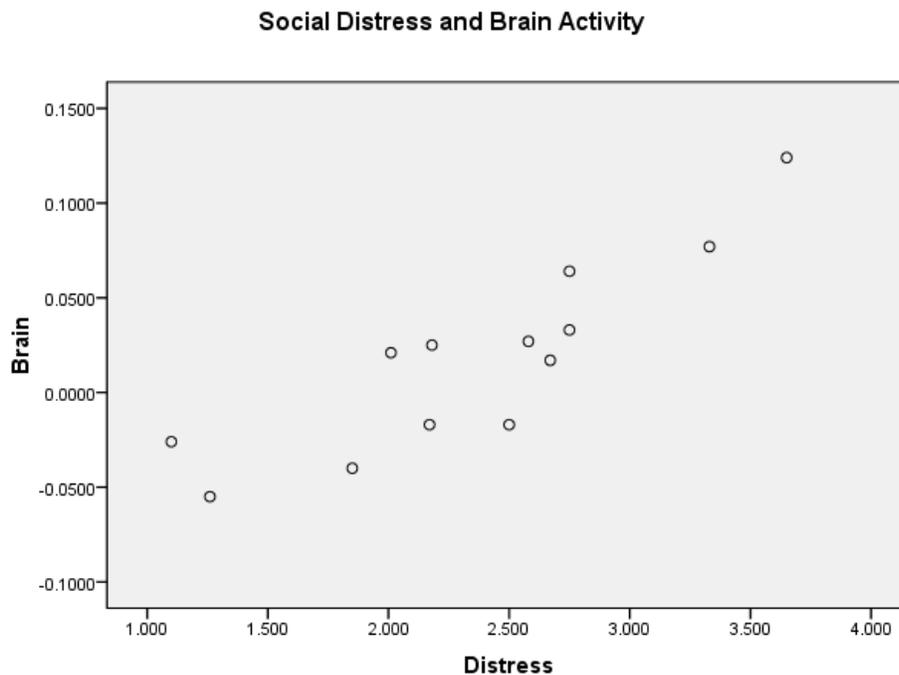
Clearly, Yellow attracts the most beetles. We can't speak of correlation here because color is not a numeric variable (our numeric values were merely chosen by convenience).

**4.43** Open data file *ex04-43*. We start by plotting both sets of data on the same graph. To make a scatterplot of these data, click **Graph, Legacy Dialogs, Scatter/Dot**. The default is **Simple**. Click **Define** to continue on to the plot definition dialog box. Enter **Gas** as the Y variable and **DD** as the X variable; **Data** is used to Set Markers by. Be sure to give your graph an appropriate **Titles**, then OK generates the initial graph. Double click in the graph for the Chart Editor, and proceed just as in Exercise 4.31 to change the plotting symbols.



The solar panels have lowered the gas usage, since the squares are below the pentagons.

**4.49** The data have been entered into two variables called **Distress** and **Brain**. To make a scatterplot of these data, click **Graph, Legacy Dialogs, Scatter/Dot**. The default is **Simple**. Click **Define** to continue on to the plot definition dialog box. Enter **Brain** as the Y variable and **Distress** as the X variable; be sure to give your graph a **Titles**, then **OK** for the plot.



To calculate the correlation, click **Analyze, Correlate, Bivariate**. Click to enter both variables (you can hold down the Shift key and highlight both names to only use the arrow key once), then **OK**.

**Correlations**

		Distress	Brain
Distress	Pearson Correlation	1.000	.878**
	Sig. (2-tailed)		.000
	N	13.000	13
Brain	Pearson Correlation	.878**	1.000
	Sig. (2-tailed)	.000	
	N	13	13.000

\*\* . Correlation is significant at the 0.01 level (2-tailed).

The data show an increase in brain activity with increased distress. The correlation is  $r = 0.878$ , which is strong and positive.

## Chapter 5 SPSS Solutions

**5.3** Open data file *ex05-03*. We can find the mean and standard deviation for both variables using **Analyze**, **Descriptive Statistics**, **Descriptives**. Click to enter both variables, then **OK**.

	N	Minimum	Maximum	Mean	Std. Deviation
Ranges	6	1	5	3.50	1.378
Days	6	4	46	31.33	16.133
Valid N (listwise)	6				

To find the correlation, use **Analyze**, **Correlate**, **Bivariate**. Again, click to enter both variables and **OK**.

		Ranges	Days
Ranges	Pearson Correlation	1.000	.962**
	Sig. (2-tailed)		.002
	N	6.000	6
Days	Pearson Correlation	.962**	1.000
	Sig. (2-tailed)	.002	
	N	6	6.000

\*\* . Correlation is significant at the 0.01 level (2-tailed).

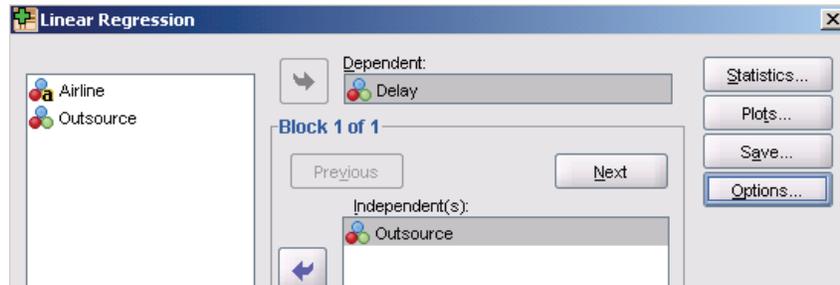
To “hand calculate” the slope, we have  $b = r \frac{s_y}{s_x} = .962 \frac{16.133}{1.378} = 11.263$ . Similarly, we find the intercept as  $a = \bar{y} - b\bar{x} = 31.33 - 11.263 * 3.5 = -8.091$ .

To check this against the software calculation, compute the regression using **Analyze**, **Regression**, **Linear**. Click to enter **Days** as the Dependent variable and **Ranges** as the Independent variable. **OK** performs the calculations. We see our results agree to within rounding.

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-8.088	5.917		-1.367	.243
	Ranges	11.263	1.591	.962	7.080	.002

a. Dependent Variable: Days

**5.11** The scatterplot was created in the solution to Exercise 4.5. See that solution for details. We'll compute the correlation and regression with all the data points, then delete Hawaiian and recomputed to see its influence. Click **Analyze, Regression, Linear**. (There is also an option of **Analyze, Correlate, Bivariate** that computes the correlation but using this would mean you still have to perform the regression.) Click to enter **Delay** as the Dependent and **Outsource** as the Independent. **OK** performs the regression.



**Model Summary**

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.476 <sup>a</sup>	.227	.163	13.413

a. Predictors: (Constant), Outsource

**Coefficients<sup>a</sup>**

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	4.731	13.893		.341	.739
	Outsource	.387	.206	.476	1.877	.085

a. Dependent Variable: Delay

The regression equation becomes  $\text{DelayPct} = 4.731 + 0.387 * \text{OutsourcePct}$ . The correlation is  $r = 0.476$ . Based on this line, an airline with 76% outsourcing should have about  $4.731 + 0.387 * 76 = 34.1\%$  delays.

Now, delete the Hawaiian Airlines data values. Use the same set of commands to find the new equation.

**Model Summary**

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.484 <sup>a</sup>	.234	.164	8.606

a. Predictors: (Constant), Outsource

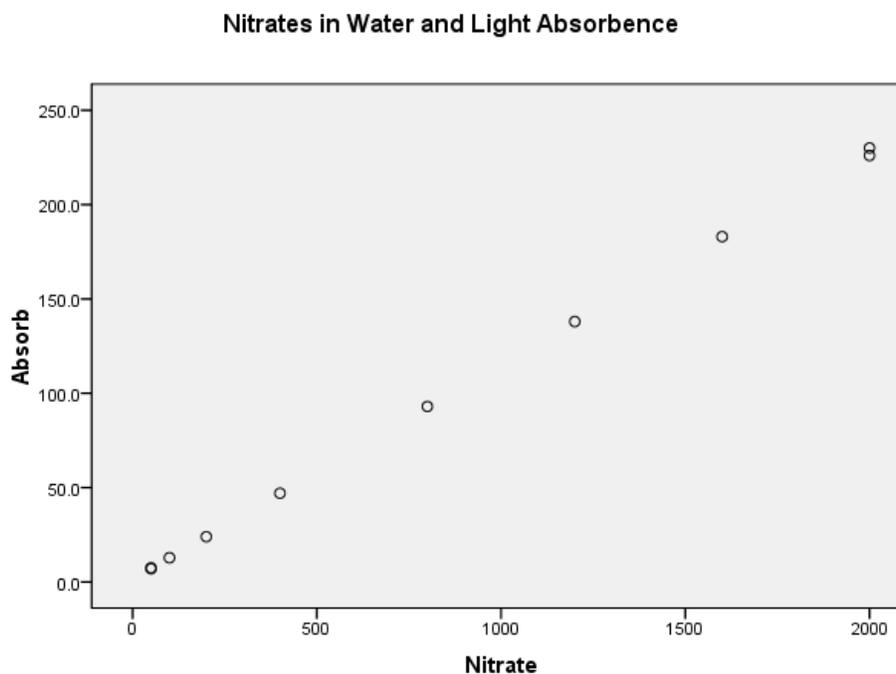
Coefficients<sup>a</sup>

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	10.878	9.030		1.205	.254
	Outsource	.249	.136	.484	1.834	.094

a. Dependent Variable: Delay

Hawaiian was influential for the regression, but not for the correlation. The correlation increased from 0.476 to 0.484. The regression equation changed from  $\text{DelayPct} = 4.73 + 0.387 * \text{OutsourcePct}$  to  $\text{DelayPct} = 10.878 + 0.250 * \text{OutsourcePct}$ . The equation that included Hawaiian Airlines predicts 34.1% delayed flights for the airline with 76% outsourcing. The prediction from the regression that did not include Hawaiian is 29.8% delayed flights – a decrease of 4.3%.

**5.35** To make a scatterplot of these data, click **Graph, Legacy Dialogs, Scatter/Dot**. The default is **Simple**. Click **Define** to continue on to the plot definition dialog box. Click to Enter **Absorb** as the Y variable and **Nitrates** as the X variable. Click **Titles** to give your graph a title, then **Continue** and **OK** to generate the graph.



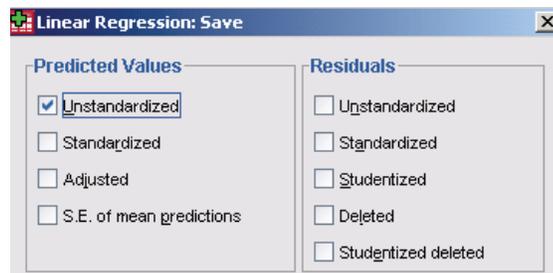
The plot is extremely linear. To find the correlation, use **Analyze, Correlate, Bivariate**. Click to enter both variable names and **OK**.

## Correlations

		Nitrate	Absorb
Nitrate	Pearson Correlation	1.000	1.000**
	Sig. (2-tailed)		.000
	N	10.000	10
Absorb	Pearson Correlation	1.000**	1.000
	Sig. (2-tailed)	.000	
	N	10	10.000

\*\* . Correlation is significant at the 0.01 level (2-tailed).

The correlation is 1.000, which is perfect; this calibration will not need to be done again. To find the equation that estimates nitrate levels, use **Analyze, Regression, Linear**. Click to enter **Nitrate** as the Response and **Absorb** as the Predictor. Since we also want to predict nitrate level for an absorbance of 40, add a new observation in absorbance of 40 without a Nitrate value (this will be ignored when computing the regression). Now, click **Save**, and put a check in the Unstandardized Predicted Values box.

Coefficients<sup>a</sup>

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-14.522	4.461		-3.255	.012
	Absorb	8.825	.034	1.000	256.550	.000

a. Dependent Variable: Nitrate

The estimated nitrate level for an absorbance of 40 is  $-14.522 + 8.825 \times 40 = 338.48$ . Predictions should be very accurate due to the strength of the linear relationship. Returning to the data worksheet, we find that the predicted nitrate level with an absorbance of 40 is 338.48.

11	.	40.0	338.47707
----	---	------	-----------

**5.37** Open data file *ex04-29*. The scatterplot was created in the solution to Exercise 4.29; see that solution if you need help recreating this. We'll calculate the regression without and with the outlier. Since the outlier seems to follow the pattern of the rest of the data (making the relationship appear stronger), we expect that point to have little impact on the regression equation, but to increase the correlation (and  $r^2$ ). We first calculate with the outlier included using **Analyze, Regression, Linear** using **Behave** as the Dependent variable and **Neural** as the Independent to find the following results.

**Model Summary**

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.849 <sup>a</sup>	.720	.700	.27973

a. Predictors: (Constant), Neural

**Coefficients<sup>a</sup>**

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	.585	.071		8.247	.000
	Neural	.009	.001	.849	6.002	.000

a. Dependent Variable: Behave

Now, delete the outlier (the last observation) and recompute the regression to find the following:

**Model Summary**

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.701 <sup>a</sup>	.492	.453	.29025

a. Predictors: (Constant), Neural

**Coefficients<sup>a</sup>**

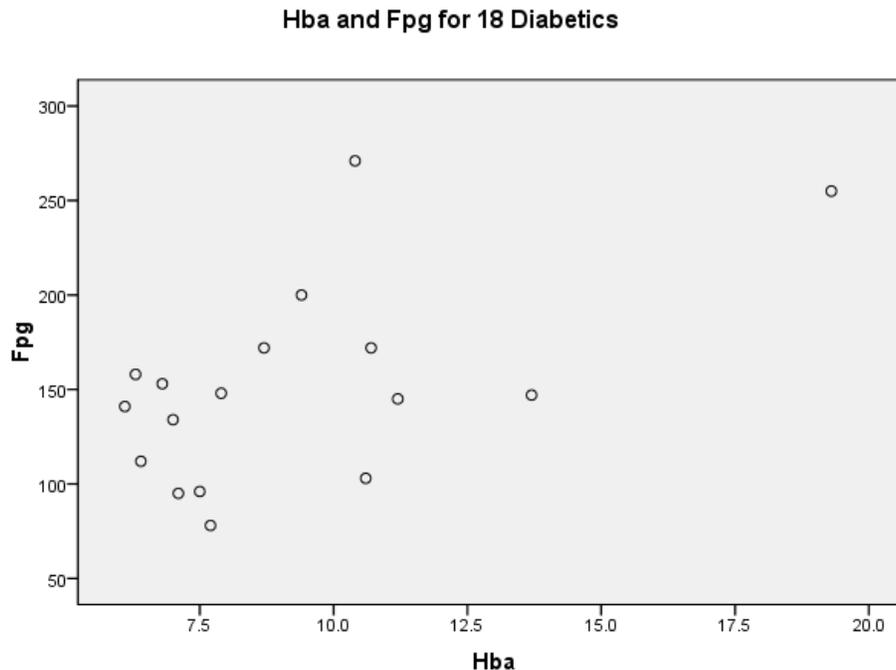
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	.586	.075		7.804	.000
	Neural	.009	.003	.701	3.549	.004

a. Dependent Variable: Behave

With the outlier included, the correlation is  $r = 0.849$  (the square root of 0.72). With the outlier deleted, the correlation has weakened to  $r = 0.701$ . Notice that the outlier has little

impact on the regression equation: with the point included, we have  $BehaviorLoss = 0.585 + 0.0088 * NeuralLoss$ ; with the point deleted, the equation is  $BehaviorLoss = 0.586 + 0.0089 * NeuralLoss$ .

**5.39** Open data file *ta05-02*. Use **Graphs**, **Legacy Dialogs**, **Scatter/Dot** to create the scatterplot, click to enter **Fpg** as the Y variable and **Hba** as the X variable. Click **Titles** to give your graph an appropriate title; **Continue** and **OK** generates the graph.



The data point furthest to the right (subject 18) and in the top center (subject 15) are outliers. We'll first calculate the regression with all the data values; then remove subject 18 and recompute. Finally, we'll add back subject 18 and remove subject 15.

Use **Analyze**, **Regression**, **Linear** to fit the model. Click to enter **Fpg** as the Dependent variable and **Hba** as the Independent variable. **OK** performs the calculations.

**Model Summary**

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.482 <sup>a</sup>	.232	.184	63.816

a. Predictors: (Constant), Hba

**Coefficients<sup>a</sup>**

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	66.429	46.523		1.428	.173
	Hba	10.408	4.731	.482	2.200	.043

a. Dependent Variable: Fpg

With all the data, we have the equation  $Fpg = 66.4 + 10.4 * Hba$ . The correlation is  $r = 0.482$ .

Delete subject 18 in the worksheet, and return to the regression dialog box to recompute.

**Model Summary**

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.384 <sup>a</sup>	.147	.090	65.715

a. Predictors: (Constant), Hba

**Coefficients<sup>a</sup>**

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	52.261	67.541		.774	.451
	Hba	12.116	7.529	.384	1.609	.128

a. Dependent Variable: Fpg

With subject 18 removed, we have  $Fpg = 52.3 + 12.1 * Hba$ . The correlation has lowered to  $r = 0.384$ . Finally, add subject 18 (19.3, 255) back in and delete subject 15 (move the cursor to each value and press the delete key).

**Model Summary**

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.568 <sup>a</sup>	.323	.278	44.716

a. Predictors: (Constant), Hba

**Coefficients<sup>a</sup>**

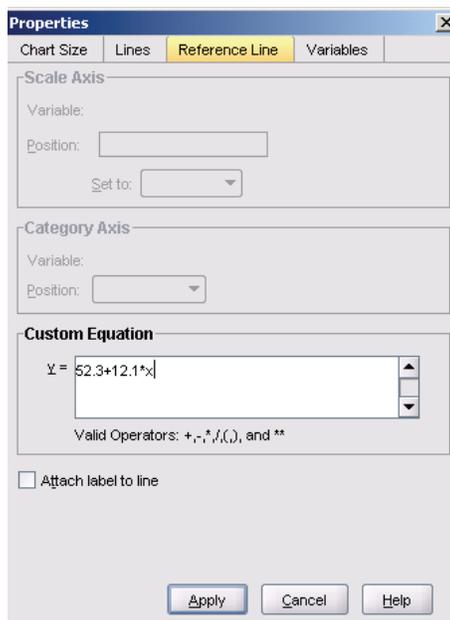
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	69.487	32.607		2.131	.050
	Hba	8.920	3.334	.568	2.676	.017

a. Dependent Variable: Fpg

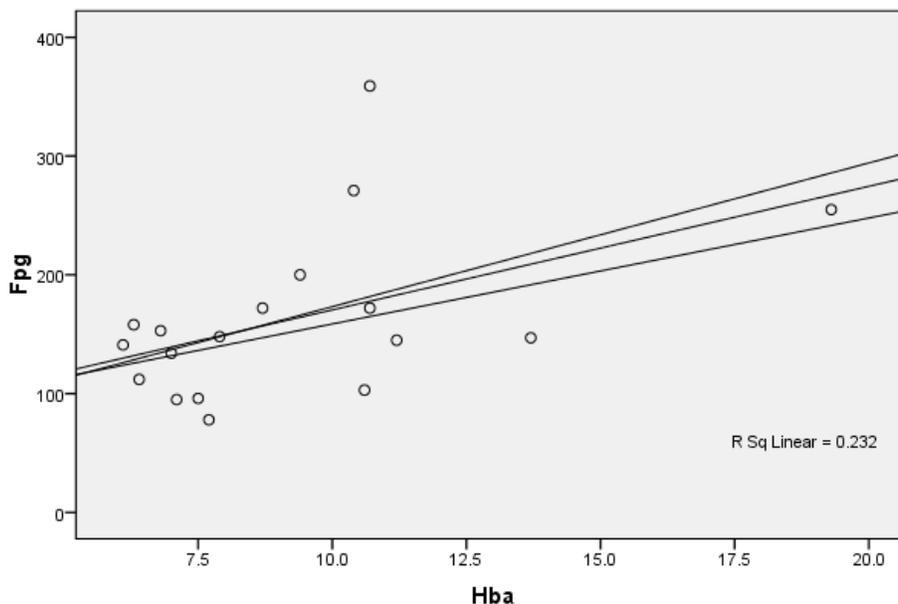
With all but subject 15, we have  $Fpg = 69.5 + 8.92 * Hba$  and the correlation has increased to  $r = 0.568$ .

Deleting subject 18 weakened the relationship. This is because this data point was far to the right in the graph (extreme in its  $x$  value); visually, one's eyes follow through to points like these – they are influential in fitting a model. Subject 15 was relatively in the middle of the  $x$  range but was clearly unusually high in his/her  $Fpg$  level. This type of point really adds only scatter to a plot, so deleting this increased the correlation.

**5.41** We can recreate the scatterplot of all the data and add the three regression lines into the plot using the Chart Editor. If you have just completed Exercise 5.39, add subject 15's data back into the worksheet; otherwise, open data file *ta05-02*. Create the scatterplot as was done in Exercise 5.39. Double-click in the graph to bring up the Chart Editor. Click **Elements, Fit Line at Total** to add the regression line for the entire data set. This also brings up a Properties Box. **Close** this box. To add the other two lines, click **Options, Reference Line from Equation**. Type one of the equations into the Custom Equation box, then **Apply** and **Close**. Repeat this to add the last equation. Close the Chart Editor.

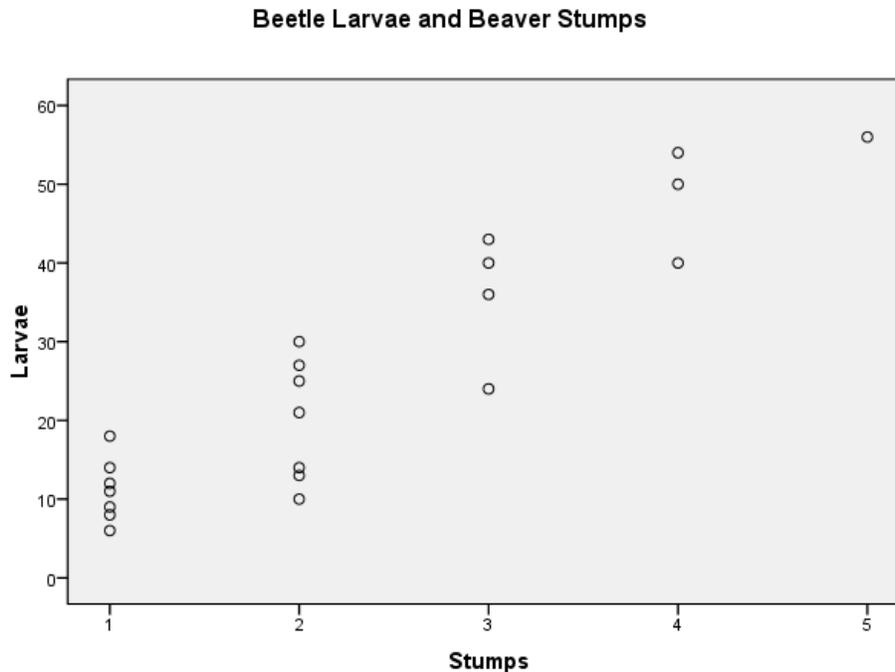


Hba and Fpg for 18 Diabetics



The separation between the lines becomes apparent at the right side of the plot. This is due to the influence of Subject 18 at the far right.

**5.51** Open data file *ex05-51*. First, create a scatterplot of the data using **Graphs, Scatter/Dot**. Enter **Stumps** as the X variable and **Larvae** as the Y variable. Give your graph an appropriate title using **Titles**.



We see that these data indicate that there are more beetle larvae with more stumps. Use **Analyze, Regression, Linear** to fit the line using **Stumps** as the Independent and **Larvae** as the Dependent.

**Model Summary**

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.918 <sup>a</sup>	.843	.835	6.455

a. Predictors: (Constant), Stumps

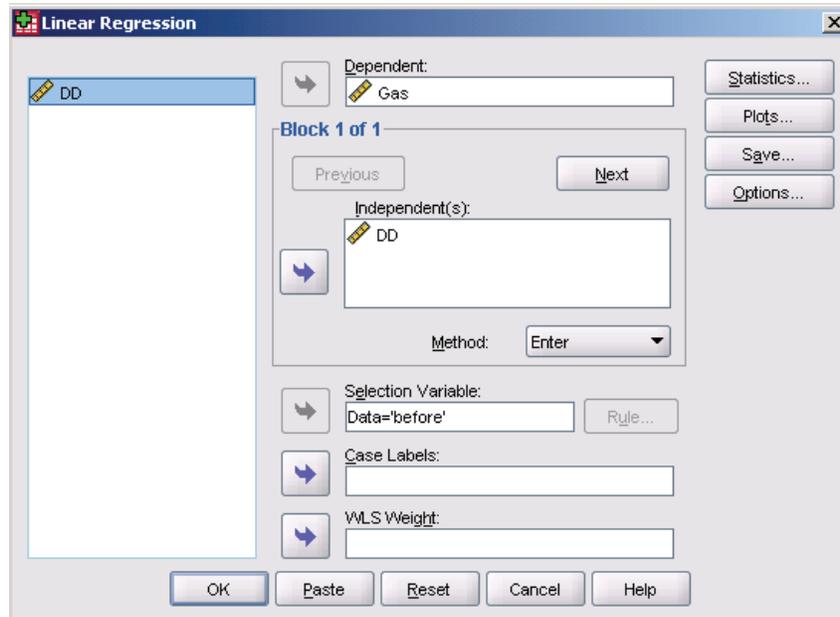
**Coefficients<sup>a</sup>**

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
1 (Constant)	-1.286	2.853		-.451	.657
stumps	11.894	1.136	.916	10.467	.000

a. Dependent Variable: larvae

The regression equation is  $Larvae = -1.28 + 11.89 * Stumps$ . The relationship is strong; the regression model explains 84.3% of the variability in larvae (the correlation is  $r = \sqrt{.843} = 0.918$ ). These data support the “beavers benefit beetles” idea.

**5.55** Open data for *ex04-43*. Since the data in this file have all the observations in each column with the before and after indicators in the variable named **Data**, we’ll modify the regression dialog box input to fit each line separately. Click **Analyze, Regression, Linear**. Enter **DD** as the Independent and **Gas** as the Dependent variables. Now, Click to enter **Data** in the Selection Variable box, then **Rule**. Type in before, then **Continue** and **OK**.



**Model Summary**

Model	R			
	Data = before (Selected)	R Square	Adjusted R Square	Std. Error of the Estimate
1	.995 <sup>a</sup>	.991	.990	.3389

a. Predictors: (Constant), DD

**Coefficients<sup>a,b</sup>**

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	1.089	.139		7.841	.000
	DD	.189	.005	.995	38.309	.000

a. Dependent Variable: Gas

b. Selecting only cases for which Data = before

The regression equation before the solar panels is  $\text{Gas} = 1.089 + 0.189 \cdot \text{DD}$ . The relationship is strong –  $r = 0.995$ . This regression predicts  $1.089 + 0.189 \cdot 45 = 9.594$  hundred cubic feet of gas for a 45 degree-day day. Now, repeat the regression, selecting the after data.

**Model Summary**

Model	R			
	Data = after (Selected)	R Square	Adjusted R Square	Std. Error of the Estimate
1	.991 <sup>a</sup>	.982	.982	.3323

a. Predictors: (Constant), DD

**Coefficients<sup>a,b</sup>**

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	.853	.098		8.732	.000
	DD	.157	.005	.991	34.251	.000

a. Dependent Variable: Gas

b. Selecting only cases for which Data = after

The after regression equation is  $\text{Gas} = 0.853 + 0.157 \cdot \text{DD}$ . This is also a very strong relationship; we have  $r = 0.991$ . This regression predicts  $0.853 + 0.157 \cdot 45 = 7.918$  hundred cubic feet of gas for a 45 degree-day day. The solar panels saved about 160 cubic feet of gas usage.

## Chapter 6 SPSS Solutions

**6.1** The table represents the opinions of  $20+7+9+29+25+43=133$  people. Of these, 36 were buyers of recycled paper coffee filters. The marginal distribution of opinions is  $49/133 = 36.8\%$  think the quality is higher,  $32/133 = 24.1\%$  think the quality is the same, and  $52/133 = 39.1\%$  think the quality is lower.  $60.1\%$  think the quality is the same or higher than other filters.

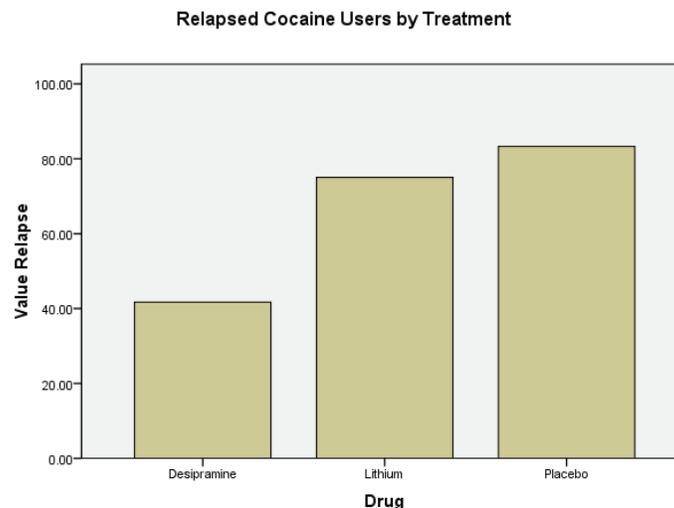
		133
49/133		.3684210526
32/133		.2406015038
52/133		.3909774436
■		

**6.3** SPSS does not like data tables already summarized. To find the conditional distributions, we'll simply divide each cell count by the total number of buyers (36) and non-buyers (97). There should be no surprises – more than half the buyers think the quality of the recycled filters is higher; almost half of the non-buyers think the quality of the recycled filters is lower.

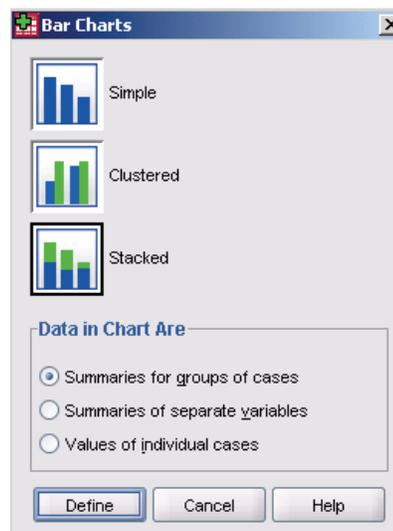
20/36	.5555555556
7/36	.1944444444
9/36	.25

29/97	.2989690722
25/97	.2577319588
43/97	.4432989691

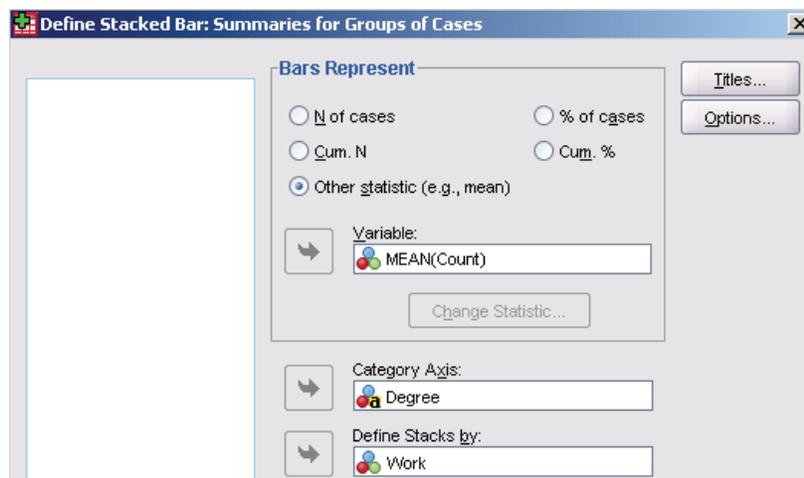
**6.19** The percents of relapsed subjects are  $10/24 = 41.7\%$  for the Desipramine group,  $18/24 = 75\%$  for the Lithium group, and  $20/24 = 83.3\%$  for the Placebo group. Enter these percents and the treatments into two variables. Then click **Graphs, Legacy Dialogs, Bar** and define a simple bar chart where Data in Chart are **Values of Individual Cases**. The bars represent the relapse percentages and the treatment type is the category variable. Give the graph a title, then click OK to generate the graph.



6.27 Open data file *ex06-27*. We'll create a stacked bar chart to display these data and examine the effect of education on freedom at work. Click **Graphs**, **Legacy Dialogs**, **Bar**. Select a **Stacked** chart with **Summaries for groups of cases**. Click **Define** to continue.

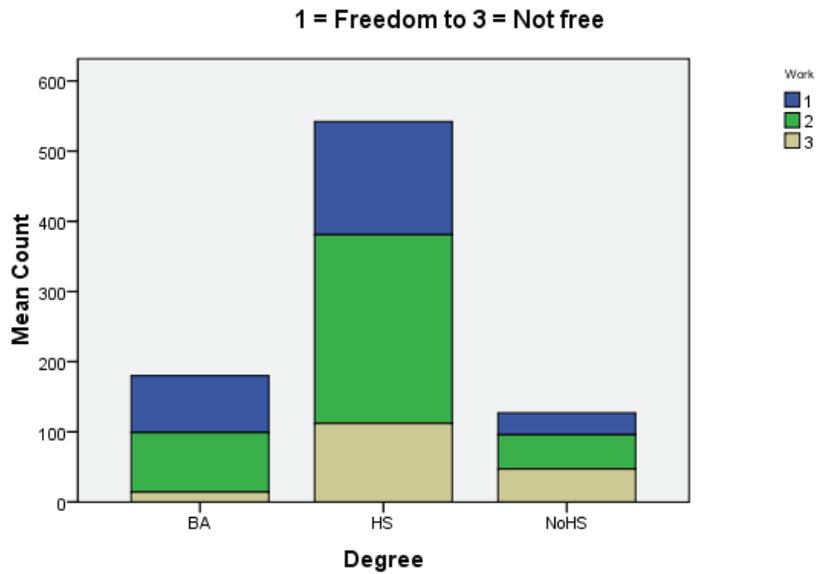


Move the button to Other statistic and enter **Count**; enter **Degree** as the Category axis and Define Stacks by **Work**. Give your graph an appropriate **Titles**.



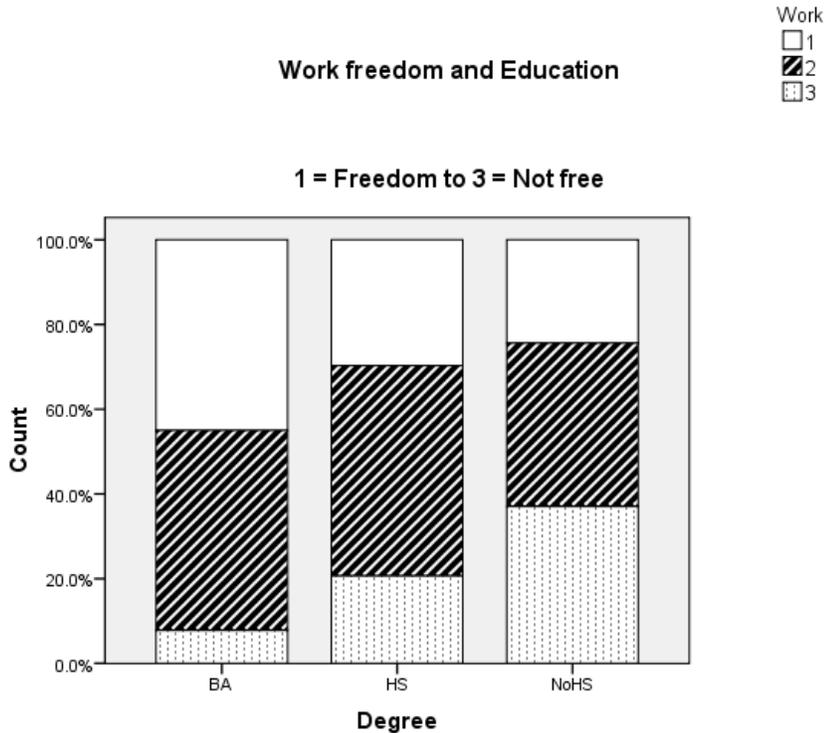
The initial graph is shown below. The bars are not all the same height because there were different numbers of individuals with a given education level. We'll use the Chart Editor to make all the bars extend to 100%.

Work freedom and Education



Double-click in a bar to bring up the Chart Editor. Click **Options**, **Scale to 100%**. You can also click on the Y axis label to change this label to Percent (type the new label in the box). To change from color fill to a black-and-white Pattern, click in a bar for a Properties window. Click the **Variables** tab and change the drop-down box entry for **Work** from Style:color to Style:pattern. **Apply** your changes and **Close** the Chart Editor.

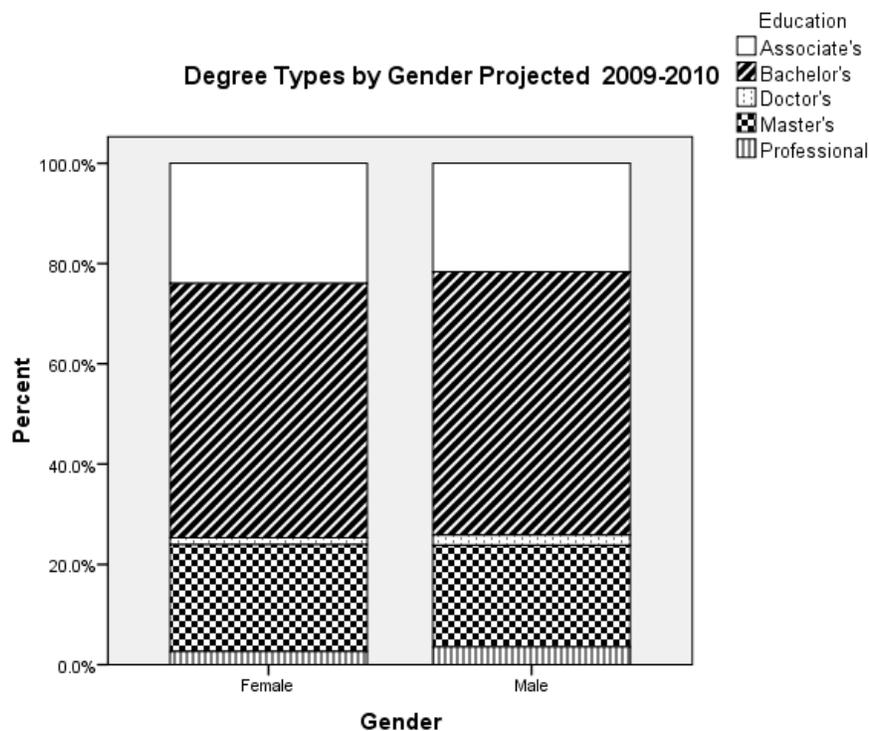
Work freedom and Education



This chart clearly shows that individuals with more education have more freedom to organize their own work. Less than 10% (7.78%) of those with Bachelor's degrees have no freedom while almost 40% (37.01%) of those with less than a high school diploma have no freedom.

**6.29** Follow the instructions given in Exercise 6.27 to create another stacked bar chart after entering the data as shown at right. Our finished chart is below.

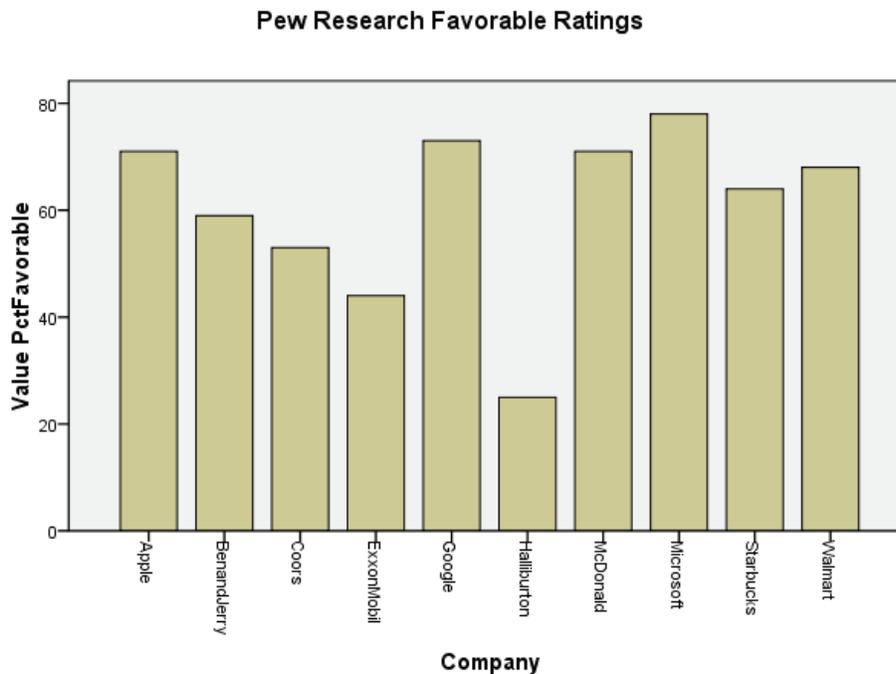
	Education	Gender	Count
1	Associate's	Male	268
2	Associate's	Female	447
3	Bachelor's	Male	651
4	Bachelor's	Female	945
5	Master's	Male	251
6	Master's	Female	397
7	Professional	Male	44
8	Professional	Female	49
9	Doctor's	Male	25
10	Doctor's	Female	26



While the projection is for women to earn far more degrees than men, the conditional distributions are remarkably similar. Men are slightly more likely to pursue professional and doctor's degrees than women.

## Chapter 7 SPSS Solutions

7.3 We'll make a bar chart of these data. Open data file *ex07-03*. Click **Graphs**, **Legacy Dialogs**, **Bar**. We want a simple bar chart where data are **Values of individual cases**. Click **Define** to continue. The Bars represent **PctFavorable**, and the Category labels are **Company**. Be sure to click **Titles** and give your graph an appropriate title. **OK** generates the graph.



This chart is in the order of the data (alphabetical). If you want to create a Pareto Chart (the bars in descending order), double-click in the output graph for the Chart Editor, then click the large X icon for X axis properties. Change the Sort by drop-down box to **Statistic**, then the order box to **Descending**. **Apply** the change and **Close** the properties box and the Chart Editor.

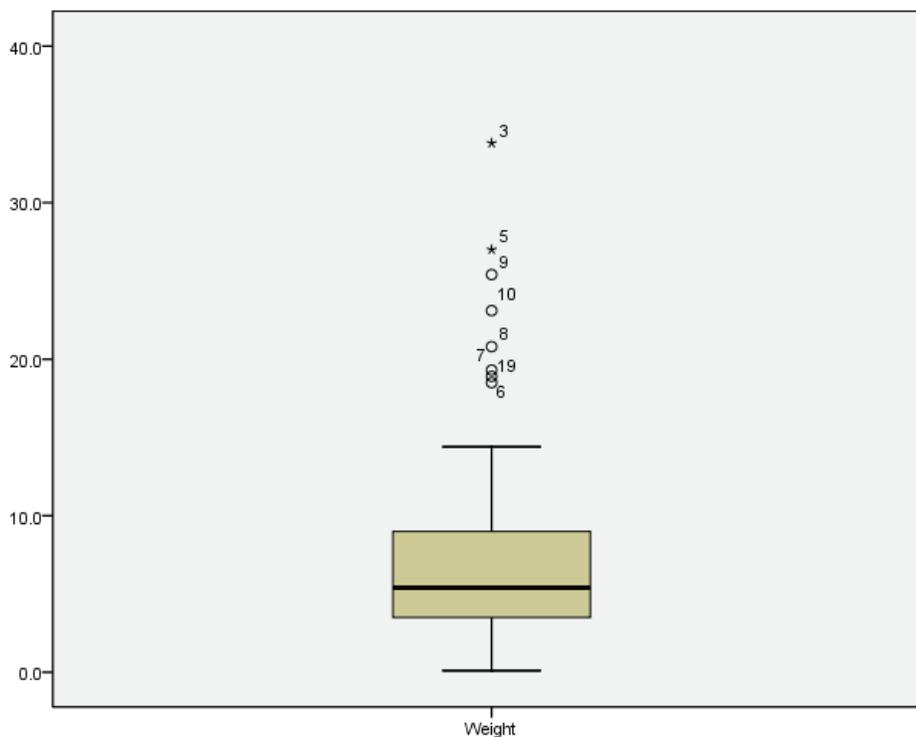
**7.5** For these data, we can use **Analyze, Descriptive Statistics, Explore** to create both graphs and compute summary statistics. Click to enter **Weight** as the Dependent, then **OK**. The stem-and-leaf plot indicates there are 8 high outliers with values greater than 19.

Weight Stem-and-Leaf Plot

Frequency	Stem &	Leaf
7.00	0 .	0001111
14.00	0 .	22222233333333
10.00	0 .	4444455555
10.00	0 .	6677777777
4.00	0 .	9999
3.00	1 .	011
1.00	1 .	3
1.00	1 .	4
8.00	Extremes	(>=19)

Stem width: 10.0  
Each leaf: 1 case(s)

The boxplot below indicates that not only are there outliers, but observations 3 and 5 are extreme outliers (indicated by the star instead of circles).



With this type of distribution, an appropriate numerical summary is the five-number summary. We read from the output below that we have Mon = 0.1, Med = 5.4, and Max

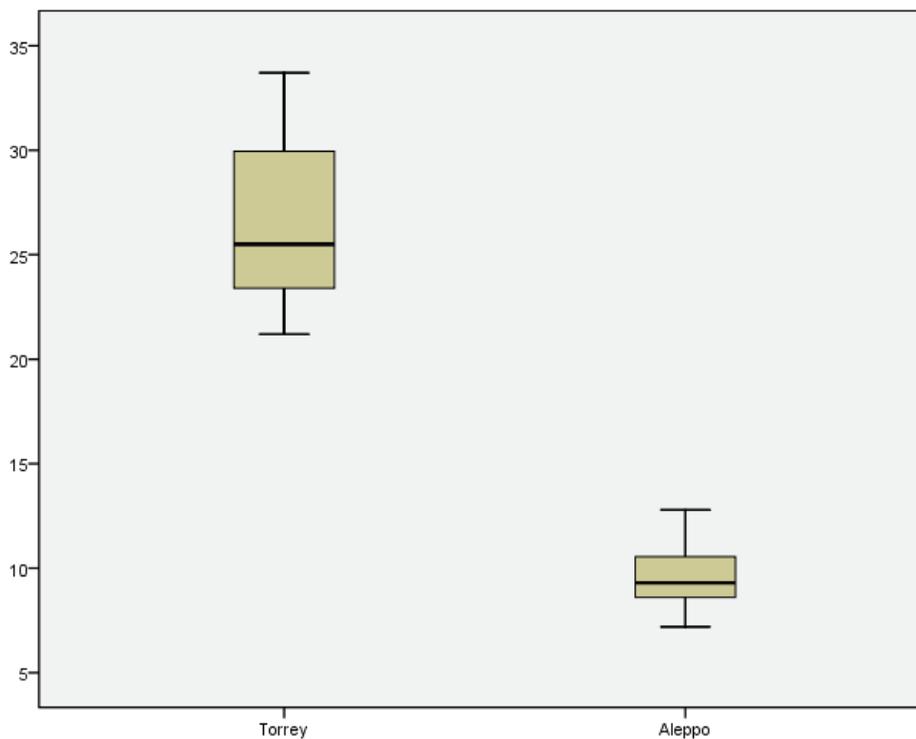
= 33.8. To find  $Q_1$  and  $Q_3$ , redo the **Explore**, but click the **Statistics** button and ask for **Percentiles**.

Descriptives			Statistic	Std. Error
Weight	Mean		7.867	.9536
	95% Confidence Interval for Mean	Lower Bound	5.958	
		Upper Bound	9.777	
	5% Trimmed Mean		7.133	
	Median		5.400	
	Variance		52.740	
	Std. Deviation		7.2622	
	Minimum		.1	
	Maximum		33.8	
	Range		33.7	
	Interquartile Range		5.7	
	Skewness		1.696	.314
	Kurtosis		2.710	.618

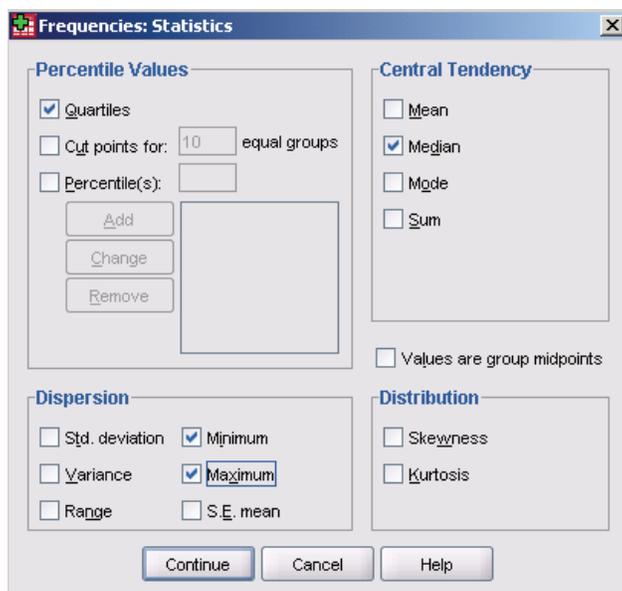
		Percentiles						
		5	10	25	50	75	90	95
Weighted Average(Definition 1)	Weight	.100	1.490	3.425	5.400	9.125	19.450	25.480
Tukey's Hinges	Weight			3.500	5.400	9.000		

We find that  $Q_1 = 3.5$  and  $Q_3 = 9.0$ .

**7.9** To create side-by-side boxplots for the needle lengths, use **Graphs, Legacy Dialogs, Boxplot**. We want a simple boxplot where Data in chart are **Summaries of separate variables**. Click **Define** to continue. Click to select both variables into Boxes represent, and **OK**.



We can clearly see that the Torrey pines have longer needles than Aleppo pines. To find the five-number summary, use **Analyze, Descriptive Statistics, Frequencies**. Enter both variable names (hold down the shift key to select both at once). Click the **Statistics** button and ask for the **Quartiles**, **Median**, **Minimum**, and **Maximum**. **Continue** and **OK** finds the results shown below.



		Torrey	Aleppo
N	Valid	18	15
	Missing	0	3
Median		26.700	9.300
Minimum		21.2	7.2
Maximum		33.7	12.8
Percentiles	25	23.550	8.500
	50	26.700	9.300
	75	29.825	10.900

**7.15** Using the 68-95-99.7 Rule, the center 95% of Aleppo pine needles should be within  $9.6 \pm 2 \times 1.6 = 9.6 \pm 3.2$ , or 6.4 to 12.8 centimeters long. About 2.5% of needles should be less than 6.4 cm long. We verify this using **Transform, Compute Variable**. The function group is **CDF & Noncentral CDF**, then locate **Cdf.Normal** at the lower right. Enter the high value (6.4), mean and standard deviation as shown. If necessary, click on the **Variable View** tab of the worksheet to increase the number of decimal places displayed. We see that according to the Normal Distribution, 2.28% of Aleppo pine needles should be shorter than 6.4 cm.

The screenshot shows the 'Compute Variable' dialog box with the following details:

- Target Variable:** Percent
- Numeric Expression:** CDF.NORMAL(6.4,9.6,1.6)
- Result:** Percent = 0.0228

**7.17** We find these percents using **Transform, Compute Variable**. The function group is **CDF & Noncentral CDF**, then locate **Cdf.Normal** at the lower right. To find the percent scoring higher than 13, we'll subtract the Normal CDF area (area to the left) from 1. 6.1% of all medical school applicants should score 13 or higher.

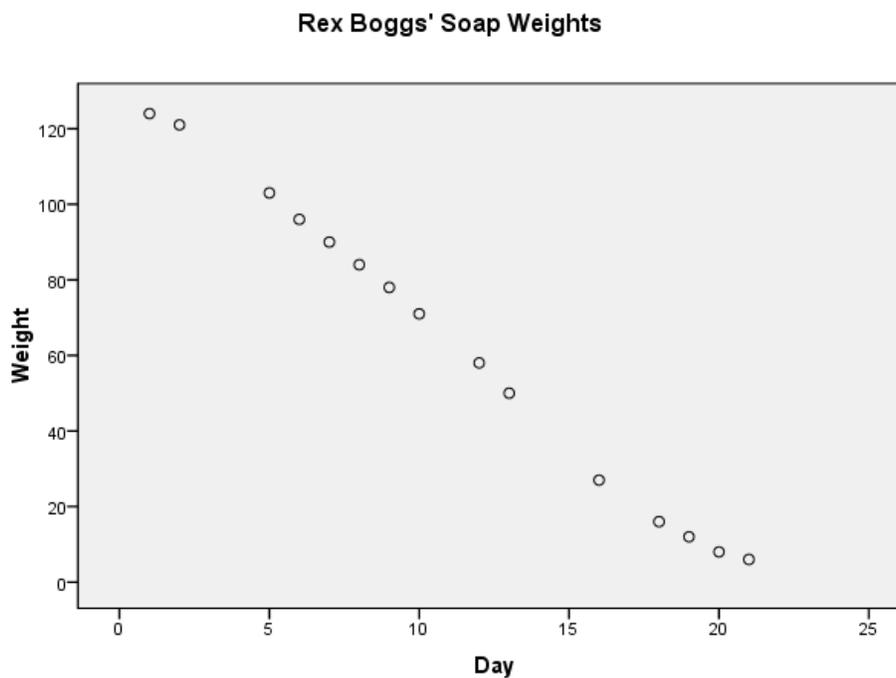
The screenshot shows the 'Compute Variable' dialog box with the following details:

- Target Variable:** Percent
- Numeric Expression:** 1-CDF.NORMAL(13,9.6,2.2)
- Result:** Percent = 0.0611

To find the proportion of those who enrolled who score between 8 and 12, we'll subtract the area to the left of 8 from the area to the left of 12 as shown below. We find that 73.2% should score between 8 and 12.

Compute Variable		Percent
Target Variable:		0.7318
Percent	=	CDF.NORMAL(12,10.6,1.7)-CDF.NORMAL(8,10.6,1.7)

7.19 Open data file *ex07-19*. Use **Graphs, Legacy Dialogs, Scatter/Dot** to plot the data. Use **Weight** as the Y variable and **Day** as the X variable. Use **Titles** to give your graph an appropriate title.

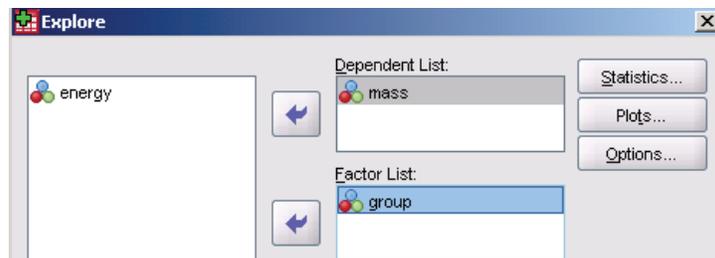


The plot is fairly linear, but there might be some curvature at the ends. There is very little scatter, so the correlation should be close to  $-1$ . Find the actual correlation using **Analyze, Correlate, Bivariate**. Click to enter both variable names and **OK**.

Correlations			
		Day	Weight
Day	Pearson Correlation	1.000	-.998**
	Sig. (2-tailed)		.000
	N	15.000	15

The correlation is a very strong  $r = -0.998$ .

7.23 Open data file *ex06\_25.por*. To compute mean mass by monkey type, use **Analyze**, **Descriptive Statistics**, **Explore**. Click to enter **mass** in the Dependent List and **group** in the Factor list. Click **OK**.

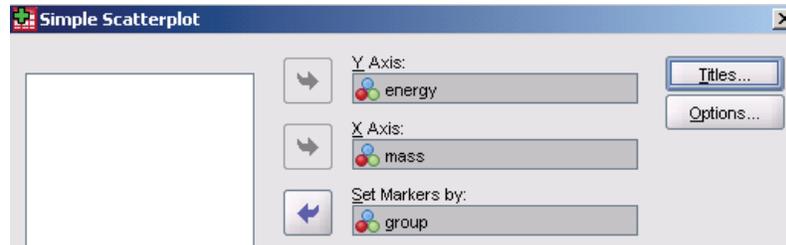


Lean monkeys had a mean lean body mass of 8.68 kilograms; obese monkeys had a mean lean body mass of 10.52 kilograms.

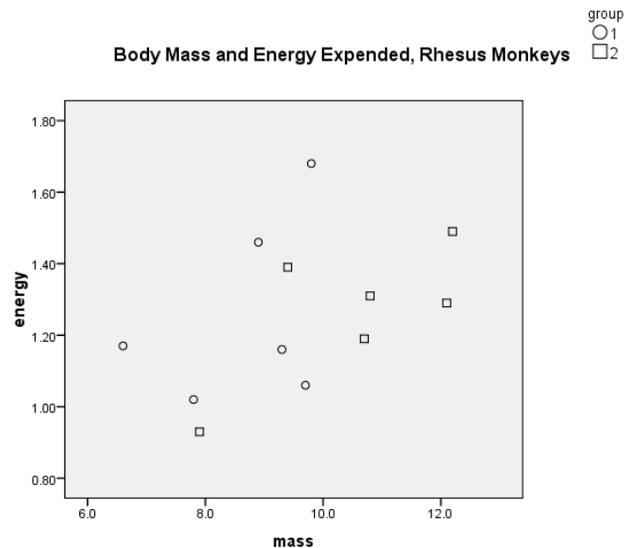
#### Descriptives

group			Statistic	Std. Error	
mass	1	Mean	8.683	.5108	
		95% Confidence Interval for Mean	Lower Bound	7.370	
			Upper Bound	9.996	
		5% Trimmed Mean	8.737		
		Median	9.100		
		Variance	1.566		
		Std. Deviation	1.2513		
		Minimum	6.6		
		Maximum	9.8		
		Range	3.2		
		Kurtosis	.094	1.741	
	2	Mean	10.517	.6720	
		95% Confidence Interval for Mean	Lower Bound	8.789	
			Upper Bound	12.244	
		5% Trimmed Mean	10.569		
		Median	10.750		
		Variance	2.710		
		Std. Deviation	1.6461		
		Minimum	7.9		
		Maximum	12.2		
		Range	4.3		
		Interquartile Range	3.1		

To define the scatterplot, use **Graphs, Legacy Dialogs, Scatter/Dot**. Click to enter **mass** as the X Axis variable and **energy** as the Y Axis variable. Use Group to Set Markers. Give the graph a title and **OK**.

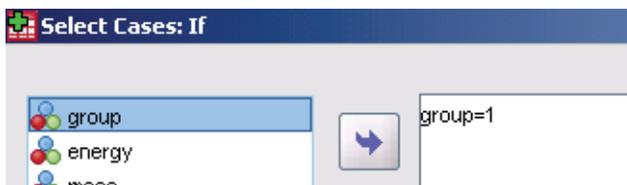


The SPSS default is to mark different groups with colors (in the initial plot, group 1 (lean monkeys) are blue and group 2 are green). If you have only a black-and-white printer, it will be impossible to distinguish these. Double click in the graph to bring up the Chart Editor, then double click in one of the legend circles for the Variables Properties box. Click to change Style: Color for **group** to Style:Shape. **Apply** the change and **Close** the Properties box. Close the Chart Editor. Our finished graph is below.



The smallest of the obese monkeys does not fit the pattern of the rest (who show basically no relationship); the lean monkeys seem to have a weak linear relationship between lean body mass and energy.

Now, we'd like to add the regression lines into the graph; we'll have to compute them first. Click **Data, Select Cases** and move the button to If condition is satisfied, then click the **If** box and enter the expression  $group=1$  to select only the lean monkeys, then **Continue** and **OK**.



Calculate the regression for the lean monkeys using **Analyze, Regression, Linear**. For these lean monkeys the relationship is very weak; we have the regression equation  $\text{energy} = 0.541 + 0.083 \cdot \text{mass}$ .

**Model Summary**

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.401 <sup>a</sup>	.161	-.049	.26393

a. Predictors: (Constant), mass

**Coefficients<sup>a</sup>**

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	.541	.826		.655	.548
	mass	.083	.094	.401	.876	.431

a. Dependent Variable: energy

Now, go back to **Data, Select Cases** and select for group 2 (the obese monkeys); compute their regression line to be  $\text{energy} = 0.371 + 0.085 \cdot \text{mass}$ . This relationship is stronger (perhaps due to the outlier).

**Model Summary**

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.726 <sup>a</sup>	.527	.408	.14865

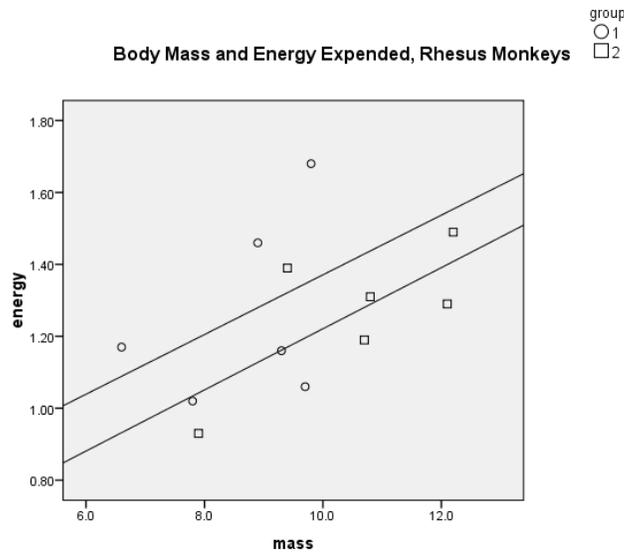
a. Predictors: (Constant), mass

**Coefficients<sup>a</sup>**

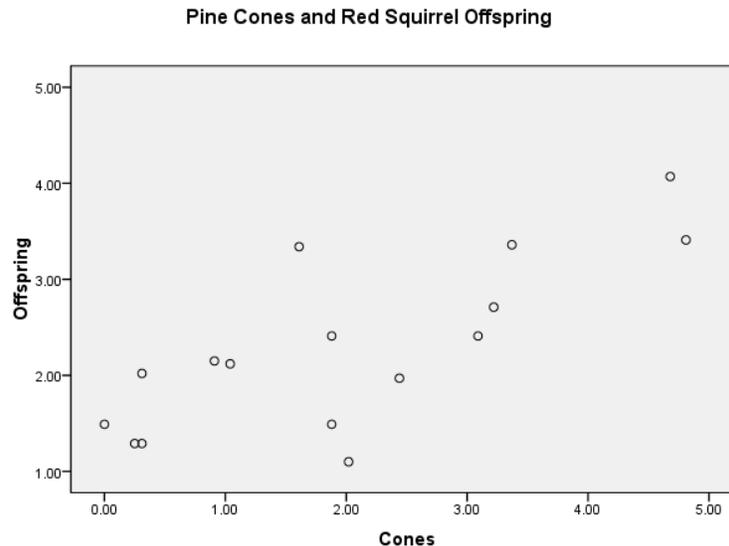
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	.371	.429		.864	.436
	mass	.085	.040	.726	2.110	.103

a. Dependent Variable: energy

To add the regression lines to the graph created earlier, go back to it in the Output window. Double click on the graph for the Chart Editor. Then **Options, Reference Line from Equation**. Enter the first equation then **Apply** and **Close** the box. Repeat to enter the second equation. We essentially see two parallel lines (the slopes are almost equal). Lean monkeys have slightly higher energy expenditures per body mass due to the larger intercept.



**7.25** We create a scatterplot of the **Cones** as the X axis variable and **Offspring** as the Y axis variable using **Graphs, Legacy Dialogs, Scatter/Dot**. The pattern is roughly linear (there is a fair amount of scatter) and increasing – more cones seem to be associated with more offspring.



Use **Analyze, Regression, Linear** to find linear regression and measures of association. The regression equation is  $\text{Offspring} = 1.415 + 0.44 * \text{Cones}$ . The relationship is fairly strong –  $r = 0.756$ ; the cone index explains  $r^2 = 57.2\%$  of the variation in offspring.

**Model Summary**

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.756 <sup>a</sup>	.572	.542	.60031

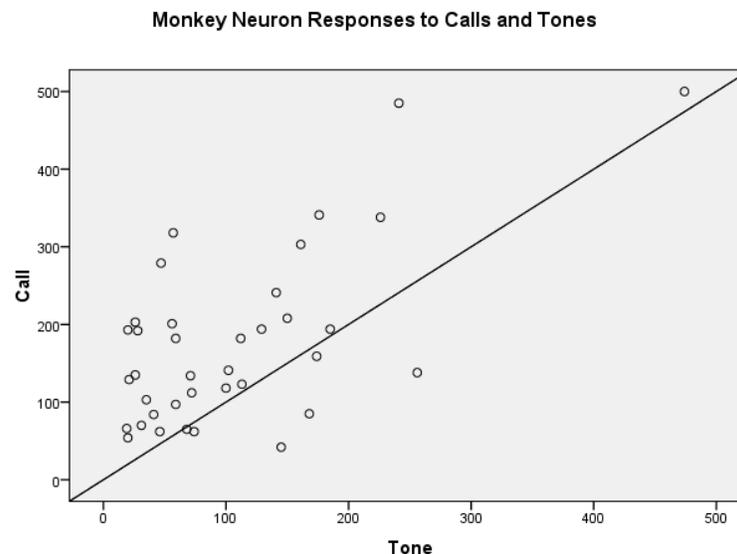
a. Predictors: (Constant), Cones

**Coefficients<sup>a</sup>**

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	1.415	.252		5.619	.000
	Cones	.440	.102	.756	4.328	.001

a. Dependent Variable: Offspring

**7.27** Open data file *ta07-01*. To find out how many neurons have a stronger call response, we'll create a scatterplot and add the line  $Y=X$  to it. Data points above the line have call response stronger than tone, while data points below the line have a stronger tone response. Use **Graphs, Legacy Dialogs, Scatter/Dot** and define the simple scatterplot with **Call** as the Y axis variable and **Tone** as the X axis variable. Be sure to give your graph an appropriate **Titles**. To add the line, double-click in the output graph to bring up the Chart Editor. Click **Options, Reference Line from Equation**. The default should be the line  $Y=1*X+0$ , but change this if needed. **Apply** the change and **Close** the Properties box and Chart Editor. We can easily count that there are 6 points below the line, so  $37 - 6 = 31$  neurons have a higher call response.



The relationship is fairly strong, but there seem to be some outliers – the point at the upper right is possibly influential; the data point at the upper center might be a large residual outlier. To find the correlation, use **Analyze, Correlate, Bivariate**. Click to enter both variable names and **OK**. The correlation is  $r = 0.639$ .

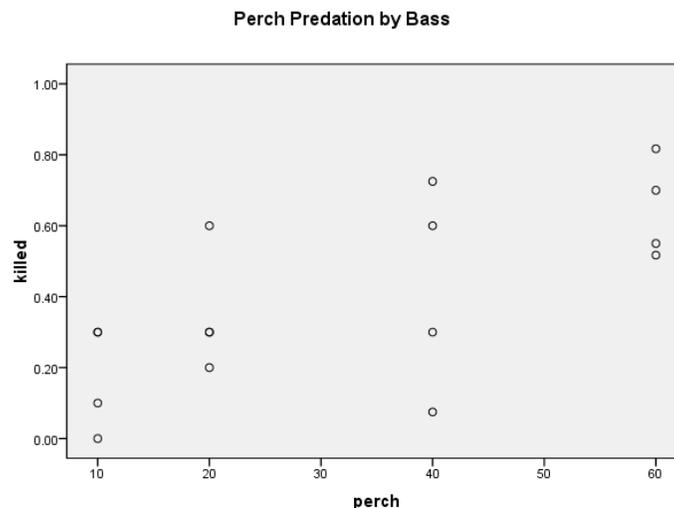
		Tone	Call
Tone	Pearson Correlation	1.000	.639**
	Sig. (2-tailed)		.000
	N	37.000	37
Call	Pearson Correlation	.639**	1.000
	Sig. (2-tailed)	.000	
	N	37	37.000

\*\* . Correlation is significant at the 0.01 level (2-tailed).

**7.37** To find the median data, use **Analyze, Basic Statistics, Frequencies**, and click the **Statistics** button to ask for the **Median**. The median is 16. Sixteen days from April 20 is May 6.

Date		
N	Valid	91
	Missing	0
	Median	16

**7.43** We want to know if there is a positive relationship between the number of prey and the proportion left at the end of two hours; in other words, do these data support the principle? We'll make a scatterplot of the data and perform a linear regression. Open data file *ex06\_19.por*. Define the plot using **Graphs, Legacy Dialogs, Scatter/Dot**. The graph seen below shows what appears to be an increasing pattern, but with lots of scatter.



How strong is the relationship? Fit a regression line using **Analyze, Regression, Linear**. Click to enter the variable names and **OK**. The regression equation is  $\text{ProportionKilled} = 0.12 + 0.0086 * \text{Perch}$ . With an  $r^2$  of 46.5%, this relationship is moderate. The slope and correlation support the principle, but predictions made using this model would not be very trustworthy.

**Model Summary**

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.682 <sup>a</sup>	.465	.427	.18861

a. Predictors: (Constant),

**Coefficients<sup>a</sup>**

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	.120	.093		1.300	.215
		.009	.002	.682	3.490	.004

a. Dependent Variable:

**7.49** Open data file *ta07-01*. Our scatterplot created in Exercise 7.27 indicated two possible outliers. We'll use **Analyze, Regression, Linear** to find regression line for all the data, then remove the first neuron (the point at the upper right) and refit the line. Finally, we'll remove the third neuron (the point at the upper center), add back in the first and again recomputed the regression. For all the data, our results are

**Model Summary**

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.639 <sup>a</sup>	.408	.391	87.297

a. Predictors: (Constant), Tone

**Coefficients<sup>a</sup>**

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	93.917	22.123		4.245	.000
	Tone	.778	.159	.639	4.909	.000

a. Dependent Variable: Call

The regression equation is  $\text{Call} = 93.917 + 0.778 * \text{Tone}$ . Deleting Neuron 1, we have

**Model Summary**

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.479 <sup>a</sup>	.230	.207	88.135

a. Predictors: (Constant), Tone

**Coefficients<sup>a</sup>**

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	101.099	25.530		3.960	.000
	Tone	.693	.218	.479	3.184	.003

a. Dependent Variable: Call

The new regression equation is  $\text{Call} = 101.099 + 0.693 \cdot \text{Tone}$ .  $r^2$  has decreased from 0.408 to 0.230. Now, add back in the first data point (474, 500) and delete the third.

**Model Summary**

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.610 <sup>a</sup>	.372	.354	80.689

a. Predictors: (Constant), Tone

**Coefficients<sup>a</sup>**

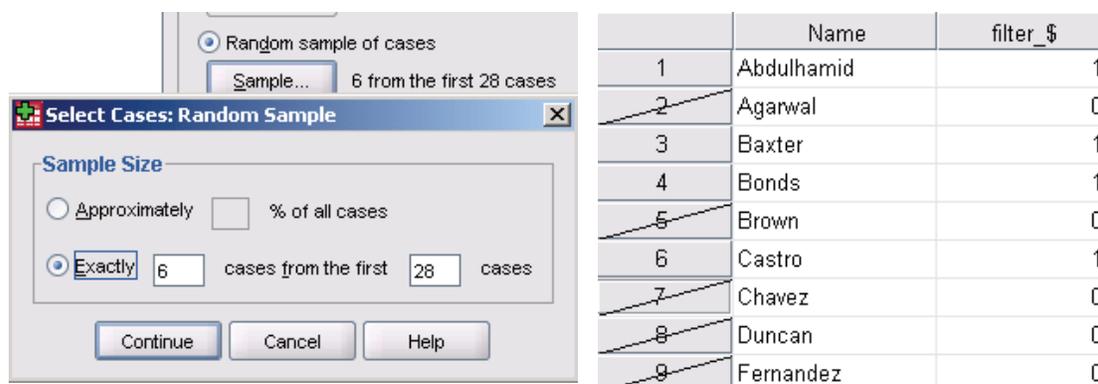
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	98.419	20.520		4.796	.000
	Tone	.679	.151	.610	4.490	.000

a. Dependent Variable: Call

This model is  $\text{Call} = 98.419 + 0.679 \cdot \text{Tone}$ . Neither point is particularly influential on the regression; both coefficients change, but less than 10% of the original values. However, the  $x$  outlier is very influential on the correlation; deleting that point lowers the strength of the relationship from  $r^2 = 40.8\%$  to 23%.

## Chapter 8 SPSS Solutions

**8.7** The names have been entered into a column in the worksheet. There are a couple of ways to perform random selections in SPSS. We'll demonstrate one of them with this solution – by filtering the data randomly. To perform this selection, click **Data, Select cases**. If the names are the only data in your worksheet, you may get an error box – click **OK** to proceed. In the dialog box, select **Random sample of cases**, then click the **Random** button to define how many you want selected. Here, we have said we want to sample 6 of the first 28 (the entire list) cases. **Continue** and **OK**, then return to the worksheet. SPSS adds a filtering variable, and a cross-out line to the case number. From the portion of the worksheet shown, we can see that Abdulhamid, Baxter, Bonds, and Castro were selected. Page on down to find the others.

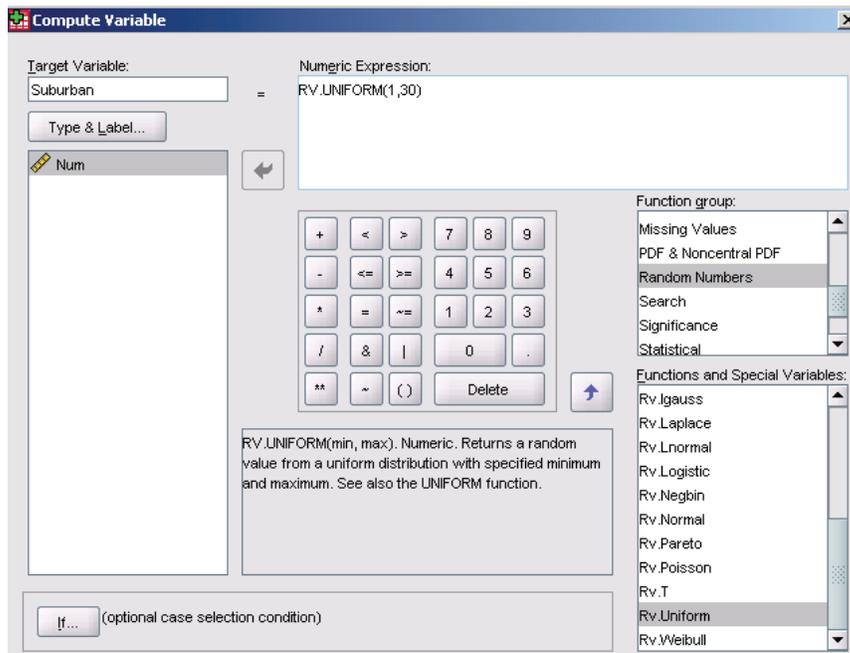


The image shows the 'Select Cases: Random Sample' dialog box in SPSS. The 'Random sample of cases' radio button is selected. Under 'Sample Size', the 'Exactly' radio button is selected, with '6' cases from the first '28' cases. The dialog box has 'Continue', 'Cancel', and 'Help' buttons.

To the right is a portion of a worksheet with three columns: 'Case #', 'Name', and 'filter\_\$'. The rows are numbered 1 through 9. Rows 2, 5, 7, and 8 are crossed out with a diagonal line. The 'filter\_\$' column contains 1 for rows 1, 3, 4, and 6, and 0 for rows 2, 5, 7, 8, and 9.

	Name	filter_\$
1	Abdulhamid	1
<del>2</del>	Agarwal	0
3	Baxter	1
4	Bonds	1
<del>5</del>	Brown	0
6	Castro	1
<del>7</del>	Chavez	0
<del>8</del>	Duncan	0
<del>9</del>	Fernandez	0

**8.11** With this solution, we illustrate another method of random selection in SPSS. We want to select six of the 30 suburban townships and four of the eight Chicago townships. We'll generate random integers – generating a few more than we really need, in case of duplicates (which would be ignored). On a new worksheet, place a 1 in the tenth row of the first column. This signals to SPSS when we use **Transform, Compute Variable**, that we'll generate ten random numbers. In the Function group box, scroll down to select **Random Numbers**, then locate **Rv.Uniform** in the Functions box. In the dialog box below, I'm creating random numbers between 1 and 30 into the variable named **Suburban**. Click on the **Variable View** tab, and set the decimal places for this variable to 0. Our random selection selects Hanover, Palos, Stickney, Calumet, Northfield, and Bremen. We'll repeat the process to generate random numbers between 1 and 8 to select the four Chicago townships. We've selected Rogers Park, South Chicago, Hyde Park, and Lake View.



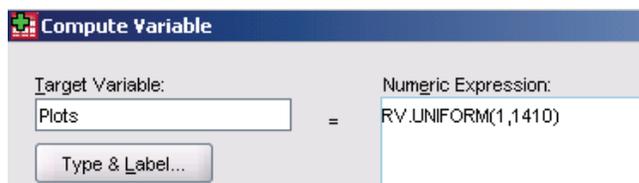
Suburban	
	9
	21
	27
	5
	16
	4

Chicago	
	6
	7
	1
	4

**8.27** With our list of 40 class members entered in a variable named **Name**, we'll use the **Data, Select Cases** to select ten of them. Follow the instructions given in Exercise 8.27 to find we have selected Fernandez, Fullmer, Husain, Molina, Percival, Prince, Shen, Velasco, Wallace and Zhao.

<del>1</del>	Anderson	<del>11</del>	Drasin	<del>21</del>	Kim	<del>31</del>	Rider
<del>2</del>	Arroyo	<del>12</del>	Eckstein	<del>22</del>	Molina	<del>32</del>	Rodriguez
<del>3</del>	Batista	<del>13</del>	Fernandez	<del>23</del>	Morgan	<del>33</del>	Samuels
<del>4</del>	Bell	<del>14</del>	Fullmer	<del>24</del>	Murphy	<del>34</del>	Shen
<del>5</del>	Burke	<del>15</del>	Gandhi	<del>25</del>	Nguyen	<del>35</del>	Tse
<del>6</del>	Cabrera	<del>16</del>	Garcia	<del>26</del>	Palmiero	<del>36</del>	Velasco
<del>7</del>	Calloway	<del>17</del>	Glaus	<del>27</del>	Percival	<del>37</del>	Wallace
<del>8</del>	Delluci	<del>18</del>	Helling	<del>28</del>	Prince	<del>38</del>	Washburn
<del>9</del>	Deng	<del>19</del>	Husain	<del>29</del>	Puri	<del>39</del>	Zabadi
<del>10</del>	De Ramos	<del>20</del>	Johnson	<del>30</del>	Richards	<del>40</del>	Zhao

**8.29** For this sampling scheme, we'll use **Transform, Compute Variable** to generate 15 (allowing for any duplicates) random numbers between 1 and 1410 (if we were using the table of random digits, we'd label the plots 0001 to 1410), setting decimal places to 0 on the **Variable View** tab. Enter a 1 in the 15<sup>th</sup> row of the first column of the worksheet (to set how many random numbers will be generated), then use a command like that shown below. There weren't any duplicates, so the first ten selected plot numbers are shown below.



Plots
36
18
1102
1251
751
280
958
622
579
704

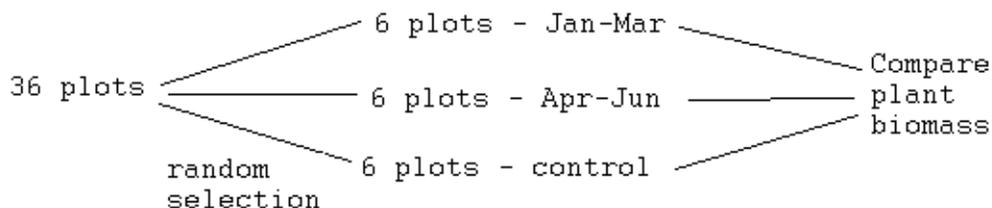
**8.39** We'll label the men from 1 to 290 (if we were using the table of random digits, this would be 001 to 290), and the women similarly from 1 to 110 (001 to 110) – possibly alphabetically, or in order of arrival? Since we don't really anticipate duplicates in selecting 3 individuals from each gender, place a 1 in the third row of the first column of a new worksheet. Use **Transform, Compute Variable** as described above to generate the random numbers. We show the command for the men, and then the selected individuals' numbers.



Men	Women
208	107
47	80
120	40

## Chapter 9 SPSS Solutions

9.9 An outline of the experiment might be like the one below.

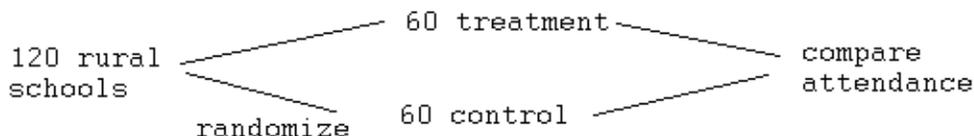


Assuming we have numbered the plots from 1 to 36, we can use **Transform, Compute Variable** to select the plots to be used. Since we may encounter duplicates, we'll place a 1 in the 36<sup>th</sup> row of the first column of a blank worksheet; this will instruct SPSS to generate 36 random numbers into the variable we called **Plots**. Locate the **RV.Uniform** command shell in the **Random Numbers** function group. The first six (nonduplicates) will be assigned to the January through March additional water (these are 22, 24, 12, 16, 19, and 2), the next six (32, 1, 3, 9, 8, and 34) to the April through June additional water, and the third six (25, 20, 21, 4, 30, and 23) to be controls.



Plots	2	25
22	32	20
24	24	21
12	1	4
16	3	9
19	2	34
2	1	30
19	9	30
2	8	23
2	34	13

9.33 Our outline of the experiment is below.

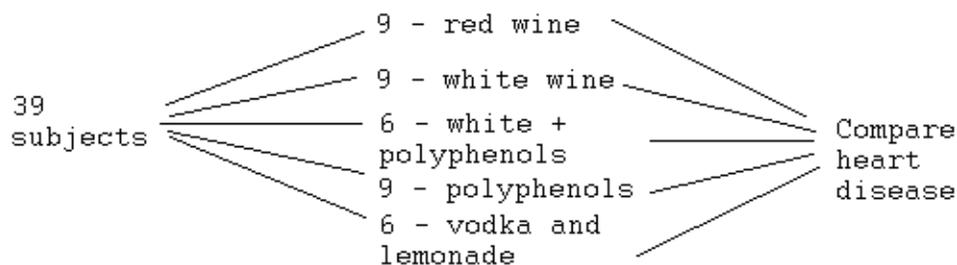


To select the first 10 schools for the treatment group, we'll again use **Transform, Compute Variable** make the selection. Since we may encounter duplicates, we'll place a 1 in the 15<sup>th</sup> row of the first column of a blank worksheet; this will instruct SPSS to generate 15 random numbers into the variable we called **Schools**. Locate the **RV.Uniform** command shell in the **Random Numbers** function group. The first ten (nonduplicates) will be assigned to the treatment. The first ten in our selection had no duplicates. Their numeric labels are shown below.

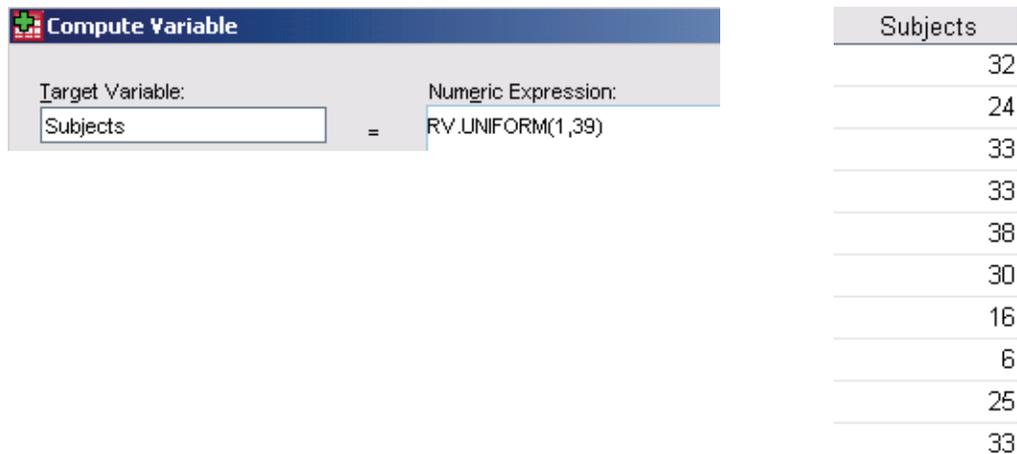


	VAR00001	Schools
1	.	30
2	.	31
3	.	109
4	.	64
5	.	52
6	.	21
7	.	111
8	.	2
9	.	51
10	.	116

9.35 Our outline of the experiment is below.



To randomize the 39 subjects, we'll use Transform, Compute Variable to generate a list of random numbers (allowing for possible duplicates) to make the assignments, as detailed above. As seen below, subjects 32, 24, 33, 38, 30, 16, 6, 25, are the first eight to be assigned to red wine. Continue down the list to complete the assignment.

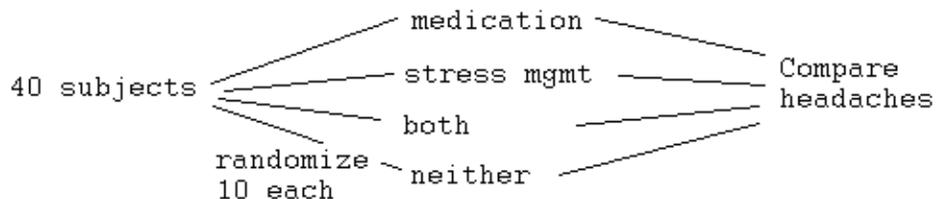


The image shows the SPSS 'Compute Variable' dialog box. The 'Target Variable' is 'Subjects' and the 'Numeric Expression' is 'RV.UNIFORM(1,39)'. To the right is a list of 10 subject numbers: 32, 24, 33, 33, 38, 30, 16, 6, 25, 33.

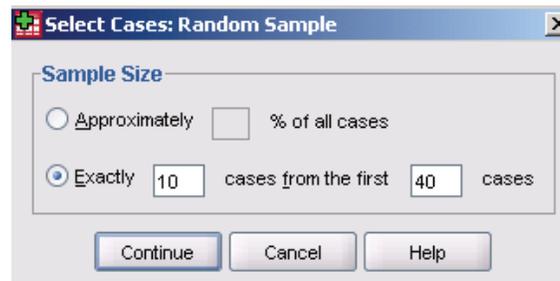
9.37 A diagram of the treatments for this two factor experiment is below.

		Stress Management	
		Yes	No
Medication	Yes	10 subjects	10 subjects
	No	10 subjects	10 subjects

A completely randomized design diagram is below.



Since we have 40 volunteers, assign 10 to each treatment. We will assign the first ten selected to Nothing. To make the selections using SPSS, enter the numbers 1 through 40 in a variable. Click **Data, Select cases**. Click to select a random sample of exactly 10 of the first 40 cases. **Continue** and **OK** makes the first selection.

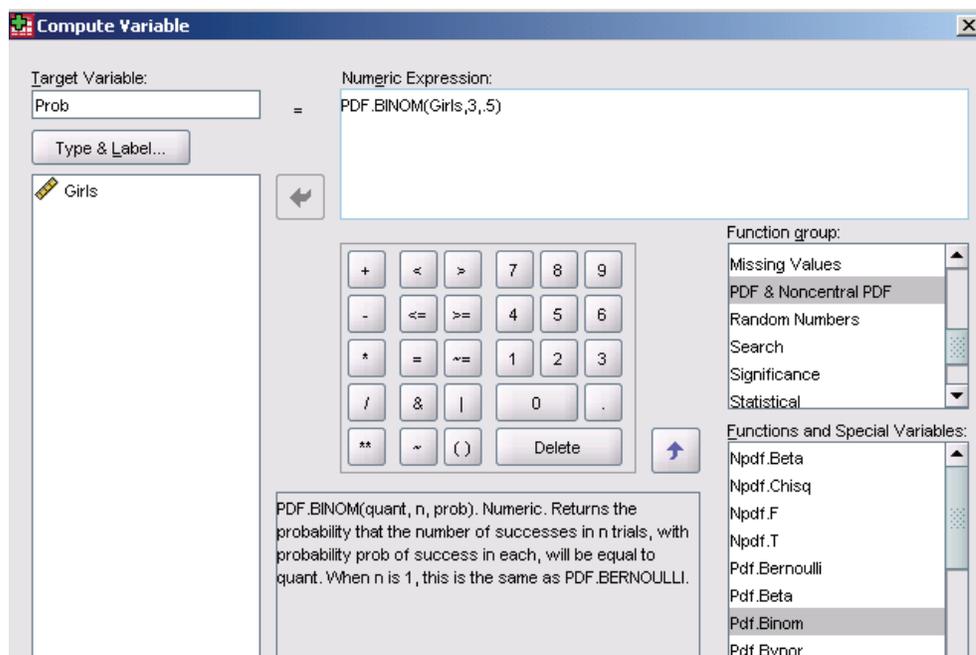




## Chapter 10 SPSS Solutions

**10.47** The possible arrangements are: BBB, GGG, BGG, GBG, GGB, BBG, BGB, GBB. If all arrangements are equally likely, 3 of the eight arrangements have two girls, so the probability of two girls becomes  $3/8 = 0.375$ . To find the probability distribution for the number of girls, we can see that one arrangement results in 0 or 3 girls, so  $P(X = 0) = P(X = 3) = 1/8$ . Similarly, there are also three arrangements that lead to one girl (two boys), so  $P(X = 1) = 3/8$ . We'll learn in Chapter 12 that this random variable is Binomial. We can create the probability distribution for this variable by entering values 0 through 3 in a column of the worksheet we have named **Girls** and using **Transform, Compute Variable** to find the probabilities as shown below.

	Girls	Prob
1	0.00	0.1250
2	1.00	0.3750
3	2.00	0.3750
4	3.00	0.1250



**10.51** We use **Transform, Compute Variable** to find the probability of between 52% and 60% of respondents claiming to have voted as shown below.



Similarly, we find the probability of at least 72% voting by subtracting the cumulative probability from 1. This is (to four decimal places) 0.

Compute Variable	
Target Variable:	Numeric Expression:
Prob	= 1-CDF.NORMAL(.72,.56,.019)

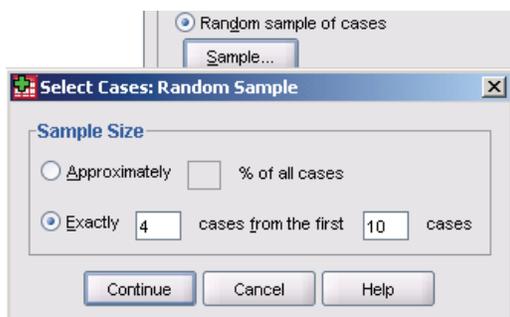
Prob
0.0000

## Chapter 11 SPSS Solutions

**11.7** The test scores have been entered into a variable named **Test**. We find the mean of this “population” using **Analyze, Descriptive Statistics, Descriptives** to be  $\mu = 69.4$ .

	N	Minimum	Maximum	Mean	Std. Deviation
Test	10	58	82	69.40	8.044
Valid N (listwise)	10				

We use **Data, Select Cases** to randomly select a sample of 4 scores.



	Test	filter_ \$
1	82	1
2	62	0
3	80	0
4	58	1
5	72	1
6	73	0
7	65	0
8	66	0
9	74	1
10	62	0

Now, use **Analyze, Descriptive Statistics, Descriptives** to find the mean of the selected cases. The mean of the first sample is 71.5. Enter this in a new variable (we called it **Samples**), and repeat the process of selecting random samples nine more times.

	N	Minimum	Maximum	Mean	Std. Deviation
Test	4	58	82	71.50	9.983
Valid N (listwise)	4				

You can create a histogram of the sample means in **Samples**, but with only ten observations, it won't show much. We can find the mean of these sample means using **Analyze, Descriptive Statistics, Descriptives**. The mean of the sample means is 68.25, 1.15 below the “population” mean.

	N	Minimum	Maximum	Mean	Std. Deviation
Samples	4	68.25	75.25	71.0625	3.05079
Valid N (listwise)	4				

**11.9** The sample of 100 men should have mean cholesterol distributed as  $\bar{x} \sim N(188, 41/\sqrt{100}) = N(188, 4.1)$ . We'll use **Transform, Compute Variable** with CDF.NORMAL as shown to find the probability of a mean cholesterol level for these 100 between 185 and 191. There is about a 53.6% chance of a mean cholesterol for a random sample of 100 men being between 185 and 191.

Compute Variable		Prob
Target Variable:	Numeric Expression:	
Prob	= CDF.NORMAL(191,188,4.1)-CDF.NORMAL(185,188,4.1)	0.5357

For the sample of 1000 men, we'll have mean cholesterol levels distributed as  $\bar{x} \sim N(188, 41/\sqrt{1000}) = N(188, 1.297)$ . We use the same procedure, changing the standard deviation, to find the chance this sample mean is within 3 of the population mean is 97.9%.

Compute Variable		Prob
Target Variable:	Numeric Expression:	
Prob	= CDF.NORMAL(191,188,1.297)-CDF.NORMAL(185,188,1.297)	0.9793

**11.13** We want to know the chance the average loss,  $L$ , is less than or equal to \$275. For 10,000 policies, we have  $\bar{x} \sim N(250, 1000/\sqrt{10,000}) = N(250, 10)$ . By the 68-95-99.7 Rule, the chance of having an average loss no more than \$275 is more than 97.5%. We use **Transform, Compute Variable** to find this is 99.4%.

Compute Variable		Prob
Target Variable:	Numeric Expression:	
Prob	= CDF.NORMAL(275,250,10)	0.9938

**11.27** We use **Transform, Compute Variable** to find the chance Sheila's glucose level is above 140. We find the chance is about 6.7% that she would be diagnosed with gestational diabetes based on one sample.

Compute Variable		Prob
Target Variable:	Numeric Expression:	
Prob	= 1-CDF.NORMAL(140,125,10)	0.0668

For an average of four days, we change the standard deviation to  $\sigma = 10/\sqrt{4} = 5$ , and again use **Transform, Compute Variable** to find the chance Sheila's glucose level is above 140. This gives a 0.13% chance.

Compute Variable		Prob
Target Variable:	Numeric Expression:	
Prob	= 1-CDF.NORMAL(140,125,5)	0.0013

**11.29** This backward Normal calculation uses **Transform, Compute Variable**, but we use the **IDF.Normal** function. The parameters are area to the *left* of the desired point, the mean, and the standard deviation. Remember that the average of four readings has  $\sigma = 10/\sqrt{4} = 5$ . We find there should be a 5% chance her average glucose level is above 133.2.

Compute Variable		Level
Target Variable:	Numeric Expression:	
Level	= IDF.NORMAL(.95,125,5)	133.22

**11.31** For 52 weeks, the distribution of the average number of accidents will be  $\bar{x} \sim N(2.2, 1.4/\sqrt{52}) = N(2.2, 0.194)$ . We use **Transform, Compute Variable** to find the probability the average is less than 2 is 15.1%.

Compute Variable		Prob
Target Variable:	Numeric Expression:	
Prob	= CDF.NORMAL(2,2.2,.194)	0.1513

If there are fewer than 100 accidents in a year, this means the mean is less than  $100/52 = 1.923$ . Using `normalcdf` again, we find this probability is 7.7%.

Compute Variable		Prob
Target Variable:	Numeric Expression:	
Prob	= CDF.NORMAL(1.923,2.2,.194)	0.0767

**11.33** For a period of 40 years, the average return should have distribution  $\bar{x} \sim N(8.7, 20.2/\sqrt{40}) = N(8.7, 3.194)$ . Use **Transform, Compute Variable** to find the probability of an average return more than 10% is 34.2%.

Compute Variable		Prob
Target Variable:	Numeric Expression:	
Prob	= 1-CDF.NORMAL(10,8.7,3.194)	0.3420

The chance of an average return less than 5% is 12.3%.

Compute Variable		Prob
Target Variable:	Numeric Expression:	
Prob	= CDF.NORMAL(5,8.7,3.194)	0.1233

**11.39** Casper's average winnings for 150,000 bets should have distribution  $\bar{x} \sim N(0.40, 18.96/\sqrt{150,000}) = N(0.40, 0.0490)$ . Use **Transform, Compute Variable** to find that the chance his average winnings for a week are between \$0.30 and \$0.50 is 95.9%.

Compute Variable		Prob
Target Variable:	Numeric Expression:	
Prob	= CDF.NORMAL(.5,.4,.049)-CDF.NORMAL(.3,.4,.049)	0.9587

## Chapter 12 SPSS Solutions

\*Note: These solutions do not really use SPSS (they're really generic exponentiation and multiplication). We provide these for any individuals needing some help with these functions.

**12.9** Since 10% of adults belong to health clubs and 40% of these go at least once a week, we have  $P(\text{health club}) = 0.10$  and  $P(\text{at least once} | \text{health club}) = 0.40$ . We want  $P(\text{health club and at least once}) = 0.10 * 0.40 = 0.04$ .

**12.15** In Exercise 12.13, we were told that 1% of Americans is allergic to peanuts or tree nuts. If  $X$  is the number of allergic people, we can find the chance of at least one allergic person in five people as  $1 - P(X = 0)$ . This is 0.049. The desired conditional probability is  $P(X = 1 | X \geq 1) = P(X = 1) / P(X \geq 1)$ . (The chance of being both equal to 1 and at least 1 is really the change of being equal to 1.) From Exercise 12.13, you should have  $P(X = 1) = 4 * .01 * .99^4 = 0.0384$ . Putting this together, the conditional probability is 78.4%.

```
1-.99^5
      .0490099501
4*.01*.99^4
      .0384238404
.0384/.049
      .7836734694
█
```

**12.27** The chance that any one bet loses is  $1 - 0.25 = 0.75$ . The chance that all eight bets lose is 10.0%.

```
.75^8
      .100112915
█
```

**12.29** The chance of winning the jackpot is  $1/20 * 9/20 * 1/20 = 9/8000 = 1.125\%$ . If we call the wheels A, B, and C (where A and C are the outside wheels), you can have two cherries if A and B have cherries, B and C have cherries, or A and C have cherries. We find  $P(A \text{ and } B \text{ and not } C) = P(B \text{ and } C \text{ and not } A) = 1/20 * 9/20 * 19/20 = 2.14\%$  and  $P(A \text{ and } C \text{ and not } B) = 1/20 * 1/20 * 11/20 = 0.14\%$ . The chance of exactly two cherries is then  $.0214 + .0214 + .0014 = 0.0442$ .

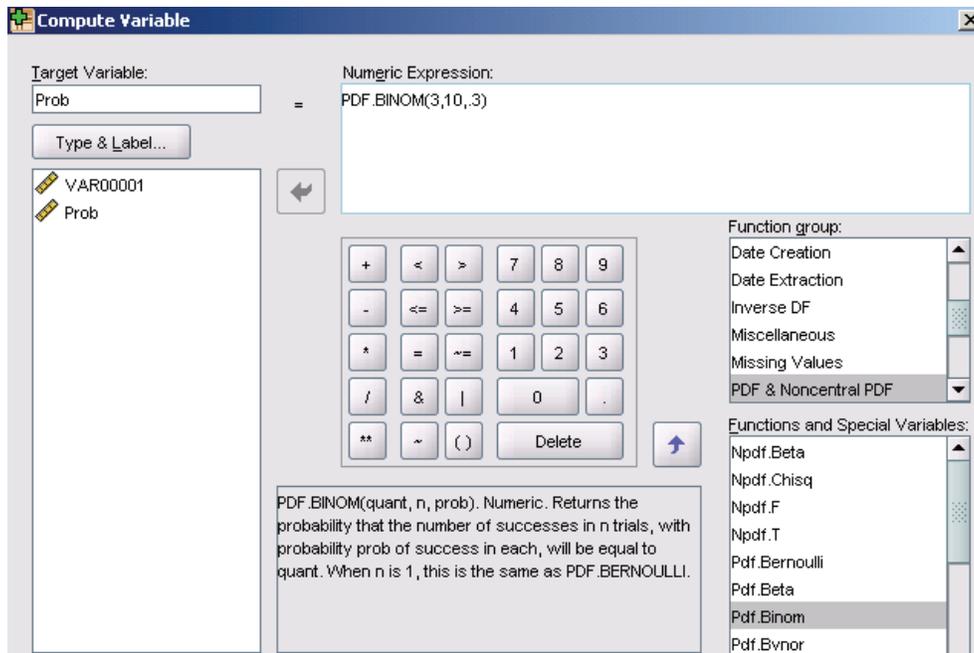
```
1/20*9/20*1/20
      .001125
1/20*9/20*19/20
      .021375
1/20*1/20*11/20
      .001375
█
```

**12.47** The chance of rolling doubles on any one throw is  $6/36 = 16.67\%$ . The probability of not rolling doubles on the first, but rolling doubles on the second is  $30/36 * 6/36 = 13.9\%$ . The probability the first two are not doubles, and the third toss is doubles is 11.6%. The general rule becomes  $P(\text{first doubles on } k^{\text{th}} \text{ toss}) = (30/36)^{k-1} (6/36)$ .

```
6/36
      .1666666667
30/36*6/36
      .1388888889
(30/36)^2*6/36
      .1157407407
█
```

## Chapter 13 SPSS Solutions

**13.5** If the student catches 70% of errors, 30% will be missed, so if  $X$  = number of errors missed,  $X$  is Binomial (assuming independence),  $n = 10$ ,  $p = 0.30$ . To find the probability of missing exactly 3 of the 10 errors, use **Transform**, **Compute Variable** and locate **PDF.Binom** in the **PDF and Noncentral PDF** function group. The parameters for the command are  $x$  (the number we're interested in – 3, here),  $n$  (the number of trials, 10), and  $p$  (the probability of a “success”, which is 0.3).



The probability is shown to be 0.2668 (about 26.7%). If you don't see enough decimals in your answer, click the Variable View tab in the worksheet and increase them using the arrows at the right side of the box (or type in the number of places your want).

Prob
0.2668

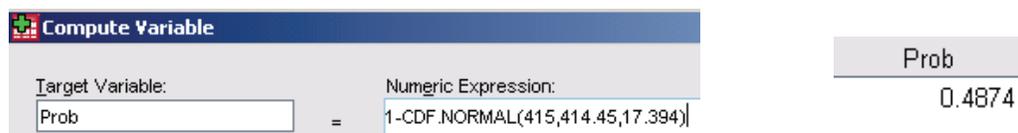
To find the probability of 3 or more, we subtract the probability of 2 or less from 1. The probability of 2 or less is found using CDF.Binom from the CDF and Noncentral CDF function group. The chance of missing at least 3 errors is 0.6172 (61.7%).



Prob
0.6172

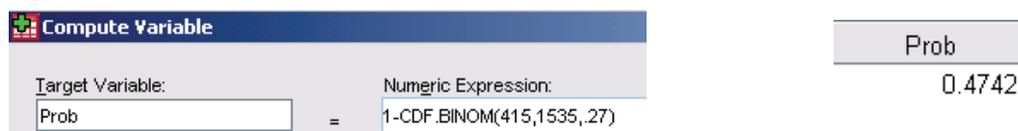
**13.11** The mean is  $\mu = np = 1535 * .27 = 414.45$ . The standard deviation is  $\sigma = \sqrt{np(1-p)} = \sqrt{1535 * .27 * .73} = 17.394$ . We're interested in the chance that more

than 415 students enroll; this is the complement of less than 415. Use **Transform, Compute Variable** and **CDF.Normal** to find this probability. Since we want the probability of more than 415, using the Normal function, we'll subtract the probability of less than or equal to 415 from 1.



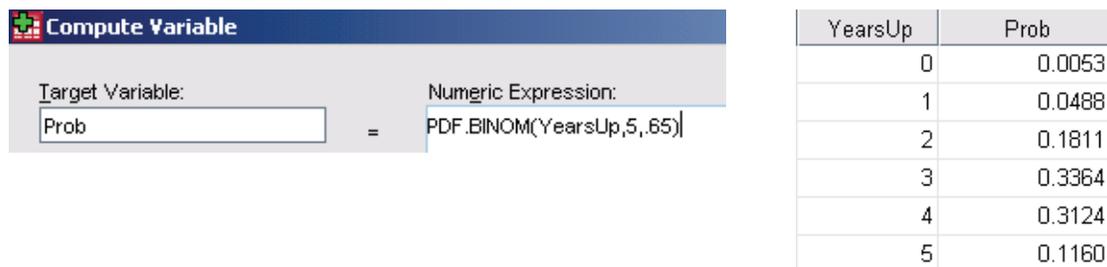
Prob
0.4874

We find that the Normal approximation for this probability is  $1 - 0.5126 = 0.4874$ . To find the exact probability, use **CDF.Binom** and subtract the probability of 415 or fewer from 1. We have 0.4742. The two calculations differ by about 1.3%.



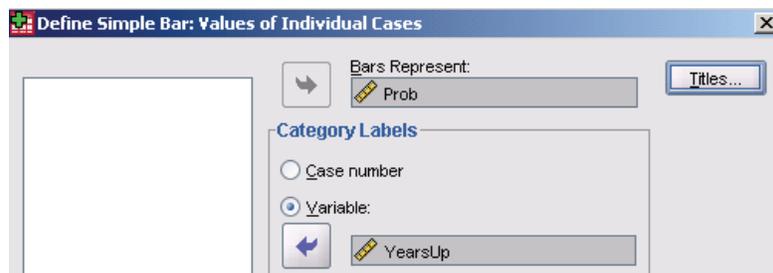
Prob
0.4742

**13.25** Since we are interested in five years,  $n = 5$ , and  $p = 0.65$ .  $X$  can take values 0, 1, 2, 3, 4, and 5. We can find the probability of each value by entering 0 through 5 in a worksheet column (we called it **YearsUp**), then use **Transform, Compute Variable** and **PDF.Binom** to find the probabilities of each possibility. Note that variable **YearsUp** has been used instead of specifying a particular number.

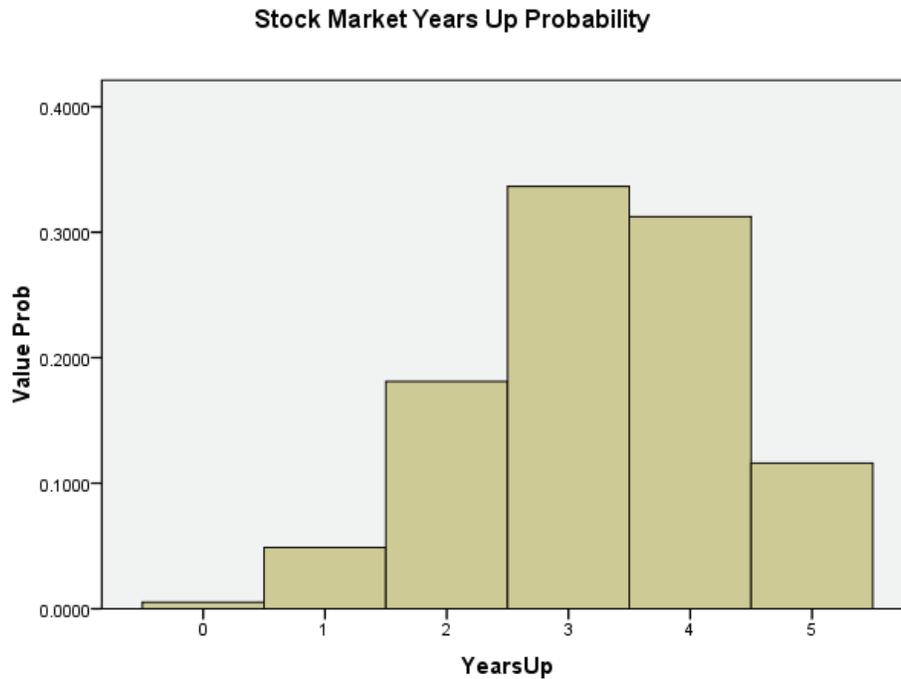


YearsUp	Prob
0	0.0053
1	0.0488
2	0.1811
3	0.3364
4	0.3124
5	0.1160

To graph the probability histogram, we'll make a bar chart of the data and then use the Chart Editor to connect the bars. Click **Graphs, Legacy Dialogs, Bar**. Select that Data in Chart are **Values of individual cases**, and proceed to **Define** the plot.



Click to enter that the bars represent the value of **Prob** (the probability and that the Category Labels are **YearsUp**. Give your graph an appropriate **Titles**, and **OK**. Now, we'll connect the bars. Double-click in the graph to bring up the Chart Editor, then double-click in any bar. Click the **Bar Options** tab, and move the slider to change Bar Width to 100%. **Apply** the change, and **Close** the Properties Box and the Chart Editor. Our finished graph is below.



The mean is  $\mu = np = 5 * .65 = 3.25$ . The standard deviation is  $\sigma = \sqrt{np(1-p)} = \sqrt{5 * .65 * .35} = 1.067$ .

**13.27** We can assume the women are independent of each other, there is an (assumed) constant probability of becoming pregnant, the women either become pregnant or not, and we have a fixed number (20) of women of interest, so this setting fulfills all the requirements for a binomial. We find the probability of at least 1 ( $X \geq 1$ ) as the complement of none.

Compute Variable	
Target Variable:	Numeric Expression:
Prob	1-PDF.BINOM(0,20,.01)

Prob
0.1821

Under ideal conditions, there is an 18.2% chance of at least one pregnancy. To find the probability under typical use, repeat the above calculation, but change the success probability to 0.05.

Compute Variable		Prob
Target Variable:	Numeric Expression:	
Prob	= 1-PDF.BINOM(0,20,.05)	0.6415

Under typical use, there is a  $1 - 0.358 = 64.2\%$  chance of at least one pregnancy in 20 women.

**13.29** The Normal approximation can be used; we have  $\mu = np = 500 * .05 = 25$ , and  $n(1 - p) = 475$ . The standard deviation is  $\sigma = \sqrt{np(1 - p)} = \sqrt{500 * .05 * .95} = 4.873$ . The Normal approximation yields a probability of 0.50 that at least 25 will become pregnant (remember, 25 is the mean for this Normal distribution). The binomial probability is (remember, 25 or more is the complement of 24 or less)  $1 - 0.4714 = 0.5286$ . These differ by 2.86%. We can't use the Normal approximation for the ideal case because  $np = 5$ .

Compute Variable		Prob
Target Variable:	Numeric Expression:	
Prob	= 1-CDF.Binom(24,500,.05)	0.5286

**13.31** We use **PDF.Binom** to find the probability that 6 of the 8 have red blossoms.

Compute Variable		Prob
Target Variable:	Numeric Expression:	
Prob	= PDF.Binom(6,8,.75)	0.3115

The probability is 31.15%. The mean for 80 plants is  $\mu = np = 80 * .75 = 60$ , and the standard deviation is  $\sigma = \sqrt{80 * .75 * .25} = 3.873$ . Using the Normal approximation, the probability of at least 60 red blossomed plants is 0.5 (60 is the mean of this distribution). To find the exact probability, we use the fact that at least 60 is the complement of 59 or less. The exact probability is  $1 - 0.4403 = 55.97\%$ ; the two differ by almost 6%.

Compute Variable		Prob
Target Variable:	Numeric Expression:	
Prob	= 1-CDF.Binom(59,80,.75)	0.5597

**13.33** The mean is  $\mu = np = 25000 * .21 = 5250$ , and the standard deviation is  $\sigma = \sqrt{25000 * .21 * .79} = 64.401$ . Use **CDF.Normal** to find the approximate probability that at least 5000 dropouts receive the flyer is 99.99%.

Compute Variable		Prob
Target Variable:	Numeric Expression:	
Prob	= 1-CDF.Normal(5000,5250,64.401)	0.9999

**13.35** Jodi's mean is  $\mu = np = 100 * .75 = 75$ , and her standard deviation is  $\sigma = \sqrt{100 * .75 * .25} = 4.330$ . We use **CDF.Normal** to find the probability of at most 80 correct, and subtract the probability of at most 70 correct. The result is 0.7518, she has about a 75% chance of scoring between 70 and 80.

Compute Variable		Prob
Target Variable:	Numeric Expression:	
Prob	= CDF.Normal(80,75,4.33)-CDF.Normal(70,75,4.33)	0.7518

For the 250 question test, the mean is  $250 * .75 = 187.5$ , and the standard deviation is 6.847. Her chance of scoring between 70% (175 correct) and 80% (200 correct) on that test is  $0.966 - 0.034 = 93.2\%$ .

Compute Variable		Prob
Target Variable:	Numeric Expression:	
Prob	= CDF.Normal(200,187.5,6.847)-CDF.Normal(175,187.5,6.847)	0.9321

**13.39** We were given the information that 80% of unvaccinated people exposed to the virus will develop the infection. We want the probability that at least 75% of their 1400 students (that's at least 1050 students) would develop the infection if not treated. The mean is  $\mu = np = 1400 * .80 = 1120$ , and the standard deviation is  $\sigma = \sqrt{1400 * .80 * .20} = 14.967$ . The Normal approximation gives us a probability of (to four decimal places) 1.0000; it is virtually guaranteed that at least 75% would get sick if not treated.

Compute Variable		Prob
Target Variable:	Numeric Expression:	
Prob	= 1-CDF.Normal(1050,1120,14.967)	1.0000

**13.41** The probability of exactly one infection in the three unvaccinated children is 9.6%. The probability of exactly one infection in the 17 vaccinated children is 37.4%. Since these are independent, we can multiply these together to find the probability of exactly one in each group,  $0.096 * 0.374 = 0.0359$  (3.6%).

Compute Variable	
Target Variable:	Numeric Expression:
Prob	PDF.Binom(1,3,.8)

Prob
0.0960

Compute Variable	
Target Variable:	Numeric Expression:
Prob	PDF.Binom(1,17,.05)

Prob
0.3741

We could also have two infections in the unvaccinated group or two in the vaccinated group. These probabilities are 0.384 for the unvaccinated group and 0.1575 for the vaccinated group. Adding all these results together, there is a  $0.0359 + 0.384 + 0.1575 = 0.5774$  (about 57.7%) chance of two whooping cough infections in this group of 20 children.

Compute Variable	
Target Variable:	Numeric Expression:
Prob	PDF.Binom(2,3,.8)

Prob
0.3840

Compute Variable	
Target Variable:	Numeric Expression:
Prob	PDF.Binom(2,17,.05)

Prob
0.1575

## Chapter 14 SPSS Solutions

**\*\*NOTE:** SPSS does not do inference based on Z distributions, nor does it perform inference on variables that are already summarized. If you really want to use SPSS for these problems or chapters, follow the instructions below (you'll be basically using **Transform, Compute Variable** as a calculator) or use another technology (such as a graphing calculator or another statistics program like Minitab or Crunchit.)

**14.3** We'll use **Transform, Compute Variable** and **IDF.Normal** from the **Inverse DF** function group. With 97.5% in the center of a standard Normal distribution, there is  $0.025/2 = 0.0125$  on each end. Due to the symmetry of the distribution, we can find either  $-z^*$  (and remove the negative sign) or  $z^*$ .  $z^* = 2.2414$ . As always, if you do not see all the decimal places you want, go to the **Variable View** and change them.



Zstar
-2.2414

**14.5** Open data file *ex14-05*. To create the stemplot, use **Analyze, Descriptive Statistics, Explore**.

IQ Stem-and-Leaf Plot

Frequency	Stem &	Leaf
2.00	Extremes	(=<74)
2.00	8 .	69
4.00	9 .	1368
9.00	10 .	023334578
10.00	11 .	1122244489
2.00	12 .	08
2.00	13 .	02

Stem width: 10  
Each leaf: 1 case(s)

This stemplot indicates there are two low outliers ( $=<74$ , namely 72 and 74). In the Descriptives block of the output, a confidence interval is given (set the confidence level using the **Statistics** button in the Explore dialog box).

Descriptives			
		Statistic	Std. Error
IQ	Mean	105.84	2.563
	99% Confidence Interval for Mean	Lower Bound	98.79
		Upper Bound	112.89

This confidence interval is based on a distribution we won't meet until Chapter 17 (the  $t$  distributions). To find the confidence interval, we'll calculate it "by hand." We will, however, make use of the mean given above.

Compute Variable	
Target Variable:	Numeric Expression:
Low	$105.84 - 2.576 * 15 / \sqrt{31}$

Low
98.90

Compute Variable	
Target Variable:	Numeric Expression:
High	$105.84 + 2.576 * 15 / \sqrt{31}$

High
112.78

Based on this sample, with 99% confidence, the average IQ score of all seventh-grade girls in this school district is between 98.9 and 112.8.

**14.13** When  $\mu = 0$ , the distribution of  $\bar{x}$  will be  $N(0, 1/\sqrt{10} = 0.316)$ . The  $P$ -value is the area to the right of  $\bar{x} = 0.3$  on this distribution. We use **Transform, Compute Variable** and **CDF.Normal** to find the  $P$ -value is  $1 - 0.8288 = 0.1712$ .

Compute Variable	
Target Variable:	Numeric Expression:
Pvalue	$\text{CDF.Normal}(0.3, 0, 0.316)$

Pvalue
0.8288

**14.17** For  $n = 6$  measurements, the standard deviation of the sampling distribution is  $\sigma = .2/\sqrt{6} = 0.0816$ . For this two-sided alternative, we'll double the area to the left of our observed sample mean (since these are less than the claimed mean). We'll find this area using **Transform, Compute Variable** and **CDF.Normal**.

Compute Variable	
Target Variable:	Numeric Expression:
Pvalue	$2 * \text{CDF.Normal}(4.98, 5, 0.0816)$

Pvalue
0.8064

Compute Variable	
Target Variable:	Numeric Expression:
Pvalue	$2 * \text{CDF.Normal}(4.7, 5, 0.0816)$

Pvalue
0.0002

A sample mean of 4.98 (very close to 5) has  $P$ -value 0.8064, while the sample mean of 4.7 (much farther away from 5) has  $P$ -value 0.0002; this is much better evidence against the null as it is farther away.

**14.19** We'll enter the data and use **Analyze, Basic Statistics, Descriptives** to find the mean of these data, then compute the test "by hand."

Descriptive Statistics					
	N	Minimum	Maximum	Mean	Std. Deviation
Conduct	6	4.73	5.32	4.9883	.23828
Valid N (listwise)	6				

Target Variable: Z = Numeric Expression:  $(4.9883-5)/(.2/\sqrt{6})$

Z

-0.14

With the "not equal" (two-tailed) alternate hypothesis, the  $P$ -value of the test is twice the area below  $z = -0.14$ .

Target Variable: Pvalue = Numeric Expression:  $2*\text{CDF.Normal}(-0.14,0,1)$

Pvalue

0.8887

These data are *not* good evidence that the mean conductivity is not 5.

**14.21** The  $P$ -value of this test is the area to the right of  $z = 1.776$ . We find this using **CDF.Normal**,

Target Variable: Pvalue = Numeric Expression:  $1-\text{CDF.Normal}(1.776,0,1)$

Pvalue

0.0379

The  $P$ -value is 0.0379, so this result is significant at the  $\alpha = 0.05$  level ( $P < \alpha$ ), but not at the 0.01 level.

**14.23** We compute the test statistic and  $P$ -value below.

Target Variable: Z = Numeric Expression:  $(0.4365-.5)/(.2887/\sqrt{100})$

Z

-2.20

Compute Variable		Pvalue
Target Variable:	Numeric Expression:	
Pvalue	= 2*CDF.Normal(-2.20,0,1)	0.0278

The test statistic is  $z = -2.20$  with  $P$ -value 0.0278. This result is significant at the  $\alpha = 0.05$  level ( $P < \alpha$ ), but not at the 0.01 level.

**14.35** Again, we compute the confidence interval “by hand.” If you don’t know the value of  $z^*$ , use **IDF.Normal** to find it.

Compute Variable		Low
Target Variable:	Numeric Expression:	
Low	= 2.35-1.645*2.5/sqrt(200)	2.06

Compute Variable		High
Target Variable:	Numeric Expression:	
High	= 2.35+1.645*2.5/sqrt(200)	2.64

Based on this sample, the mean “muscle gap” for American young men (this is where the sample was from) is between 2.06 and 2.64 kg/m<sup>2</sup>, with 90% confidence.

**14.41** This is a continuation of Exercise 14.35 (above). If we assume that  $\mu$  is the difference women’s preference minus what they have, we have hypotheses  $H_0 : \mu = 0$ ,  $H_a : \mu > 0$ . Since the alternative is “greater than” the  $P$ -value will be the area above the computed test statistic.

Compute Variable		Z
Target Variable:	Numeric Expression:	
Z	= (2.35-0)/(2.5/sqrt(200))	13.29

The test statistic is  $z = 13.29$  with  $P$ -value essentially 0. We know this  $P$ -value should be very small because the 68-95-99.7 Rule states that being more than 3 standard deviations above or below the mean is extremely unusual.

**14.51** Open data file *ex14-51* and use **Analyze, Descriptive Statistics, Explore** to make the stemplot.

Change Stem-and-Leaf Plot

Frequency	Stem &	Leaf
1.00	-8 .	3
2.00	-7 .	08
5.00	-6 .	25588
8.00	-5 .	12233679
7.00	-4 .	0347799
5.00	-3 .	01368
9.00	-2 .	011223557
3.00	-1 .	008
2.00	-0 .	38
3.00	0 .	234
1.00	1 .	7
1.00	2 .	2

Stem width: 1.0  
Each leaf: 1 case(s)

Based on the graph above, there are no strong departures from Normality, so proceeding with inference is reasonable. We use the mean computed by SPSS in calculating the confidence interval.

Descriptives			
		Statistic	Std. Error
Change	Mean	-3.587	.3655
	95% Confidence Interval for Mean		
	Lower Bound	-4.323	
	Upper Bound	-2.852	

**Compute Variable**

Target Variable:  = Numeric Expression:

Low  
-4.53

**Compute Variable**

Target Variable:  = Numeric Expression:

High  
-2.65

We are 99% confident the mean bone loss of all breast-feeding mothers is between 4.53% and 2.65%, based on this sample of 47 mothers.

**14.53** If you did Exercise 14.51, you should have noted that the interval does not contain 0, indicating that breast-feeding mothers *do* lose bone mineral, on average, and that this result is statistically significant. We'll compute the  $z$  test statistic first for this test.

**Compute Variable**

Target Variable:  = Numeric Expression:

Z  
-9.84

We have a test statistic of  $z = -9.837$  with  $P$ -value 0 (being almost 10 standard deviations below the mean and essentially no chance of happening). This confirms that breast-feeding mothers lose bone mineral, on average.

**14.55** If  $\mu$  is the mean difference in sensitivity, (with – without grease), we have hypotheses  $H_0 : \mu = 0$ ,  $H_a : \mu > 0$ , since the question is if grease increases sensitivity. We use **Analyze, Descriptive Statistics, Descriptives** to find the mean of the data, then compute the test and  $P$ -value.

	N	Minimum	Maximum	Mean	Std. Deviation
Diff	16	-.18	.64	.1012	.22633
Valid N (listwise)	16				

**Compute Variable**

Target Variable:  = Numeric Expression:

Z  
1.84

**Compute Variable**

Target Variable:  = Numeric Expression:

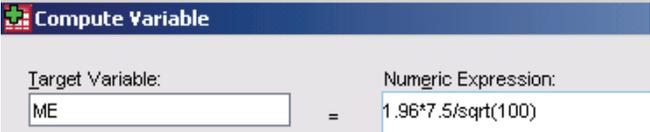
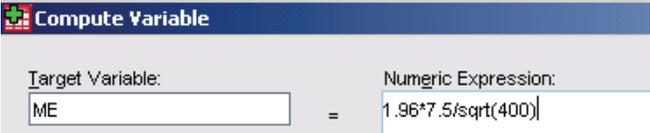
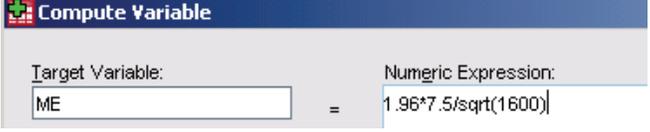
Pvalue  
0.0329

Our test statistic is  $z = 1.84$  with  $P$ -value 0.0329. At the 5% level, these results indicate that eye grease *does* increase sensitivity, on average.

## Chapter 15 SPSS Solutions

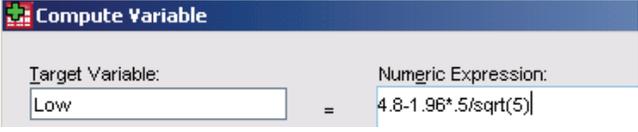
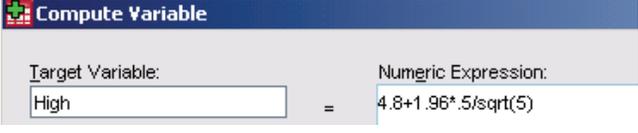
**\*\*NOTE:** SPSS does not do inference based on Z distributions, nor does it perform inference on variables that are already summarized. If you really want to use SPSS for these problems or chapters, follow the instructions below (you'll be basically using **Transform, Compute Variable** as a calculator) or use another technology (such as a graphing calculator or another statistics program like Minitab or Crunchit.)

**15.5** We'll compute the confidence intervals for each sample size using Transform, Compute Variable. Note that the only thing that changes each time is the sample size.

 <p>Target Variable: ME = Numeric Expression: <math>1.96 * 7.5 / \sqrt{100}</math></p>	<table border="1"> <tr><td>ME</td></tr> <tr><td>1.47</td></tr> </table>	ME	1.47
ME			
1.47			
 <p>Target Variable: ME = Numeric Expression: <math>1.96 * 7.5 / \sqrt{400}</math></p>	<table border="1"> <tr><td>ME</td></tr> <tr><td>0.74</td></tr> </table>	ME	0.74
ME			
0.74			
 <p>Target Variable: ME = Numeric Expression: <math>1.96 * 7.5 / \sqrt{1600}</math></p>	<table border="1"> <tr><td>ME</td></tr> <tr><td>0.37</td></tr> </table>	ME	0.37
ME			
0.37			

As sample size gets larger, the margin of error decreases; specifically, if we quadruple the sample size, the margin of error is cut in half.

**15.9** We compute confidence intervals for each sample size, the only thing that changes, so after the first interval, the expressions are not given.

 <p>Target Variable: Low = Numeric Expression: <math>4.8 - 1.96 * 5 / \sqrt{5}</math></p>	<table border="1"> <tr><td>Low</td></tr> <tr><td>4.36</td></tr> </table>	Low	4.36
Low			
4.36			
 <p>Target Variable: High = Numeric Expression: <math>4.8 + 1.96 * 5 / \sqrt{5}</math></p>	<table border="1"> <tr><td>High</td></tr> <tr><td>5.24</td></tr> </table>	High	5.24
High			
5.24			

Low	High	Low	High
4.55	5.05	4.65	4.95
$n = 15$		$n = 40$	

We can see clearly that the intervals become narrower (and the margin of error smaller) with increasing sample size.

**15.11** To compute the required sample size, we use the formula

$$n = \left( \frac{z^* \sigma}{ME} \right)^2$$

Use **IDF.Normal** to find  $z^*$  if we do not already know it. For 95% confidence there is 0.025 (2.5%) to the left of  $-z^*$ . The correct value is  $z^* = 1.96$ . We find we need a sample of at least 217 (since the *smallest* that will work is 216.09).

Compute Variable		N
Target Variable:	Numeric Expression:	
N	$(1.96 * 7.51)^2$	216.09

**15.41** We entered the data and used **Analyze, Descriptive Statistics, Descriptives** to find the mean. We then compute the test and its  $P$ -value.

Descriptive Statistics					
	N	Minimum	Maximum	Mean	Std. Deviation
Samples	5	6.47	13.63	9.5240	2.77142
Valid N (listwise)	5				

Compute Variable		Z
Target Variable:	Numeric Expression:	
Z	$(9.524 - 8) / (2 / \sqrt{5})$	1.70

Compute Variable		Pvalue
Target Variable:	Numeric Expression:	
Pvalue	$2 * (1 - \text{CDF.Normal}(1.70, 0, 1))$	0.0891

We have  $z = 1.70$  with  $P$ -value 0.0884. This result is *not* significant at the 5% level. This result is not surprising; this very small sample will have low power.

**15.49** The test statistic is

$$z = \frac{\bar{x} - 5}{.2/\sqrt{6}}$$

To reject  $H_0$  at the 5% level, we need a test statistic  $|z| \geq 1.96$ . Solving for  $\bar{x}$ , this gives  $z < 4.84$  or  $z > 5.16$ . To find the probability of rejecting the null hypothesis, use **CDF.Normal** as shown below.

Compute Variable	
Target Variable:	Numeric Expression:
Problow	CDF.Normal(4.84,5.1,0.0816)

Problow
0.0007

Compute Variable	
Target Variable:	Numeric Expression:
Probhigh	1-CDF.Normal(5.16,5.1,0.0816)

Probhigh
0.2311

Add the two to get power = 0.2318.

## Chapter 16 SPSS Solutions

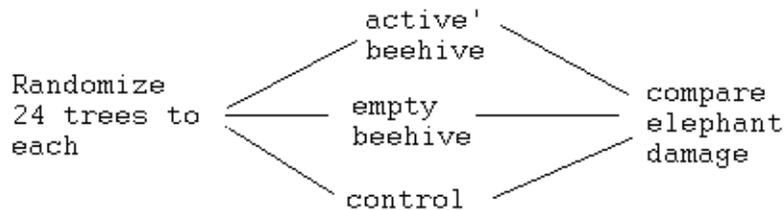
**16.3** Using technology, we will number the students from 1 to 3,478 (if we were using the random digits table, we'd use 0001 to 3478). Since we want five students randomly selected, place a 1 in row 5 of the first variable (column) in a new worksheet. We'll use **Transform, Compute Variable** with RV.Uniform as shown below to generate random numbers in our target range. Ignoring the decimals (truncating our selections), we have individuals numbered 486, 1500, 2129, 1011, and 542.

Compute Variable		
Target Variable:	Numeric Expression:	
Selected	= RV.UNIFORM(1,3478)	

	VAR00001	Selected
1	.	486.54
2	.	1500.64
3	.	2129.55
4	.	1011.95
5	1.00	542.48

**16.5** An outline of the experiment might look like that below. The response variable in this experiment is tree damage caused by elephants.



Similar to the procedures used in Exercise 16.3 above, we'll place a 1 in the fourth row of a new worksheet and use RV.UNIFORM to select numbers between 1 and 72. Our selection is (again ignoring the decimals) trees 50, 25, 32, and 4.

Compute Variable		
Target Variable:	Numeric Expression:	
Selected	= RV.UNIFORM(1,72)	

	VAR00001	Selected
1	.	50.66
2	.	25.59
3	.	32.64
4	1.00	4.72

**16.13** We "hand compute" the confidence interval, using  $z^* = 1.645$  for 90% confidence.

Compute Variable		
Target Variable:	Numeric Expression:	
Low	= $172 - 1.645 * 41 / \sqrt{14}$	

Low
153.97

Compute Variable		High
Target Variable:	Numeric Expression:	190.03
High	= 172+1.645*41/sqrt(14)	

Based on this sample, we're 90% confident the mean cholesterol reading for cross-country runners is between 154.0 and 190.0 mg/dl (rounding to one more decimal place than the data given).

**16.15** The margin of error in Exercise 16.13 is  $190 - 172 = 18$ . To cut this in half, we'll use the formula

$$n = \left( \frac{z^* \sigma}{ME} \right)^2$$

with  $ME = 9$ . We'll need a sample of at least 57 runners to have a margin of error 9.

$$\frac{(1.645 * 41 / 9)^2}{}$$

56.15837068

**16.17** For 95% confidence,  $z^* = 1.96$ . We use this in computing the ends of the interval.

Compute Variable		Low
Target Variable:	Numeric Expression:	322.35
Low	= 357-1.96*50/sqrt(8)	

Compute Variable		High
Target Variable:	Numeric Expression:	391.65
High	= 357+1.96*50/sqrt(8)	

Based on this sample, we estimate the mean level of pesticides in minke whales in the West Greenland area is between 322.4 and 391.7 ng/g, with 95% confidence.

**16.19** We use the same computations as in Exercise 16.17 above, changing  $z^*$  from 1.96 (95% confidence) to 1.28 (80% confidence) and 1.645 (90% confidence).

Low	High	Low	High
334.37	379.63	327.92	386.08

The 80% interval is from 334.4 to 379.7 ng/g; the 90% interval is from 327.9 to 386.1 ng/g. We see that as confidence levels increase, the intervals become wider.

**16.21** From Exercise 16.20, we have  $n = 113$ ,  $\bar{x} = 87.6$ , and  $\sigma = 15$ . Our hypotheses are  $H_0 : \mu = 100$ ,  $H_a : \mu < 100$ . We compute our test statistic and its P-value as shown below.

Compute Variable	
Target Variable:	Numeric Expression:
z	(87.6-100)/(15/sqrt(113))

z
-8.79

Compute Variable	
Target Variable:	Numeric Expression:
Pvalue	CDF.Normal(-8.79,0,1)

Pvalue
0.0000

We have  $z = -8.79$  with  $P$ -value 0. This is overwhelming evidence that very-low-birth-weight children have lower than average IQs.

**16.27** For 15 children,  $\bar{x}$  should have mean 445 (the same as the population mean), and standard deviation  $\sigma = 82/\sqrt{15} = 21.172$ . For 150 children, the mean of  $\bar{x}$  will be the same, but the standard deviation will be  $\sigma = 82/\sqrt{150} = 6.695$ . A sample size of  $n = 15$  is not large enough to consider the mean Normally distributed by the Central Limit Theorem with a strongly skewed distribution; with  $n = 150$ , this is much more reasonable. To find the probability that the mean reaction time for a sample of 150 is greater than 450 ms, use **CDF.Normal**. We find this probability is about 22.8%.

Compute Variable	
Target Variable:	Numeric Expression:
Pvalue	1-CDF.Normal(450,445,6.695)

Pvalue
0.2276

**16.45** Use **Analyze, Descriptive Statistics, Explore** to create the stemplot and compute the mean.

Temp Stem-and-Leaf Plot

Frequency	Stem &	Leaf
1.00	96 .	8
3.00	97 .	344
6.00	97 .	888889
4.00	98 .	0133
4.00	98 .	5789
.00	99 .	
1.00	99 .	6
1.00	Extremes	(>=100.3)

Stem width: 1.00

The stemplot does indicate one mild outlier (100.3) as a Extreme, but this values is not unreasonable for these data. We compute the test statistic and its P-value as shown below (having noted in the Explore output that the mean of these data is 98.203).

<b>Compute Variable</b>		Z
Target Variable:	Numeric Expression:	-2.54
Z	$(98.203 - 98.6) / (.7 / \sqrt{20})$	
<b>Compute Variable</b>		Pvalue
Target Variable:	Numeric Expression:	0.0111
Pvalue	$2 * \text{CDF.Normal}(-2.54, 0, 1)$	

With test statistic  $z = -2.54$  and  $P$ -value 0.0111, our conclusion depends on the  $\alpha$ -level of the test. At the 5% level, we'd say these data indicate the average body temperature is *not* 98.6°, at the 1% level, we have not shown there is a difference.

**16.47** We compute the ends of the interval using  $z^* = 1.645$  for 90% confidence.

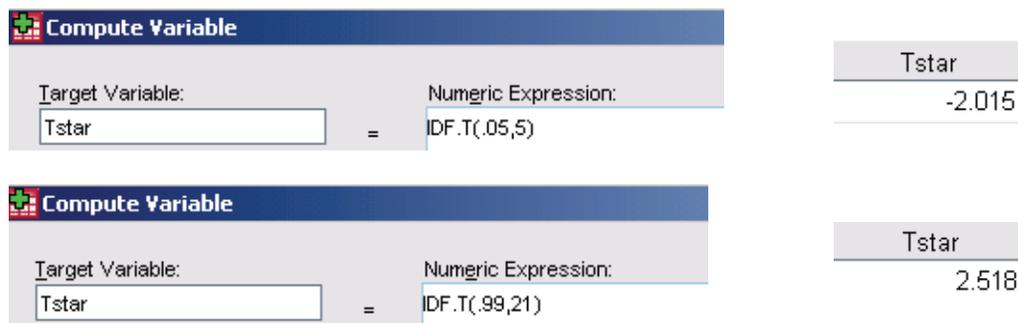
<b>Compute Variable</b>		Low
Target Variable:	Numeric Expression:	97.95
Low	$98.203 - 1.645 * .7 / \sqrt{20}$	
<b>Compute Variable</b>		High
Target Variable:	Numeric Expression:	98.46
High	$98.203 + 1.645 * .7 / \sqrt{20}$	

With 90% confidence, the mean body temperature of healthy adults should be between 97.95° and 98.46°. Note that both ends of this interval are below the assumed “normal” value of 98.6°.

## Chapter 17 SPSS Solutions

**\*\*NOTE:** SPSS does not perform inference on variables that are already summarized. If you really want to use SPSS for these problems or chapters, follow the instructions below (you'll be basically using **Transform, Compute Variable** as a calculator) or use another technology (such as a graphing calculator or another statistics program like Minitab or Crunchit.)

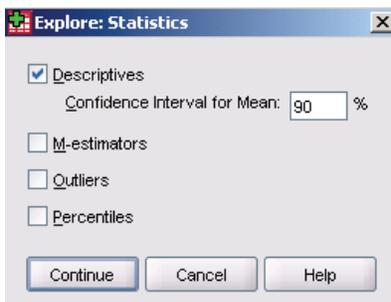
**17.3** We use **IDF.T** from the Inverse DF Function group to find the critical values.

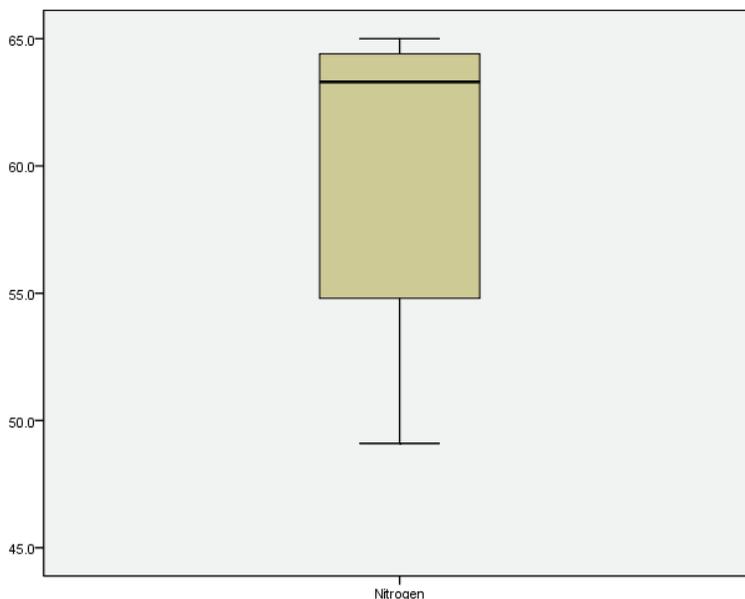


The first screenshot shows the 'Compute Variable' dialog box with 'Target Variable:' set to 'Tstar' and 'Numeric Expression:' set to 'IDF.T(.05,5)'. To the right, a small table shows 'Tstar' with the value '-2.015'.

The second screenshot shows the 'Compute Variable' dialog box with 'Target Variable:' set to 'Tstar' and 'Numeric Expression:' set to 'IDF.T(.99,21)'. To the right, a small table shows 'Tstar' with the value '2.518'.

**17.7** The data have been entered in a column we named **Nitrogen**. If  $\mu$  is the mean nitrogen content of Cretaceous era air, we'd like a 90% confidence interval estimate. First, check the conditions: we're assuming our data come from a SRS; can we believe these data came from an (approximately) Normal distribution? With only 9 data values, a histogram will not show the distribution very well. We'll use **Analyze, Descriptive Statistics, Explore** to create the confidence interval and a boxplot of the data. Click to enter the variable name in the Dependent list, then click **Statistics**. Change the confidence level to 90%, then **Continue** and **OK**.





There are no outliers, but the distribution is definitely skewed; observe the median far to the high end in the box. Use of  $t$  procedures might not be valid, we can only proceed with caution.

The confidence interval is included in the Descriptives table, as shown below (some of the table has been omitted).

Descriptives			
		Statistic	Std. Error
Nitrogen	Mean	59.589	2.0851
	90% Confidence Interval for Mean		
	Lower Bound	55.712	
	Upper Bound	63.466	

Based on these samples, we estimate that Cretaceous era air had between 55.7% and 63.5% nitrogen, with 90% confidence (assuming the distribution is really approximately Normal).

**17.9** With  $n = 25$ , there are  $25 - 1 = 24$  degrees of freedom. Using technology, we'll find an exact  $P$ -value for this test. Since the test is two-tailed, we'll use the symmetry of the distribution and double the area to the *left* of  $t = -1.12$ . Use **Transform**, **Compute Variable** and **CDF.T** from the **CDF and Noncentral CDF** function group. As always, if you want more decimal places than the default 2, increase them on the **Variable View**.

Compute Variable	
Target Variable:	Numeric Expression:
Pvalue	2*CDF.T(-1.12,24)

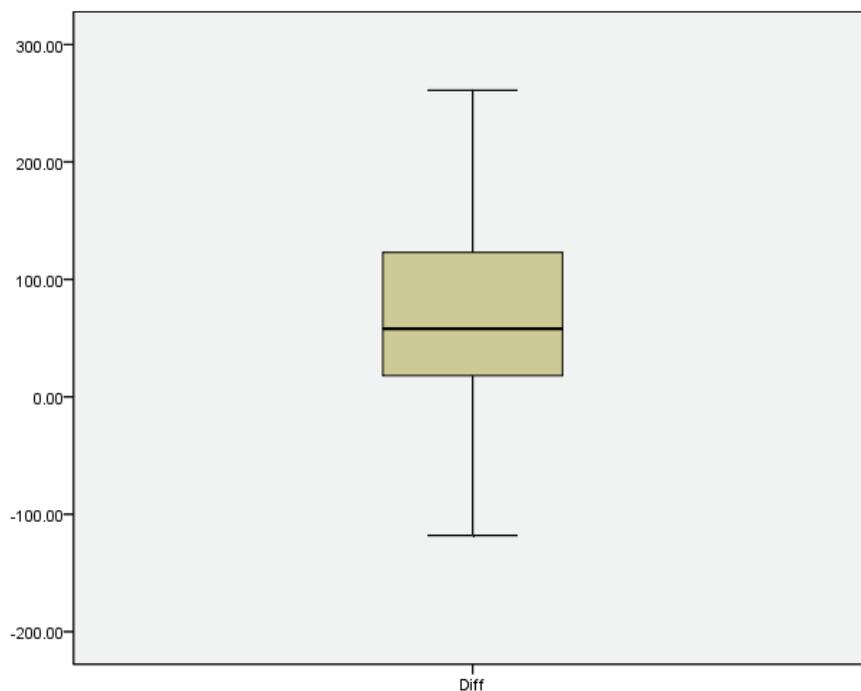
Pvalue
0.2738

Our P-value is 0.2738. This test is not significant at any normal alpha level.

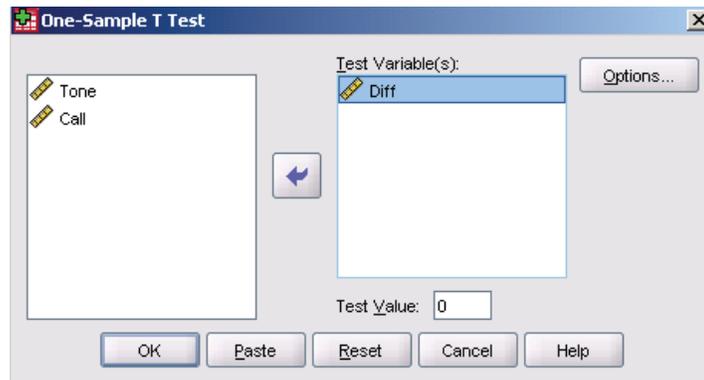
**17.11** This is a matched pairs situation (we have the response to both tone and call for each neuron). If  $\mu_D$  is the mean of the differences (call – tone), we'll have hypotheses  $H_0 : \mu_D = 0$  and  $H_a : \mu_D > 0$ . SPSS has a built-in matched pairs t test procedure, but it doesn't create any graphs to check assumptions (such as the data come from an approximately Normal distribution). Use **Transform, Compute Variable** to create a variable named **Diff** as shown below.



We'll now create a boxplot of these differences using **Graphs, Legacy Dialogs, Boxplot**. Use the **Summaries of Separate Variables** option for a **Simple** Boxplot.



Our boxplot looks symmetric (with no outliers), so  $t$  procedures are justified. At this point, we have formed the differences, so we can use **Analyze, Compare Means, 1-Sample T Test** to perform our test.

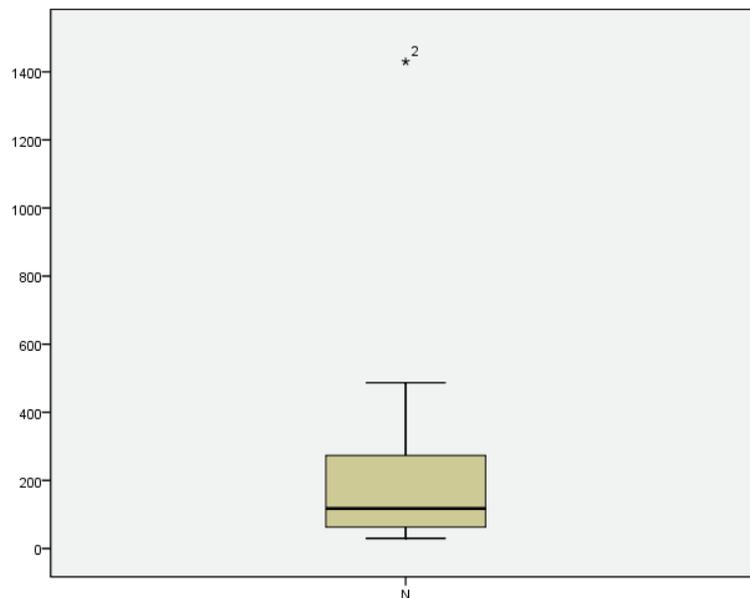


Since our null hypothesis is that the mean difference is 0, the Test Value is set to 0 (the default).

One-Sample Test						
Test Value = 0						
	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
					Lower	Upper
Diff	4.840	36	.000	70.37838	40.8888	99.8679

SPSS gives only two-sided  $P$ -values (in the Sig column). To find the one-sided  $P$ -value, divide by 2; however  $.000/2 = .000$ , so this test rejects the null hypothesis. With a test statistic of  $t = -4.84$  and  $P$ -value of 0.000, this is extremely strong evidence that the call response is stronger than the tone response, on average, in macaque monkeys.

**17.13** We use **Graphs, Legacy Dialogs, Boxplot** to create a boxplot of the Nitrogen data (labeled N in data file *ta17-03*). We are looking for skewness and outliers. The graph shows this data set has an extreme outlier; further, the main portion seems rather skewed right (toward the high end). We can't trust  $t$  procedures for these data.



**17.25** We'll recompute the  $t$  statistics and find the correct  $P$ -values using **Transform, Compute Variable**. We also use the symmetry of the  $t$  distributions; the area to the right of a positive value is the same as the area to the left of the negative of that value. For the student group, we have

Compute Variable	
Target Variable: T	Numeric Expression: =.08*(.37/sqrt(12))

T
0.75

Compute Variable	
Target Variable: Pvalue	Numeric Expression: =2*CDF.T(-.75,11)

Pvalue
0.4690

For the student group, we have  $t = 0.75$  with  $P$ -value 0.469 (the conclusion was correct, however, there is no significant effect here). Repeating for the non-student group, we find have  $t = 3.28$  with  $P$ -value 0.0073; there is a significant effect in this group.

Compute Variable	
Target Variable: T	Numeric Expression: =.35*(.37/sqrt(12))

T
3.28

Compute Variable	
Target Variable: Pvalue	Numeric Expression: =2*CDF.T(-3.28,11)

Pvalue
0.0073

**17.27** With a sample size of  $n = 1470$ , this is certainly large enough to appeal to the Central Limit Theorem, and call  $\bar{x}$  approximately Normal. To find the confidence interval, we'll first find the critical value  $t^*$  using **IDF.T** and then find the confidence interval as  $\bar{x} \pm t^* SE$ .

Compute Variable	
Target Variable: Tstar	Numeric Expression: =IDF.T(.005,1469)

Tstar
-2.58

Compute Variable	
Target Variable: Clow	Numeric Expression: =240-2.58*1.1

Clow
237.16

Compute Variable	
Target Variable:	Numeric Expression:
Chigh	= 240+2.58*1.1

Chigh
242.84

Based on this information, we're 99% confident Atlanta eighth-graders should have a mean TUDA score between 237.2 and 242.8. Since the high end of this interval is below 243 (basic), our indications are that Atlanta eighth-graders, on average, perform below this level.

**17.29** Since each patient was given both treatments, we use a matched pairs  $t$  test to compare the treatments to control variability among the subjects. We are given the summary statistics, so we'll use SPSS as a calculator to compute the test statistic and  $P$ -value. With a test statistic of  $-4.41$  and a  $P$ -value of  $0.0070$ , there is evidence of a significant difference between treatment and control.

Compute Variable	
Target Variable:	Numeric Expression:
T	= (-.326-0)/( .181 /sqrt(6))

T
-4.41

Compute Variable	
Target Variable:	Numeric Expression:
P	= 2*CDF.T(-4.41,5)

P
0.0070

**17.31** We'll use **Analyze, Descriptive Statistics, Explore** to create the stemplot and the confidence interval. Be sure to check Statistics for the correct confidence level (90%).

Count Stem-and-Leaf Plot

Frequency	Stem &	Leaf
3.00	0 .	699
5.00	1 .	01124
2.00	1 .	55
2.00	2 .	02

Stem width: 10  
Each leaf: 1 case(s)

The histogram shows no overt skewness nor any outliers, so we'll check the confidence interval in the Descriptives table.

Descriptives				
			Statistic	Std. Error
Count	Mean		12.83	1.342
	90% Confidence Interval for Mean	Lower Bound	10.42	
		Upper Bound	15.24	

We're 90% confident the average count of correct Blissymbols among children using this program will be between 10.4 and 15.2.

**17.33** The parameter of interest is  $\mu_D$ , the mean difference between the experimental and control limbs. We want to know whether the electrical field slows healing, so forming differences as Experimental – Control, we have hypotheses  $H_0 : \mu_D = 0$  and  $H_a : \mu_D > 0$ . Open data file *ta17\_03.por*. Compute the differences using **Transform, Compute Variable**.



To create the stemplot of the differences, use **Analyze, Descriptive Statistics, Explore**.

diff Stem-and-Leaf Plot

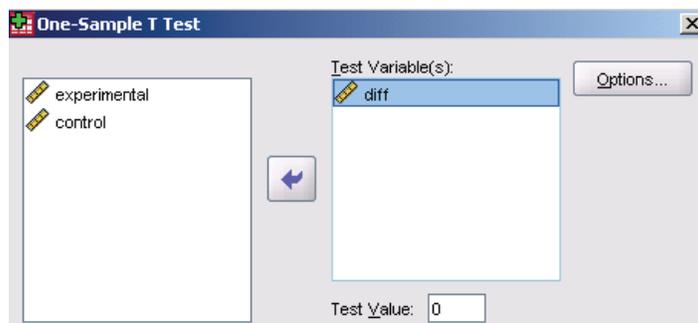
```

Frequency      Stem & Leaf
      1.00 Extremes      (<=-13)
      1.00      -0 . 6
       .00      -0 .
      2.00       0 . 12
      4.00       0 . 5789
      3.00       1 . 012
      1.00 Extremes      (>=31)

Stem width:      10.00
Each leaf:      1 case(s)

```

Since we've actually computed the differences, we use **Analyze, Compare Means, One-Sample T Test** using **diff** as our variable. The test value is 0 (no difference between the control and experimental limbs).



**One-Sample Test**

	Test Value = 0					
	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
					Lower	Upper
diff	2.076	11	.062	6.41667	-.3859	13.2192

With a one-sided  $P$ -value of 0.0310 (divide the SPSS two-tailed by 2), we conclude at the 5% significance level that the electrical field does slow healing, on average. Now, delete the high outlier (31) from **diff** and recalculate the test.

**One-Sample Test**

	Test Value = 0					
	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
					Lower	Upper
diff	1.788	10	.104	4.18182	-1.0291	9.3927

We now have a  $P$ -value of 0.0520; this is insufficient evidence at the 0.05 level that the electrical field (or lack thereof) made a difference in average healing rates.

**17.35** We'll use **Analyze, Descriptive Statistics, Explore** to create the stemplot and the confidence interval. Be sure to check Statistics for the correct confidence level (90%).

Doubling Stem-and-Leaf Plot

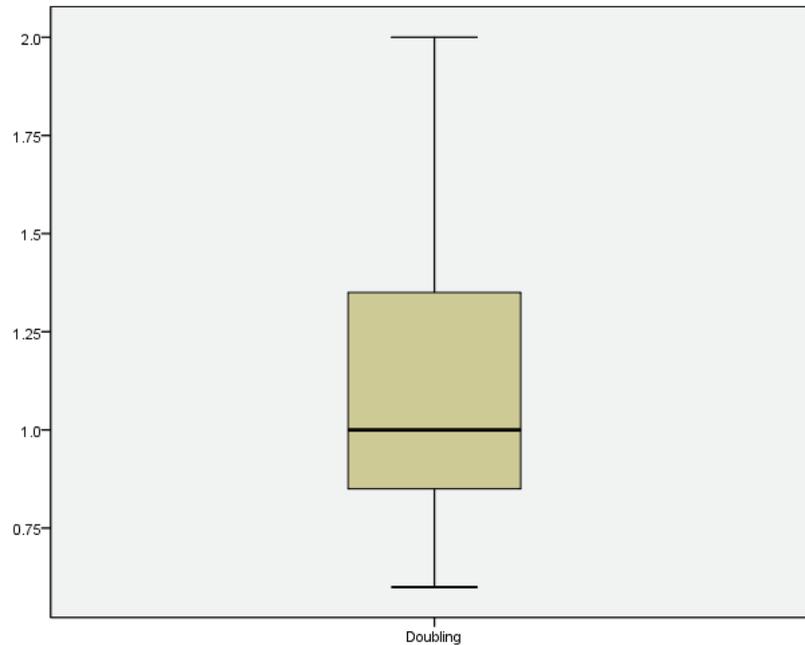
```

Frequency      Stem & Leaf
  4.00          0 . 6789
  5.00          1 . 00334
  1.00          1 . 9
  1.00          2 . 0

Stem width:    1.0
Each leaf:     1 case(s)

```

The distribution is rather skewed right (the high hand side on the boxplot is twice as long as the low from the median out), but there are no outliers; further, all the data values are reasonably close to one another.

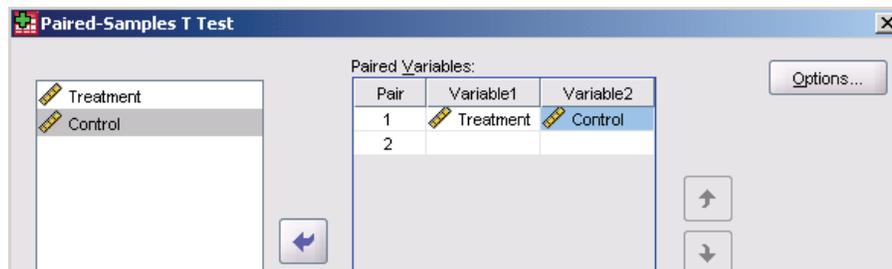


**Descriptives**

		Statistic	Std. Error
Doubling	Mean	1.173	.1389
	90% Confidence Interval for Mean		
	Lower Bound	.921	
	Upper Bound	1.424	
	Kurtosis	-.357	1.279

Based on this sample, the mean doubling time is between 0.92 and 1.42 days (with 90% confidence); however, based on the shape of the data distribution, we'd hesitate to use this for inference about all possible similar patients.

**17.37** Because the investigators believed that extra CO<sub>2</sub> would cause the trees to grow faster, the hypotheses are  $H_0: \mu_D = 0$  and  $H_a: \mu_D > 0$ , where  $\mu_D$  is the mean difference, treatment – control. We enter the Treatment values in a column and the Control in another, then use **Analyze, Compare Means, Paired Samples t test** to perform the test.



Paired Samples Test

		Paired Differences							
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference		t	df	Sig. (2-tailed)
					Lower	Upper			
Pair 1	Treatment - Control	1.916333	1.050494	.606503	-.693238	4.525905	3.160	2	.087

The test statistic is  $t = 3.16$  with  $P$ -value  $0.087/2 = 0.0435$ . While significant at the 5% level, a sample of only  $n = 3$  is not very convincing, and risky because we do not have a good idea of the real variation that might occur.

**17.39** We can use Analyze, Descriptive Statistics, Explore to examine a stemplot (and boxplot) for these data.

Seeds Stem-and-Leaf Plot

```

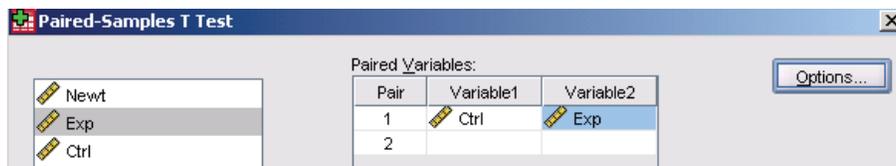
Frequency      Stem & Leaf
 4.00          0 . 1123
 3.00          0 . 788
 4.00          1 . 0011
 7.00          1 . 5677899
 6.00          2 . 011124
 2.00          2 . 58
 2.00 Extremes (>=5973)

Stem width: 1000
Each leaf: 1 case(s)

```

This stemplot indicates 2 high extremes (5973 and larger). With two high outliers, these data are not suitable for  $t$  procedures.

**17.43** If you haven't done Exercise 17.42, you can find the confidence interval for the difference using **Analyze, Compare Means, Paired Samples T Test**. If you did Exercise 17.42, the confidence interval is given on the output from the test. Use **Options** to change the confidence level to 90%.



Paired Samples Test								
Paired Differences								
	Mean	Std. Deviation	Std. Error Mean	90% Confidence Interval of the Difference		t	df	Sig. (2-tailed)
				Lower	Upper			
Pair 1 Ctrl - Exp	5.714	10.564	2.823	.714	10.714	2.024	13	.064

Based on these data, we are 90% confident the difference in mean healing rate will be between 0.71 and 10.71. This indicates the control limb has the faster healing rate.

**17.45** Since all subjects will use both instruments, we'll flip a coin for each to see which hand is used first. So that we can examine the shape of the distribution of differences for skewness and outliers we'll actually compute them using **Transform, Compute Variable**.



We now examine a stemplot of the differences using **Analyze, Descriptive Statistics, Explore**.

Diff Stem-and-Leaf Plot

Frequency	Stem &	Leaf
1.00	-5 .	2
3.00	-4 .	358
3.00	-3 .	115
2.00	-2 .	49
5.00	-1 .	12666
5.00	-0 .	13347
2.00	0 .	02
1.00	1 .	1
2.00	2 .	03
1.00	3 .	8

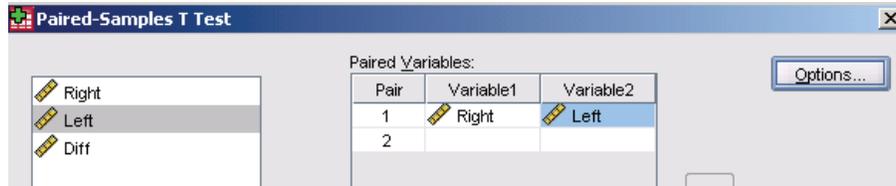
Stem width: 10.00  
Each leaf: 1 case(s)

This distribution is symmetric and shows no outliers (so does the boxplot). Since we believe the right hand times should be faster, we will test

$$H_0 : \mu_D = 0$$

$$H_a : \mu_D < 0$$

(if you subtract the other way, the direction of the alternate hypothesis will change). We'll use **Analyze, Compare Means, Paired Samples T Test**. Since we want the confidence interval in Exercise 17.47, use Options to change the confidence level to 90%.



**Paired Samples Test**

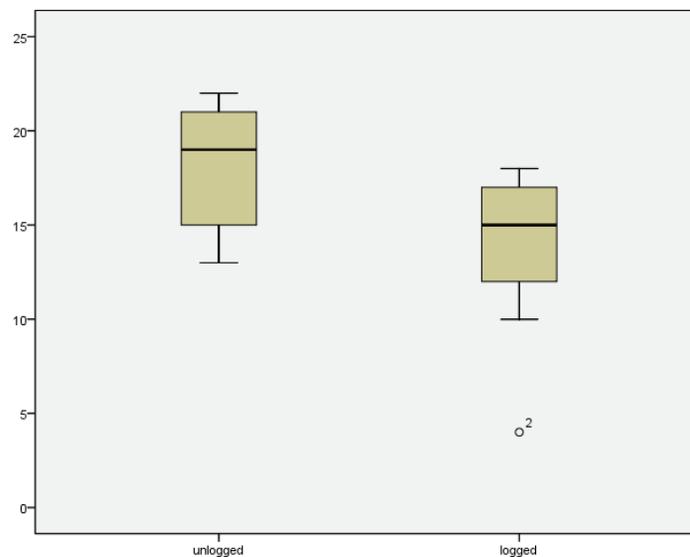
	Paired Differences							
	Mean	Std. Deviation	Std. Error Mean	90% Confidence Interval of the Difference		t	df	Sig. (2-tailed)
				Lower	Upper			
Pair 1 Right - Left	-13.320	22.936	4.587	-21.168	-5.472	-2.904	24	.008

With a test statistic of  $t = -2.90$  and  $P$ -value  $0.008/2 = 0.004$ , we reject the null hypothesis of no difference. This experiment does show that people find right-hand threads easier to use.

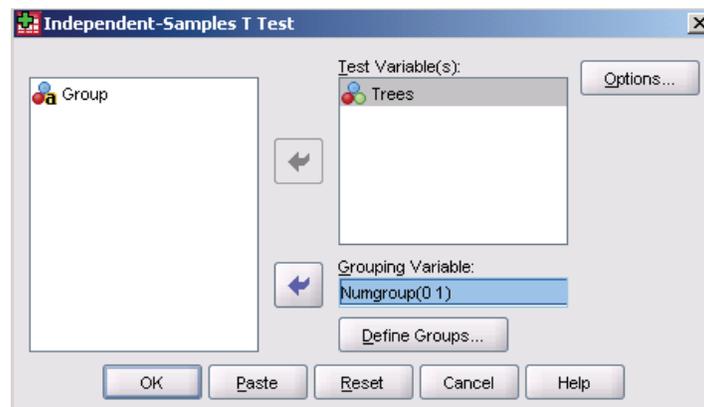
**17.47** If you followed our instructions in the solution to Exercise 17.45, you already have the confidence interval for the difference. We are 90% confident that the right-hand thread will save between 5.5 and 21.2 seconds. This could be of great importance if a task were performed over and over – a minute might be saved for every three repetitions.

## Chapter 18 SPSS Solutions

18.5 Open data file *ex18\_07.por*. If  $\mu_1$  is the mean number of species for unlogged plots and  $\mu_2$  is the mean number of species for logged plots, we want to test hypotheses  $H_0 : \mu_1 = \mu_2$  against  $H_a : \mu_1 > \mu_2$ . We first examine side-by-side boxplots of the data to check for skewness and outliers. Use **Graphs, Legacy Dialogs, Boxplot**. We want simple boxplots where Data in Chart are **Summaries of separate variables**. Click to enter both variable names and OK. Neither data set has outliers; the data for logged plots seems to be left skewed due to the plot with only 4 species; however with a sample size of only 9 we'll cautiously proceed.



Use **Analyze, Compare Means, Independent Samples T Test** to perform the test. However, the data in this file are not suited for what SPSS requires (numeric group identifiers instead of alphabetic). Create another variable where 0 = unlogged and 1 = logged (we called this **numgroup**). Now you can perform the test. Click to enter variable **trees** and **group** as the grouping variable. Click **Define Groups** and define the groups as 0 and 1. Looking ahead to Exercise 18.7, use **Options** to change the confidence level to 90%. **OK** performs the test.



Group Statistics

group	N	Mean	Std. Deviation	Std. Error Mean
trees 0	12	17.50	3.529	1.019
1	9	13.67	4.500	1.500

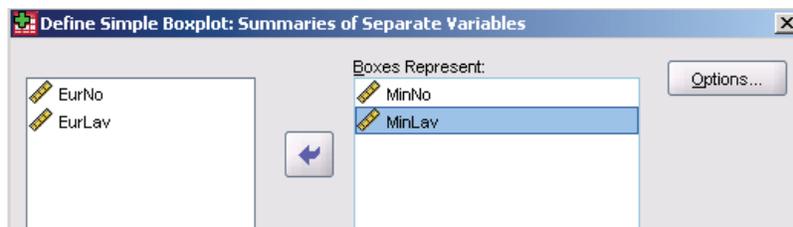
Independent Samples Test

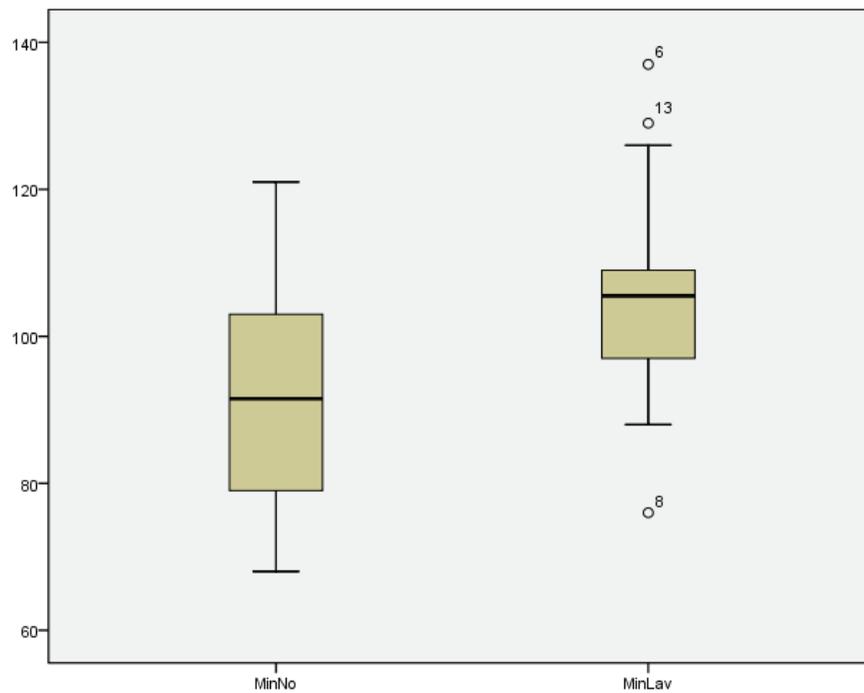
	Levene's Test for Equality of Variances		t-test for Equality of Means						
	F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	90% Confidence Interval of the Difference	
								Lower	Upper
Tree Equal variances assumed	.072	.791	2.191	19	.041	3.833	1.749	.809	6.858
Equal variances not assumed			2.114	14.793	.052	3.833	1.813	.652	7.015

Since we have no reason to believe the populations should have the same standard deviation, look at the “Equal variances not assumed” row of the results.. With a test statistic of  $t = 2.114$  and  $P$ -value  $0.052/2 = 0.26$ , we will reject  $H_0$  and conclude that logging does reduce species diversity (in Borneo.)

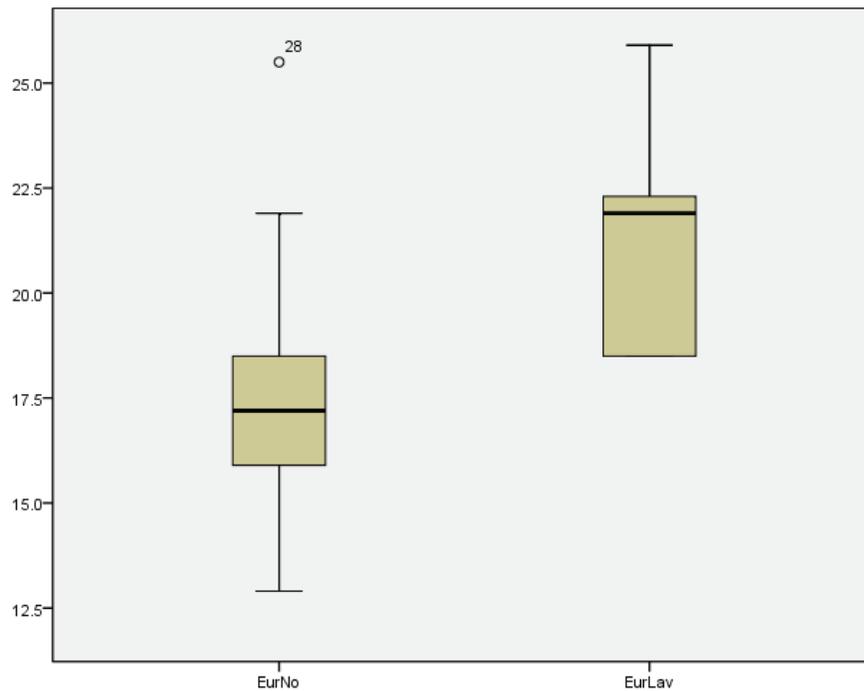
**18.7** SPSS gives the confidence interval for the difference in its output. We are therefore 90% confident that unlogged plots in Borneo will have between 0.7 and 7.0 more tree species than logged plots. Note that 0 is not included in the interval, which is confirmation that it is highly unlikely for there to be the same number of species.

**18.9** We'll create side-by-side boxplots to examine the distributions of the data, then use **Analyze, Compare Means, Independent Samples T Test** for each pair (time and money spent).



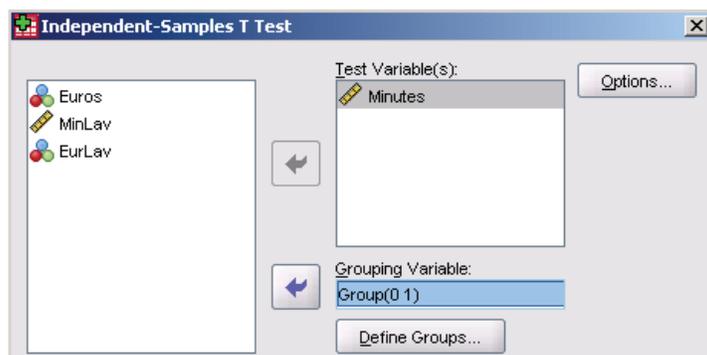


The no odor data is symmetric with no outliers; while there are three outliers in the data for lavender odor, they are not out of line with the other data and help make the distribution more symmetric. With a sample size of  $n = 30$ , we do have similar shapes, so these data are suitable for the two-sample  $t$  test. We see the median for the lavender odor is larger than  $Q_3$  for no odor; is it enough larger to be significant?



The bulk of the no odor spending data is symmetric, but also has a high outlier; the lavender odor data has no lower tail – it is skewed right. The medians are even more different than for the time data; the median of the lavender odor is much larger than  $Q_3$ . We'll rely on robustness with our  $n = 30$  sample.

Unfortunately, the data in file *ta18-02* is not in a format that SPSS will like – it wants all the sample data in a single column with integer group identifiers. Use copy and paste to copy the lavender data for each variable below the no odor data, then create a new variable for subscripts. We renamed the variables into **Minutes**, **Spending** and **Group**. Now we're ready for the tests (one for each).



		t-test for Equality of Means						
		t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
							Lower	Upper
Minutes	Equal variances assumed	-3.980	58	.000	-14.433	3.627	-21.693	-7.173
	Equal variances not assumed	-3.980	57.041	.000	-14.433	3.627	-21.696	-7.171
Spending	Equal variances assumed	-5.945	58	.000	-3.6100	.6073	-4.8256	-2.3944
	Equal variances not assumed	-5.945	57.998	.000	-3.6100	.6073	-4.8256	-2.3944

Our hypotheses for both tests are  $H_0 : \mu_{NO} = \mu_L$  and  $H_a : \mu_{NO} < \mu_L$ , since the question is if the lavender odor encourages customers to stay longer. With  $P$ -values of 0.000 (really half of this), we reject the null in both cases, and conclude that the lavender odor does encourage customers to stay longer and spend more.

**18.25** With summarized data, we'll have to use **Transform, Compute Variable** to compute the test statistic and its  $P$ -value. Since the women are conjectured to talk more, the alternate hypothesis is  $H_a : \mu_w > \mu_m$ . We'll look at Study 1 first. First, we notice

that the mean for women in this study is *less* than the mean for the men. We'll have a  $P$ -value greater than 0.5; this study does not support the conjecture. (The actual  $t$  statistic is -0.31 and the  $P$ -value will be 0.598.)

The second study will have conservative degrees of freedom  $20 - 1 = 19$ . We'll compute the test statistic and conservative  $P$ -value below.

The first screenshot shows the 'Compute Variable' dialog with 'Target Variable' set to 'T' and 'Numeric Expression' set to  $(16496-12867)/\sqrt{(7914^{**2}/27+8343^{**2}/20)}$ . To the right, a box displays 'T' with the value 1.51.

The second screenshot shows the 'Compute Variable' dialog with 'Target Variable' set to 'Pvalue' and 'Numeric Expression' set to  $1-\text{CDF.T}(1.51,19)$ . To the right, a box displays 'Pvalue' with the value 0.0737.

We find a  $P$ -value of 0.0737. At the 5% level, these data also fails to reject the null hypothesis. These studies fail to show that women talk more than men, on average.

**18.27** We are given standard errors, so we first multiply these by the square root of the sample sizes to find the standard deviations. The summary table is filled in below

Group	Location	$n$	$\bar{x}$	$s$
1	Oregon	26.9	6	3.821
2	California	11.9	7	7.091

For the conservative approach we will have  $6 - 1 = 5$  df. We compute the test statistic and find the conservative  $P$ -value below.

The first screenshot shows the 'Compute Variable' dialog with 'Target Variable' set to 'T' and 'Numeric Expression' set to  $(26.9-11.9)/\sqrt{(3.821^{**2}/6+7.091^{**2}/7)}$ . To the right, a box displays 'T' with the value 4.84.

The second screenshot shows the 'Compute Variable' dialog with 'Target Variable' set to 'Pvalue' and 'Numeric Expression' set to  $2*(1-\text{CDF.T}(4.84,5))$ . To the right, a box displays 'Pvalue' with the value 0.0047.

The test statistic is 4.84, so we'll double the area to the right of  $t = 4.84$  for the conservative approach. The conservative  $P$ -value will be 0.0047. Our conclusion matches theirs (reject the null, there is a difference in whelks), but our  $P$ -value is different.

**18.29** Placebos are “inert, sugar pills” – they have no medicinal value, but are used so that all subjects are treated alike; also, there may be a “placebo effect” in that giving this pill may help people feel (or perform) better. Double-binding an experiment means that neither the subjects nor any experimenters who interact with them know the treatment. This can be valuable in preventing bias. We’ll test  $H_0 : \mu_{PL} = \mu_G$  against  $H_a : \mu_{PL} \neq \mu_G$ .

Compute Variable		T
Target Variable: T	Numeric Expression: (0.06383-.05342)/sqrt(.01462**2/21+.01549**2/18)	2.15

Compute Variable		Pvalue
Target Variable: Pvalue	Numeric Expression: 2*(1-CDF.T(2.15,17))	0.0462

Using the conservative degrees of freedom we have  $t = 2.15$  with  $P$ -value 0.0462. We reject the null hypothesis at the 5% level; there is a difference in the number of misses between the two groups. Interestingly, it seems there are more misses with ginkgo than placebo, on average.

**18.31** We again test using the two sided alternative (is there are difference in mean stress levels) by computing the test statistic and conservative  $P$ -value.

Compute Variable		T
Target Variable: t	Numeric Expression: (1.92-1.74)/SQRT(.6**2/12+.57**2/9)	0.70

Compute Variable		Pvalue
Target Variable: Pvalue	Numeric Expression: 2*(1-CDF.T(.70,8))	0.5038

With test statistic  $t = 0.70$  and  $P$ -value 0.5038, we fail to reject the null hypothesis; these data do not show a difference in mean stress levels based on the origin of the mother.

**18.33** We use the values for Gain in this test with hypotheses  $H_0 : \mu_C = \mu_{UN}$  and  $H_a : \mu_C > \mu_{UN}$ .

Compute Variable		T
Target Variable: T	Numeric Expression: (29-21)/sqrt(59**2/427+52**2/2733)	2.65

The conservative  $P$ -value is

Compute Variable	
Target Variable:	Numeric Expression:
Pvalue	= 1-CDF.T(2.65,426)

Pvalue
0.0042

We reject the null hypothesis; apparently, coaching does raise the average gain on retaking the SAT. How much is the average increase? We'll again use the conservative degrees of freedom to find  $t^*$ .

Compute Variable	
Target Variable:	Numeric Expression:
Tstar	= IDF.T(.005,426)

Tstar
-2.59

Compute Variable	
Target Variable:	Numeric Expression:
Clow	= 8-2.59*sqrt(59**2/427+52**2/2733)

Clow
0.17

Compute Variable	
Target Variable:	Numeric Expression:
Chigh	= 8+2.59*sqrt(59**2/427+52**2/2733)

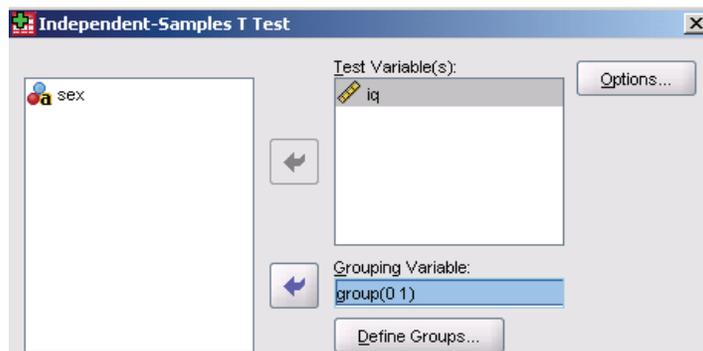
Chigh
15.83

We're (conservatively) 99% confident the average gain will be between 0.17 and 15.83 points more with coaching. Given that it's possible there is a very minimal (almost 0) additional gain with coaching, it's probably not worth the money.

**18.35** Open data file *ex18\_25.por*. To create stemplots for each gender, use **Analyze**, **Descriptive Statistics**, **Explore**. Click to enter **iq** as the Dependent and **sex** as the factor. You can click to display **Plots** only, then **OK**.

Stem-and-Leaf Plot for sex= F			Stem-and-Leaf Plot for sex= M		
Frequency	Stem &	Leaf	Frequency	Stem &	Leaf
2.00	Extremes	(=<74)	2.00	Extremes	(=<79)
2.00	8 .	69	2.00	9 .	03
4.00	9 .	1368	2.00	9 .	77
9.00	10 .	023334578	4.00	10 .	0234
10.00	11 .	1122244489	9.00	10 .	556667779
2.00	12 .	08	11.00	11 .	00001123334
2.00	13 .	02	6.00	11 .	556899
Stem width:	10		5.00	12 .	03344
Each leaf:	1 case(s)		5.00	12 .	67788
			.00	13 .	
			1.00	13 .	6

Before we can do the test, we need to create a numeric grouping variable. Name a new variable **group** and enter 0 for males and 1 for females. Use **Analyze, Compare Means, Independent Samples T Test** to perform the test. Click to enter variable **iq** and **group** as the grouping variable. Click **Define Groups** and define the groups as 0 and 1. **OK** performs the test.



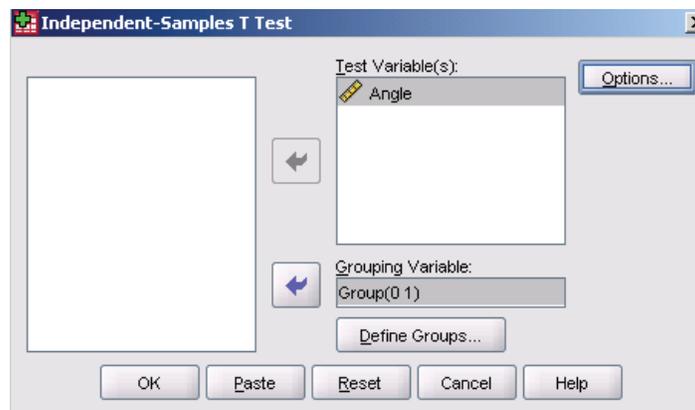
**Independent Samples Test**

	Levene's Test for Equality of Variances		t-test for Equality of Means						
	F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
								Lower	Upper
Equal variances assumed	.908	.344	1.700	76	.093	5.119	3.011	-878	11.115
Equal variances not assumed			1.644	56.932	.106	5.119	3.114	-1.117	11.354

Since we have no reason to believe the populations should have the same standard deviation, look at the “Equal variances not assumed” row of output. With a test statistic of  $t = 1.64$  and  $P$ -value 0.106, we will do reject  $H_0$  and conclude there is not a significant difference in IQ scores for boys and girls in this school district.

**18.37** If you have not done the exercise, look at the solution to Exercise 18.35. The output from the Independent Samples T Test includes the confidence interval. Based on the information given, we estimate with 95% confidence that the difference in mean IQ score between seventh grade boys and girls in this school district is between  $-1.12$  and  $11.35$ . Since 0 is included in the interval, no difference in mean IQ between boys and girls is reasonable.

**18.39** Looking at the data, it might appear that Hylite has a higher wrinkle recovery angle than Permafresh, so it would have better wrinkle resistance. With samples of  $n = 5$ , any graph might be suspect. We can see there are no obvious outliers in either data set. We'll use **Analyze, Compare Means, Independent Samples T Test** to perform the test, after entering the data into a single column and creating a grouping variable (0 = Permafresh, 1 = Hylite). Looking ahead, we want a 90% confidence interval for the difference in Exercise 18.41, so click **Options** to set the level.



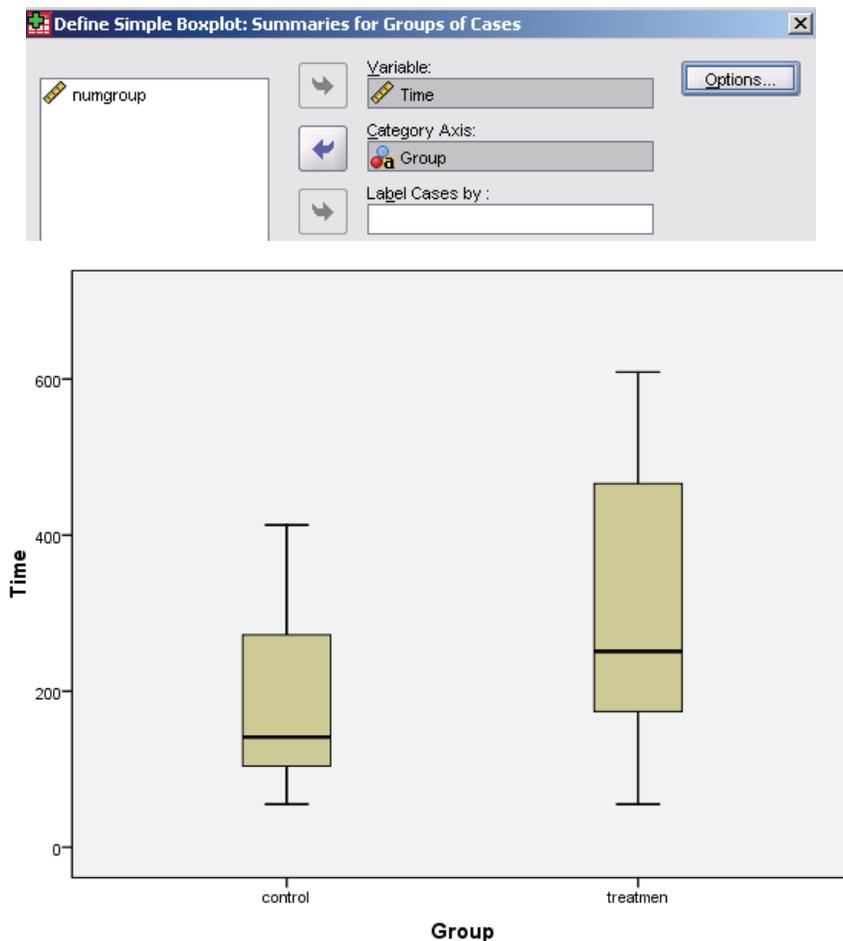
		t-test for Equality of Means						
		t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	90% Confidence Interval of the Difference	
							Lower	Upper
Angle	Equal variances assumed	-6.296	8	.000	-8.400	1.334	-10.881	-5.919
	Equal variances not assumed	-6.296	7.779	.000	-8.400	1.334	-10.890	-5.910

With test statistic  $t = -6.296$  and  $P$ -value 0.000, we reject the null hypothesis and conclude that Hylite is better at reducing wrinkles.

**18.41** We can read the confidence interval for the difference in the output above. Our 90% interval for the difference in mean wrinkle recovery angle is  $-10.89$  to  $-5.91$ . Since we are looking at Permafresh – Hylite, this means that Hylite's mean is between 5.91 and 10.89 higher than Permafresh's.

**18.45** Since researchers suspect that the treatment group will ask for help less quickly, we have hypotheses  $H_0: \mu_C = \mu_T$  and  $H_a: \mu_C < \mu_T$ . Data file ex18-45 has groups labeled as "treatment" and "control;" SPSS requires integer grouping variables, so we've created a new variable called numgroup with 0 = treatment and 1 = control. We use **Graphs, Legacy Dialogs, Boxplot** to create side-by-side boxplots for the two groups,

using **Data are summaries for groups of cases** option (we can use the alpha labels here).



Neither group shows any outliers, and both distributions are rather right-skewed. We'll rely on robustness of  $t$  procedures and the fact that these distributions have the same general shape. We use **Analyze, Compare Means, Independent Samples T Test** to perform the test.

		t-test for Equality of Means						
		t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
							Lower	Upper
Time	Equal variances assumed	2.521	32	.017	127.941	50.760	24.546	231.336
	Equal variances not assumed	2.521	28.270	.018	127.941	50.760	24.008	231.874

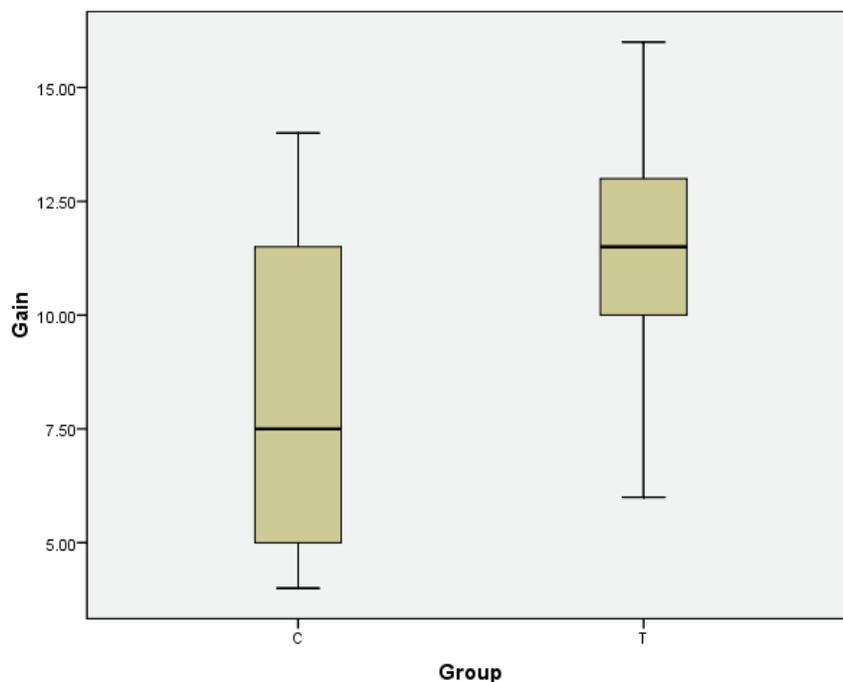
Our  $P$ -value is small ( $0.018/2 = 0.009$ ), so we can reject the null hypothesis and conclude the suspicion is correct; the treatment group (the ones dealing with money phrases) were less likely to ask for help.

**18.47** These are matched pairs data within two different treatments. We'll need the differences to use as the data for the hypothesis test.



The screenshot shows the 'Compute Variable' dialog box. The 'Target Variable' field is set to 'Gain' and the 'Numeric Expression' field is set to 'After-Before'.

We create boxplots to check the distribution assumption.



Both distributions are rather symmetric, with no outliers. It seems the median of the treatment group is higher than the control group. Before computing the test and the confidence interval, we need to create a numeric grouping variable. We've called ours **numgroup**, where 0 = control and 1 = treatment.

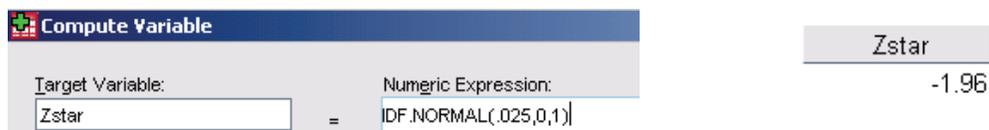
		t-test for Equality of Means						
		t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
							Lower	Upper
Gain	Equal variances assumed	-1.948	16	.069	-3.15000	1.61685	-6.57758	.27758
	Equal variances not assumed	-1.914	13.919	.076	-3.15000	1.64615	-6.68257	.38257

With  $P$ -value  $0.076/2 = 0.038$ , we reject the idea of no difference at the 5% level, and conclude the positive messages must have helped. The confidence interval tells how much these messages might have helped – the treatment will improve the mean math skills score somewhere between 0.38 and 6.68 points, with 90% confidence. It's very possible that the significant difference is very unmeaningful in practical terms.

## Chapter 19 SPSS Solutions

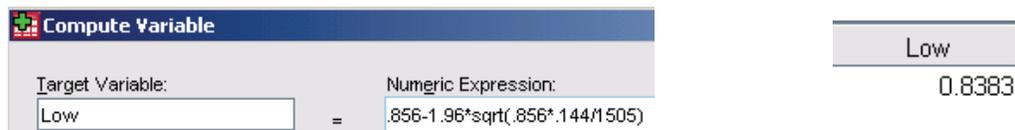
**\*\*NOTE:** SPSS does not do inference based on  $Z$  distributions, nor does it perform inference on variables that are already summarized. If you really want to use SPSS for these problems, follow the instructions below (you'll be basically using **Transform, Compute Variable** as a calculator) or use another technology (such as a graphing calculator or another statistics program like Minitab or Crunchit.)

**19.5** The biggest weakness might be that people refuse to cooperate. Another is that not everyone has a telephone (landline). Time-of-day could also be a factor – when were the calls made? To create the confidence interval, we have  $\hat{p} = 1288/1505 = 0.856$ . To refresh our memories, we use **IDF.Normal** to find  $z^*$  for the interval as shown below; this is  $z^* = 1.96$ .



The screenshot shows the 'Compute Variable' dialog box. The 'Target Variable' is 'Zstar' and the 'Numeric Expression' is 'IDF.NORMAL(.025,0,1)'. To the right, a small table shows 'Zstar' with the value '-1.96'.

Now, find the endpoints of the interval as  $\hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ . Click the Variable View tab and increase the number of decimal places that will display, if desired (we increased them to four places in each of **Low** and **High**).



The screenshot shows the 'Compute Variable' dialog box. The 'Target Variable' is 'Low' and the 'Numeric Expression' is '.856-1.96\*sqrt(.856\*.144/1505)'. To the right, a small table shows 'Low' with the value '0.8383'.



The screenshot shows the 'Compute Variable' dialog box. The 'Target Variable' is 'High' and the 'Numeric Expression' is '.856+1.96\*sqrt(.856\*.144/1505)'. To the right, a small table shows 'High' with the value '0.8737'.

Based on this survey, we're 95% confident that between 83.8% and 87.4% of Canadians support registration of all firearms.

**19.7** Since only 9 of 98 whelks drilled into mussels, we can't use a large-sample confidence interval. Add 2 to the successes (that makes 11) and 4 to the number of trials (that makes 102) and compute the interval as  $\tilde{p} \pm z^* \sqrt{\tilde{p}(1-\tilde{p})/n+4}$ . We have  $\tilde{p} = 11/102 = 0.108$ . The interval becomes  $0.108 \pm 1.645 \sqrt{.108*(1-.108)/102}$ , or  $0.108 \pm 0.051$ . With 90% confidence, the proportion of Oregon whelks that will spontaneously drill in to mussels is between 5.7% and 15.9%, based on this sample.

**19.9** The sample proportion is  $\hat{p} = 20/20 = 1.0$ . Since the confidence interval is  $\hat{p} \pm z^* \sqrt{\hat{p}(1-\hat{p})/n}$  and  $(1-\hat{p})$  is 0, the interval is from 1 to 1 (not very interesting, nor can we be sure *every* bill is tainted with cocaine). To form the plus four interval add 4 to the trials (this becomes 24) and 2 to the “successes” (this becomes 22). The plus four estimate is  $\tilde{p} = 91.7\%$ , and based on this, we’re 95% confident that between 80.6% and 100% (it can’t be more than 100%) of Spanish currency is tainted with cocaine.

Compute Variable	
Target Variable:	Numeric Expression:
Low	.917-1.96*sqrt(.917*.083/24)

Compute Variable	
Target Variable:	Numeric Expression:
High	.917+1.96*sqrt(.917*.083/24)

Low	0.8066
High	1.0274

**19.11** To find the sample size needed, we’ll use the formula

$$n \geq \left( \frac{z^*}{ME} \right)^2 p^*(1-p^*)$$

where  $p^*$  is the “guessed”  $p$ . We find  $z^*$  for 90% confidence using `invNorm` with half the leftover area to be 1.645. We’ll need a sample of at least 318 Americans who have at least one Italian grandparent.

```
invNorm(.05,0,1)
      -1.644853626
(1.645/.04)^2*.75
*.25
      317.1123047
```

**19.13** We let  $p$  be the proportion of candidates with the more competent face who win. If the face doesn’t influence voters, we should have  $p = 0.5$ ; this becomes the null hypothesis. We seek evidence in support of an alternative that the more competent face should win more elections, in other words, we have  $H_a : p > 0.5$ . From the data given,

we have  $\hat{p} = 22/32 = 0.6875$ . Our test statistic is  $z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$ . We then use

**CDF.Normal** to find the  $P$ -value of this test; since this is a one-sided test, the  $P$ -value is the area to the right of our test statistic.

Compute Variable	
Target Variable:	Numeric Expression:
Z	(.6875-.5)/sqrt(.5*.5/32)

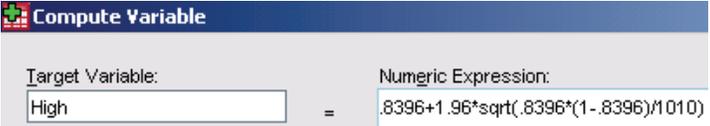
  

Compute Variable	
Target Variable:	Numeric Expression:
Pvalue	1-CDF.Normal(2.12,0,1)

Z	2.12
Pvalue	0.0170

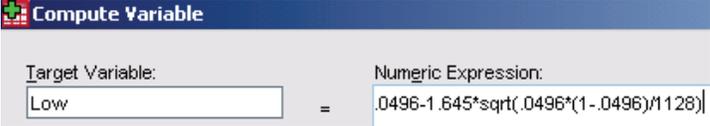
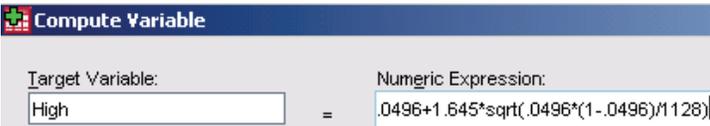
With  $z = 2.12$  and  $P$ -value 0.0170, we reject the null at the 5% level and conclude that the more competent face tends to win elections.

**19.25** As with any telephone poll, the biggest weakness might be that people refuse to cooperate; in this case, they may lie about being smokers. Another is that not everyone has a telephone (landline). Time-of-day could also be a factor – when were the calls made? The estimate from the sample is  $\hat{p} = 848/1010 = 0.8396$ ; we use this in computing the low and high ends of the interval.

 <p>Target Variable: Low          Numeric Expression: <math>0.8396 - 1.96 * \sqrt{0.8396 * (1 - 0.8396) / 1010}</math></p>	<table border="1" style="margin-left: auto; margin-right: auto;"> <tr><td>Low</td></tr> <tr><td>0.8170</td></tr> </table>	Low	0.8170
Low			
0.8170			
 <p>Target Variable: High          Numeric Expression: <math>0.8396 + 1.96 * \sqrt{0.8396 * (1 - 0.8396) / 1010}</math></p>	<table border="1" style="margin-left: auto; margin-right: auto;"> <tr><td>High</td></tr> <tr><td>0.8622</td></tr> </table>	High	0.8622
High			
0.8622			

Based on the Harris survey, between 81.7% and 86.2% of smokers believe smoking will probably shorten their lives, with 95% confidence.

**19.29** As with any telephone poll, the biggest weakness might be that people refuse to cooperate. Another is that not everyone has a telephone (landline). Time-of-day could also be a factor – when were the calls made? The large sample estimate is  $\hat{p} = 56/1128 = 0.0496$ .

 <p>Target Variable: Low          Numeric Expression: <math>0.0496 - 1.645 * \sqrt{0.0496 * (1 - 0.0496) / 1128}</math></p>	<table border="1" style="margin-left: auto; margin-right: auto;"> <tr><td>Low</td></tr> <tr><td>0.0390</td></tr> </table>	Low	0.0390
Low			
0.0390			
 <p>Target Variable: High          Numeric Expression: <math>0.0496 + 1.645 * \sqrt{0.0496 * (1 - 0.0496) / 1128}</math></p>	<table border="1" style="margin-left: auto; margin-right: auto;"> <tr><td>High</td></tr> <tr><td>0.0602</td></tr> </table>	High	0.0602
High			
0.0602			

The large sample interval has between 3.9% and 6.0% of American women dissatisfied with their life. To find the plus four interval, we have  $\tilde{p} = 58/1132 = 0.0512$ . Redo the computations using this new estimate and be sure to use the “updated” value of  $n$ .

Low	High
0.0404	0.0620

The plus four interval is from 4.0% to 6.2% - a shift slightly higher toward the center 50%.

**19.31** The conditions for the large sample methods are not met – there were only 23 facilities; 18 detected the GM beans but 5 did not (there are fewer “failures” than we need). Using 20 “successes” and 27 trials, we have  $\tilde{p} = 20/27 = 0.741$ . We’re 90% confident that between 60.2% and 88.0% of these facilities will be able to detect genetically modified beans.

<div style="border: 1px solid #ccc; padding: 5px; margin-bottom: 10px;"> <div style="background-color: #4a7ebb; color: white; padding: 2px;"><b>Compute Variable</b></div> <table border="0" style="width: 100%; border-collapse: collapse;"> <tr> <td style="width: 30%; border-bottom: 1px solid #ccc;">Target Variable:</td> <td style="width: 10%; text-align: center;">=</td> <td style="border-bottom: 1px solid #ccc;">Numeric Expression:</td> </tr> <tr> <td style="border: 1px solid #ccc;">Low</td> <td></td> <td style="border: 1px solid #ccc;">.741-1.645*sqrt(.741*(1-.741)/27)</td> </tr> </table> </div>	Target Variable:	=	Numeric Expression:	Low		.741-1.645*sqrt(.741*(1-.741)/27)	<div style="border: 1px solid #ccc; padding: 5px; margin-bottom: 10px; background-color: #f0f0f0;">Low</div> <div style="border: 1px solid #ccc; padding: 5px; margin-bottom: 10px; background-color: #f0f0f0;">0.6023</div>
Target Variable:	=	Numeric Expression:					
Low		.741-1.645*sqrt(.741*(1-.741)/27)					
<div style="border: 1px solid #ccc; padding: 5px;"> <div style="background-color: #4a7ebb; color: white; padding: 2px;"><b>Compute Variable</b></div> <table border="0" style="width: 100%; border-collapse: collapse;"> <tr> <td style="width: 30%; border-bottom: 1px solid #ccc;">Target Variable:</td> <td style="width: 10%; text-align: center;">=</td> <td style="border-bottom: 1px solid #ccc;">Numeric Expression:</td> </tr> <tr> <td style="border: 1px solid #ccc;">High</td> <td></td> <td style="border: 1px solid #ccc;">.741+1.645*sqrt(.741*(1-.741)/27)</td> </tr> </table> </div>	Target Variable:	=	Numeric Expression:	High		.741+1.645*sqrt(.741*(1-.741)/27)	<div style="border: 1px solid #ccc; padding: 5px; margin-bottom: 10px; background-color: #f0f0f0;">High</div> <div style="border: 1px solid #ccc; padding: 5px; margin-bottom: 10px; background-color: #f0f0f0;">0.8797</div>
Target Variable:	=	Numeric Expression:					
High		.741+1.645*sqrt(.741*(1-.741)/27)					

**19.37** We have only 12 specimens that were torched; of these 5 resprouted. With these small numbers, we must use the plus four method to construct the confidence interval. The 5 becomes 7 and the 12 becomes 16, with  $\tilde{p} = 7/16 = 0.4375$ . We’re 90% confident that between 23.4% and 64.2% of *Krameria cytisoides* shrubs will resprout after a fire.

<div style="border: 1px solid #ccc; padding: 5px; margin-bottom: 10px;"> <div style="background-color: #4a7ebb; color: white; padding: 2px;"><b>Compute Variable</b></div> <table border="0" style="width: 100%; border-collapse: collapse;"> <tr> <td style="width: 30%; border-bottom: 1px solid #ccc;">Target Variable:</td> <td style="width: 10%; text-align: center;">=</td> <td style="border-bottom: 1px solid #ccc;">Numeric Expression:</td> </tr> <tr> <td style="border: 1px solid #ccc;">Low</td> <td></td> <td style="border: 1px solid #ccc;">.4375-1.645*sqrt(.4375*(1-.4375)/16)</td> </tr> </table> </div>	Target Variable:	=	Numeric Expression:	Low		.4375-1.645*sqrt(.4375*(1-.4375)/16)	<div style="border: 1px solid #ccc; padding: 5px; margin-bottom: 10px; background-color: #f0f0f0;">Low</div> <div style="border: 1px solid #ccc; padding: 5px; margin-bottom: 10px; background-color: #f0f0f0;">0.2335</div>
Target Variable:	=	Numeric Expression:					
Low		.4375-1.645*sqrt(.4375*(1-.4375)/16)					
<div style="border: 1px solid #ccc; padding: 5px;"> <div style="background-color: #4a7ebb; color: white; padding: 2px;"><b>Compute Variable</b></div> <table border="0" style="width: 100%; border-collapse: collapse;"> <tr> <td style="width: 30%; border-bottom: 1px solid #ccc;">Target Variable:</td> <td style="width: 10%; text-align: center;">=</td> <td style="border-bottom: 1px solid #ccc;">Numeric Expression:</td> </tr> <tr> <td style="border: 1px solid #ccc;">High</td> <td></td> <td style="border: 1px solid #ccc;">.4375+1.645*sqrt(.4375*(1-.4375)/16)</td> </tr> </table> </div>	Target Variable:	=	Numeric Expression:	High		.4375+1.645*sqrt(.4375*(1-.4375)/16)	<div style="border: 1px solid #ccc; padding: 5px; margin-bottom: 10px; background-color: #f0f0f0;">High</div> <div style="border: 1px solid #ccc; padding: 5px; margin-bottom: 10px; background-color: #f0f0f0;">0.6415</div>
Target Variable:	=	Numeric Expression:					
High		.4375+1.645*sqrt(.4375*(1-.4375)/16)					

**19.39** We can use the large sample method with the results of this survey. We have  $\hat{p} = 594/1484 = 0.4003$ . Based on this survey, between 37.5% and 42.5% of American adults believe humans developed from earlier animals.

<div style="border: 1px solid #ccc; padding: 5px; margin-bottom: 10px;"> <div style="background-color: #4a7ebb; color: white; padding: 2px;"><b>Compute Variable</b></div> <table border="0" style="width: 100%; border-collapse: collapse;"> <tr> <td style="width: 30%; border-bottom: 1px solid #ccc;">Target Variable:</td> <td style="width: 10%; text-align: center;">=</td> <td style="border-bottom: 1px solid #ccc;">Numeric Expression:</td> </tr> <tr> <td style="border: 1px solid #ccc;">Low</td> <td></td> <td style="border: 1px solid #ccc;">.4003-1.96*sqrt(.4003*(1-.4003)/1484)</td> </tr> </table> </div>	Target Variable:	=	Numeric Expression:	Low		.4003-1.96*sqrt(.4003*(1-.4003)/1484)	<div style="border: 1px solid #ccc; padding: 5px; margin-bottom: 10px; background-color: #f0f0f0;">Low</div> <div style="border: 1px solid #ccc; padding: 5px; margin-bottom: 10px; background-color: #f0f0f0;">0.3754</div>
Target Variable:	=	Numeric Expression:					
Low		.4003-1.96*sqrt(.4003*(1-.4003)/1484)					
<div style="border: 1px solid #ccc; padding: 5px;"> <div style="background-color: #4a7ebb; color: white; padding: 2px;"><b>Compute Variable</b></div> <table border="0" style="width: 100%; border-collapse: collapse;"> <tr> <td style="width: 30%; border-bottom: 1px solid #ccc;">Target Variable:</td> <td style="width: 10%; text-align: center;">=</td> <td style="border-bottom: 1px solid #ccc;">Numeric Expression:</td> </tr> <tr> <td style="border: 1px solid #ccc;">High</td> <td></td> <td style="border: 1px solid #ccc;">.4003+1.96*sqrt(.4003*(1-.4003)/1484)</td> </tr> </table> </div>	Target Variable:	=	Numeric Expression:	High		.4003+1.96*sqrt(.4003*(1-.4003)/1484)	<div style="border: 1px solid #ccc; padding: 5px; margin-bottom: 10px; background-color: #f0f0f0;">High</div> <div style="border: 1px solid #ccc; padding: 5px; margin-bottom: 10px; background-color: #f0f0f0;">0.4252</div>
Target Variable:	=	Numeric Expression:					
High		.4003+1.96*sqrt(.4003*(1-.4003)/1484)					

**19.41** The upper end of our confidence interval in Exercise 19.39 was below 50%, so we have good evidence that less than half believe humans developed from lower animals. We'll go ahead and compute a test statistic and  $P$ -value for this test anyway. With a test statistic of  $z = -7.68$  there is no real need to find the  $P$ -value exactly; using the 68-95-99.7 rule, we know the chance of being more than 7 standard deviations below a mean is essentially 0.

Compute Variable	
Target Variable:	Numeric Expression:
Z	$(.4003 - .5) / \text{sqrt}(.5 * .5 / 484)$

Z
-7.68

## Chapter 20 SPSS Solutions

**\*\*NOTE:** SPSS does not do inference based on  $Z$  distributions, nor does it perform inference on variables that are already summarized. If you really want to use SPSS for these problems, follow the instructions below (you'll be basically using **Transform, Compute Variable** as a calculator) or use another technology (such as a graphing calculator or another statistics program like Minitab or Crunchit.)

**20.1** We'll find a 95% confidence interval for  $p_1 - p_2$ , where  $p_1$  is the proportion of IM primary users in the 18 to 27 age group and  $p_2$  is the proportion of IM primary users in the 28 to 39 age group. There are enough "successes" IM primary users in each age group (more than 10), so we can use large-sample methods. From the data given, we have  $\hat{p}_1 = 73/158 = 0.462$  and  $\hat{p}_2 = 26/143 = 0.182$ . The observed difference is  $\hat{p}_1 - \hat{p}_2 = 0.462 - 0.182 = 0.28$ . The confidence interval formula is given by

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

Compute Variable	
Target Variable:	Numeric Expression:
Low	= .28-1.96*sqrt(.462*.538/158+.182*.818/143)

Low
0.1798

Compute Variable	
Target Variable:	Numeric Expression:
High	= .28+1.96*sqrt(.462*.538/158+.182*.818/143)

High
0.3802

The interval for the difference is 18.0% to 38.0%. We're 95% confident, based on this information that between 18% and 28% more people in the 18 to 27 age group use IM more often than email, compared to people in the 28 to 39 age group. This difference is statistically significant because the interval does not contain zero (and we have a  $P$ -value of 0.000 for the test).

**20.3** Let  $p_F$  be the proportion of all female high school students who meet the physically activity recommendations and  $p_M$  be the proportion of males. We want a confidence interval for  $p_F - p_M$ . From the data given, we have  $\hat{p}_F = 1915/6889 = 0.2780$  and  $\hat{p}_M = 3078/7028 = 0.438$ . The observed difference is  $\hat{p}_F - \hat{p}_M = 0.278 - 0.438 = -0.16$ .

Compute Variable	
Target Variable:	Numeric Expression:
Low	= -.16-2.576*sqrt(.278*.722/6889+.438*.562/7028)

Low
-0.1806

Compute Variable		High
Target Variable:	Numeric Expression:	
High	$-.16+2.576*\sqrt{(.278*.722/6889+.438*.562/7028)}$	-0.1394

Both ends of the interval are negative, which means males are more likely to meet the physical activity recommendations than females. Between 13.9% and 18.1% more males meet the recommendations than females, with 99% confidence.

**20.5** We need the plus four method here because none of the microwave group of crackers had checking. With two samples, add 1 to the “successes” in each group and 2 to the sample size in each group. We have  $\tilde{p}_M = 1/67 = 0.0149$  and  $\tilde{p}_{NM} = 17/67 = 0.2537$ . The observed difference is  $-0.2388$ .

Compute Variable		Low
Target Variable:	Numeric Expression:	
Low	$-.2388-1.96*\sqrt{(.0149*(1-.0149)/67+.2537*(1-.2537)/67)}$	-0.3470

Compute Variable		High
Target Variable:	Numeric Expression:	
High	$-.2388+1.96*\sqrt{(.0149*(1-.0149)/67+.2537*(1-.2537)/67)}$	-0.1306

Based on this study, between 13.1% and 34.7% fewer crackers will have checking when microwaved, with 95% confidence. In the actual crackers used, the sample proportions are  $\hat{p}_M = 0$  and  $\hat{p}_{NM} = 16/65 = 24.6\%$ .

**20.7** We want to know if helmet use is less common among those who have had head injuries. We'll test  $H_0 : p_{HI} = p_{NHI}$  against  $H_a : p_{HI} < p_{NHI}$  where  $p$  is the proportion in each group who use helmets. From the data, we compute the estimates needed for the hypothesis test as  $\hat{p}_{HI} = 96/578 = 0.1661$ ,  $\hat{p}_{NHI} = 656/2992 = 0.2193$ , and  $\hat{p} = (96 + 656)/(578 + 2992) = 0.2106$ . We now compute the test statistic and its  $P$ -value.

Compute Variable		Z
Target Variable:	Numeric Expression:	
Z	$(.1661-.2193)/\sqrt{(.2106*.7894*(1/578+1/2992))}$	-2.87

Compute Variable		Pvalue
Target Variable:	Numeric Expression:	
Pvalue	$\text{CDF.Normal}(-2.87,0,1)$	0.0021

The test statistic is  $z = -2.87$  with  $P$ -value 0.0021. This study does show that skiers and snowboarders who have had head injuries are less likely to use helmets.

**20.17** These samples satisfy the large sample guidelines – there were more than 10 each of successes and failures in each group. The sample proportions are  $\hat{p}_1 = 117/170 = 0.6882$  for the younger group and  $\hat{p}_2 = 152/317 = 0.4795$  for the older group. Both ends of the interval are positive; younger teens are more likely to have false information in their profiles – between 12.0% and 29.7% more of young teens have false information included than older teens, with 95% confidence.

Compute Variable		Low
Target Variable:	Numeric Expression:	0.1205
Low	$2088 - 1.95 * \sqrt{.6882 * (1 - .6882) / 170 + .4795 * (1 - .4795) / 317}$	

Compute Variable		High
Target Variable:	Numeric Expression:	0.2971
High	$2088 + 1.95 * \sqrt{.6882 * (1 - .6882) / 170 + .4795 * (1 - .4795) / 317}$	

**20.19** Since none in the control group of 18 mice developed tumors, we can't use large-sample methods. After adding 1 to each “success, we have 24 mice with tumors in the group with lowered levels of DNA methylation, and 1 case of a tumor in the control group; there are now 35 mice in the group with lowered levels of DNA methylation and 20 in the control group. If we call the group with lowered levels group 1, we have  $\tilde{p}_1 = 24/35 = 0.688$ . In the other group, we have  $\tilde{p}_2 = 1/20 = 0.05$ . To find the 99% confidence interval, we compute the margin of error as  $z^* \sqrt{\frac{\tilde{p}_1(1-\tilde{p}_1)}{n_1+2} + \frac{\tilde{p}_2(1-\tilde{p}_2)}{n_2+2}}$ , or

$2.576 \sqrt{\frac{.688 * .312}{35} + \frac{.05 * .95}{20}} = 0.238$ . The difference in the proportions of mice who develop tumors is between 39.8% and 87.4% higher in mice with lowered levels of DNA methylation. Since this interval does not contain 0, the difference is significant at the 1% level.

**20.21** We wish to test  $H_0 : p_H = p_{NH}$  against  $H_a : p_H \neq p_{NH}$  where  $p$  is the proportion of papers rejected without review in each group. The proportion of rejected papers that did have help is  $\hat{p}_H = 293/514 = 0.57$ , and the proportion rejected who did not have help is  $\hat{p}_{NH} = 135/190 = 0.711$ . The pooled proportion is  $\hat{p} = (293 + 135) / (514 + 190) = 0.608$ .

Compute Variable		Z
Target Variable:	Numeric Expression:	-3.40
Z	$(.57 - .711) / \sqrt{.608 * .392 * (1/514 + 1/190)}$	

Compute Variable		Pvalue
Target Variable: Pvalue	Numeric Expression: = 2*CDF.Normal(-3.40,0,1)	0.0007

We reject the null hypothesis because the test statistic is  $z = -3.40$  with  $P$ -value 0.0007. The observed difference in papers rejected without review is 14%; papers without statistical help are more likely to be rejected.

**20.23** We compute the confidence interval for how much more often papers without statistical help are rejected after rejecting equality of rejection rates in Exercise 20.21.

Compute Variable		Low
Target Variable: Low	Numeric Expression: = -.141-1.96*sqrt(.57*.43/514+.711*.289/190)	-0.2184
Compute Variable		High
Target Variable: High	Numeric Expression: = -.141+1.96*sqrt(.57*.43/514+.711*.289/190)	-0.0636

Both ends of the interval are negative; with 95% confidence, papers who don't have statistical help are rejected between 6.4% and 21.8% more often than papers that did have statistical help.

**20.27** We want to know if there is a difference in success by gender. We'll test  $H_0: p_M = p_W$  against  $H_a: p_M \neq p_W$ . From the data given, we have  $\hat{p}_W = 23/34 = 0.6765$  and  $\hat{p}_M = 60/89 = 0.6742$ . The pooled estimate is  $\hat{p} = (23 + 60)/(34 + 89) = 0.6748$ .

Compute Variable		Z
Target Variable: Z	Numeric Expression: = (.6765-.6742)/sqrt(.6748*(1-.6748)*(1/34+1/89))	0.02
Compute Variable		Pvalue
Target Variable: Pvalue	Numeric Expression: = 2*(1-CDF.Normal(0.02,0,1))	0.9840

These data do not show a difference in success by gender. 67.6% of females succeed and 67.4% of males succeed. The test statistic is  $z = 0.02$  with  $P$ -value 0.9840. (We really didn't need to compute the  $P$ -value – a test statistic of  $z = 0.02$  will *never* be significant).

**20.29** For the patch only (control) group, we have  $\hat{p}_C = 40/244 = 0.1639$ , and for the treatment group, we have  $\hat{p}_T = 87/245 = 0.3551$ . The confidence interval becomes

Compute Variable		Low
Target Variable:	Numeric Expression:	0.0916
Low	$.1912 - 2.576 * \text{sqrt}(.1639 * (1 - .1639) / 244 + .3551 * (1 - .3551) / 245)$	

Compute Variable		High
Target Variable:	Numeric Expression:	0.2908
High	$.1912 + 2.576 * \text{sqrt}(.1639 * (1 - .1639) / 244 + .3551 * (1 - .3551) / 245)$	

We're convinced the drug helps smokers quit; between 9.2% and 29.1% more will successfully quit with the drug, with 99% confidence.

**20.31** We'll test  $H_0: p_C = P_H$  against  $H_a: p_H > P_C$  where  $p$  is the proportion of offers rejected in each group. Be careful – we'll have to add together the offers accepted and rejected for the total number of offers. From the data, we have  $\hat{p}_C = 6/38 = 0.1579$  and  $\hat{p}_H = 18/38 = 0.4737$ . The pooled estimate is  $\hat{p} = (6 + 18) / (38 + 38) = 0.3158$ .

Compute Variable		Z
Target Variable:	Numeric Expression:	-2.97
Z	$(.1579 - .4747) / \text{sqrt}(.3158 * (1 - .3158) * (1/38 + 1/38))$	

Compute Variable		Pvalue
Target Variable:	Numeric Expression:	0.0015
Pvalue	$\text{CDF.Normal}(-2.97, 0, 1)$	

Computer offers are less likely to be rejected; the test statistic is  $z = 2.97$  with  $P$ -value 0.0015.

**20.35** Calling the group who made impulse purchases group 1, we have the alternate hypothesis  $p_1 \neq p_2$  because we merely want to know if there is a difference in credit card use. Add the Yes and No answers to find the total number of shoppers for each type of purchase. We have  $\hat{p}_1 = 13/31 = 0.419$  and  $\hat{p}_2 = 35/66 = 0.530$ . For the hypothesis test, we also need the “blended”  $p = (13 + 35) / (31 + 66) = 0.495$ . We compute the test statistic and its  $P$ -value below. With a test statistic of  $z = -1.02$  and  $P$ -value of 0.3077, we cannot say there is a difference in credit card use between planned and impulse purchases.

Compute Variable	
Target Variable:	Numeric Expression:
z	(.419-.530)/sqrt(.495*.505*(1/31+1/66))

z
-1.02

Compute Variable	
Target Variable:	Numeric Expression:
P	2*CDF.Normal(-1.02,0,1)

P
0.3077

To find the 95% confidence interval, we compute the margin of error as

$$z^* \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}, \text{ or } 1.96 \sqrt{\frac{.419*(1-.419)}{31} + \frac{.530*(1-.530)}{66}} = 0.211. \quad \text{The}$$

95% confidence interval for the difference in the proportion of credit card purchases is – 32.2% to 10.0%; since this interval includes 0, we again cannot say there is a difference in credit card use for the different types of purchases.

## Chapter 21 SPSS Solutions

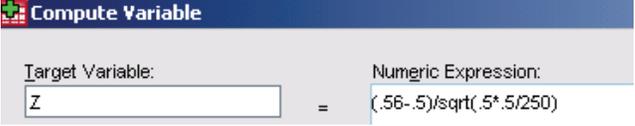
**\*\*NOTE:** SPSS does not do inference based on  $Z$  distributions, nor does it perform inference on variables that are already summarized. If you really want to use SPSS for these problems, follow the instructions below (you'll be basically using **Transform, Compute Variable** as a calculator) or use another technology (such as a graphing calculator or another statistics program like Minitab or Crunchit.)

**21.1** The survey found 36% of adult Internet users using Wikipedia. We calculate the confidence interval as shown below.

 <p>Target Variable: Low = Numeric Expression: <math>.36 - 1.96 * \text{sqrt}(.36 * .64 / 1497)</math></p>	<table border="1" style="margin-left: auto; margin-right: auto;"> <tr><td>Low</td></tr> <tr><td>0.3357</td></tr> </table>	Low	0.3357
Low			
0.3357			
 <p>Target Variable: High = Numeric Expression: <math>.36 + 1.96 * \text{sqrt}(.36 * .64 / 1497)</math></p>	<table border="1" style="margin-left: auto; margin-right: auto;"> <tr><td>High</td></tr> <tr><td>0.3843</td></tr> </table>	High	0.3843
High			
0.3843			

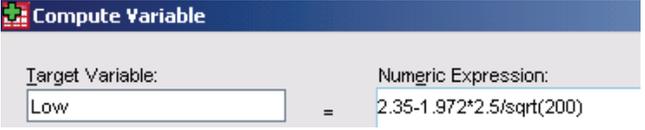
Based on this sample, between 33.6% and 38.4% of adult Internet users consult Wikipedia, with 95% confidence.

**21.3** If the coin is balanced, we should have half heads, so we have  $H_0 : p = 0.5$ . The question is if the coin is unbalanced, so we also have  $H_a : p \neq 0.5$ . The observed proportion of heads is  $\hat{p} = 140 / 250 = 0.56$ . We compute the test statistic and its  $P$ -value (doubling the area above the test statistic) below.

 <p>Target Variable: Z = Numeric Expression: <math>(.56 - .5) / \text{sqrt}(.5 * .5 / 250)</math></p>	<table border="1" style="margin-left: auto; margin-right: auto;"> <tr><td>Z</td></tr> <tr><td>1.90</td></tr> </table>	Z	1.90
Z			
1.90			
 <p>Target Variable: Pvalue = Numeric Expression: <math>2 * (1 - \text{CDF.Normal}(1.90, 0, 1))</math></p>	<table border="1" style="margin-left: auto; margin-right: auto;"> <tr><td>Pvalue</td></tr> <tr><td>0.0574</td></tr> </table>	Pvalue	0.0574
Pvalue			
0.0574			

We have  $z = 1.90$  with  $P$ -value 0.0574. At the 0.05 level, this is not sufficient evidence to conclude that the coin is unbalanced (although it might be close).

**21.5** We need a 95% confidence interval for the mean muscle gap perceived by American/European men. We first need the correct  $t^*$  for a sample of  $n = 200$  (199 df), then compute the lower and upper ends of the interval.

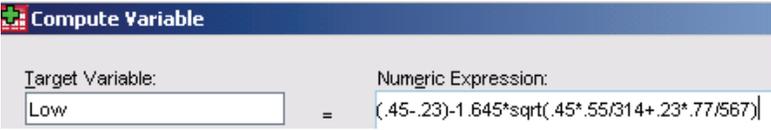
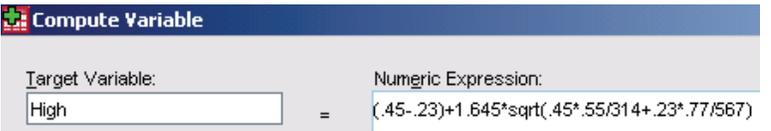
 <p>Target Variable: Tstar Numeric Expression: IDF.T(.975,199)</p>	Tstar 1.972
 <p>Target Variable: Low Numeric Expression: 2.35-1.972*2.5/sqrt(200)</p>	Low 2.0014
 <p>Target Variable: High Numeric Expression: 2.35+1.972*2.5/sqrt(200)</p>	High 2.6986

With 95% confidence, these men think they need between 2.001 and 2.699 kilograms more muscle to make them attractive to women.

**21.7** We'll manually compute a two-sample  $t$  test to determine whether the difference in mean time between matings is significantly different. With a  $t$  statistic of 10.42, the  $P$ -value will be essentially 0, there is a significant difference in time between matings; butterflies given the large spermatophore wait longer between matings.

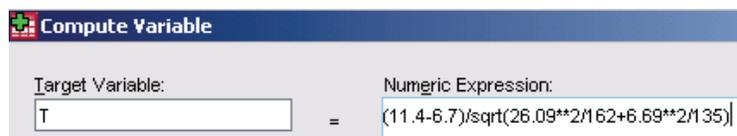
 <p>Target Variable: T Numeric Expression: (5.15-4.33)/sqrt(.18**2/20+.31**2/21)</p>	T 10.42
--	------------

**21.9** From the material before Exercise 21.8, 45% of the 314 Hispanics and 23% of the 567 whites listen to rap every day. We use this information to compute the interval.

 <p>Target Variable: Low Numeric Expression: (.45-.23)-1.645*sqrt(.45*.55/314+.23*.77/567)</p>	Low 0.1654
 <p>Target Variable: High Numeric Expression: (.45-.23)+1.645*sqrt(.45*.55/314+.23*.77/567)</p>	High 0.2746

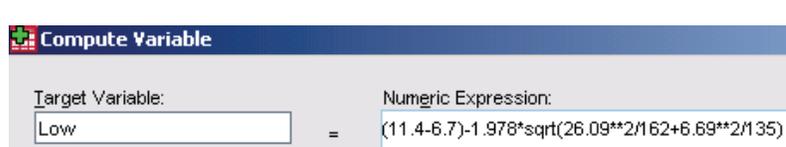
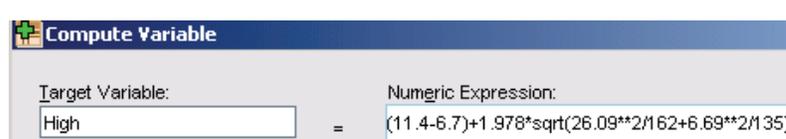
Based on our information, between 16.5% and 27.4% more Hispanics will listen to rap every day, as compared to whites, with 90% confidence.

**21.11** T procedures will be reasonably accurate for these data due to the large sample sizes (the Central Limit theorem), even though the distributions are skewed to the right. We want to know if female mice have significantly higher endurance, on average, so we compute a test statistic and find the one-sided  $P$ -value, using the conservative degrees of freedom.

 <p>Target Variable: T          Numeric Expression: <math>(11.4-6.7)/\sqrt{(26.09^{**}2/162+6.69^{**}2/135)}</math></p>	<table border="1" data-bbox="1071 567 1271 703"> <tr><td>T</td></tr> <tr><td>2.21</td></tr> </table>	T	2.21
T			
2.21			
 <p>Target Variable: Pvalue          Numeric Expression: <math>1-CDF.T(2.21,134)</math></p>	<table border="1" data-bbox="1071 724 1271 877"> <tr><td>Pvalue</td></tr> <tr><td>0.0144</td></tr> </table>	Pvalue	0.0144
Pvalue			
0.0144			

The conclusion will depend on the alpha level of the test. At the 0.05 level, we conclude that female mice *do* have more endurance, on average. At the 0.01 level, we have failed to show a difference in mean endurance between genders of mice.

**21.13** We first find the conservative  $t^*$ , then compute the interval.

 <p>Target Variable: Tstar          Numeric Expression: <math>IDF.T(.975,134)</math></p>	<table border="1" data-bbox="1071 1157 1271 1291"> <tr><td>Tstar</td></tr> <tr><td>1.978</td></tr> </table>	Tstar	1.978
Tstar			
1.978			
 <p>Target Variable: Low          Numeric Expression: <math>(11.4-6.7)-1.978*\sqrt{(26.09^{**}2/162+6.69^{**}2/135)}</math></p>	<table border="1" data-bbox="1071 1312 1271 1459"> <tr><td>Low</td></tr> <tr><td>0.4885</td></tr> </table>	Low	0.4885
Low			
0.4885			
 <p>Target Variable: High          Numeric Expression: <math>(11.4-6.7)+1.978*\sqrt{(26.09^{**}2/162+6.69^{**}2/135)}</math></p>	<table border="1" data-bbox="1071 1480 1271 1619"> <tr><td>High</td></tr> <tr><td>8.9115</td></tr> </table>	High	8.9115
High			
8.9115			

With 95% confidence, female mice have endurance between 0.49 and 8.91 minutes more than male mice, on average.

**21.15** From Exercise 21.14, 5617 students had at least one parent who graduated from college in a sample of 17,554 students., this is a sample proportion of  $\hat{p} = 5617/17554 = 0.32$ . We compute the confidence interval below.

<b>Compute Variable</b> Target Variable: Low = Numeric Expression: $.32 - 2.576 * \text{sqrt}(.32 * .68 / 17554)$		Low
		0.3109
<b>Compute Variable</b> Target Variable: High = Numeric Expression: $.32 + 2.576 * \text{sqrt}(.32 * .68 / 17554)$		High
		0.3291

The information in this survey indicates that between 31.1% and 32.9% of 17-year-old students had at least one parent who graduated from college, with 99% confidence.

**21.19** This is an observational study. It would be unethical to randomly assign babies to be born early and have very low birth weights. We first compute hypothesis test for a null hypothesis of no difference in graduation rates against the alternate  $H_a: p_{VLBW} < p_{NBW}$ . The observed proportions are  $\hat{p}_{VLBW} = 179/242 = 0.7397$  and  $\hat{p}_{NBW} = 193/233 = 0.8283$ . Further, the pooled estimate of the proportion is  $\hat{p} = (179 + 193)/(242 + 233) = 0.7832$ .

<b>Compute Variable</b> Target Variable: Z = Numeric Expression: $(.7397 - .8283) / \text{sqrt}(.7832 * (1 - .7832) * (1/242 + 1/233))$		Z
		-2.34
<b>Compute Variable</b> Target Variable: Pvalue = Numeric Expression: $\text{CDF.Normat}(-2.34, 0, 1)$		Pvalue
		0.0096

With test statistic  $z = -2.34$  and P-value 0.0096, we will conclude that very low birth weight babies are less likely to have graduated from high school by age 20 than normal weight babies.

**21.21** The first of these questions is a test of proportions. We want to test  $H_0: p_{VLBW} = p_{NBW}$  against the two-tailed alternate. From the data, we have  $\hat{p}_{VLBW} = 37/126 = 0.2937$  and  $\hat{p}_{NBW} = 52/124 = 0.4194$ . We also have the pooled estimate  $\hat{p} = (37 + 52)/(126 + 124) = 0.356$ . We compute the test statistic and its P-value below.

Compute Variable		Z
Target Variable:	Numeric Expression:	-2.08
Z	$(.2936-.4194)/\text{sqrt}(.356*.644*(1/126+1/124))$	

Compute Variable		Pvalue
Target Variable:	Numeric Expression:	0.0375
Pvalue	$2*\text{CDF.Normal}(-2.08,0,1)$	

For the question about drug use, it seems the VLBW women are less likely to use illegal drugs;  $z = -2.08$  with  $P$ -value 0.0375.

The question about IQ needs a two-sample  $t$  test. We compute the test statistic and the conservative  $P$ -value below. The results of this test are very similar. The VLBW women have significantly lower mean IQ (at the 5% level) than women of normal birth weight.

Compute Variable		T
Target Variable:	Numeric Expression:	-2.08
T	$(86.2-89.8)/\text{sqrt}(13.4**2/126+14.0**2/124)$	

Compute Variable		Pvalue
Target Variable:	Numeric Expression:	0.0396
Pvalue	$2*\text{CDF.T}(-2.08,123)$	

**21.23** We want to test  $H_0 : p_{LF} = p_N$  against  $H_a : p_{LF} < p_N$ , where  $p_{LF}$  is the proportion of women who eat low fat diets that will develop breast cancer and  $p_N$  is the proportion of women who eat normal diets that will develop breast cancer. From the data, we have  $\hat{p}_{LF} = 655/19541 = 0.0335$  and  $\hat{p}_N = 1072/29294 = 0.0366$ . We also have the pooled estimate  $\hat{p} = (655 + 1072)/(19541 + 29294) = 0.0354$ .

Compute Variable		Z
Target Variable:	Numeric Expression:	-1.82
Z	$(.0335-.0366)/\text{sqrt}(.0354*(1-.0354)*(1/19541+1/29294))$	

Compute Variable		Pvalue
Target Variable:	Numeric Expression:	0.0344
Pvalue	$\text{CDF.Normal}(-1.82,0,1)$	

The difference is significant at the 5% level with a  $P$ -value of 0.036. The indication is that a low fat diet will reduce breast cancer.

**21.25** We'll manually compute a two-sample  $t$  test to determine whether the pets have a higher mean cholesterol level than clinic dogs. With a test statistic of 1.17 and conservative  $P$ -value 0.1273, these data do not show a difference in mean cholesterol levels.

Compute Variable	
Target Variable:	Numeric Expression:
T	$(193-174)/\sqrt{68^{**}2/26+44^{**}2/23}$

T
1.17

Compute Variable	
Target Variable:	Numeric Expression:
P	$1-\text{CDF.T}(1.17,22)$

P
0.1273

**21.27** With only summary statistics, we'll manually compute the interval using the conservative degrees of freedom. The interval is  $(\bar{x}_1 - \bar{x}_2) \pm t^* \sqrt{s_1^2/n_1 + s_2^2/n_2}$ . This becomes  $(193-174) \pm 2.074 \sqrt{68^2/26 + 44^2/23} = 19 \pm 33.571$ . We are 95% confident the difference in mean cholesterol levels is between  $-14.57$  and  $52.57$ ; since 0 is included in the interval, clinic and pet dogs may have the same mean cholesterol level.

**21.39** If a rat is successful in 80 of 80 trials, its success rate is  $\hat{p} = 80/80 = 1$ . A large sample confidence interval will be from 1 to 1, since  $(1 - p) = 0$  (in other words, there is no variability). To find the plus four estimate, add four to the number of trials (this becomes 84) and two to the number of successes (this becomes 82). The plus four estimate is then  $\tilde{p} = 82/84 = 0.9762$ .

Compute Variable	
Target Variable:	Numeric Expression:
Low	$.9762-1.96*\sqrt{.9762*.0238/84}$

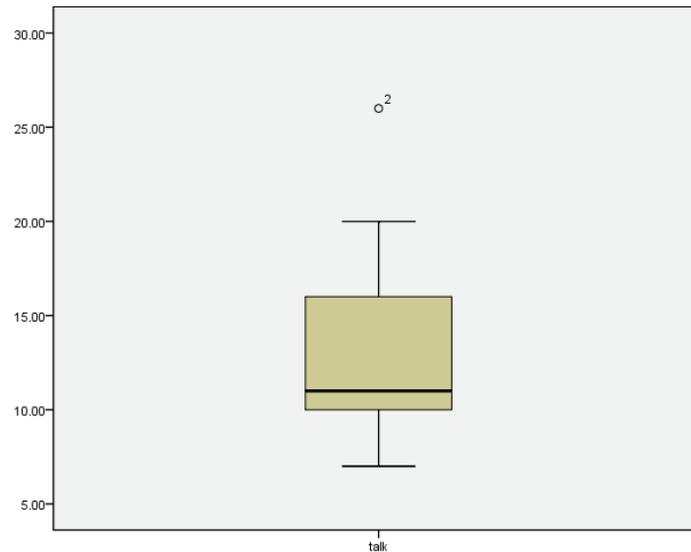
Low
0.944

Compute Variable	
Target Variable:	Numeric Expression:
High	$.9762+1.96*\sqrt{.9762*.0238/84}$

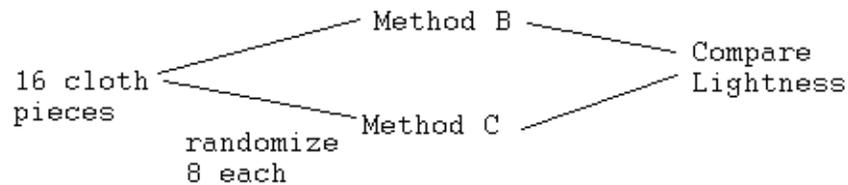
High
1.009

We'll estimate such a rat would be successful at least 94.4% of the time (it can't be right more than 100% of the time) with 95% confidence.

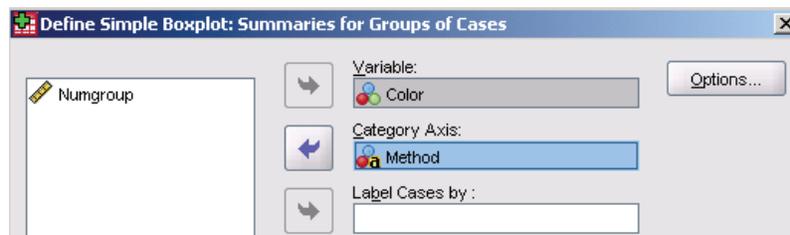
**21.41** Open data file *ex21\_41.por* (or enter the data). This is a small sample ( $n = 20$ ), so we check a boxplot for skewness and outliers. Use **Graphs, Legacy Dialogs, Boxplot**. Click to enter the variable name and **OK**. This boxplot does show that the child whose first word was at 26 months is an outlier. The distance from the median to the right side indicates a skew. Inference using t distributions is still not appropriate for this data.

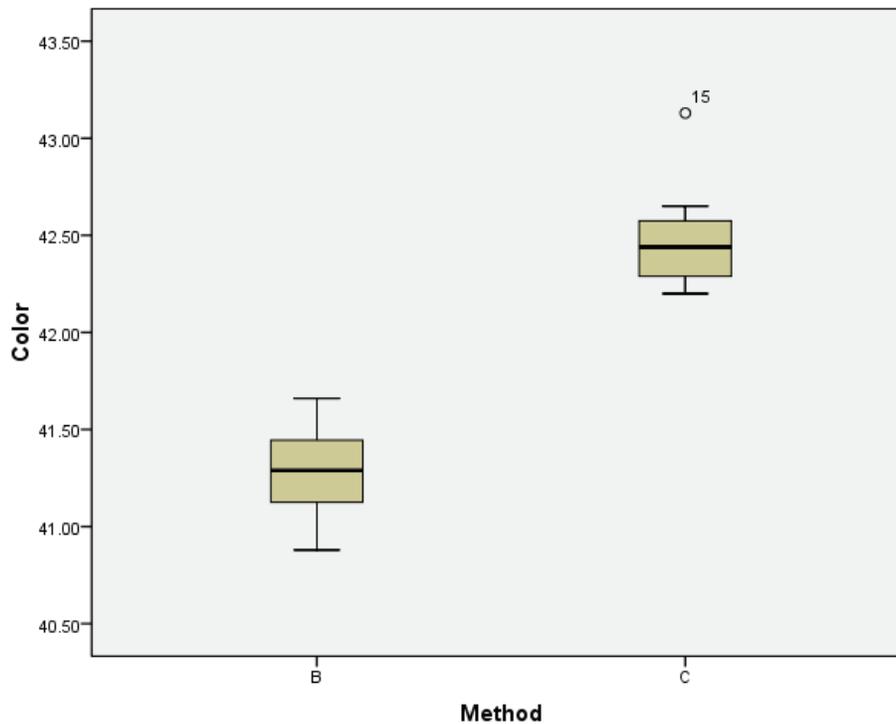


**21.45** An outline of the experiment might be as below.

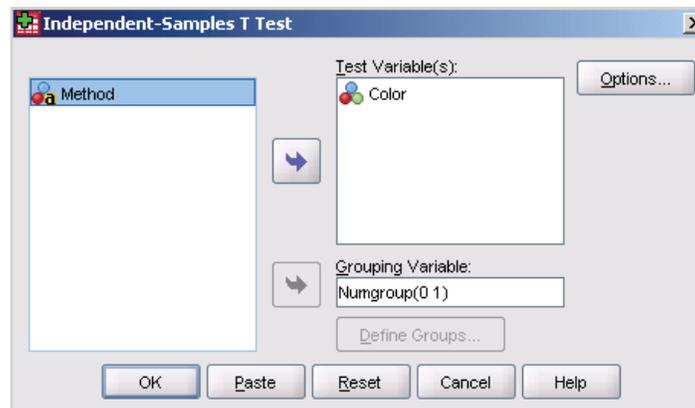


We want to know which gives the lower score (darker color) on average. Open data file *ex21-45*. This file has one column for the method and another for the color score. Before doing a test, we should check to see that our data are (roughly) Normal, since these are small samples. We create side-by-side boxplots using **Graphs, Legacy Dialogs, Boxplot** with a **Simple** chart of data for **Summaries for groups of cases**.





Method C has a high outlier (43.13). However, with sample sizes of  $n = 8$ , the  $1.5 \times \text{IQR}$  criterion is not always reliable. Since that value is not particularly extreme, and it certainly appears that Method B gives the darker color (has lower values), we'll proceed to the test. The methods listed in the data file are B and C; SPSS requires integer values for groups. We created a new variable called **Numgroup** where 0 = B and 1 = C. To perform the test, use **Analyze, Compare Means, Independent Samples T-Test**.



		t-test for Equality of Means						
		t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
							Lower	Upper
Color	Equal variances assumed	-8.794	14	.000	-1.21000	.13759	-1.50510	-.91490
	Equal variances not assumed	-8.794	13.727	.000	-1.21000	.13759	-1.50565	-.91435

The test statistic is  $t = -8.79$  with  $P$ -value 0.000. We have clear evidence of a difference between the two methods; method B gives darker colors (on average).

**21.47** The data in Table 21.1 give the number of drinks per session by female students whose parents do or do not allow them to drink. In the table, there are 65 students whose parents allow them to drink and 33 who do not (making a total of 98 female students represented); this makes  $\hat{p} = 65/98 = 0.6633$ .

<b>Compute Variable</b>		Low
Target Variable:	Numeric Expression:	0.570
Low	= .6633-1.96*sqrt(.6633*.3367/98)	
<b>Compute Variable</b>		High
Target Variable:	Numeric Expression:	0.757
High	= .6633+1.96*sqrt(.6633*.3367/98)	

Assuming these sophomore students are representative of females (sophomore) students in general, we estimate that between 57.0% and 75.7% have at least one parent who “allows” them to drink, with 95% confidence.

**21.49** For the experiment discussed in part (a), we’re interested in the proportion of crackers with cracking. Since there were only 3 microwaved crackers that showed visible cracking, we’ll use a “plus 4” confidence interval to estimate the difference in proportions. We add 1 “success” (a cracked cracker) to each group, and 2 trials to each group, and find  $\tilde{p}_1 = 4/67 = 0.060$  and  $\tilde{p}_2 = 58/67 = 0.866$ . To find the 95% confidence

interval, we compute the margin of error as  $z^* \sqrt{\frac{\tilde{p}_1(1-\tilde{p}_1)}{n_1+2} + \frac{\tilde{p}_2(1-\tilde{p}_2)}{n_2+2}}$ , or

$1.96 \sqrt{\frac{.06 \cdot .94}{67} + \frac{.866 \cdot .134}{67}} = 0.099$ . Based on this information, microwaving crackers results in between 70.6% and 90.5% fewer crackers with cracking, with 95% confidence.

For the experiment in part (b), we'll analyze the data with a two-sample  $t$  test, since we're interested in mean pressure to break the crackers. Since the intent of the experiment is to prove that microwaving improves resistance to breaking, we have selected the greater than alternate hypothesis. The test statistic is  $t = 6.91$  with  $P$ -value 0.0000. These data clearly show that microwaving crackers improve their resistance to breaking.

Compute Variable		T
Target Variable:	Numeric Expression:	6.91
T	$(139.6-77)/\sqrt{(33.6^2/20+22.6^2/20)}$	

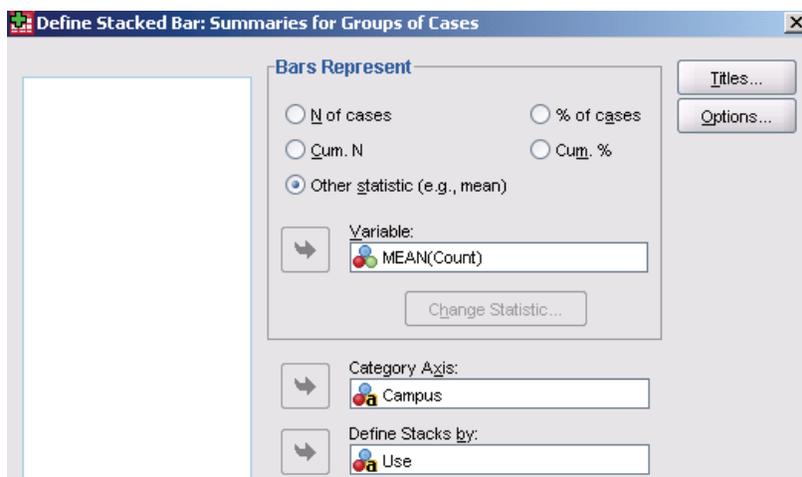
Compute Variable		P
Target Variable:	Numeric Expression:	0.0000
P	$1-\text{CDF.T}(6.91,19)$	

## Chapter 22 SPSS Solutions

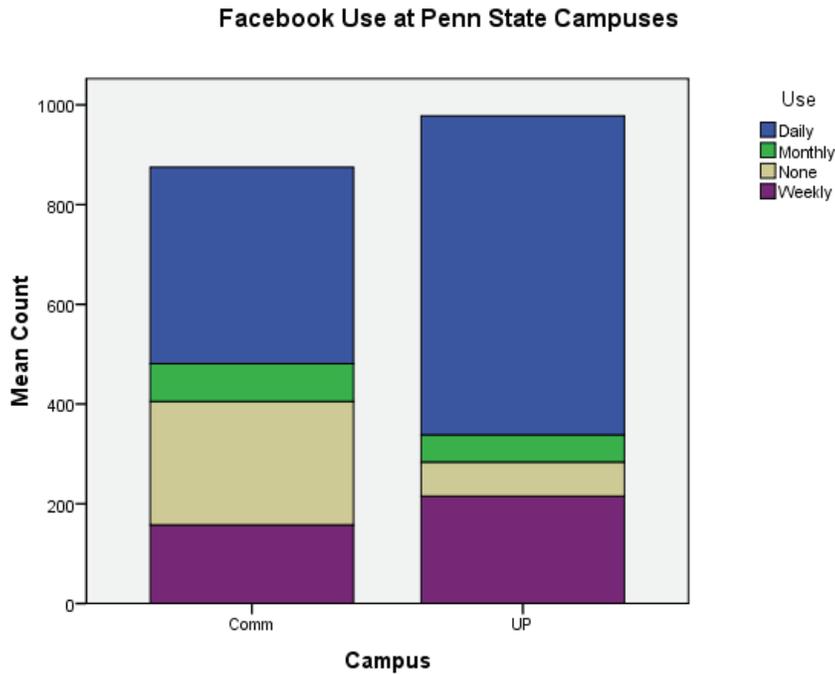
**22.1** To find the percent of University Park students who fall in each Facebook category, add the values given for University Park ( $68 + 55 + 215 + 640 = 978$ ). Then, divide each category's number by the total. We see that about 7% of the University Park students do not use Facebook and about 5.6% use it several times per month or less. Continue with the other two categories, to find that about 22% use Facebook at least once a week and 65.4% use it at least once a day.

```
68+55+215+640      978
68/978             .0695296524
55/978             .0562372188
```

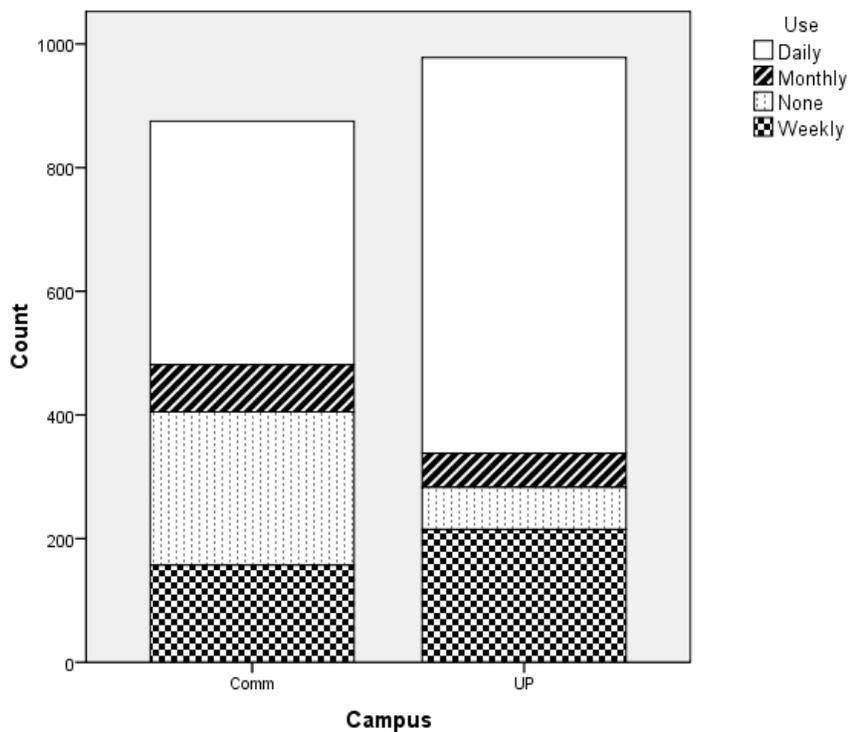
To compare the distributions, we'll make a stacked bar chart of the data with one bar for each of the University Park and Commonwealth students. Open data file *ex22-01*. To create the chart, click **Graphs, Legacy Dialogs, Bar**. We want the **Stacked** bar chart where data are **Summaries for groups of cases**. We'll create an initial chart (and modify it later with the Chart Editor) by defining it as below (don't forget to give your graph a **Titles**).



It's hard to compare the two distributions in the initial graph, because there were different numbers of students surveyed at the different campuses.



Click in the graph to bring up the Chart Editor, then click **Options**, **Scale to 100%**. You can also click in the y-axis label and remove it by unchecking the **Display axis title** box (with percents showing, this is not needed). If you wish, click the **Variables** tab and change the **Style** for **Use** from Color to pattern. **Apply** and **Close** the Chart Editor.



It is clear that University Park students are much more likely to be daily Facebook users; Commonwealth students are more likely to not use it at all; the “occasional” users seem similar.

**22.3** Parts (a) and (b) want us to compute tests for a difference in proportions. We first compute the test for those who do not use Facebook. There were  $68/978 = 0.0695$  University Park students who do not use it and  $248/875 = 0.2834$  Commonwealth students who do not. The pooled proportion is  $(68+248)/(978+875) = 0.1705$ .

Compute Variable		Z
Target Variable:	Numeric Expression:	-12.22
Z	$(.0695-.2834)/\text{sqrt}(.1705*.8295*(1/978+1/875))$	

With a test statistic of  $z = -12.22$ , we do not really need to compute the  $P$ -value, as this will be (essentially) 0. There is a difference. University Park students are definitely more likely to use Facebook.

We repeat the computation for those using Facebook at least once a week. The observed proportions are: University Park,  $215/978 = 0.2198$  and Commonwealth,  $157/875 = 0.1794$ . The pooled proportion is  $(215+157)/(978+875) = 0.2008$ .

Compute Variable		Z
Target Variable:	Numeric Expression:	-2.17
Z	$(.1794-.2198)/\text{sqrt}(.2008*.7992*(1/978+1/875))$	

Compute Variable		Pvalue
Target Variable:	Numeric Expression:	0.0300
Pvalue	$2*\text{CDF.Normal}(-2.17,0,1)$	

The difference is not quite as significant, but is still significant at the 0.05 level ( $P$ -value 0.030). Again, University Park students are more likely to use Facebook at least once a week.

These two  $P$ -values can't tell us about the two distributions for all four outcomes because they don't represent all the categories. Further, they are really dependent – if a student is in one category, they can't be in another, but we don't know which other category.

**22.5** If there is no relationship, the expected counts are  $(R \times C)/T$ , where  $R$  is the row total,  $C$  is the column total, and  $T$  is the grand total. The grand total for the table is  $910 + 627 = 1537$ . There were a total of  $55 + 76 = 131$  students who use Facebook several times a month or less. The expected count of these for Commonwealth students is 53.44. Similarly, the Commonwealth expected count for at least weekly users is 151.75 and for at least once a day users, the expected count is 421.81. The expected counts should total 627; we see they do.

910+627 55+76 131*627/1537 53.43981783	215+157 372*627/1537 151.7527651 627*1034/1537 421.807417	53.44+151.75+421.81 627
---	---	----------------------------

The general trend for these older Commonwealth students is that they are more likely to be occasional Facebook users than daily users; other claims on their time is most likely the reason.

**22.13** The expected counts are  $53 \times 1/3 = 17.6667$ , since if the tilts made no difference, there should be an equal number of strikes on each type of window. We enter the observed and expected counts in two variables and compute the components of the chi-square statistic as shown below. Sum the components to find  $\chi^2 = 16.11$ .

observe	expect	chis
31	17.6667	10.06
14	17.6667	0.76
8	17.6667	5.29

Pvalue
0.0003

Target Variable: chis = Numeric Expression: (observe-expect)**2/expect
Target Variable: Pvalue = Numeric Expression: 1-CDF.Chis(16.11,2)

**22.15** We entered the data as shown at right. Our null hypothesis is that the counts agree with the population proportions; the alternate is that they do not agree. SPSS still doesn't like summarized data. We add the number of observations to find that  $401 + 480 + 20 = 901$  citations represented. We compute the test statistic entries (and then sum them) to find  $\chi^2 = 79.3$ .

Proportion	Count	Age
0.328	401	16 to 29
0.534	480	30 to 59
0.078	20	60 up

Compute Variable	
Target Variable:	Numeric Expression:
Chis	(Count-Proportion*901)**2/(Proportion*901)

Chis
37.64
5.69
35.97

Compute Variable	
Target Variable:	Numeric Expression:
Pvalue	1-CDF.Chis(79.3,2)

Pvalue
0.0000

With a test statistic of  $\chi^2 = 119.84$  and  $P$ -value of 0.000, we conclude that the actual citations do not match the population distributions. It is clear from the above the the largest contributions come from the youngest and oldest age groups. The younger ones are cited much more than expected, the older ones much less.

**22.17** If births are equally spread throughout the year, each sign should have 1/12 of them. We have  $H_0$ : all signs have probability 1/12.  $H_A$  is that  $H_0$  is false. We will perform a  $\chi^2$  goodness-of-fit test with the given data. (It is reasonable to assume the GSS is a random survey of all US adults.) The data given represent 4344 individuals. Under the null hypothesis, we expect  $4344/12 = 362$  individuals in each sign. We omit details (see Exercises 22.113 and 22.15 above), and find  $\chi^2 = 19.76$  with  $P$ -value 0.049, barely significant at the 5% level. We reject  $H_0$  and conclude births are not equally spread through the year. We can see that Aries and Virgo make the largest contributions to the statistic – Aries (a winter month) has a lower than expected count and Virgo (a fall month) has a higher than expected count.

**22.29** If we combine the races, we have  $140 + 976 + 121 = 1237$  individuals who would let the racist speak and  $129 + 480 + 131 = 740$  who would not, making a total sample of size  $n = 1977$ . The observed proportion who would allow a racist to speak is  $\hat{p} = 1237/1977 = 0.6257$ .

Compute Variable	
Target Variable:	Numeric Expression:
Low	.6257-2.576*sqrt(.6257*.3743/1977)

Low
0.598

Compute Variable	
Target Variable:	Numeric Expression:
High	.6257+2.576*sqrt(.6257*.3743/1977)

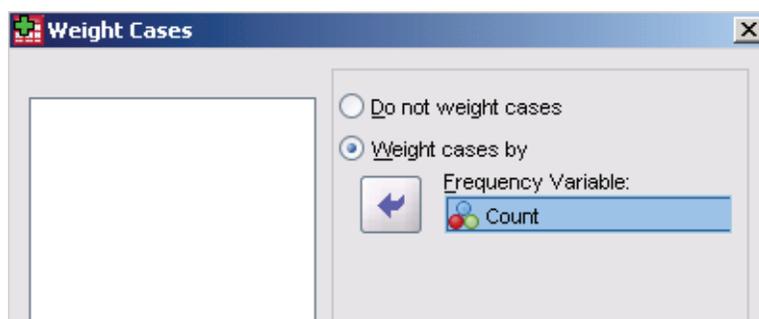
High
0.654

Based on this GSS survey, between 59.8% and 65.4% of U.S. adults think a racist should be allowed to speak, with 99% confidence.

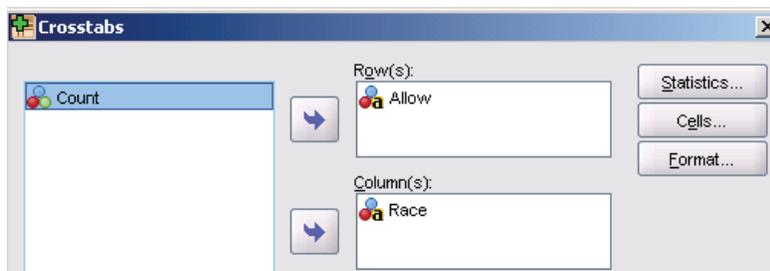
There were 269 Blacks, of whom  $140/269 = 52.0\%$  thought racists should be allowed to speak. For Whites, the percent is  $976/(976+480) = 67.0\%$ ; for Others we have  $121/252 = 48.0\%$ . Both the Blacks and Others have percentages much less than Whites, but there were more Whites in the sample. To perform the chi-square test, enter the data as below.

Race	Allow	Count
black	yes	140
black	no	129
white	yes	976
white	no	480
other	yes	121
other	no	131

Click **Data**, **Weight Cases**. Click to weight cases by **Count**, then **OK**.



Now, click **Analyze**, **Descriptive Statistics**, **Crosstabs**. Click to enter **Allow** as the row and **Race** as the column. Now, click the **Statistics** button and check the box to ask for the Chi-square. **Continue** and click the **Cells** button. Click to ask for the observed and expected counts. **Continue** and **OK** computes the test.



We have the table below with both observed and expected counts.

Allow \* Race Crosstabulation

			Race			
			black	other	white	Total
Allow	no	Count	129	131	480	740
		Expected Count	100.7	94.3	545.0	740.0
	yes	Count	140	121	976	1237
		Expected Count	168.3	157.7	911.0	1237.0
Total		Count	269	252	1456	1977
		Expected Count	269.0	252.0	1456.0	1977.0

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	47.899 <sup>a</sup>	2	.000
Likelihood Ratio	46.952	2	.000
N of Valid Cases	1977		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 94.32.

The *P*-value of the test is 0.000. We have overwhelming evidence that more whites would allow a racist to speak than Blacks or people of other ethnicities. Note that the largest contributions to the test statistic are from the Other column.

**22.43** We're using the data layout from file *ex22-43*. This file has race in a column, school opinion in one, and the counts in a third. We again use the variable Count to weight cases, then use **Analyze**, **Descriptive Statistics**, **Crosstabs** as described in Exercise 22.29.

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	22.426 <sup>a</sup>	8	.004
Likelihood Ratio	22.897	8	.003
N of Valid Cases	605		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 21.26.

Schools \* Race Crosstabulation

			Race			
			black	white	hispanic	Total
Schools	don't	Count	22	14	28	64
		Expected Count	21.4	21.3	21.4	64.0
	excel	Count	12	22	34	68
		Expected Count	22.7	22.6	22.7	68.0
	fair	Count	75	60	61	196
		Expected Count	65.4	65.1	65.4	196.0
	good	Count	69	81	55	205
		Expected Count	68.4	68.1	68.4	205.0
	poor	Count	24	24	24	72
		Expected Count	24.0	23.9	24.0	72.0
Total	Count		202	201	202	605
		Expected Count	202.0	201.0	202.0	605.0

The differences in the distributions are statistically significant ( $P = 0.004$ ). To see the departures from the null hypothesis, examine the expected counts. Blacks are less likely to call schools Excellent than expected (12 observed versus 22.7 expected) while Hispanics are more likely to call them Excellent (34 observed and 22.7 expected) and less likely to call them Good (55 versus 68). Blacks are more likely to call them Good (75 versus 65.4). There seems to be no real differences among the ethnicities on calling the schools Poor.

**22.45** We've used the data in *ex22-45*. As in the last two exercises, we use Data, Weight Cases to weight the results by Count. We then use **Analyze, Descriptive Statistics, Crosstabs** to recreate the table and add the expected counts (click **Cells, Expected**).

Newpref \* Group Crosstabulation

			Group				
			hardhot	hardwarm	softhot	softwarm	Total
Newpref	no	Count	30	42	27	53	152
		Expected Count	30.9	47.2	24.0	49.8	152.0
	yes	Count	42	68	29	63	202
		Expected Count	41.1	62.8	32.0	66.2	202.0
Total	Count		72	110	56	116	354
		Expected Count	72.0	110.0	56.0	116.0	354.0

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	2.058 <sup>a</sup>	3	.560
Likelihood Ratio	2.062	3	.560
N of Valid Cases	354		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 24.05.

There is no significant difference between the person's laundry practice and their preference for the new product ( $P = 0.560$ ), although it appears that the people with soft water seem to prefer the standard product (their expected counts are somewhat smaller than the observed) and the people with hard water seem to prefer the new product (their expected counts are also a bit smaller than observed).

**22.47** The new table will be as shown below.

	None	High School	Jr. college	Bachelor	Graduate
Democrat leaning	279	996	156	313	218
Republican leaning	135	731	129	336	128

To see if support differs by level of education, we enter the data as shown below. As in the last exercises, we weight cases by Count and use Analyze, Descriptive Statistics, Crosstabs to compute the test. Do not forget to ask for the Chi-squared **Statistic** and the **Cell Expected** values.

Leaning	Education	Count
Democrat	None	279
Democrat	HS	996
Democrat	JC	156
Democrat	Bachelor	313
Democrat	Graduate	218
Republican	None	135
Republican	HS	731
Republican	JC	129
Republican	Bachelor	336
Republican	Graduate	128

Leaning \* Education Crosstabulation

			Education					Total
			Bachelor	Graduate	HS	JC	None	
Leaning Democrat	Count		313	218	996	156	279	1962
	Expected Count		372.2	198.4	990.5	163.5	237.4	1962.0
Republican	Count		336	128	731	129	135	1459
	Expected Count		276.8	147.6	736.5	121.5	176.6	1459.0
Total	Count		649	346	1727	285	414	3421
	Expected Count		649.0	346.0	1727.0	285.0	414.0	3421.0

Chi-Square Tests

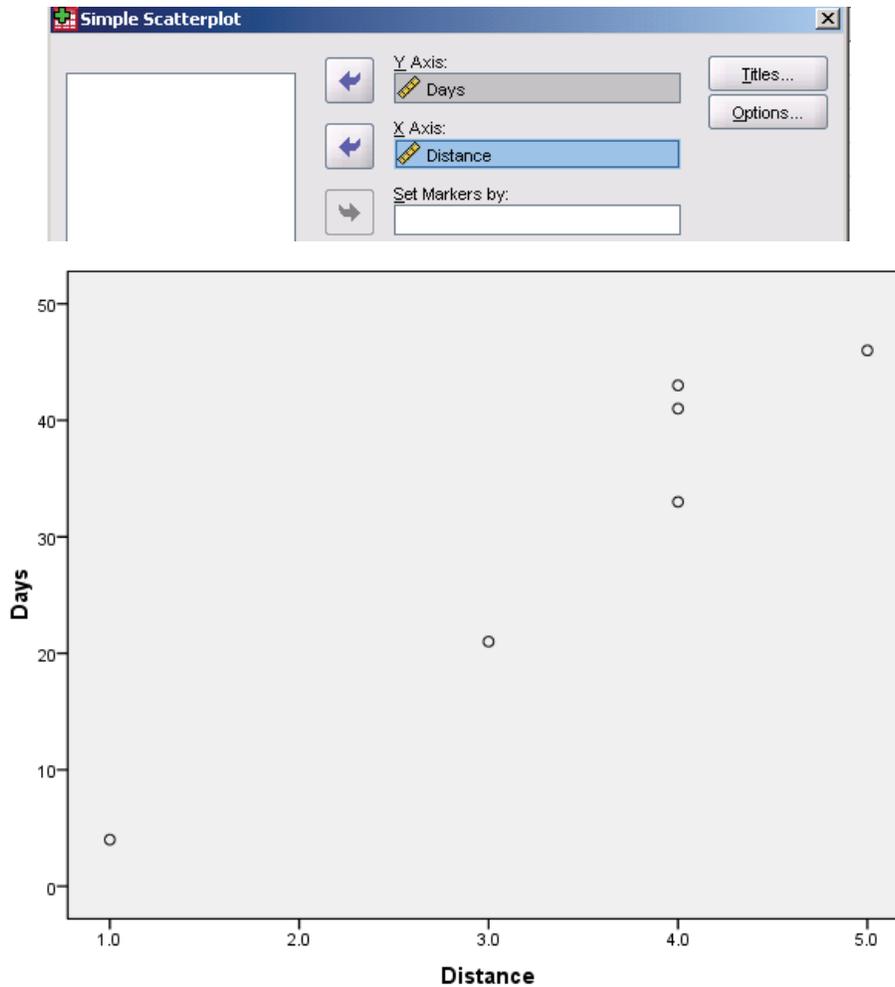
	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	44.539 <sup>a</sup>	4	.000
Likelihood Ratio	44.806	4	.000
N of Valid Cases	3421		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 121.55.

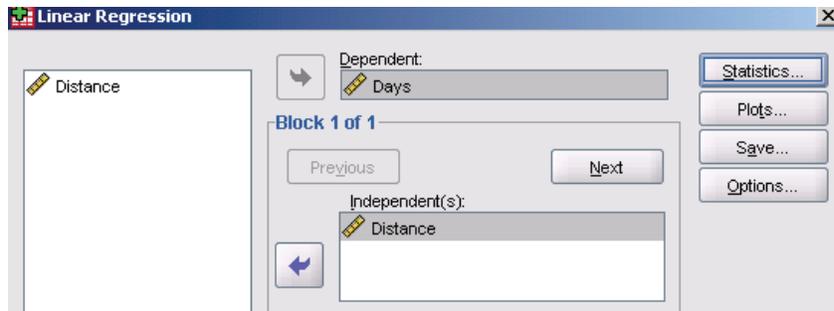
With a 0.000 *P*-value, we conclude there is a difference in political leaning with education level. People with no high school education are more likely to lean Democrat as are people with either a Bachelor's or graduate degree; in other words, the Democrats seem to draw support from either people with little or a lot of education.

## Chapter 23 SPSS Solutions

23.1 Use **Graphs**, **Legacy Dialogs**, **Scatter/Dot** to create a plot of the data.



The plot is strongly linear and increasing. We could use **Analyze**, **Correlate**, **Bivariate** to find the correlation, but we also want to find the regression equation, so use **Analyze**, **Regression** to compute the regression equation (we'll use the square root of  $r^2$  as the correlation). We can have SPSS find the residuals for us by clicking **Save** and checking the box for **Unstandardized** residuals.

**Model Summary<sup>a</sup>**

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.962 <sup>a</sup>	.926	.908	4.903

a. Predictors: (Constant), Distance

b. Dependent Variable: Days

**Coefficients<sup>a</sup>**

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-8.088	5.917		-1.367	.243
	Distance	11.263	1.591	.962	7.080	.002

a. Dependent Variable: Days

The correlation is  $r = 0.962$  – a very strong relationship. Our estimated slope of  $b = 11.263$  says the virus takes about 11.263 days for each additional home range it must travel. The estimated intercept is  $a = -8.088$ . The standard deviation around the regression line is  $s = 4.90345$  (labeled as Std. Error of the Estimate). To sum (or find the mean of) the residuals (created as variable **RES\_1**) use **Analyze, Descriptive Statistics, Descriptives**. With a mean of 0.0000000, the sum must be 0.

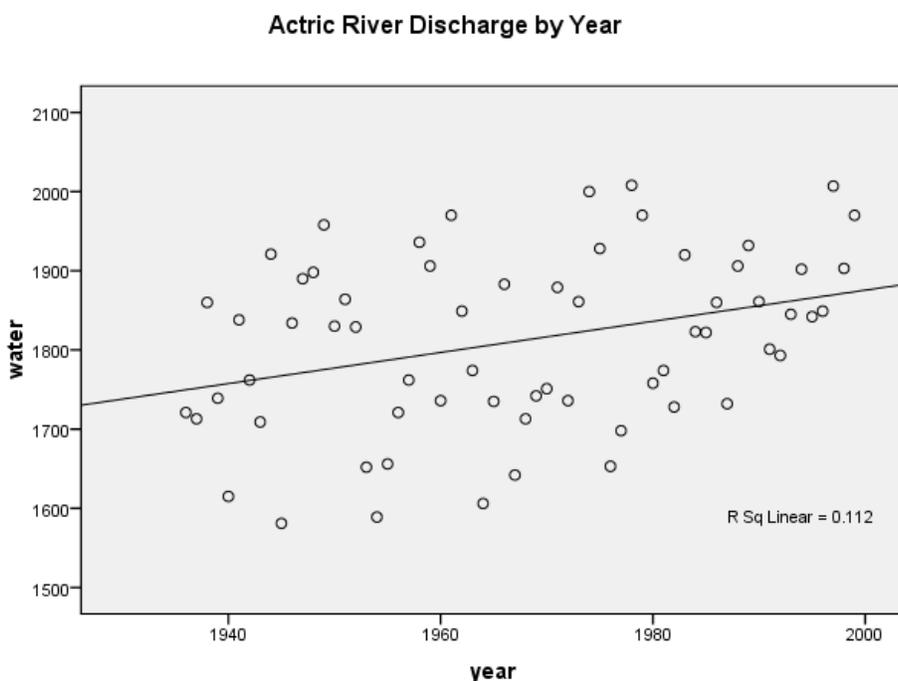
**Descriptive Statistics**

	N	Minimum	Maximum	Mean	Std. Deviation
Unstandardized Residual	6	-4.70175	6.03509	.0000000	4.38578245
Valid N (listwise)	6				

23.3 We define a scatterplot of the data in *ta23-02*.



To add the regression line in the graph, double-click for the Chart Editor, then click **Elements, Fit line at total**. There is an increasing trend in the graph, with lots of scatter. SPSS gives  $r^2 = 0.112$ ; the relationship is fairly weak. Only 11% of the variation in discharge is explained by time (year); there certainly are other factors involved.



Use **Analyze, Regression, Linear** to fit the regression. Looking ahead, we have asked for confidence intervals for the coefficients using the **Statistics** button.

Coefficients <sup>a</sup>								
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	-2056.769	1384.687		-1.485	.143	-4824.720	711.181
	Year	1.966	.704	.334	2.794	.007	.559	3.373

a. Dependent Variable:

**Model Summary**

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.334 <sup>a</sup>	.112	.097	104.003

a. Predictors: (Constant),

The regression equation is Discharge =  $-2057 + 1.97 \cdot \text{Year}$ . The regression standard error is  $s = 104.003$ .

**23.5** From the SPSS output in Exercise 23.3, the test statistic is  $t = 2.794$  with (two-sided)  $P$ -value 0.007. The one-sided  $P$ -value is 0.0035. Since this  $P$ -value is less than any standard  $\alpha$ , we reject a null hypothesis of no relationship and conclude that these data do show an increase in Arctic river discharge (supporting the global warming hypothesis).

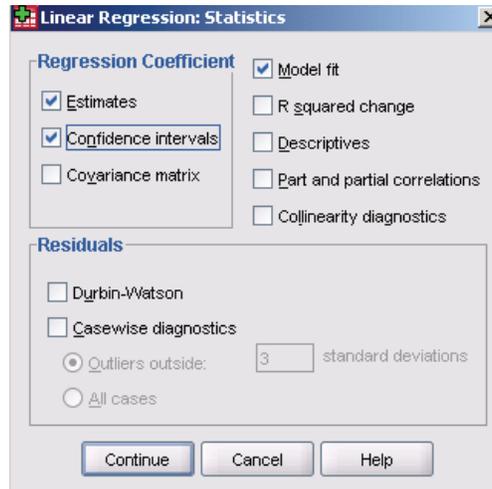
**23.7** Refer to the solution for Exercise 23.1 In the SPSS results, we were given  $t = 7.08$  and  $P = 0.002$ . Since this is a two-sided  $P$ -value, divide by 2. The one-sided  $P$ -value is 0.001. Minitab gives the same (two-sided)  $P$ -value for the correlation if you use **Stat, Basic Statistics, Correlation**. If we use **Analyze, Correlate, Bivariate**, and ask for the one-sided  $P$ -value, we have the same result.

**Correlations**

		Distance	Days
Distance	Pearson Correlation	1.000	.962**
	Sig. (1-tailed)		.001
	N	6.000	6
Days	Pearson Correlation	.962**	1.000
	Sig. (1-tailed)	.001	
	N	6	6.000

\*\* . Correlation is significant at the 0.01 level (1-tailed).

**23.9** SPSS will find the 95% confidence interval if you redo the regression and click **Statistics**, then check the box to ask for confidence intervals for the coefficients.

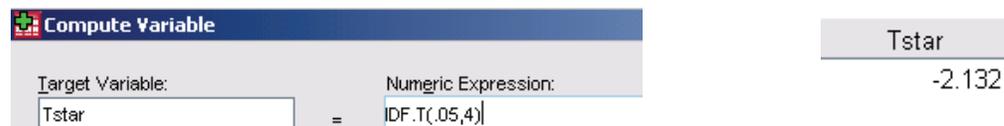


**Coefficients<sup>a</sup>**

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	-8.088	5.917		-1.367	.243	-24.516	8.341
	Distance	11.263	1.591	.962	7.080	.002	6.846	15.680

a. Dependent Variable: Days

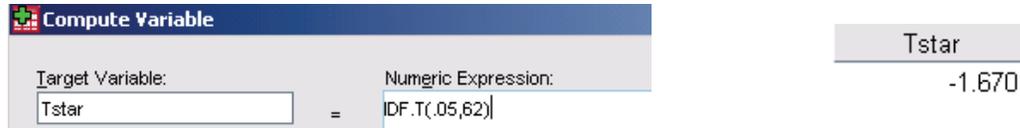
We have the same output as in Exercise 23.1, with the addition of confidence bounds at the right side. Based on this data, Ebola takes between 6.85 and 15.68 days to travel one home range, with 95% confidence. However, the problem asked for 90% confidence. For this, we use **Transform, Compute Variable** to find  $t^*$ , then compute the interval “by hand.”



The interval is  $11.263 \pm 2.132 * 1.591 = (7.871, 14.655)$ . Based on this data, Ebola takes between 7.87 and 14.66 days to travel one home range, with 90% confidence.

**23.11** SPSS gives only 95% confidence intervals for regression parameters. We saw in Exercise 23.3 that a 95% confidence interval for the slope is from 0.559 to 3.373.

However, this question asks for a 90% confidence interval. We use **Transform, Compute Variable** to find  $t^*$  (degrees of freedom are  $n - 2$ ), then compute the interval “by hand.”



The confidence interval is calculated as  $b \pm t^* SE(b)$ , giving  $1.9662 \pm 1.670 * 0.7037$ , or (0.791, 3.141). We are 90% confident that arctic river discharge increases between 0.791 and 3.141 cubic kilometers per year. Since the low end is positive, we’re convinced that discharge is increasing over time.

**23.29** To make the stemplot, use **Analyze, Descriptive Statistics, Explore**.

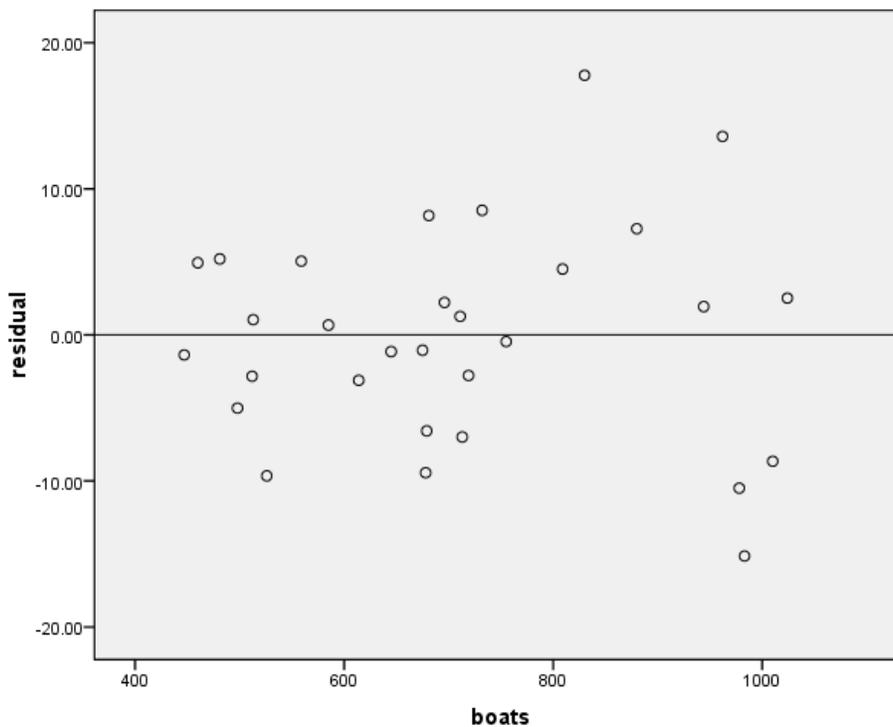
Stem-and-Leaf Plot

Frequency	Stem &	Leaf
1.00	-1 .	5
1.00	-1 .	0
6.00	-0 .	566899
7.00	-0 .	0111223
8.00	0 .	01112244
5.00	0 .	55788
1.00	1 .	3
1.00	1 .	7

Stem width: 10.00  
Each leaf: 1 case(s)

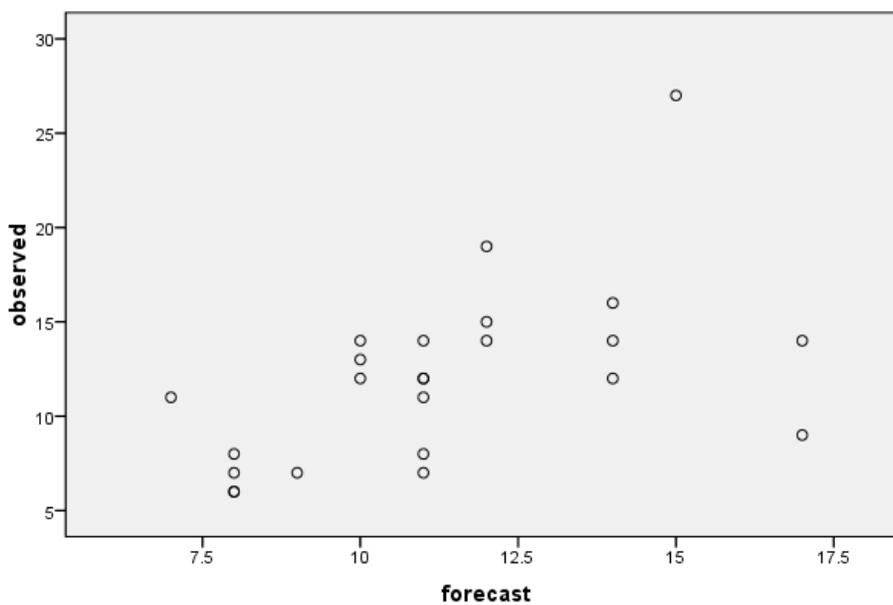
This plot is pretty symmetric and bell-shaped with no outliers. The Normal assumption is reasonable for these residuals. To make the scatterplot, use **Graphs, Legacy Dialogs, Scatter/Dot**. Use **Residual** on the y axis and **Boats** on the x axis. To add the “residual = 0” line, double click in the graph for the Chart Editor, then click **Options, Y axis reference line**. **Close** the properties window and the Chart editor.

The plot is random (no discernable pattern), so the regression model is reasonable. While pollution may have caused some manatee deaths, the data are labeled as manatees killed by boats, so pollution would not explain more of these deaths.

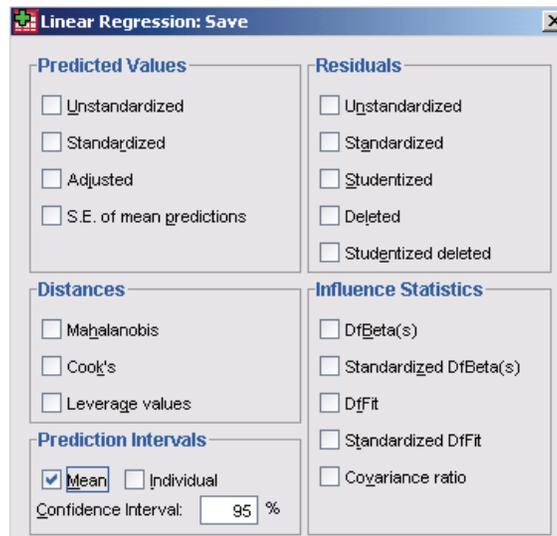


**23.33** We create a scatterplot of Dr. Gray's predictions against actual storms and compute the regression.

**Dr Gray's Atlantic Hurricane Forecasts**



There is a positive relationship seen in the graph; however, there are a couple of years in which he predicted a large number of storms and the actual number was much less. Part (b) asks for a 95% confidence interval for the mean number of storms when Dr. Gray predicts 16 storms. To do this, add a forecast value of 16 in the spreadsheet (SPSS will only create prediction and confidence intervals for values in the spreadsheet), then in the **Analyze, Regression, Linear** dialog box, click **Save** and check the box for **Mean Prediction Intervals**.



**Model Summary<sup>a</sup>**

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.529 <sup>a</sup>	.280	.247	4.086

a. Predictors: (Constant),

b. Dependent Variable:

**Coefficients<sup>a</sup>**

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	1.803	3.587		.503	.620
	Forecast	.903	.309	.529	2.923	.008

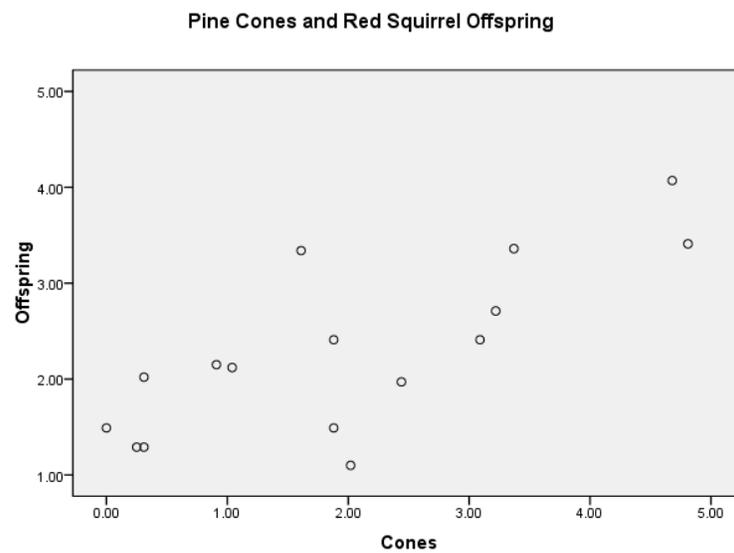
a. Dependent Variable:

The regression equation is  $\text{ActualStorms} = 1.80 + 0.90 \cdot \text{Predicted}$ . With a  $t$  statistic of 2.923 and (two-sided)  $P$ -value of 0.008 (so, the one-sided  $P$ -value is 0.004), the relationship is significantly positive. Return to the data spreadsheet. SPSS has created 95% confidence intervals for each value of **Forecast**. At the bottom, we see the values for the interval of interest.

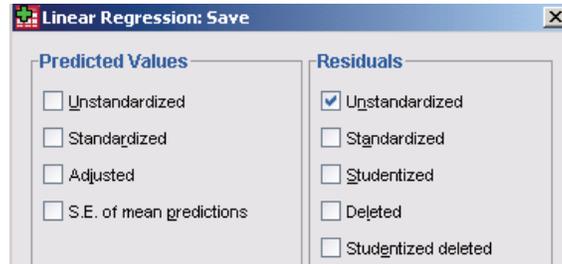
16	.	12.77479	19.72914
----	---	----------	----------

We predict the *mean* number of actual storms for years when Professor Gray predicts 16 will be between 12.78 and 19.73, with 95% confidence. If you wanted values for a particular year, you would have checked the box for **Individual** Prediction Intervals in the Save dialog box.

**23.41** We create a scatterplot of the **Cones** as the X axis variable and **Offspring** as the Y axis variable using **Graphs, Legacy Dialogs, Scatter/Dot**. The pattern is roughly linear (there is a fair amount of scatter) and increasing – more cones seem to be associated with more offspring.



Use **Analyze, Regression, Linear** to find linear regression and measures of association. We will want to examine the residuals for adequacy of the regression, so click **Save** and check the box for **Unstandardized** Residuals. You can also ask for a histogram of the standardized residuals – these are  $z$ -scores – (and a Normal plot of them) in the **Plots** box.



The regression equation is  $\text{Offspring} = 1.415 + 0.44 \cdot \text{Cones}$ . The relationship is fairly strong –  $r = 0.756$ ; the cone index explains  $r^2 = 57.2\%$  of the variation in offspring.

**Model Summary**

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.756 <sup>a</sup>	.572	.542	.60031

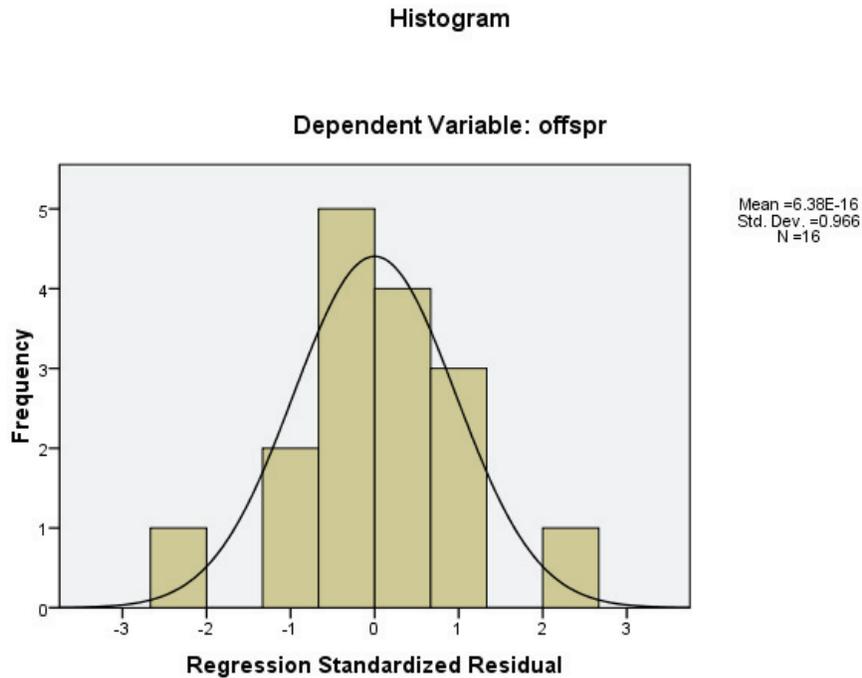
a. Predictors: (Constant), Cones

**Coefficients<sup>a</sup>**

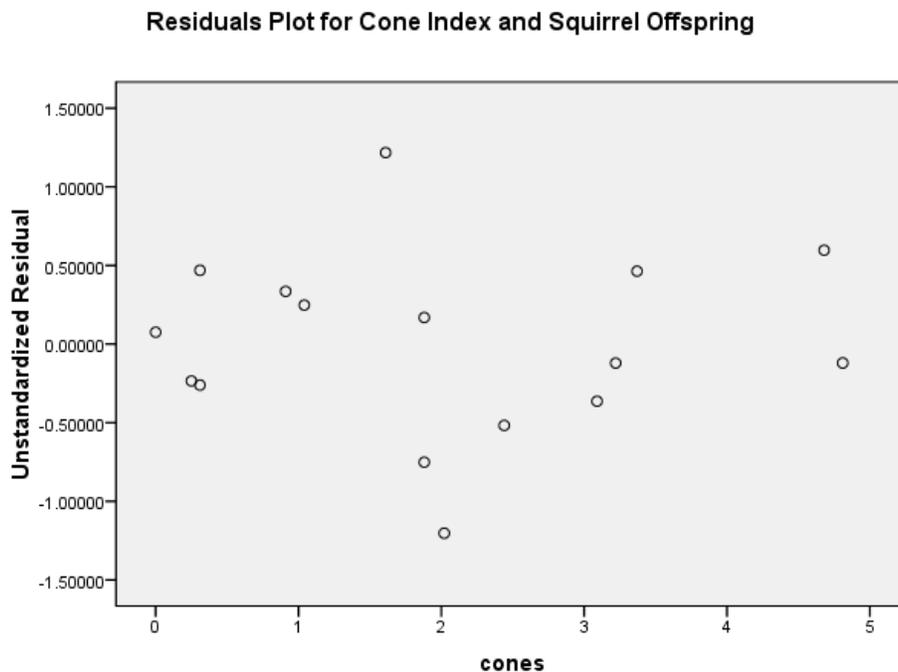
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	1.415	.252		5.619	.000
	Cones	.440	.102	.756	4.328	.001

a. Dependent Variable: Offspring

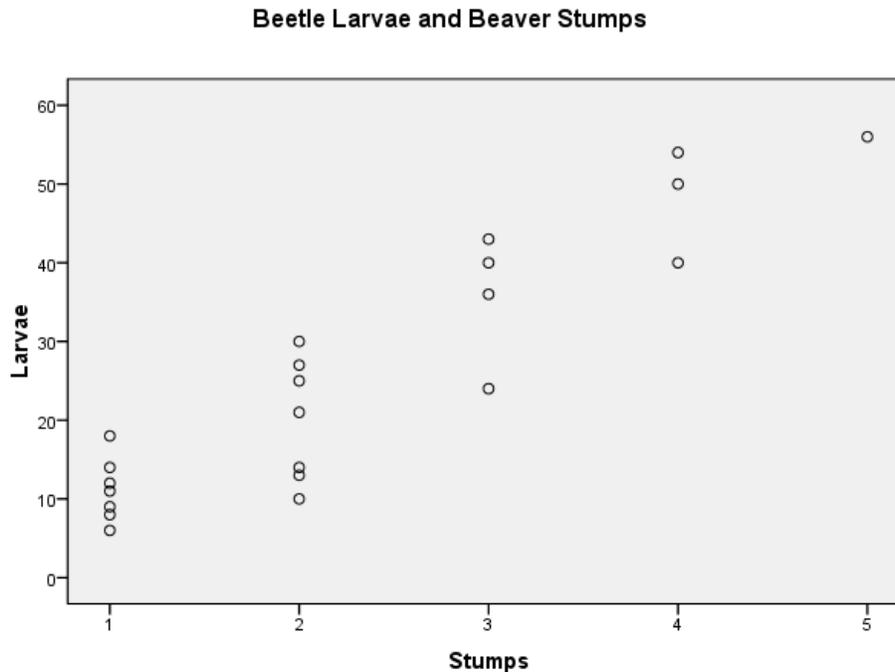
The relationship is indeed statistically significant; we have  $t = 4.328$  with (two-sided)  $P = 0.001$ , so the one-sided  $P$ -value is 0.0005.



There are some gaps in the histogram, but with the imposed density curve, the Normal assumption seems reasonable. Note the mean of these is (essentially) 0. Create a scatterplot of the saved residuals against the cone index using **Graphs, Legacy Dialogs, Scatter/Dot**. This plot shows no definite pattern, so our inference is reliable.



**23.43** Open data file *ex05-51*. First, create a scatterplot of the data using **Graphs**, **Legacy Dialogs**, **Scatter/Dot**. Enter **Stumps** as the X variable and **Larvae** as the Y variable. Give your graph an appropriate title using **Titles**.



We see that these data indicate that there are more beetle larvae with more stumps. Use **Analyze**, **Regression**, **Linear** to fit the line using **Stumps** as the Independent and **Larvae** as the Dependent. We'd like a 95% confidence interval for the slope (how many more clusters accompany each additional stump), so click **Statistics**, and check the box for Confidence Intervals.

**Model Summary**

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.918 <sup>a</sup>	.843	.835	6.455

a. Predictors: (Constant), Stumps

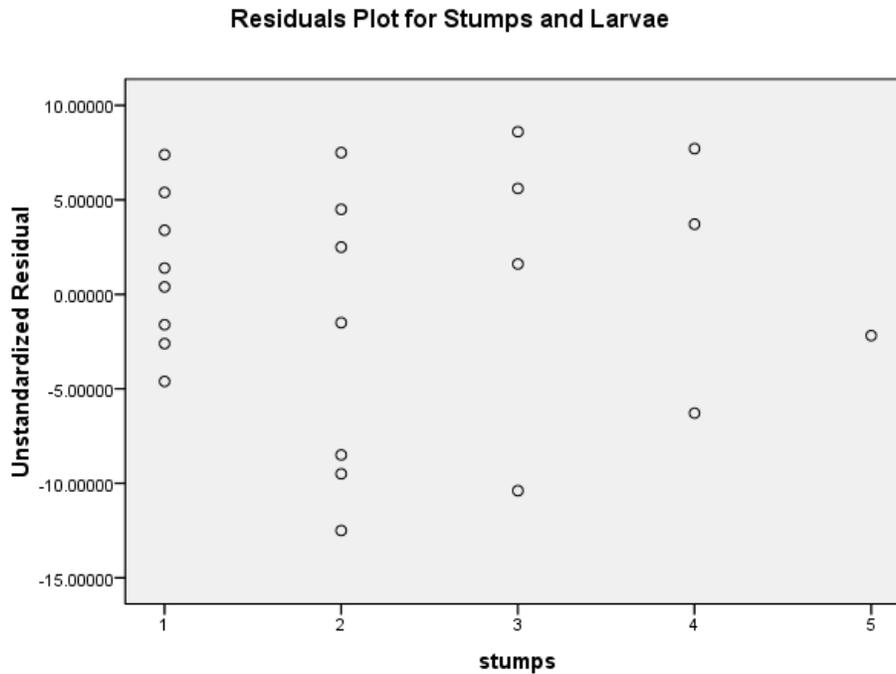
**Coefficients<sup>a</sup>**

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	-1.286	2.853		-.451	.657	-7.220	4.647
	Stumps	11.894	1.136	.916	10.467	.000	9.531	14.257

a. Dependent Variable:

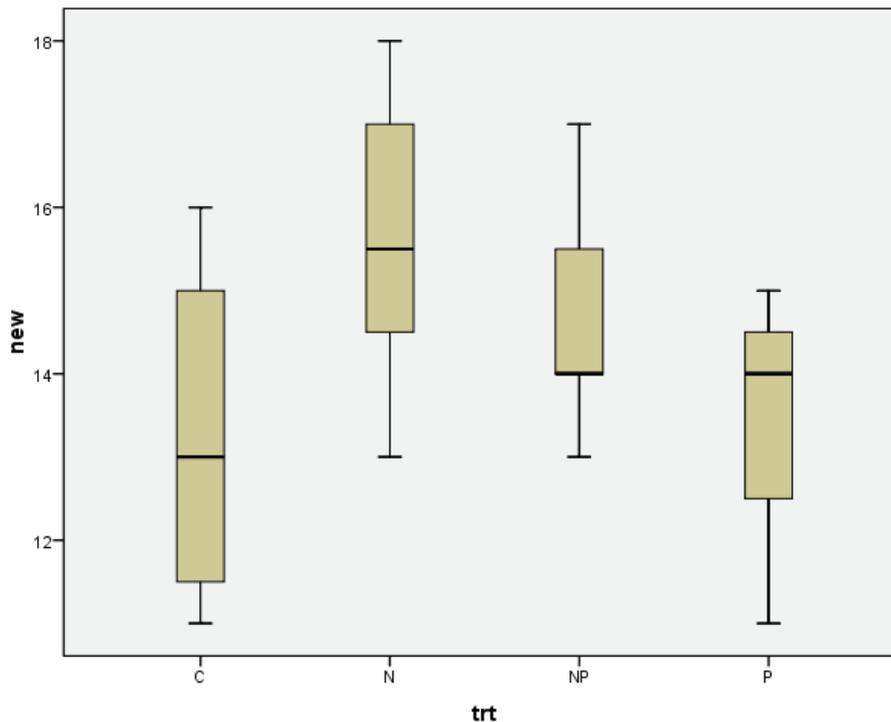
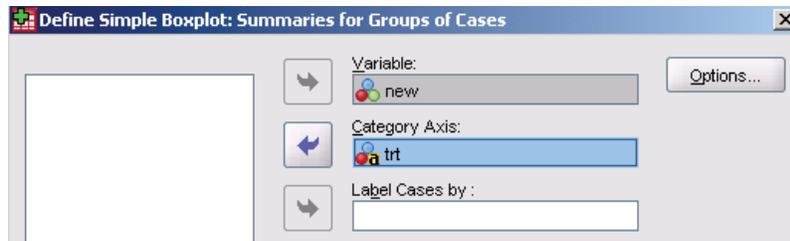
The regression equation is  $Larvae = -1.29 + 11.89 * Stumps$ . The relationship is strong; the regression model explains 84.3% of the variability in larvae (the correlation is  $r = \sqrt{.843} = 0.918$ ). We are confident that more stumps lead to more larvae because the 95% confidence for the slope is between 9.53 and 14.26 which is well above 0.

Our scatterplot of the residuals against Stumps (the predictor variable) indicates no discernable pattern, so this regression model is reasonable.



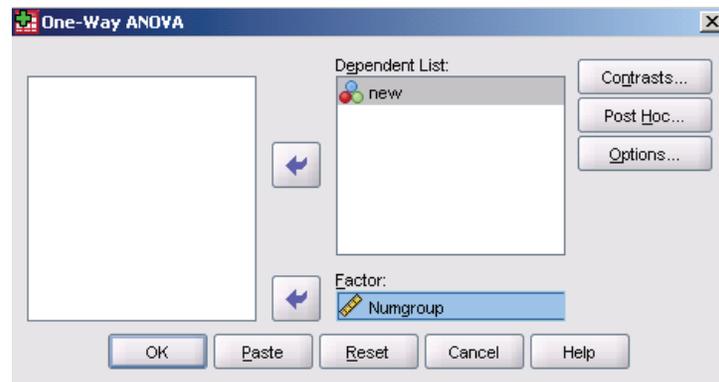
## Chapter 24 SPSS Solutions

**24.9** We'd like to know which treatment of nutrients is most effective in promoting new leaf growth. We have  $H_0: \mu_N = \mu_P = \mu_B = \mu_{NONE}$  (there is no difference) and  $H_a$ : at least one is different from the rest. To examine the data, we'll create side-by-side boxplots using **Graphs, Legacy Dialogs, Boxplot** with summaries for groups of cases.



None of the distributions has outliers, but they are not all symmetric, either. The control group has a very short lower whisker and the “both” group (NP) has the median equal to  $Q_1$ . It appears that Nitrogen has the highest median number of new leaves.

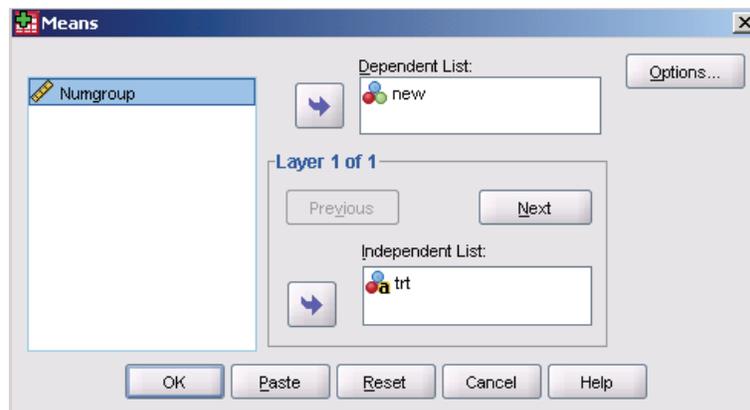
We'll perform the ANOVA using **Analyze, Compare Means, One-Way ANOVA**, but this requires that the factor values (the treatment labels) be integer-valued. We've created a new variable (called Numgroup) with values 0 for Control, through 3 for the both Nitrogen and Phosphorus group.



### ANOVA

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	27.209	3	9.070	3.440	.031
Within Groups	71.179	27	2.636		
Total	98.387	30			

With  $F = 3.44$  and  $P$ -value 0.031, we conclude that the means are not all the same. To see the actual group means and standard deviations, you can use **Analyze, Compare Means, Means**.



### Report

	Mean	N	Std. Deviation
C	13.29	7	2.059
N	15.62	8	1.685
NP	14.62	8	1.302
P	13.50	8	1.414
Total	14.29	31	1.811

Judging from our original plot and these results, Nitrogen only is the superior treatment for new leaves. All the others seem fairly close to one another (certainly the Control and Phosphorus only are similar).

**24.13** With only summary statistics, we'll have to "manually" calculate the ANOVA. First, note that the standard deviations are close enough (between 4.2 and 5.2) to satisfy the rule of thumb. The overall mean response is

$$\bar{\bar{x}} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2 + n_3\bar{x}_3}{n_1 + n_2 + n_3} = \frac{37*10.2 + 36*9.3 + 42*10.2}{37 + 36 + 42} = 9.918.$$

The mean square for groups is

$$\begin{aligned} MSG &= \frac{n_1(\bar{x}_1 - \bar{\bar{x}})^2 + n_2(\bar{x}_2 - \bar{\bar{x}})^2 + n_3(\bar{x}_3 - \bar{\bar{x}})^2}{k-1} \\ &= \frac{37(10.2 - 9.918)^2 + 36(9.3 - 9.918)^2 + 42(10.2 - 9.918)^2}{3-1} = 10.0158. \end{aligned}$$

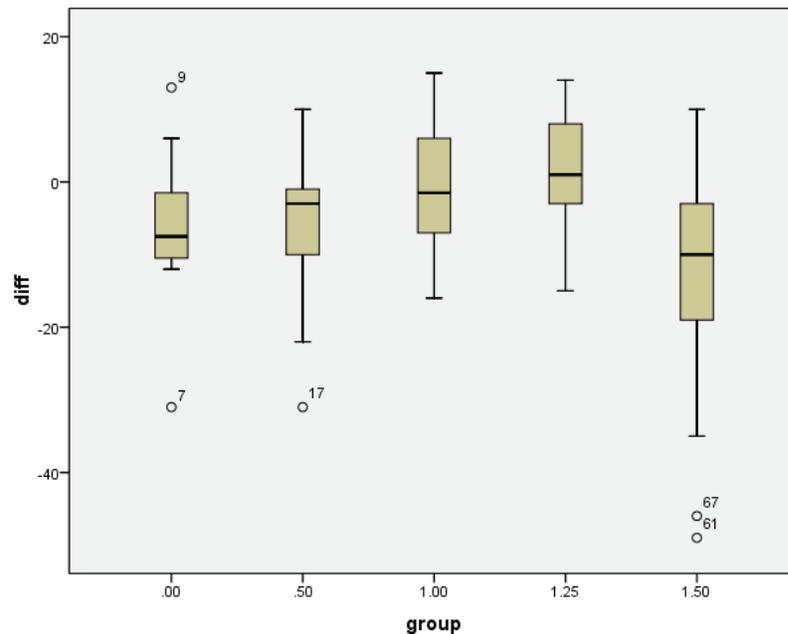
The mean square for error is

$$\begin{aligned} MSE &= \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + (n_3 - 1)s_3^2}{N - k} \\ &= \frac{(37 - 1)*4.2^2 + (36 - 1)*4.5^2 + (42 - 1)*5.2^2}{(37 + 36 + 42) - 3} = 21.897. \end{aligned}$$

The  $F$  statistic is  $F = MSG/MSE = 10.0158/21.897 = 0.457$ . This has 2 and 112 degrees of freedom. We find the  $P$ -value using **CDF.F**. With  $P$ -value 0.6344; there is insufficient evidence to show a difference in weight loss with the different exercise programs.

Compute Variable		P
Target Variable:	Numeric Expression:	0.6344
P	= 1-CDF.F(.457,2,112)	

**24.33** We first create side-by-side boxplots of the differences, examining them for outliers and any indications that the data are not Normal.



Groups 0 and 0.5 and 1.5 have outliers. Group 1.25 seems to have a smaller spread than the others. We'll proceed, but use caution. Based on the plots, the median differences are all negative (meaning a slower healing rate than natural). We test  $H_0$ : all treatments have the same mean against  $H_a$ : at least one is different using **Analyze, Compare Means, One-Way ANOVA**, but this requires that the factor values (the treatment labels) be integer-valued. We've created a new variable (called Numgroup) with values 0 for Control, through 4 for the 1.5 group. We follow up with **Analyze, Compare Means, Means**.

## ANOVA

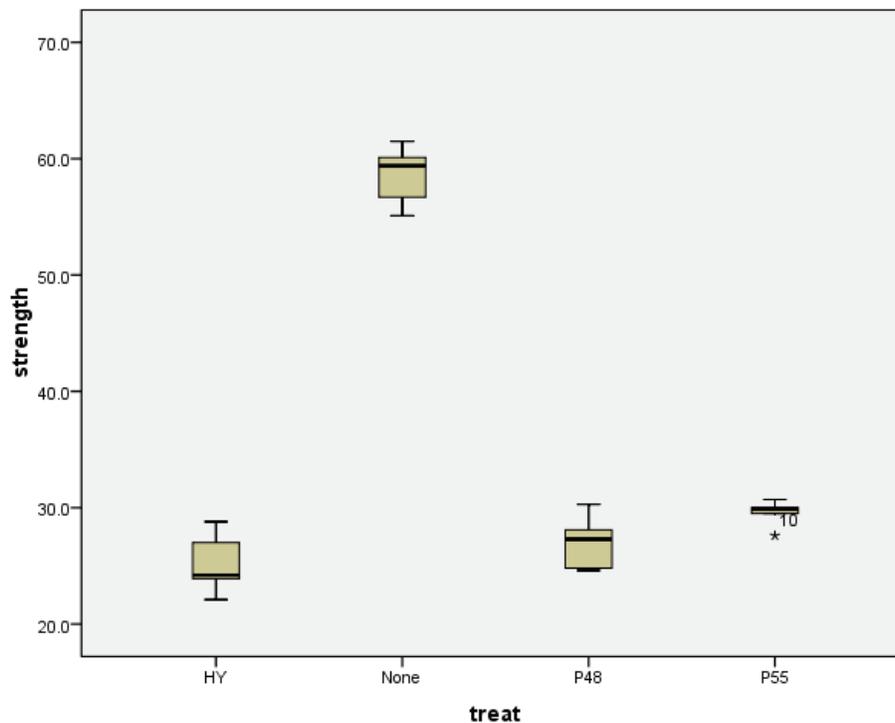
	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	2232.147	4	558.037	4.041	.005
Within Groups	9528.407	69	138.093		
Total	11760.554	73			

## Report

	Mean	N	Std. Deviation
0	-6.42	12	10.706
0.5	-5.71	14	10.564
1	-.17	18	9.345
1.25	1.47	15	8.863
1.5	-13.80	15	17.387
Total	-4.66	74	12.693

With  $F = 4.04$  and  $P$ -value 0.005, we conclude that the means are not all the same. The conjecture is that nature (group 1) heals best and that changing the field slows healing. The mean difference in Group 1 is  $-0.17$ . This small difference indicates little difference between the control and “experimental” legs. The mean difference in group 0 is  $-6.42$ ; group 0.5 has mean  $-5.71$ ; group 1.25 has mean  $1.47$ ; group 1.5 has the largest mean difference:  $-13.80$ . The 1.25 times natural field is close to natural; it appears that all the others slow healing. Further, looking at the intervals, it is clear that the 1.5 group has slower healing than either the “natural” (1.0) or 1.25 groups.

**24.35** The problem states (and it is clear from the data) that the untreated fabric is strongest. Our question refers to the three treatments – which of these will result in the strongest fabric? With only five observations per treatment, any graph should not really be trusted. We do, however, show side-by-side boxplots of the data.



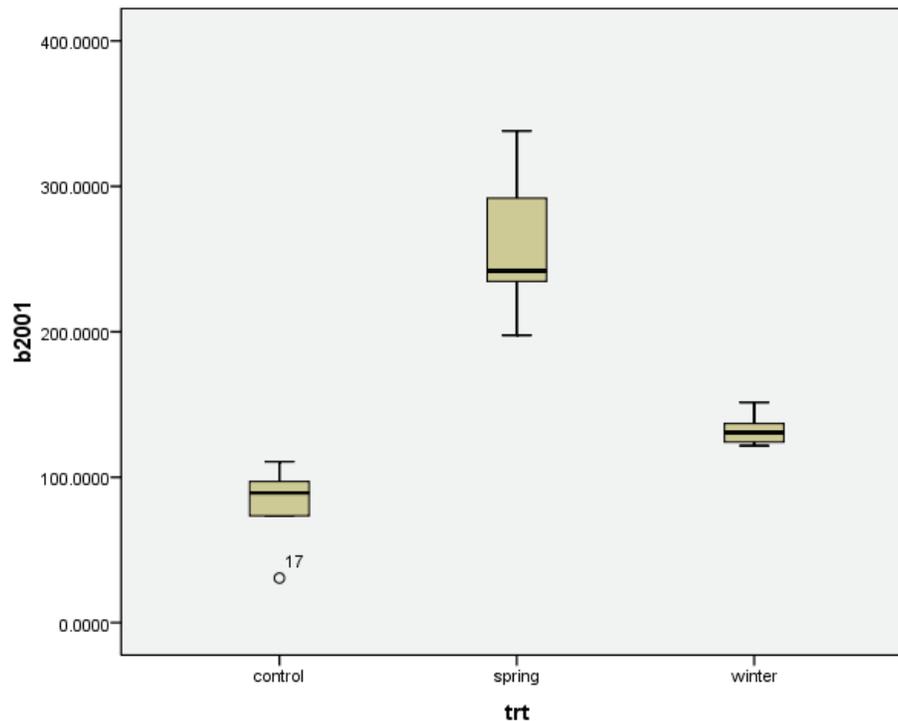
Based on these, it appears that Permafresh 55 gives the strongest fabric, and is least variable. We use ANOVA to test  $H_0$ : all treatments have the same mean against  $H_a$ : at least one is different. We first a numeric group variable for the three treatments (omitting the no treatment group), as our question really revolves around the three treatments. Use the mouse to highlight the cells representing the “None” treatment and delete them.

ANOVA					
	Sum of Squares	df	Mean Square	F	Sig.

Between Groups	47.497	2	23.749	5.023	.026
Within Groups	56.740	12	4.728		
Total	104.237	14			

With  $F = 5.023$  and  $P$ -value 0.026, we reject  $H_0$ , and conclude there is a difference in mean breaking strength among the three treatments. If we had to make a recommendation, we'd recommend Permafresh 55.

**24.39** We first display side-by-side boxplots of the 2001 data in *ta24-06*.



It appears that added water in Spring results in the most biomass; this distribution is the most variable, however. Symmetry (Normal distributions) of the data is also questionable. Our sample sizes are very small, however. We'll again need to create a numeric grouping variable before computing the ANOVA. We also follow up using **Analyze, Compare Means, Means** to check the group means and standard deviations.

#### ANOVA

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	98450.521	2	49225.261	43.787	.000
Within Groups	16863.088	15	1124.206		
Total	115313.610	17			

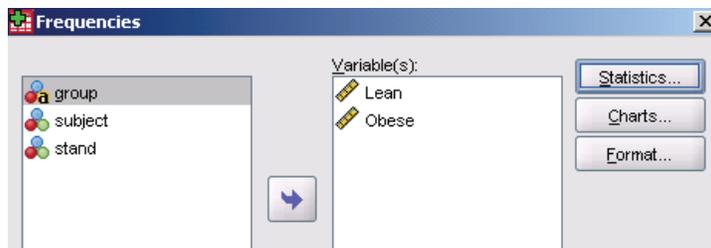
### Report

	Mean	N	Std. Deviation
control	81.669630	6	28.0673133
spring	257.686267	6	49.5874144
winter	132.580267	6	11.2219402
Total	157.312054	18	82.3599023

Computing the ANOVA, we have  $F = 43.787$  with  $P = 0.000$ . These results bear out our intuition from the plots. Additional water in the Spring dry season will result in more biomass. We see that the standard deviation for winter) is 11.22 and for spring, the standard deviation is 49.59. This difference is considerably more than our rule-of-thumb (the largest being no more than twice the smallest) – an indication of a potential problem with this analysis.

## Chapter 25 SPSS Solutions

**25.1** Having opened data file *ex25-01.por*, the easiest way to find the median for each group is to use copy and paste to move each group's data (in variable **stand**) to new columns (we named them **Lean** and **Obese**). You can then use **Analyze, Descriptive Statistics, Frequencies**. Click to enter the variables, then click **Statistics**, and check the box for the Median (you can also ask for quartiles).



		Lean	Obese
N	Valid	10	10
	Missing	10	10
Median		549.5220	388.8845
Percentiles	25	472.2328	327.3672
	50	549.5220	388.8845
	75	590.4642	418.9872

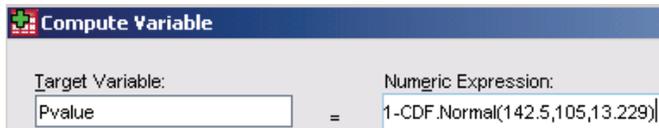
Based on these statistics, it appears that lean subjects are much more active than obese subjects. Use **Data, Sort Cases** to sort the data in ascending order by **stand**. Assign ranks to each observation from 1 to 20.

	group	subject	stand	Rank
1	Obese	11	260.244	1
2	Obese	18	267.344	2
3	Lean	3	319.212	3
4	Obese	15	347.375	4
5	Obese	17	358.650	5
6	Obese	13	367.138	6
7	Lean	9	374.831	7
8	Obese	19	410.631	8

Summing the lean ranks gives  $W = 142$ . If the null hypothesis (no difference between the two groups) is true, we should have  $\mu_W = 10 * 21 / 2 = 105$  and

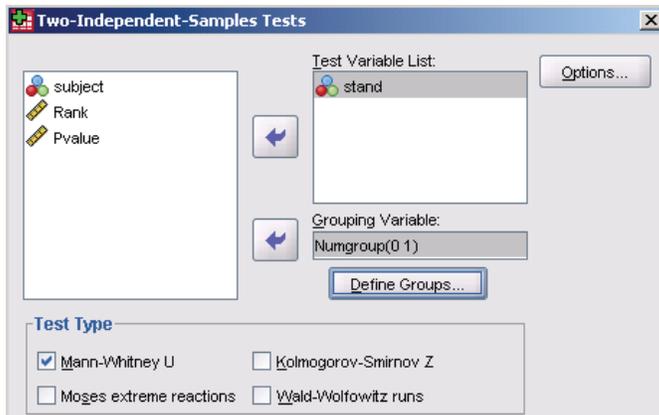
$\sigma_w = \sqrt{10 * 10 * 21 / 12} = 13.229$ . Our statistic is almost three standard deviations above this mean.

**25.3** The probability expression, using the continuity correction, will be  $P(W > 142.5)$ . We can use **CDF.Normal** to find the  $P$ -value for this statistic as 0.0023. This  $P$ -value supports our intuition that lean subjects are more active.



Pvalue
0.0023

**25.7** To compute the rank sum test “automatically,” use **Analyze, Nonparametric Tests, 2 Independent Samples**. As usual, SPSS wants an integer valued grouping variable, so we’ve created **Numgroup** with values 0 = Obese and 1 = Lean. The Mann-Whitney U is equivalent to the Wilcoxon test.



Mann-Whitney U	13.000
Wilcoxon W	68.000
Z	-2.797
Asymp. Sig. (2-tailed)	.005
Exact Sig. [2*(1-tailed Sig.)]	.004 <sup>a</sup>

a. Not corrected for ties.

b. Grouping Variable: Numgroup

The asymptotic one-sided (Normal)  $P$ -value given is  $0.005/2 = 0.0025$ .

**25.11** Open data set *ex18\_08.por*. The null hypothesis is that the two distributions are the same (additional time doesn’t result in significant decay). The alternate is that one of the distributions is systematically higher. For a two-sample  $t$  test the hypotheses are  $H_0 : \mu_2 = \mu_{16}$  and  $H_a : \mu_2 > \mu_{16}$ . The pair tied at 110 receives ranks of 2.5 (the average of 2 and 3); the pair tied at 126 receive ranks of 7.5. Click **Analyze, Nonparametric Tests, 2 Independent Samples**. Note that you must click on **Define Groups** and specify the values.

	N	Mean Rank	Sum of Ranks
2	5	6.60	33.00
16	5	4.40	22.00
Total	10		

Mann-Whitney U	7.000
Wilcoxon W	22.000
Z	-1.156
Asymp. Sig. (2-tailed)	.248
Exact Sig. [2*(1-tailed Sig.)]	.310 <sup>a</sup>

a. Not corrected for ties.

b. Grouping Variable:

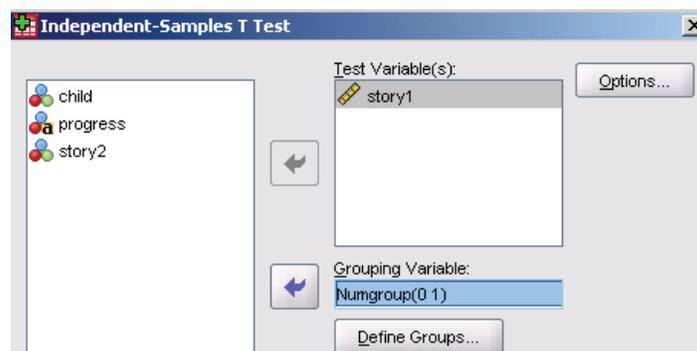
The observed sum of ranks for the two week data is  $W = 33$ . The exact  $P$ -value for the one-sided alternate is  $0.310/2 = 0.155$ ; this is actually a bit smaller than the  $P = 0.1857$  obtained with the two-sample  $t$  test. These data do not indicate substantial decay in polyester between 2 and 16 weeks.

**25.13** Most statistical packages won't create a back-to-back stemplot. SPSS is not different (it will create separate stemplots for each level of Progress using **Analyze, Descriptive Statistics, Explore**). We've created our own below. We only have five observations per group, but it certainly appears that there is a low outlier in the low progress group.

Stem-and-leaf of Story1  
Leaf Unit = 0.010

Low	High
0	0
	1
	2
6	3
0	4
5	5 57
	6
2	7 02
8	4

Before doing the two sample test using **Analyze, Compare Means, Independent Samples T Test**, we've created variable **Numgroup** with values 1 = High and 0 = Low.



For a two-sample  $t$  test, the hypotheses are  $H_0 : \mu_{High} = \mu_{Low}$  and  $H_a : \mu_{High} > \mu_{Low}$ . As can be seen below, the mean for the high progress group was 0.676 while the mean for the low progress group was 0.406. The test statistic is  $t = 2.06$  with (one-sided)  $P$ -value 0.0445. Our conclusion is that the mean story telling ability of high progress children is greater than that for low progress children.

		t-test for Equality of Means						
		t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
							Lower	Upper
story1	Equal variances assumed	-2.062	8	.073	-.27000	.13093	-.57192	.03192
	Equal variances not assumed	-2.062	5.520	.089	-.27000	.13093	-.59724	.05724

We now use **Analyze, Nonparametric Tests, 2 Independent Samples** to test  $H_0$ : the two distributions are the same. The (one-sided)  $P$ -value is similar (a bit larger) and we reach the same conclusion at the 0.05 level – namely, high progress children are better able to retell the story read to them without pictures.

Ranks				
	Numgroup	N	Mean Rank	Sum of Ranks
story1	0	5	3.80	19.00
	1	5	7.20	36.00
	Total	10		

Test Statistics <sup>b</sup>	
	story1
Mann-Whitney U	4.000
Wilcoxon W	19.000
Z	-1.786
Asymp. Sig. (2-tailed)	.074
Exact Sig. [2*(1-tailed Sig.)]	.095 <sup>a</sup>

a. Not corrected for ties.

b. Grouping Variable: Numgroup

**25.15** We ask whether fertilizing with dead cicadas increases the seed mass, so we would have hypotheses

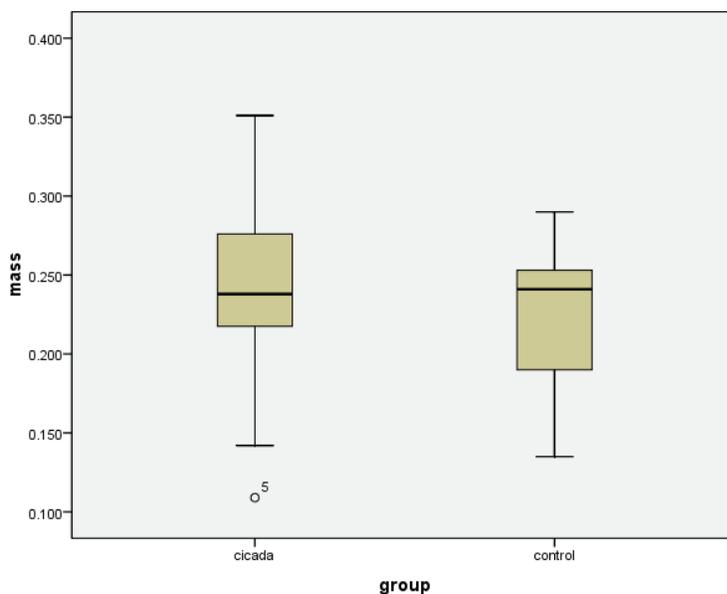
$$H_0 : \mu_{Cicada} = \mu_{Control}$$

and

$$H_a : \mu_{Cicada} > \mu_{Control}$$

for the two-sample  $t$  test. We've used

**Graphs, Legacy Dialogs, Boxplot** to create a graph for



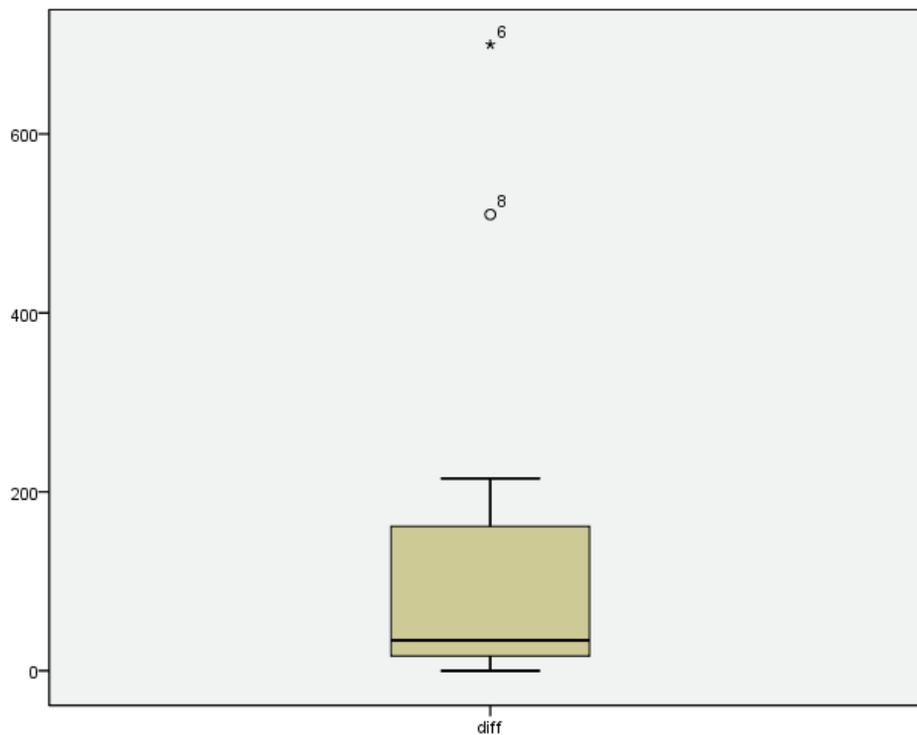
**Summaries of Groups.** These indicate both distributions are left-skewed and there is a low outlier in the cicada fertilized group. Testing with the Wilcoxon (Mann-Whitney) test changes the hypotheses to  $H_0$ : there is no difference in the distributions and  $H_a$ : yields are higher in the cicada fertilized group. As always, create a numeric grouping variable (we coded the cicada fertilized plants as 0) before using **Analyze, Nonparametric Tests, 2 Independent Samples**. We find  $W = 1567$  for the fertilized group. With  $P$ -value 0.0525, there is insufficient evidence to reject  $H_0$ . These data do *not* show (at the 0.05 level) that fertilizing with dead cicadas will increase the seed mass.

Numgroup	N	Mean Rank	Sum of Ranks
0	39	40.18	1567.00
1	33	32.15	1061.00
Total	72		

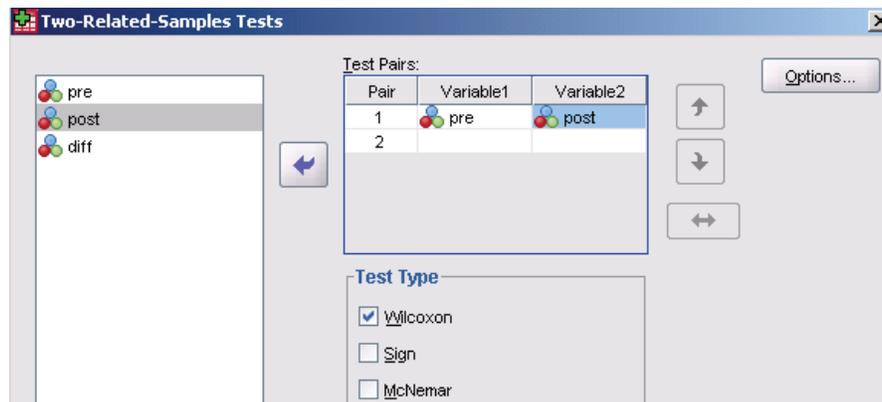
Mann-Whitney U	500.000
Wilcoxon W	1061.000
Z	-1.622
Asymp. Sig. (2-tailed)	.105

a. Grouping Variable: Numgroup

**25.19** We create a boxplot of variable **diff**. This distribution is clearly right skewed with two high outliers. With only  $n = 11$  observations,  $t$  procedures are not valid.



We use **Analyze, Nonparametric Tests, 2 Related Samples** to perform the test using the **pre** and **post** values as the variables.



	N	Mean Rank	Sum of Ranks
- Negative Ranks	0 <sup>a</sup>	.00	.00
Positive Ranks	10 <sup>b</sup>	5.50	55.00
Ties	1 <sup>c</sup>		
Total	11		

	-
Z	-2.803 <sup>a</sup>
Asymp. Sig. (2-tailed)	.005

a. Based on negative ranks.

b. Wilcoxon Signed Ranks Test

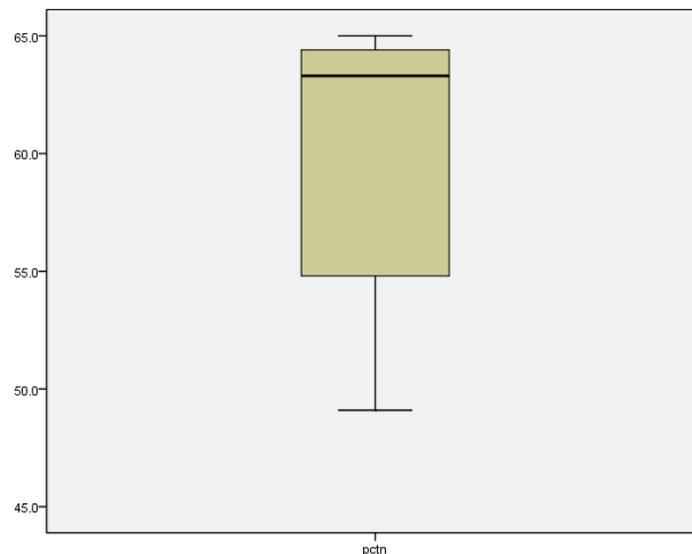
a. <

b. >

c. =

The sum of positive ranks is  $W^+ = 55$ . With  $P$ -value  $0.005/2 = 0.0025$ , we conclude the infusions definitely had a beneficial effect in raising immune response levels. The mean of  $W^+$  would be  $\mu = 11(12)/4 = 33$ , and the standard deviation is  $\sigma = \sqrt{11(12)(23)/24} = 11.247$ .

**25.23** Use **Graphs**, **Legacy Dialogs**, **Boxplot** to graph the distribution of **pctn**. There are no outliers, but the distribution is definitely left skewed.



Add a second variable (we called it **constant**) with values 78.1. To perform the test, click **Analyze, Nonparametric Tests, 2 Related Samples**. Click to enter the two variables and **OK**.

Ranks			
	N	Mean Rank	Sum of Ranks
constant - Negative Ranks	0 <sup>a</sup>	.00	.00
Positive Ranks	9 <sup>b</sup>	5.00	45.00
Ties	0 <sup>c</sup>		
Total	9		

a. constant <

b. constant >

c. constant =

Test Statistics <sup>b</sup>	
	constant -
Z	-2.666 <sup>a</sup>
Asymp. Sig. (2-tailed)	.008

a. Based on negative ranks.

b. Wilcoxon Signed Ranks Test

The sum of negative ranks for differences ( $78.1 - data$ ) is 0 (all the observations were less than 78.1). The  $P$ -value for the test is 0.008; we have overwhelming evidence that there was less nitrogen in the ancient air than there is now.

**27.21** Since the data are already differences, we have hypotheses  $H_0 : median = 0$  and  $H_a : median > 0$ , since if the cola loses sweetness the before rating should be higher than the rating after storage. Open data file *ex27\_21.por*. Enter another variable (we called it **zero**) with all 0's. To perform the test, click **Analyze, Nonparametric Tests, 2 Related Samples**. Click to enter the two variables and **OK**.

Ranks			
	N	Mean Rank	Sum of Ranks
zero - Negative Ranks	8 <sup>a</sup>	5.94	47.50
Positive Ranks	2 <sup>b</sup>	3.75	7.50
Ties	0 <sup>c</sup>		
Total	10		

a. zero <

b. zero >

c. zero =

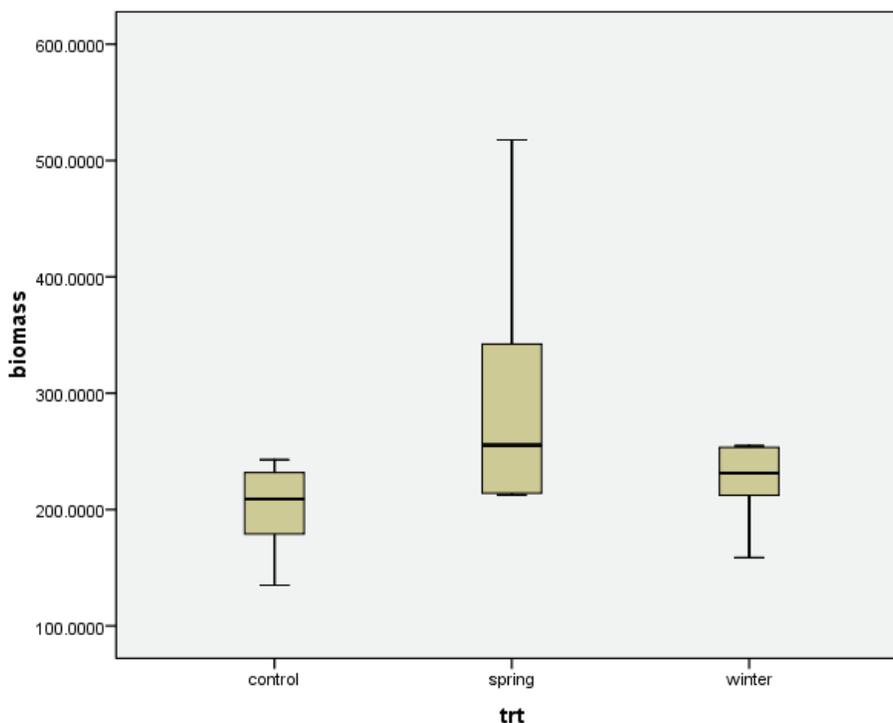
Test Statistics <sup>b</sup>	
	zero -
Z	-2.041 <sup>a</sup>
Asymp. Sig. (2-tailed)	.041

a. Based on positive ranks.

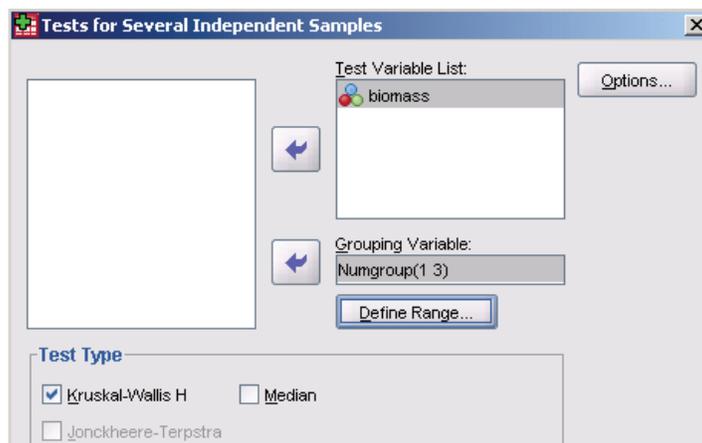
b. Wilcoxon Signed Ranks Test

With a one-sided  $P$ -value of 0.0205, we conclude that the cola does lose sweetness with storage.

**25.27** It is clear just looking at the data that the Spring values have a high (possible) outlier (517.665); Our side-by-side boxplots show that all the distributions are skewed (two to the left and one to the right). Note that  $1.5 \cdot \text{IQR}$  criterion doesn't quite work for these small data sets. It clearly appears that additional water in the Spring is helpful; additional water in the winter doesn't seem to make a difference over the control. ANOVA tests  $H_0$ : all groups have the same mean. The Kruskal-Wallis null hypothesis is that all populations have the same distribution. These data have  $I = 3$  groups, with  $n_i = 6$  observations per group, and  $N = 18$  total observations.



Use Analyze, Nonparametric Tests, K Independent Samples to compute the test, after creating a numeric grouping variable (we used 1 = winter through 3 = control).



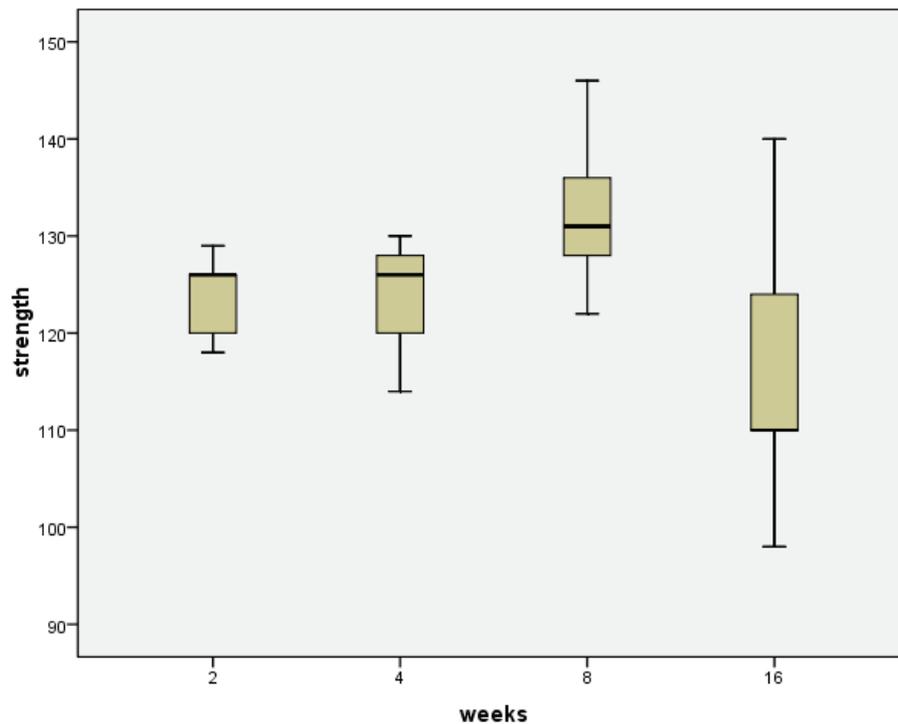
Numgroup	N	Mean Rank
1	6	9.33
2	6	13.00
3	6	6.17
Total	18	

Chi-Square	4.924
df	2
Asymp. Sig.	.085

a. Kruskal Wallis Test  
b. Grouping Variable: Numgroup

The test statistic is  $H = 4.924$  with  $P$ -value 0.085. At the 5% level, these data do not show a difference in biomass with different watering programs. The test statistic has  $3 - 1 = 2$  degrees of freedom.

**25.29** Our boxplots (unreliable for these small sample sizes) indicate the variability in the four samples is very different; in addition, the distributions are not symmetric.



Since the variable **weeks** is already numeric, use **Analyze, Nonparametric Tests, K Independent Samples** to compute the test. The null hypothesis is that all the data sets

have the same distribution; the alternative is that they are not the same. With a  $P$ -value of 0.131, these data do not show a significant change in strength (decay).

	N	Mean Rank
2	5	9.70
4	5	10.20
8	5	15.40
16	5	6.70
Total	20	

Chi-Square	5.631
df	3
Asymp. Sig.	.131

a. Kruskal Wallis Test

b. Grouping Variable:

**25.45** Enter the data into variables **species** and **group**, where **group** = 1 is the unlogged. SPSS cannot make back-to-back stemplots, but a sketch of one is below. It appears that the species counts for unlogged plots are higher than for logged plots. The hypotheses for the test are  $H_0$ : species count distributions are the same versus  $H_A$ : one of the distributions is systematically higher.

	0	4
	0	
333	1	024
99855	1	55788
2210	2	
Unlogged		Logged

To perform this test, use Analyze, Nonparametric Tests, 2 Independent Samples. Click to enter the variable **species**, then **group** as the grouping variable. Define Groups and enter 1 and 2 for the groups. Continue and OK to compute the test. We find the  $P$ -value of the test is  $0.053/2 = 0.0265$  and conclude that there is more species diversity in unlogged areas of Borneo than in logged areas.

group	N	Mean Rank	Sum of Ranks
species 1	12	13.25	159.00
2	9	8.00	72.00
Total	21		

	species
Mann-Whitney U	27.000
Wilcoxon W	72.000
Z	-1.931
Asymp. Sig. (2-tailed)	.053
Exact Sig. [2*(1-tailed Sig.)]	.058 <sup>a</sup>

a. Not corrected for ties.

b. Grouping Variable: group

**25.51** We want to compare the downstream locations to the upstream locations for each tributary. This is paired data. To perform this test, use **Analyze, Nonparametric Tests, 2 Related Samples**. Click to enter the variable names (**up** and **down**), and **OK**. With a (one-sided) *P*-value of 0.032, we conclude that there are more species downstream than upstream (The sum of those ranks was 62.5 with expected sum 39).

	N	Mean Rank	Sum of Ranks
- Negative Ranks	2 <sup>a</sup>	7.75	15.50
Positive Ranks	10 <sup>b</sup>	6.25	62.50
Ties	1 <sup>c</sup>		
Total	13		

a. <

b. >

c. =

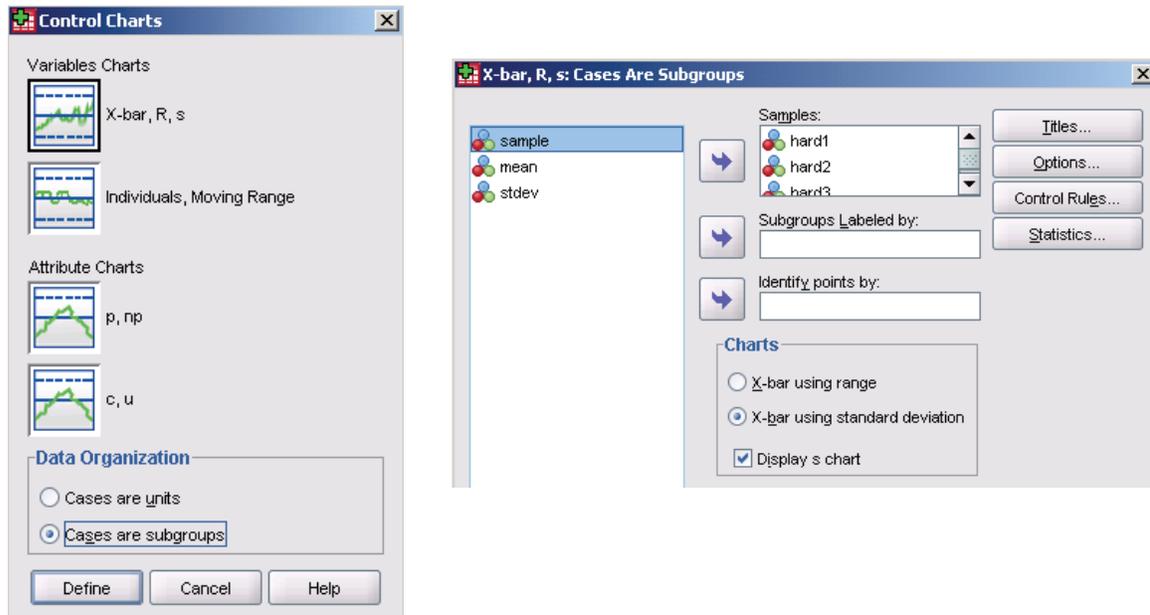
	-
Z	-1.850 <sup>a</sup>
Asymp. Sig. (2-tailed)	.064

a. Based on negative ranks.

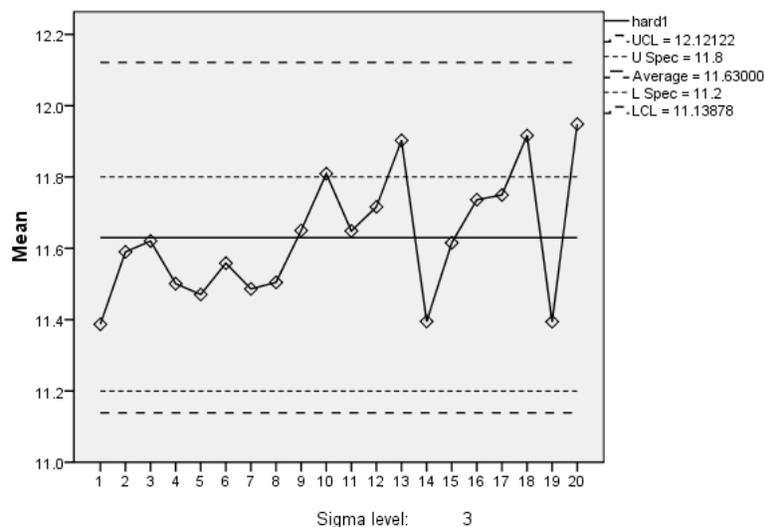
b. Wilcoxon Signed Ranks Test

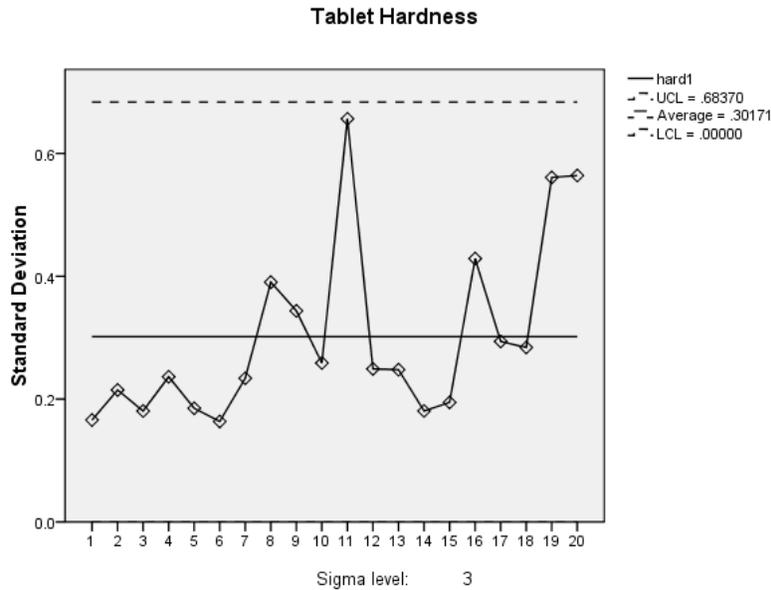
## Chapter 26 SPSS Solutions

**26.13** To make the charts, use **Analyze, Quality Control, Control Charts**. Each observation (row) of the data spreadsheet represents a sample, so we use **Cases are Subgroups**. We've entered the actual data from each sample, and chosen to construct the chart using the standard deviation. The SPSS default is to make charts based on historic data. We have a specified target mean and standard deviation, so we compute the limits as  $\mu \pm 3\sigma/\sqrt{n} = 11.5 \pm 3 * .2/\sqrt{4}$ , or 11.2 to 11.8. Click **Statistics** to enter these limits. Click to give your charts **Titles**, then **OK**.



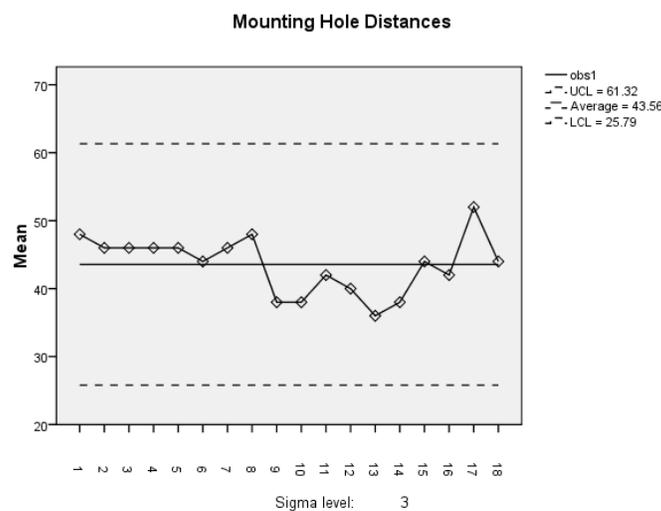
Tablet Hardness

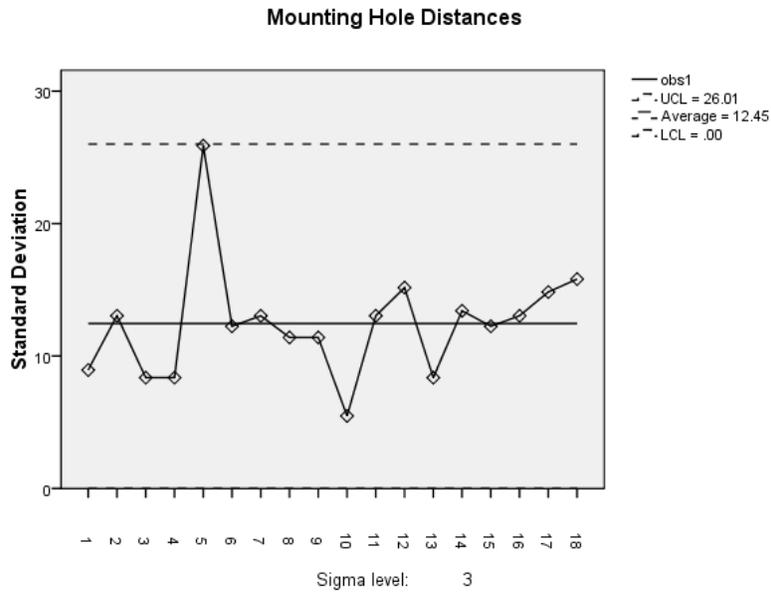




The process standard deviation changed about sample number 8. After that point, essentially all samples had standard deviations above the center line; we also see a large variation in sample means after that point. Both the standard deviation and mean changing are reflected in means mostly above the center line (and several above the UCL). Both charts are signaling an out-of-control situation.

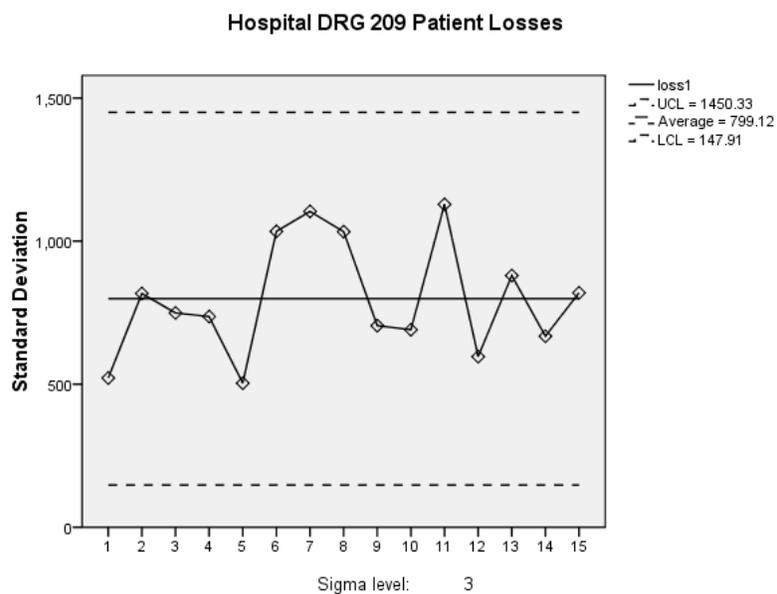
**26.15** Open data file *ex26\_15.por*. This data file already has means and standard deviations for each sample. Click **Analyze, Quality Control, Control Charts**. Since our data have one row for each sample, move the button to **Cases are subgroups** on the initial dialog box. Click **Define** to proceed. Click to enter the five observations in variables **obs1** through **obs5** as the samples and that subgroups are identified by **sample**. We also want the chart using the standard deviation. Make sure the **Display s chart** box is checked.

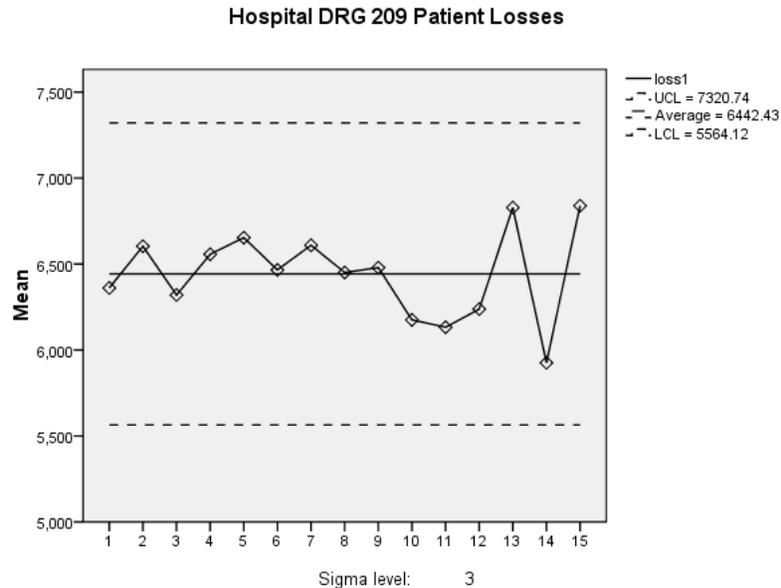




Based on these plots, the mean mounting distance is well in control. Sample 5 had a standard deviation above the upper limit; after that, this also seems in control.

**26.21** The SPSS default is to make control charts based on past data. Use **Analyze, Quality Control, Control Charts** with **Cases are subgroups**. Click to enter the eight loss values as the samples, and to create the chart using standard deviation. Give your charts **Titles**, then **OK**. The *s* chart is clearly in control.





The  $\bar{x}$  chart is also in control, although we are seeing some increased variability in process mean at the end of the time period.

**26.27** The old specifications were 100 to 400 mV. Use **CDF.Normal** to find the percent that met those is 99.94%. We then find out what percent will meet the new specifications of between 150 and 350 mV. In both calculations the mean is 275 mV and the standard deviation is 38.4 mV. About 97.4% of all monitors should meet the new specifications.

Compute Variable		Meets
Target Variable:	Numeric Expression:	0.9994
Meets	$\text{CDF.Normal}(400,275,38.4) - \text{CDF.Normal}(100,275,38.4)$	

Compute Variable		Meets
Target Variable:	Numeric Expression:	0.9740
Meets	$\text{CDF.Normal}(350,275,38.4) - \text{CDF.Normal}(150,275,38.4)$	

**26.29** The natural tolerances are  $\bar{x} \pm 3s$ , where  $s$  is estimated from all the remaining data. We can find  $\bar{x}$  and  $s$  by copying the data for all five observations per sample (after deleting the Sample 5 data) into a new variable (we called it **New**) and then using **Analyze, Descriptive Statistics, Descriptives**. The new (overall) mean is 43.412 and  $s = 11.583$ .

Descriptive Statistics

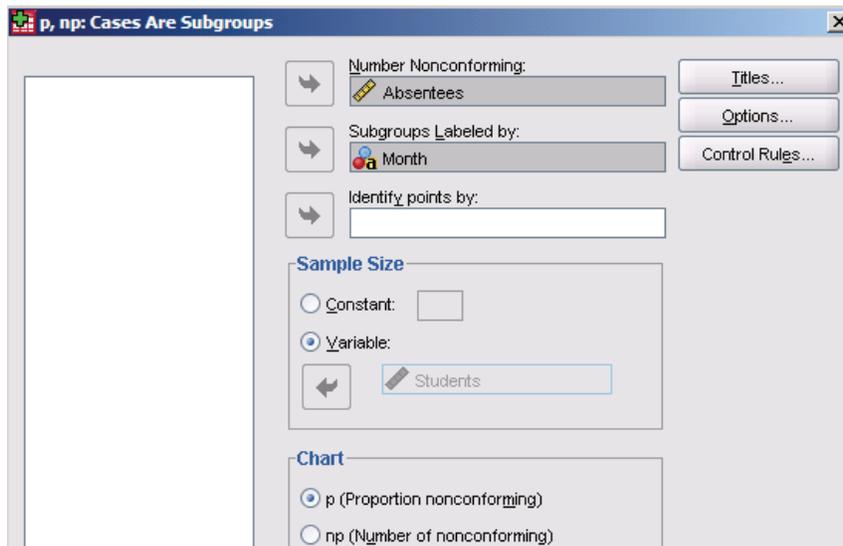
	N	Minimum	Maximum	Mean	Std. Deviation
New	85	24.00	74.00	43.4118	11.58334
Valid N (listwise)	85				

The natural tolerances then become  $43.412 \pm 3 * 11.583$  or 8.663 to 78.161.

### 26.35 Enter the Months, Student numbers, and Absentees.

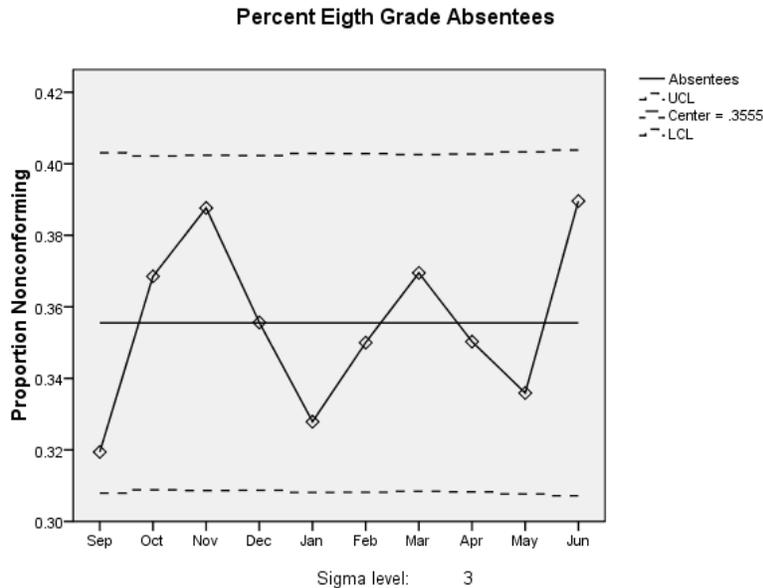
Month	Students	Absentees
Sep	911.00	291.00
Oct	947.00	349.00
Nov	939.00	364.00
Dec	942.00	335.00
Jan	918.00	301.00
Feb	920.00	322.00
Mar	931.00	344.00
Apr	925.00	324.00
May	902.00	303.00
Jun	883.00	344.00

Click **Analyze, Quality Control, Control Charts**. We want a p,np chart where data are organized as Cases are subgroups. Click **Define**.



The **Number Nonconforming** are the **Absentees**, the **Subgroups** are the **Months**, and total **Sample Size** is **Students** (this variable does not show well in the screen capture)

above; it immediately is dimmed by SPSS). Notice that we can select for either the proportion nonconforming, or the number nonconforming. Give your plot a **Title** and click **OK**.

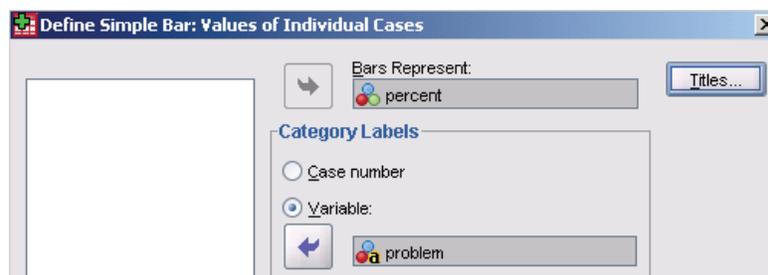


We find  $\bar{p}$  is 0.3555. Since SPSS calculates these proportions based on the total number of students each month, the upper and lower control limits change somewhat. To find  $\bar{n}$ , use **Analyze, Descriptive Statistics, Descriptives**. The average total enrolled students for this school for this year is 921.8.

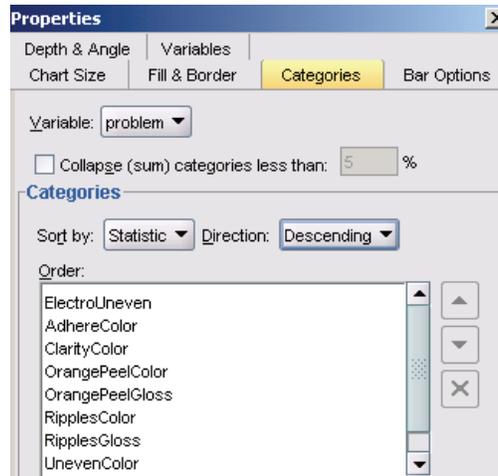
**Descriptive Statistics**

	N	Minimum	Maximum	Mean	Std. Deviation
Students	10	883.00	947.00	921.8000	19.62312
Valid N (listwise)	10				

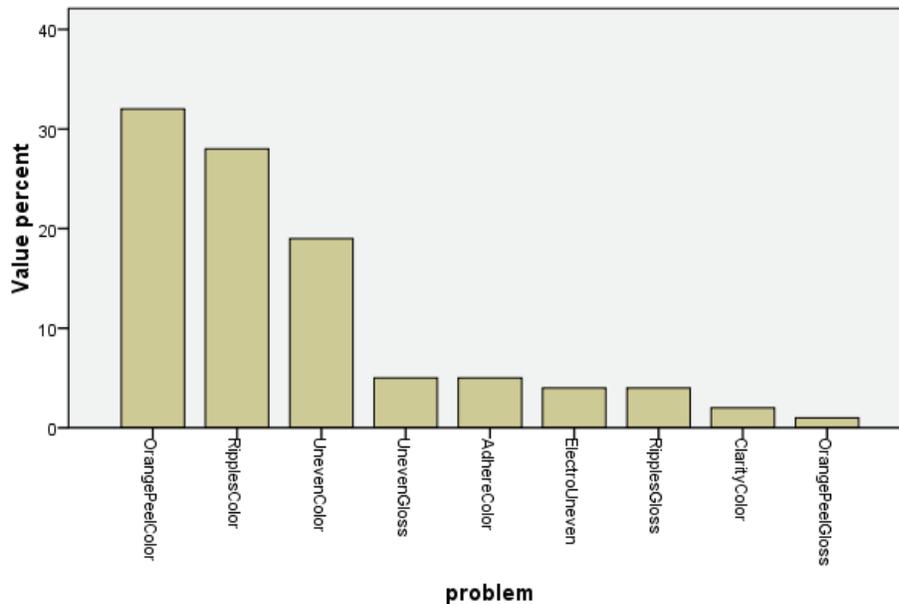
**26.43** To make the Pareto Chart, we start with a Bar Chart, then we'll order the bars in descending order using the Chart Editor. For our Bar Chart, the data are **Values of individual cases**. Be sure to give your graph **Titles**.



The initial chart has the bars in the order of their entry in the data spreadsheet. Double-click in the Chart for the Chart Editor, then double-click in a bar for the Properties Window. Change Categories to be sorted by **Statistic** in **Descending** order. **Apply** the change and **Close** the Properties box and Chart Editor.



**Major Causes of Auto Paint Defects**

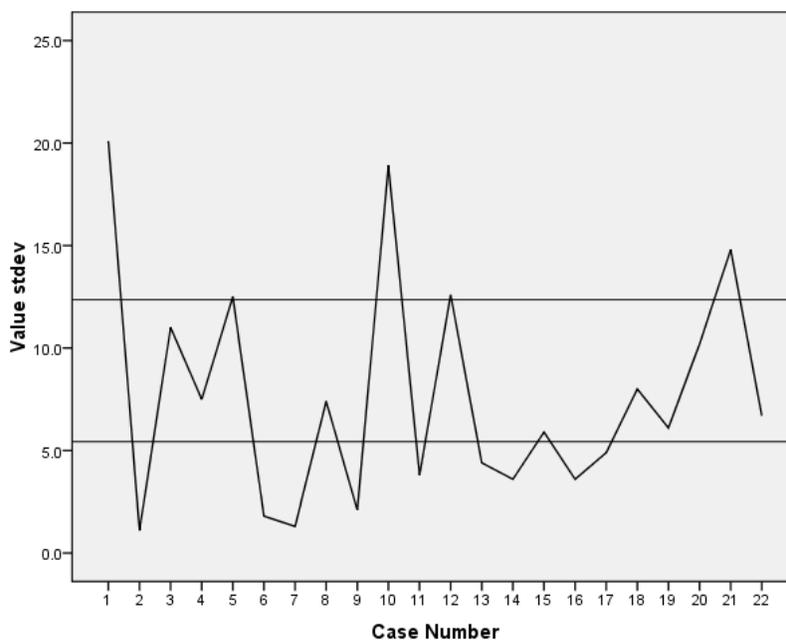


Clearly, from the chart, the largest percent of problems are due to “orange peel” texture in the color. All three of the largest frequency problems have to do with the color coat – this is where they should look first.

**26.51** The data file has only the mean and standard deviation of each sample. We must compute the control limits as  $B_5\hat{\sigma}$ ,  $\bar{s} = \hat{\sigma}$ , and  $B_6\hat{\sigma}$ . To find  $\bar{s}$ , use **Analyze, Descriptive Statistics, Descriptives** for the variable **stdev**. From Table 26.3, there is no  $B_5$  for a sample of size 3, so the lower control limit will be 0. From the output below,  $\bar{s} = 5.4293$ , and the upper limit becomes  $2.276 * 5.4293 = 12.357$ .

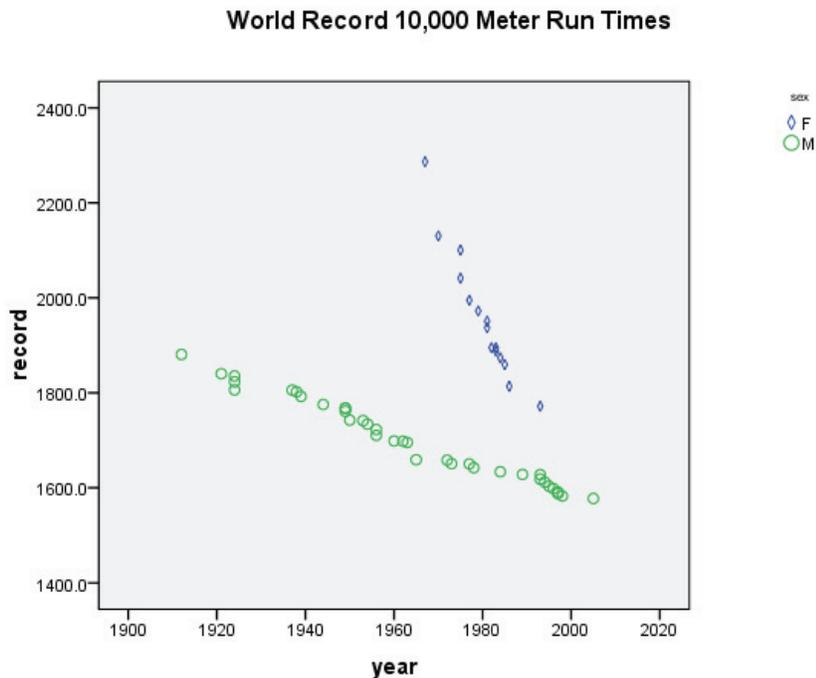
	N	Minimum	Maximum	Mean	Std. Deviation
Valid N (listwise)	22	1.1	20.1	7.650	5.4293

Click **Graphs, Legacy Dialogs, Line**. We want a simple Line chart where data in the chart are **Values of individual cases**. For the first chart, use **stdev** as the variable. You can use case number (sample number) as the category labels. To add the control limits, double-click the graph to bring up the Chart Editor, then click **Options, Y axis reference line**. Add the upper control limit at 12.357. Repeat this process to add the process centerline at 5.4293. Variation in the process fluctuated greatly at first; it seems to be settling down.



## Chapter 27 SPSS Solutions

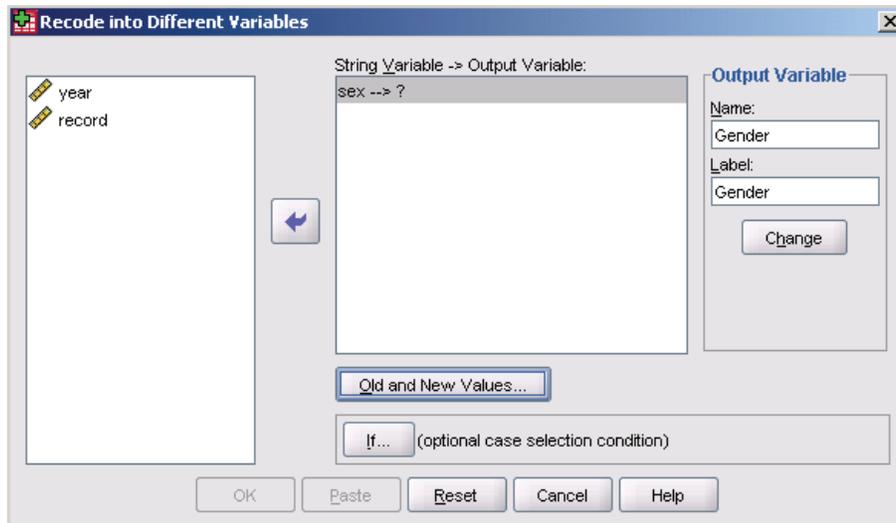
**27.15** To make the scatterplot, use **Graphs, Legacy Dialogs, Scatter/Dot**. The Y Axis variable is **record**, the X axis variable is **year** and we use **sex** as the variable to set markers by. Be sure to give your graph **Titles**.



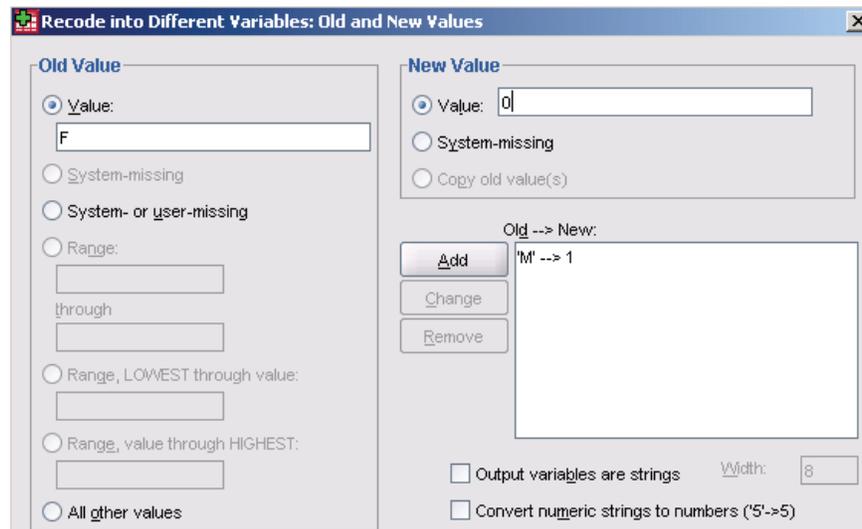
The SPSS default is to use different colors for the two variables. We have used the Chart Editor and changed the symbol type (click on the legend at the right of the graph) so that the symbols will look different if printed in black-and-white. Both genders show the decrease in record times, but women have had a steeper rate of descent.

To fit the model with two regression lines, we first need to create a numeric variable for gender. We've used 1 = men and 0 = women. You can enter the values yourself, or use **Transform, Recode into Different Variables**. We're changing the variable **sex** into

**Gender.** Once you've entered the new variable name and its label, click **Change**. Click **Old and New Values**.



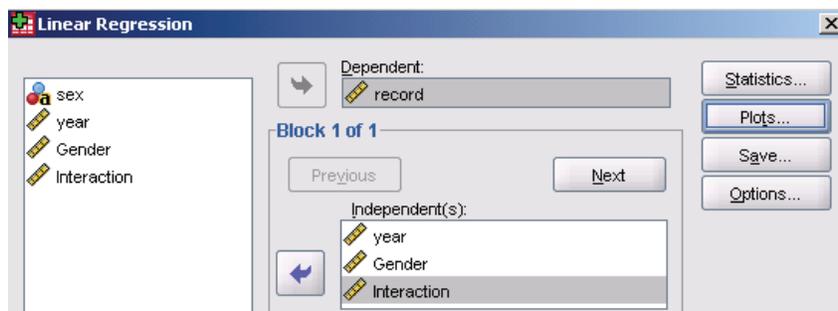
Type the old and new values in the boxes, and click **Add**. When finished, click **Continue**.



We also need to create a variable for the possible interaction (change in slope). We use **Transform, Compute Variable** to multiple **Gender** by **year** and save this as **Interaction**. So you can make sense of output, check the **Variable View** tab to make sure each variable has a label.



We're now ready to compute the regression (we think). Click **Analyze, Regression, Linear**. Click to enter **record** as the dependent variable and **year, Gender, and Interaction** as the Independent variables. (You can plot residuals and get a histogram of them using the **Plots** dialog).



If you try the regression at this point, Gender (the offset in the intercept for females) will not enter into the equation. SPSS has rules about sensitivity to changes in coefficient values when new variables are “added” into the equation. The reason for the problem in this situation is that our years are so far from the origin. We'll fix this by subtracting 1900 from variable **year** (use **Transform, Compute Variable**) and then recomputing the variable **Interaction** as shown above. Use **Analyze, Regression, Linear** as shown above.

**Model Summary**

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.991 <sup>a</sup>	.983	.982	21.1788

a. Predictors: (Constant), Interaction, Year, Gender

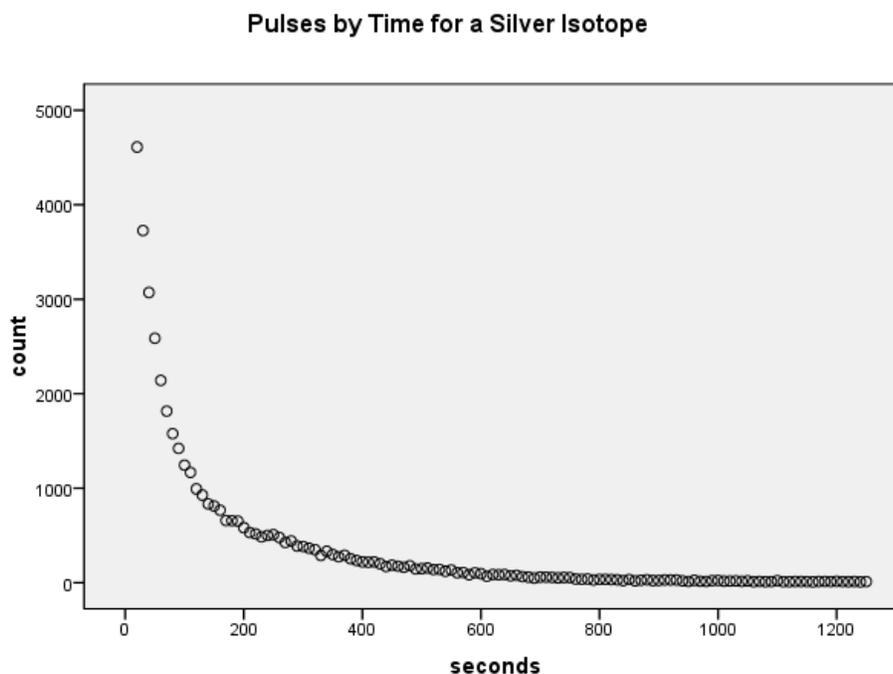
**Coefficients<sup>a</sup>**

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	3554.421	69.505		51.139	.000
	Year	-19.905	.865	-3.007	-23.001	.000
	Gender	-1644.710	70.124	-4.851	-23.454	.000
	Interaction	16.634	.876	3.880	18.987	.000

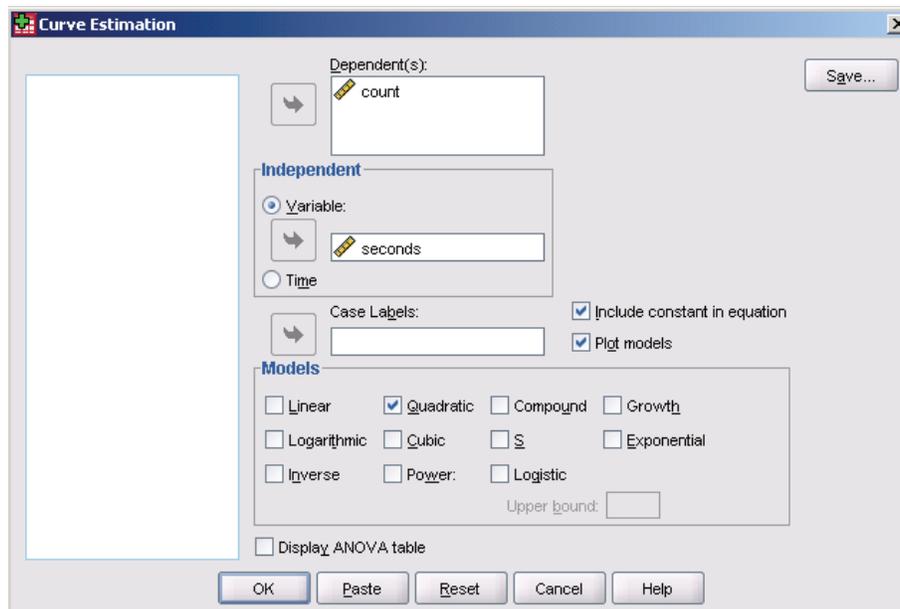
a. Dependent Variable: Record

$Time = 3554.421 - 19.905*Year - 1644.710*Gender + 16.634*Interaction$  is the final fitted regression equation. With our coding of 0 = female, that means their regression is  $Time = 3554.421 - 19.905*Year$ . For the males, the regression is  $Time = 1909.711 - 3.271*Year$ . The data certainly support that women's improvements have been faster than men's (the P-values for all coefficients are 0); they make no statement about sprints versus distance running (we have no information about this difference).

**27.19** We first create a scatterplot using **Graphs, Legacy Dialogs, Scatter/Dot**. There is little scatter, but definite curvature present. Clearly, a linear model will not adequately describe these data.



To fit a quadratic model, use **Analyze, Regression, Curve Estimation**. We've asked to fit a quadratic model by checking that box. The check in the **Plot models** box will add the fitted regression function to a scatterplot of the data.



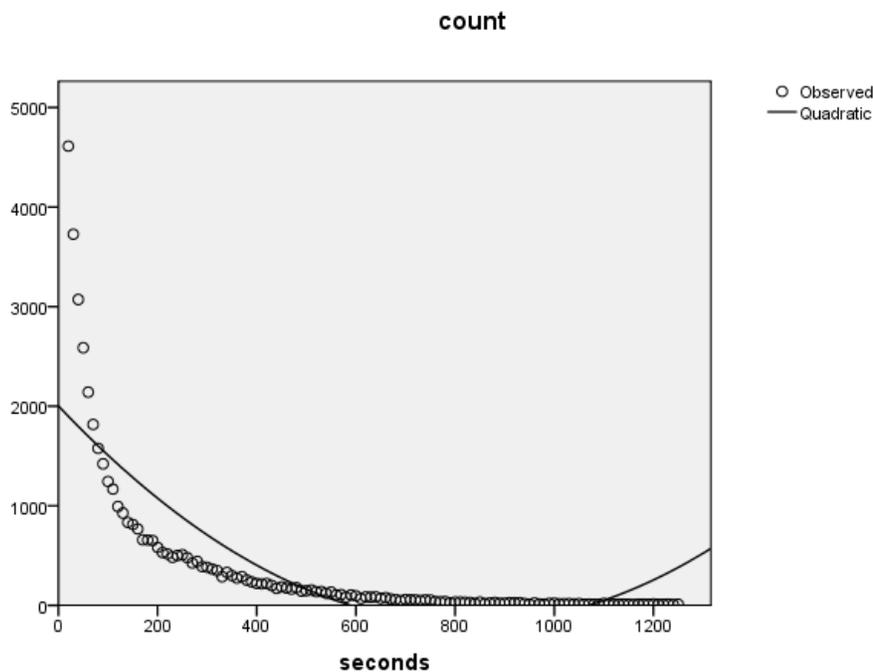
### Model Summary and Parameter Estimates

Dependent Variable:

Equation	Model Summary					Parameter Estimates		
	R Square	F	df1	df2	Sig.	Constant	b1	b2
Quadratic	.668	121.656	2	121	.000	2002.882	-5.273	.003

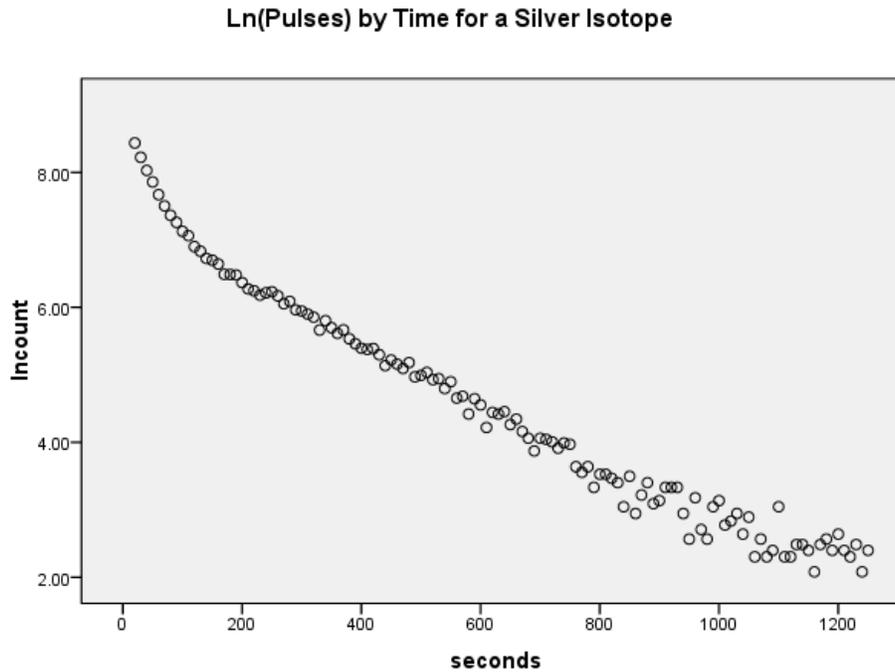
The independent variable is .

The regression equation is  $Counts = 2002.88 - 5.27*time + 0.003*time^2$ . The plot below indicates that this model does not fit the data well at all.



We use **Transform, Compute Variable** to find the natural logarithm of the counts., and then create a scatterplot of the new (transformed) data against **seconds**.

Compute Variable	
Target Variable:	Numeric Expression:
lncount	= LN(count)



This plot is clearly more linear, although there may be more scatter at the lower right end of the plot than at the upper end. There is also an indication of curvature at the upper right. We fit the new linear regression using **Analyze, Regression, Linear**. (You can add the linear regression into the original data plot using the Chart Editor and **Elements, Fit Line at Total**). If you click **Plots**, you can ask for a histogram of the residuals. Finally, click **Save** and check the box for **Unstandardized Residuals**.

**Model Summary<sup>b</sup>**

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.984 <sup>a</sup>	.967	.967	.29976

a. Predictors: (Constant),

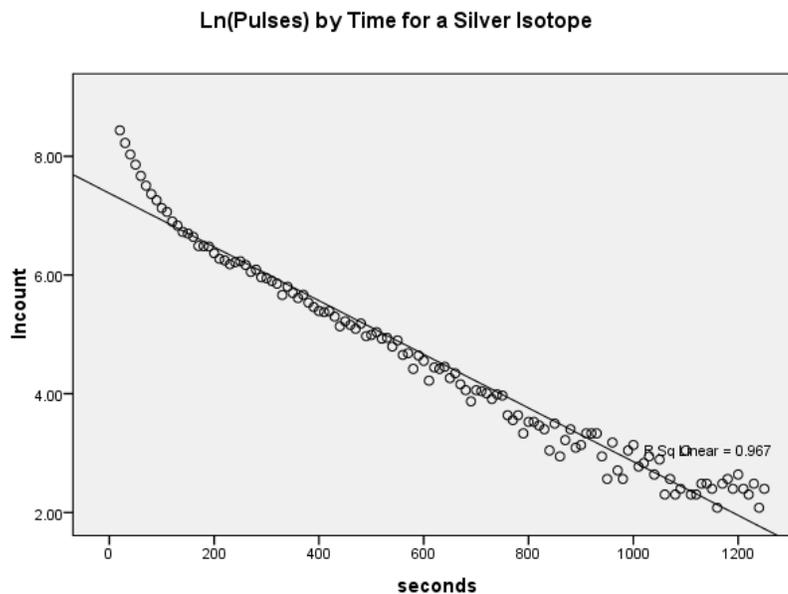
b. Dependent Variable: Incount

**Coefficients<sup>a</sup>**

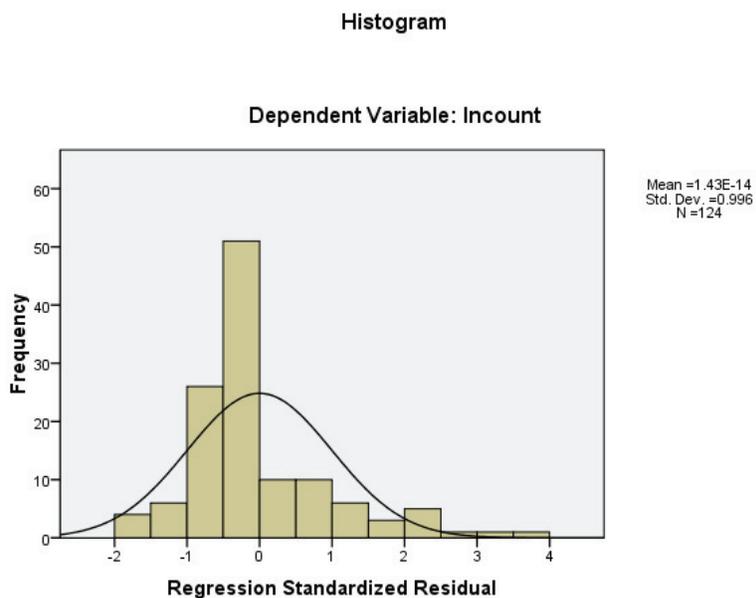
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	7.377	.055		134.564	.000
	seconds	-.005	.000	-.984	-60.117	.000

a. Dependent Variable: Incount

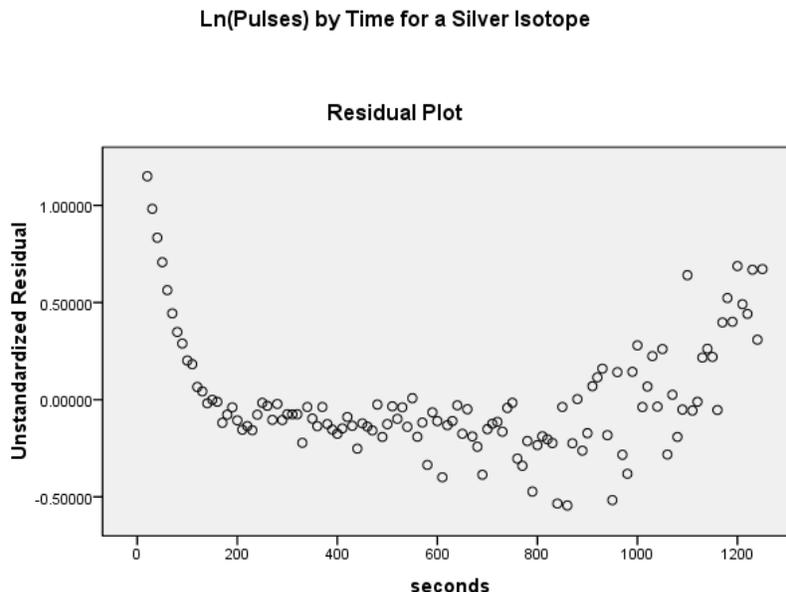
This regression fits much better – we have  $r^2 = 96.7\%$  as opposed to  $66.8\%$  for the quadratic model. The equation is  $\ln(\text{count}) = 7.377 - 0.005 * \text{seconds}$ . Let's examine the plots for diagnostic purposes. First, we added the regression equation into the data plot.



This seems to magnify a curve at the upper left end. The histogram of the residuals indicates they are skewed right (not Normal).



Finally, we create a Scatterplot of the Residuals against **seconds** (the predictor variable). The curvature in the data plot is magnified in the residuals scatterplot. Although this model fits better, it is still not really “correct.”



**27.23** To fit the quadratic model (and make a plot of the regression function with the data), use **Analyze, Regression, Curve Estimation**. (For details on this dialog box, see Exercise 27.19 above).

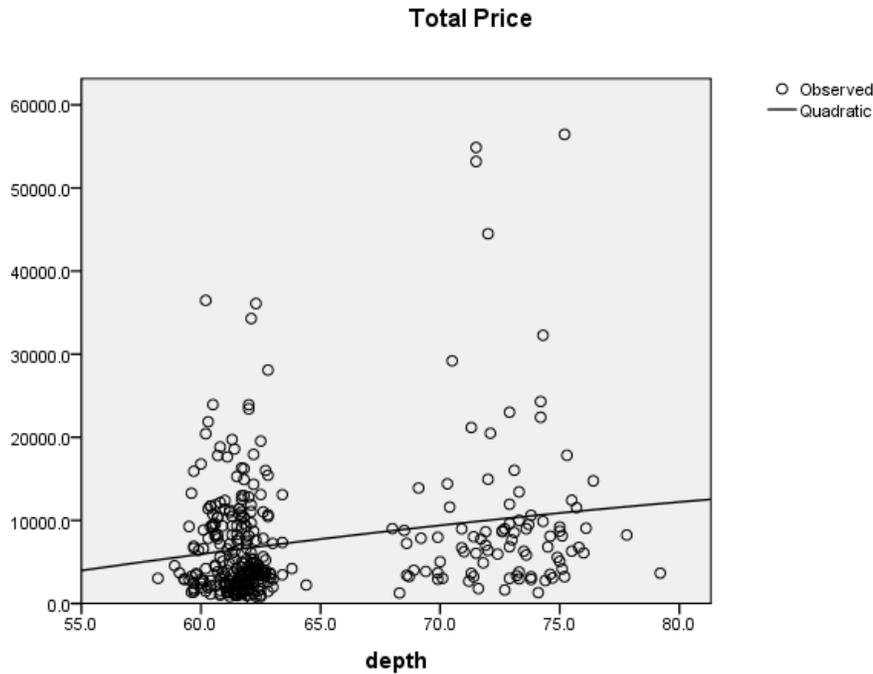
#### Model Summary and Parameter Estimates

Dependent Variable: Total Price

Equation	Model Summary					Parameter Estimates		
	R Square	F	df1	df2	Sig.	Constant	b1	b2
Quadratic	.047	8.673	2	348	.000	-28406.783	766.369	-3.233

The independent variable is .

The quadratic equation is  $Price = -28406.783 + 766.369 * Depth - 3.233 * Depth^2$ . This relationship is clearly not as useful as the one in Example 27.15 – that had  $r^2 = 92.6\%$  (even considering increasing variability with larger stones), while this has  $r^2 = 4.7\%$ . The data plot with the fitted regression explains why this relationship is so weak. The relationship between total price and depth is not linear, nor quadratic. There are two distinct clumps of data points, with considerable variability in price in each clump.



**27.25** We first fit the multiple regression using Analyze, Regression, Linear. So that variables in the output are properly labeled, check the Label column on the Variable View tab.

**Model Summary**

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.968 <sup>a</sup>	.937	.935	88.6760

a. Predictors: (Constant), ,

**Coefficients<sup>a</sup>**

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-578.758	43.667		-13.254	.000
	length	14.307	5.659	.392	2.528	.014
	width	113.500	30.265	.582	3.750	.000

a. Dependent Variable: weight

The regression equation is  $Weight = -578.8 + 14.307 * Length + 113.5 * Width$ . The model explains 93.7% of the variability in weight.

ANOVA<sup>b</sup>

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	6229332.323	2	3114666.161	396.095	.000 <sup>a</sup>
	Residual	416761.931	53	7863.433		
	Total	6646094.254	55			

a. Predictors: (Constant), width, length

b. Dependent Variable: weight

The ANOVA test has  $F = 396.095$  with  $P$ -value 0, so at least one variable is significantly non-zero. The individual  $t$  tests for the two variables had  $t = 2.528$  for Length ( $P$ -value 0.014) and  $t = 3.750$  for width. We can conclude that both variables are useful in predicting weight of the fish, and are significantly non-zero.

Use **Transform, Compute Variable** to create the interaction variable, the refit the regression, adding this new predictor into the model.



Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.992 <sup>a</sup>	.985	.984	44.2381

a. Predictors: (Constant), interaction, length, width

ANOVA<sup>b</sup>

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	6544329.589	3	2181443.196	1114.680	.000 <sup>a</sup>
	Residual	101764.664	52	1957.013		
	Total	6646094.254	55			

a. Predictors: (Constant), interaction, length, width

b. Dependent Variable: weight

Coefficients<sup>a</sup>

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	113.935	58.784		1.938	.058
	length	-3.483	3.152	-.095	-1.105	.274
	width	-94.631	22.295	-.485	-4.244	.000
	interaction	5.241	.413	1.561	12.687	.000

a. Dependent Variable: weight

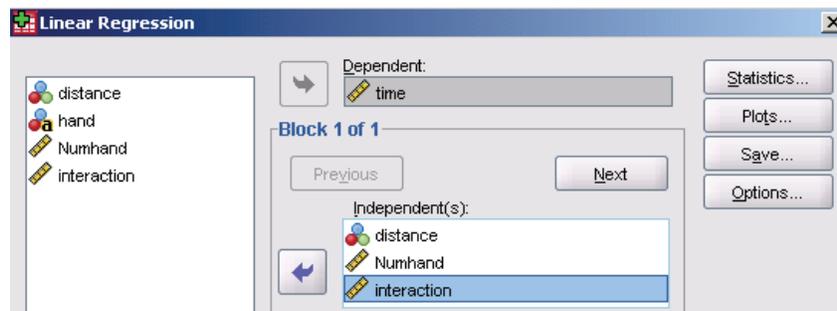
The new regression equation is  $Weight = 113.935 - 3.483*Length - 94.631*Width + 5.241*Length*Width$ . This new model explains 98.5% of the variability in weight (about an extra 5%). Again, at least one variable coefficient is non-zero because  $F = 1114.680$  with  $P$ -value 0. The coefficients of length and width changed to become negative (so their  $t$ -statistics changed sign). If we look at the  $P$ -values for the individual coefficients, we see that length is no longer significant to the model ( $P = 0.274$ ).

**27.27** To find the confidence and prediction intervals, recomputed the linear regression as described in Exercise 27.25, but click **Save**. Check both **Prediction Intervals** boxes (for Mean and Individual). We can read the values for the 10<sup>th</sup> perch in the data spreadsheet. This fish had length 21 and width 2.8 (and weight 85.0).

	LMCI_1	UMCI_1	LICI_1	UICI_1
9	67.44644	101.86444	-5.76756	175.07844
10	63.12768	104.90556	-7.17832	175.21156

The actual regression predicts fish with these dimensions will weigh about 84.02 grams (close to the observed value of 85). The mean weight of all fish with these dimensions will be between 63.13 and 104.91 grams, and any one particular fish with these dimensions should weigh between 0 (weight cannot be negative) and 175.2 grams. These statements are made with 95% confidence.

**27.41** The population model will be  $Time = \beta_0 + \beta_1 Dist + \beta_2 Hand + \beta_3 Hand * Dist$ , with hand coded as 0 or 1 (it really doesn't matter which hand is designated as which). If we code Hand = 1 as right,  $\beta_0$  is the intercept, which might be considered as some overall reaction time attributable to the left hand,  $\beta_1$  is the rate of travel for the left hand (distance per millisecond),  $\beta_2$  is an offset in the "reaction time" for the right hand, and  $\beta_3$  is a difference in rate of travel between left and right hands. We've used **Transform, Recode into Different Variables** (see Exercise 27.15) to code right as 1 and left as 0 into variable **numhand**, then multiplied this by **distance** to form the variable **interaction**. We then find the multiple regression equation.



### Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.773 <sup>a</sup>	.598	.564	50.607

a. Predictors: (Constant), interaction, , Numhand

### ANOVA<sup>b</sup>

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	136948.495	3	45649.498	17.825	.000 <sup>a</sup>
	Residual	92197.505	36	2561.042		
	Total	229146.000	39			

a. Predictors: (Constant), interaction, , Numhand

b. Dependent Variable:

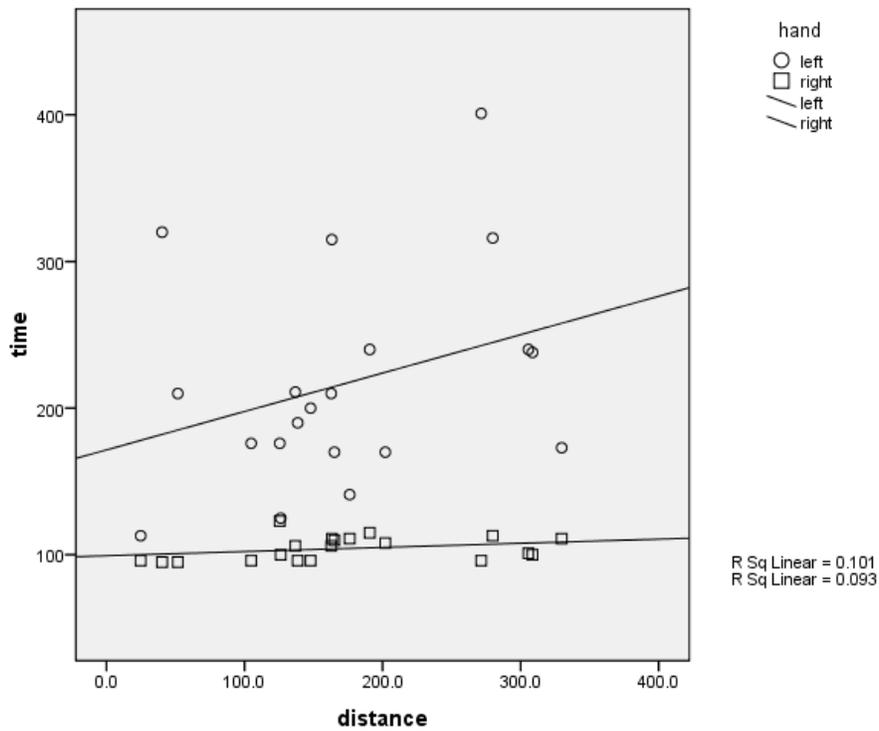
### Coefficients<sup>a</sup>

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	171.548	25.254		6.793	.000
	distance	.262	.131	.299	2.002	.053
	Numhand	-72.184	35.714	-.477	-2.021	.051
	interaction	-.234	.185	-.326	-1.263	.215

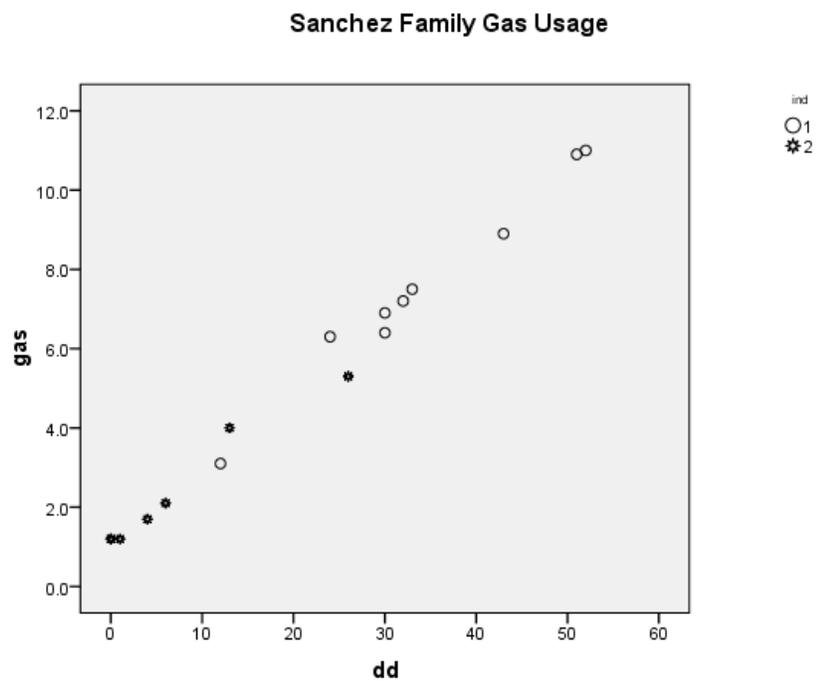
a. Dependent Variable:

The regression equation is  $Time = 171.548 + 0.262*Distance - 72.184*Right - 0.234*Right*Distance$ . This model explains 59.8% of the variability in time. Separating the equation into one for each hand, we have  $LeftHandTime = 171.548 + 0.262*Distance$  and  $RightHandTime = 99.364 + 0.028*Distance$ .

To visually see the difference in the fitted lines, we've created a scatterplot of **time** against **distance** with markers by **hand**. Double-click in the graph for the Chart Editor. We changed the marker types from color to shape by clicking in the legend and changing marker **Type** and **Border** (apply each). To add the two regression lines, click **Elements, Fit Line at Subgroups**. Close the Chart Editor Windows. It is clear that right hand times do not seem to vary with distance, while left hand times do.



**27.45** If you use dataset *ex27-45.por*, the indicator variable has already been created for you. Use **Graphs, Legacy Dialogs, Scatter/Dot** to create the plot, using **ind** to Set Markers By. We have used the Chart Editor to change the plot symbols for each value if **ind** and their colors, to make them more visible.



The graph is clearly linear. The “winter” months have the highest usages (at the low end, we have some overlap of points).

To fit the regression, we’ll need to change the values of **ind** from 1 and 2 to 0 and 1. We’ve simply changed the 2 to 0. We then create the interaction variable and fit the regression. (Be sure to check the **Variable View** so that all variables have Labels that will show in the output).

Compute Variable

Target Variable: interaction = Numeric Expression: ind\*dd

**Model Summary**

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.996 <sup>a</sup>	.992	.990	.3346

a. Predictors: (Constant), interaction, ind, dd

**ANOVA<sup>b</sup>**

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	168.846	3	56.282	502.615	.000 <sup>a</sup>
	Residual	1.344	12	.112		
	Total	170.189	15			

a. Predictors: (Constant), interaction, ind, dd

b. Dependent Variable: gas

**Coefficients<sup>a</sup>**

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	1.176	.152		7.721	.000
	dd	.168	.013	.883	12.568	.000
	ind	.302	.462	.046	.655	.525
	interaction	.013	.018	.078	.736	.476

a. Dependent Variable:

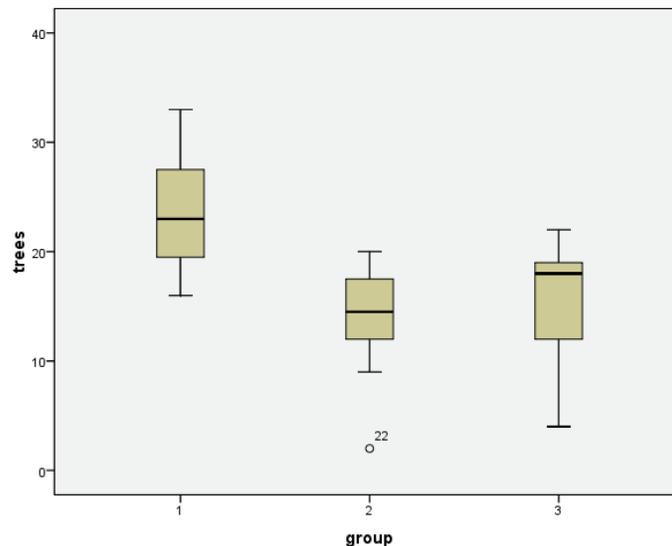
With  $r^2 = 99.2\%$ , this model appears to fit these data well. The regression equation is:  $GasUsed = 1.176 + 0.168 * DegreeDay + .302 * Winter + 0.013 * Winter * DegreeDay$ . Note the small  $t$  for the coefficient of **ind** (Winter). Winter season does not appreciably change the intercept of the model. We can also note that the  $t$  for the change in slope is 0.736 with  $P$ -value 0.476. Based on this additional information, the winter season does

not require an additional line – one regression line for the entire data set should be sufficient.

## Chapter 28 SPSS Solutions

**28.1** Open data file *ta24\_02.por*. Use **Analyze**, **Descriptive Statistics**, **Explore** with **trees** as the dependent and **group** as the factor to compute summary statistics and get boxplots of the data.

Descriptives			
		Statistic	Std. Error
1	Mean	23.75	1.462
	Std. Deviation	5.065	
2	Mean	14.08	1.438
	Std. Deviation	4.981	
3	Mean	15.78	1.920
	Std. Deviation	5.761	

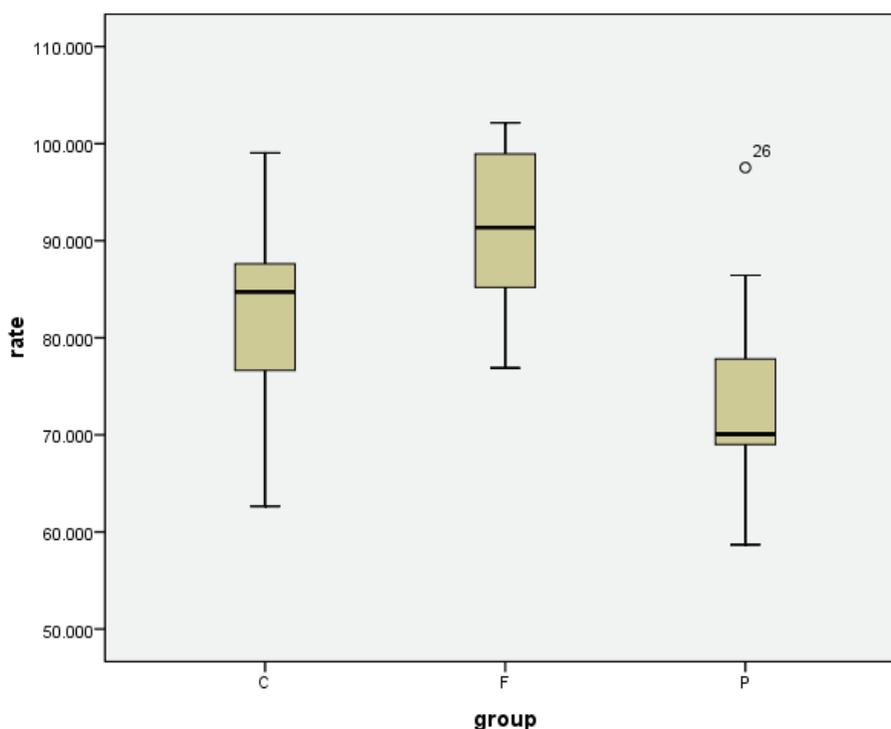


There is a low outlier (2) in the data for plots logged one year ago; however, since it is not unreasonable, we will proceed with caution. Looking at the medians, it appears that the plots with the most trees are those that have never been logged.

We want to test hypotheses  $H_0$ : all plots have the same mean number of trees against  $H_a$ : not all the means are equal. Use **Analyze**, **Compare Means**, **One-Way ANOVA** to compute the test. With a test statistic of  $F = 11.26$  and  $P$ -value 0.000, we reject the null hypothesis and conclude that logging does affect the mean number of trees per plot.

ANOVA					
	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	625.157	2	312.578	11.426	.000
Within Groups	820.722	30	27.357		
Total	1445.879	32			

**28.3** We first examine the data using side-by-side boxplots where data are Summaries for groups of cases. The variable is **rate** and the category axis variable is **group**.



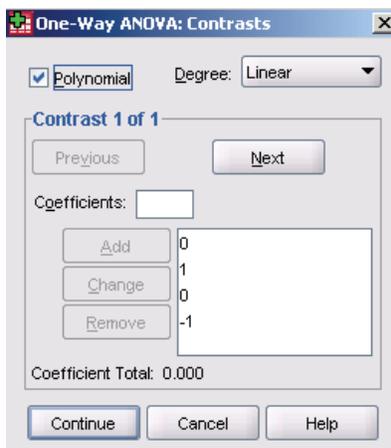
None of the distributions is perfectly symmetric and the pet data has an outlier (which helps symmetry). Looking at the medians, it appears that pets do lower the mean rate during the task, while friends raise it. We'll test  $H_0$ : all conditions have the same mean against  $H_a$ : at least one mean is different from the others. Before doing the test, we'll have to create a numeric grouping variable (called Numgroup with values 0 = Control to 2 = Friend). Use Analyze, Compare Means, One-Way ANOVA.

ANOVA

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	2387.689	2	1193.844	14.080	.000
Within Groups	3561.299	42	84.793		
Total	5948.988	44			

With a  $P$ -value this small (0.000), we'll reject the null hypothesis – friends and pets to make a difference in stress levels when performing a difficult task. It seems that pets lower the heart rate and friends raise it.

**28.7** Open data file *ex28\_02.por*. The contrast of interest is  $L_2 = \mu_G - \mu_Y$ . We first create a numeric variable called **numgroup** where **numgroup** is 1 (blue) to 4 (Yellow). Use **Analyze**, **Compare Means**, **One-Way ANOVA**. Click to enter **beetles** as the dependent and **numgroup** as the factor. Click the **Contrasts** button. Enter the coefficients in order of the group (0, 1, 0, -1); click **Add** for each. **Continue** and **OK** to compute the ANOVA and contrast.



Contrast Tests

	Contra st	Value of Contrast	Std. Error	t	df	Sig. (2-tailed)
Assume equal variances	1	-16.00	3.274	-4.886	20	.000
Does not assume equal variances	1	-16.00	3.784	-4.228	9.945	.002

The  $P$ -value for the contrast is 0.000; green and yellow attract different numbers of beetles. We are given the standard error of the contrast and its value; the confidence

interval is  $-16 \pm t^*3.274$ . From Table C,  $t^* = 2.086$ . The 95% confidence interval for the difference in green and yellow is  $-22.829$  to  $-9.171$ ; we have clear evidence that cereal leaf beetles are more attracted to green than yellow.

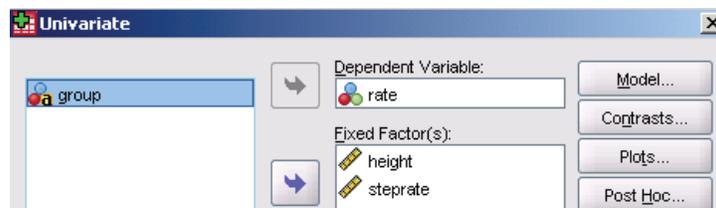
**28.11** Open data file *ta28\_03.por*. The treatments can be displayed in a two-way layout like the one below.

	Step Height	
	Low	High
Slow	Group 1	Group 2
Medium	Group 3	Group 4
Fast	Group 5	Group 6

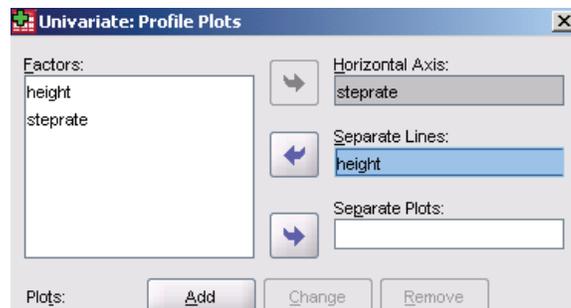
To find the means for each treatment, use **Analyze, Descriptive Statistics, Explore** with **rate** as the dependent and **group** as the factor. Our table of the means is below.

	Step Height	
	Low	High
Slow	7.8	23.4
Medium	18.6	28.2
Fast	31.8	54.6

We can create the means plot (and do the two-way ANOVA) using **Analyze, General Linear Model, Univariate**. We first create numeric factor variables **height** (1=low and 2 = high) and **steprate** (1 = slow to 3 = fast). Click to enter **rate** as the dependent and **height** and **steprate** as the fixed factors.



Click the **Plots** button. Click to enter **steprate** for the horizontal axis and **height** for the separate lines. Click **Add** and **Continue**, then **OK** to perform the test and create the plot.

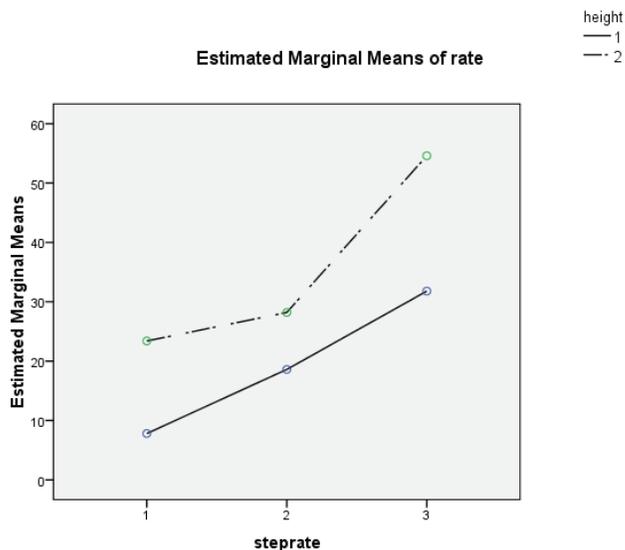


### Tests of Between-Subjects Effects

Dependent Variable:

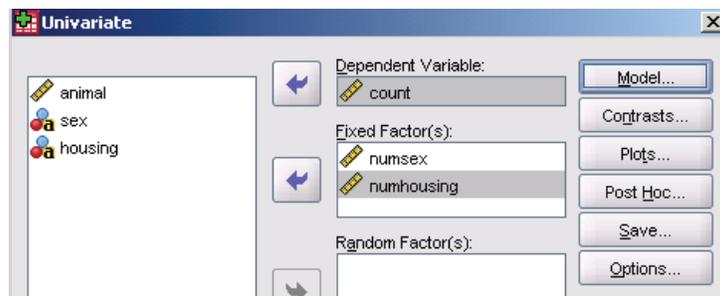
Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	6187.200 <sup>a</sup>	5	1237.440	10.999	.000
Intercept	22522.800	1	22522.800	200.203	.000
height	1920.000	1	1920.000	17.067	.000
steprate	4048.800	2	2024.400	17.995	.000
height * steprate	218.400	2	109.200	.971	.393
Error	2700.000	24	112.500		
Total	31410.000	30			
Corrected Total	8887.200	29			

a. R Squared = .696 (Adjusted R Squared = .633)



From the means plot, it appears there is interaction because the lines are not parallel; however, with a  $P$ -value of 0.393, this is not significant.

**28.13** Our null hypothesis is that neither gender nor housing affect the mean number of social play episodes. The alternate is that at least one of these affects social play. We'll use **Analyze, General Linear Model, Univariate** to do the two-way ANOVA, but we first need to create numeric variables for **sex** and **housing**. We've coded males as 1 in variable **numsex** and isolated rates as 1 in **numhousing**. (Do this by entering the values manually or using **Transform, Recode into Different Variables**).



For the models discussed in this book, click **Model** and uncheck the box to include the intercept.

#### Tests of Between-Subjects Effects

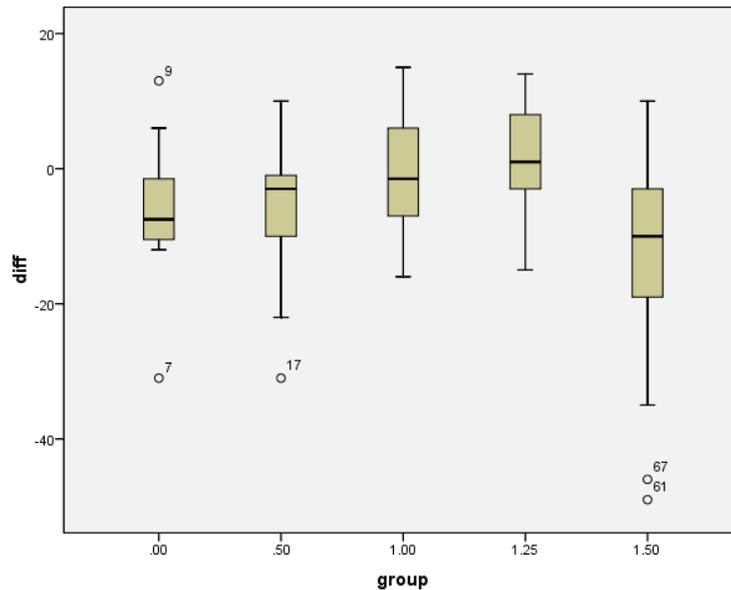
Dependent Variable:count

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Model	44012.250 <sup>a</sup>	4	11003.062	54.004	.000
numsex	3024.187	1	3024.187	14.843	.000
numhousing	346.687	1	346.687	1.702	.199
numsex * numhousing	99.188	1	99.188	.487	.489
Error	8964.750	44	203.744		
Total	52977.000	48			

a. R Squared = .831 (Adjusted R Squared = .815)

The indications are that gender makes a difference in social play for hooded rats ( $P = 0.000$ ), but there is no interaction ( $P = 0.489$ ), nor does housing make a significant difference ( $P = 0.199$ ).

**28.29** We first create side-by-side boxplots of the differences, examining them for outliers and any indications that the data are not Normal.



Groups 0 and 0.5 and 1.5 have outliers. Group 1.25 seems to have a smaller spread than the others. We'll proceed, but use caution. Based on the plots, the median differences are all negative (meaning a slower healing rate than natural). We test  $H_0$ : all treatments have the same mean against  $H_a$ : at least one is different using **Analyze, Compare Means, One-Way ANOVA**, but this requires that the factor values (the treatment labels) be integer-valued. We've created a new variable (called Numgroup) with values 0 for Control (normal 1.0 group), through 4 for the 1.5 group. We follow up with **Analyze, Compare Means, Means**. For the Tukey intervals, click **Post-Hoc**, then check the **LSD** box.

## ANOVA

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	2232.147	4	558.037	4.041	.005
Within Groups	9528.407	69	138.093		
Total	11760.554	73			

## Report

	Mean	N	Std. Deviation
0	-6.42	12	10.706
0.5	-5.71	14	10.564
1	-.17	18	9.345
1.25	1.47	15	8.863
1.5	-13.80	15	17.387
Total	-4.66	74	12.693

With  $F = 4.04$  and  $P$ -value 0.005, we conclude that the means are not all the same. The conjecture is that nature (group 1) heals best and that changing the field slows healing. The mean difference in Group 1 is  $-0.17$ . This small difference indicates little difference between the control and “experimental” legs. The mean difference in group 0 is  $-6.42$ ; group 0.5 has mean  $-5.71$ ; group 1.25 has mean  $1.47$ ; group 1.5 has the largest mean difference:  $-13.80$ . The 1.25 times natural field is close to natural; it appears that all the others slow healing. Further, looking at the intervals, it is clear that the 1.5 group has slower healing than either the “natural” (1.0) or 1.25 groups. The Multiple Comparisons output compares each group with all the others.

### Multiple Comparisons

LSD

(I) VAR00 001	(J) VAR00 001	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
0	1	6.250	4.379	.158	-2.49	14.99
	2	5.548	4.188	.190	-2.81	13.90
	3	-1.633	4.108	.692	-9.83	6.56
	4	13.633*	4.108	.001	5.44	21.83
1	0	-6.250	4.379	.158	-14.99	2.49
	2	-.702	4.623	.880	-9.92	8.52
	3	-7.883	4.551	.088	-16.96	1.20
	4	7.383	4.551	.109	-1.70	16.46
2	0	-5.548	4.188	.190	-13.90	2.81
	1	.702	4.623	.880	-8.52	9.92
	3	-7.181	4.367	.105	-15.89	1.53
	4	8.086	4.367	.068	-.63	16.80
3	0	1.633	4.108	.692	-6.56	9.83
	1	7.883	4.551	.088	-1.20	16.96
	2	7.181	4.367	.105	-1.53	15.89
	4	15.267*	4.291	.001	6.71	23.83
4	0	-13.633*	4.108	.001	-21.83	-5.44
	1	-7.383	4.551	.109	-16.46	1.70
	2	-8.086	4.367	.068	-16.80	.63
	3	-15.267*	4.291	.001	-23.83	-6.71

\*. The mean difference is significant at the 0.05 level.

In these intervals, any comparisons that span 0 are considered similar. The only comparisons that do not span 0 are groups 1 and 1.5 and groups 1.25 and 1.5. Our final “underline” diagram is (start by ordering the sample means from smallest to largest)

1.50	0.00	0.50	1.00	1.25
-13.8	-6.40	-5.71	-0.17	1.47
<hr style="width: 10%; margin: 0 auto;"/> <hr style="width: 20%; margin: 0 auto;"/> <hr style="width: 30%; margin: 0 auto;"/>				

**28.31** With 6 populations, there will be  $\binom{6}{2} = 15$  pairwise comparisons. To do the two-way ANOVA, we recoded **height** to **numheight** with 1 = low and **freq** to **numfreq** with slow = 1 to fast = 3. We use **Analyze, General Linear Model, Univariate** to compute the two-way ANOVA. Click **Model** and uncheck the box to fit an intercept.

#### Tests of Between-Subjects Effects

Dependent Variable:

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Model	192362.400 <sup>a</sup>	6	32060.400	367.876	.000
numheight	235.200	1	235.200	2.699	.113
numfreq	11.400	2	5.700	.065	.937
numheight * numfreq	115.800	2	57.900	.664	.524
Error	2091.600	24	87.150		
Total	194454.000	30			

a. R Squared = .989 (Adjusted R Squared = .987)

Neither factor nor the interaction is significant as all have *P*-values at least 0.113. We should find that all Tukey intervals span 0. To compute these, add another grouping variable – this one from 1 = low/slow to 6 = high/fast. Use **Analyze, Compare Means, One-way ANOVA** and click **Post-Hoc**. Check the box to ask for the **LSD** intervals. The resulting table is rather long, but we can see that all the intervals span zero (reach from negative to positive values). The randomization worked well.

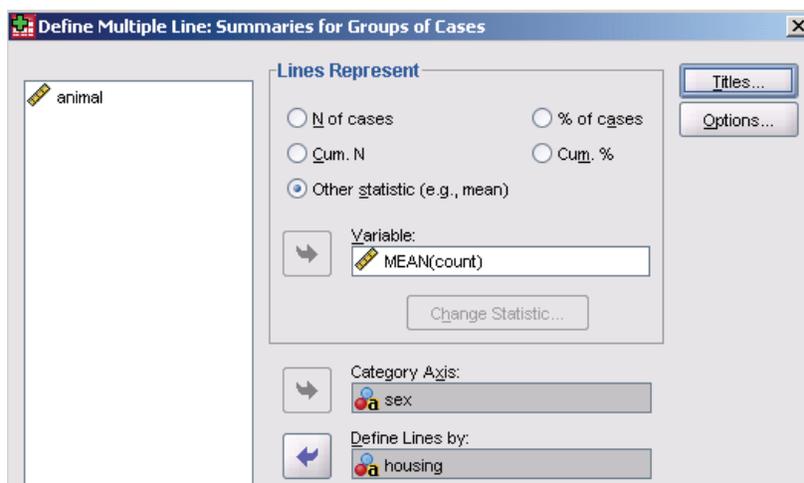
#### Multiple Comparisons

LSD

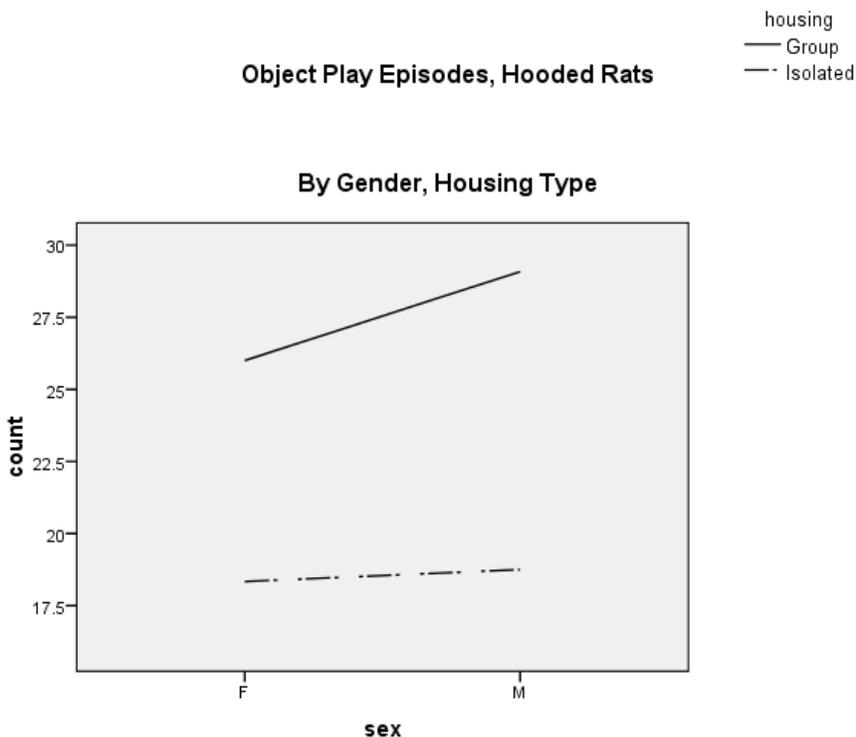
(I) numgro up	(J) numgro up	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
1	2	4.200	5.904	.484	-7.99	16.39
	3	3.600	5.904	.548	-8.59	15.79
	4	-.600	5.904	.920	-12.79	11.59
	5	-6.000	5.904	.320	-18.19	6.19
	6	-2.400	5.904	.688	-14.59	9.79
2	1	-4.200	5.904	.484	-16.39	7.99
	3	-.600	5.904	.920	-12.79	11.59
	4	-4.800	5.904	.424	-16.99	7.39
	5	-10.200	5.904	.097	-22.39	1.99
	6	-6.600	5.904	.275	-18.79	5.59
3	1	-3.600	5.904	.548	-15.79	8.59
	2	.600	5.904	.920	-11.59	12.79

	4	-4.200	5.904	.484	-16.39	7.99
	5	-9.600	5.904	.117	-21.79	2.59
	6	-6.000	5.904	.320	-18.19	6.19
4	1	.600	5.904	.920	-11.59	12.79
	2	4.800	5.904	.424	-7.39	16.99
	3	4.200	5.904	.484	-7.99	16.39
	5	-5.400	5.904	.370	-17.59	6.79
	6	-1.800	5.904	.763	-13.99	10.39
5	1	6.000	5.904	.320	-6.19	18.19
	2	10.200	5.904	.097	-1.99	22.39
	3	9.600	5.904	.117	-2.59	21.79
	4	5.400	5.904	.370	-6.79	17.59
	6	3.600	5.904	.548	-8.59	15.79
6	1	2.400	5.904	.688	-9.79	14.59
	2	6.600	5.904	.275	-5.59	18.79
	3	6.000	5.904	.320	-6.19	18.19
	4	1.800	5.904	.763	-10.39	13.99
	5	-3.600	5.904	.548	-15.79	8.59

**28.33** We begin by graphing a means plot to look for any individual factor or interaction effects. Click **Graphs, Legacy Dialogs, Line**. Select a **Multiple** graph with data from **Summaries for groups of cases**.



In our graph below, we used the Chart Editor to change the variable attribute for **housing** to be dash instead of the default color. There may be interaction in these data. Rats in isolated housing do not seem to differ in episodes of object play by gender, but for those in group housing, males seem to show more object play than females.



To perform the test, we recode to numeric values for **sex** (1 = male) and **housing** (1 = isolated). Uncheck the box for fitting the intercept in the **Model** dialog box.

#### Tests of Between-Subjects Effects

Dependent Variable:

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Model	26514.167 <sup>a</sup>	4	6628.542	57.257	.000
numsex	36.750	1	36.750	.317	.576
numhousing	972.000	1	972.000	8.396	.006
numsex * numhousing	21.333	1	21.333	.184	.670
Error	5093.833	44	115.769		
Total	31608.000	48			

a. R Squared = .839 (Adjusted R Squared = .824)

Housing is significant ( $P = 0.006$ ), but the interaction and gender variables are not. A boxplot of the residuals by group (use Clustered boxplots) shows that these distributions are not perfectly symmetric. They also seem to vary a bit with the animals in group housing having larger spreads. To obtain summary statistics for each treatment group, we'll need to create another variable for the four combinations. (**Analyze, Descriptive Statistics, Explore** says we are within the multiple of 2 for the standard deviation rule).

