

**JMP<sup>®</sup>ing**  
**into**

**Moore, McCabe, and Craig's**  
**Introduction to the Practice of Statistics**  
*Sixth Edition*

**Thomas F. Devlin**  
*Montclair State University*



**W. H. Freeman and Company**  
**New York**

JMP® is a registered trademark of SAS Institute Inc.

© 2009 by W. H. Freeman and Company

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

# Contents

<b>PREFACE</b>	<b>III</b>
<b>CHAPTER 0 INTRODUCTION TO <i>JMP</i> STATISTICAL SOFTWARE</b>	<b>1</b>
<b>CHAPTER 1 LOOKING AT DATA: EXPLORING DISTRIBUTIONS</b>	<b>25</b>
<b>CHAPTER 2 LOOKING AT DATA: EXPLORING RELATIONSHIPS</b>	<b>45</b>
<b>CHAPTER 3 PRODUCING DATA</b>	<b>67</b>
<b>CHAPTER 4 PROBABILITY</b>	<b>72</b>
<b>CHAPTER 5 SAMPLING DISTRIBUTIONS</b>	<b>74</b>
<b>CHAPTER 6 INTRODUCTION TO INFERENCE</b>	<b>77</b>
<b>CHAPTER 7 INFERENCE FOR DISTRIBUTIONS</b>	<b>90</b>
<b>CHAPTER 8 INFERENCE FOR PROPORTIONS</b>	<b>108</b>
<b>CHAPTER 9 INFERENCE FOR TWO-WAY TABLES</b>	<b>117</b>
<b>CHAPTER 10 INFERENCE FOR REGRESSION</b>	<b>123</b>
<b>CHAPTER 11 MULTIPLE REGRESSION</b>	<b>132</b>
<b>CHAPTER 12 ONE-WAY ANALYSIS OF VARIANCE</b>	<b>140</b>
<b>CHAPTER 13 TWO-WAY ANALYSIS OF VARIANCE</b>	<b>153</b>
<b>CHAPTER 14 BOOTSTRAP METHODS AND PERMUTATION TESTS</b>	<b>158</b>
<b>CHAPTER 15 NONPARAMETRIC TESTS</b>	<b>161</b>
<b>CHAPTER 16 LOGISTIC REGRESSION</b>	<b>166</b>
<b>CHAPTER 17 STATISTICS FOR QUALITY</b>	<b>173</b>
<b>CHAPTER 18 TIME SERIES FORECASTING</b>	<b>181</b>
<b>EXERCISES</b>	<b>201</b>

The following JMP script, BootstrappingTheMean.jsl, is needed to complete problems in chapters 6 and 16. To access this file please link to [www.whfreeman.com/ips6e](http://www.whfreeman.com/ips6e) and find the file in the category “JMP Script”.



# Preface

This manual is intended to guide students in their use of *JMP* Software to automate the statistical graphics and analyses found in his or her introductory statistics textbook.

John Sall created *JMP* in 1989 as a tool for discovering information in data through visualization and graphics. *JMP* is designed to be a point-and-click, walk-up-and-use product that enables a user to discover more, interact more, and understand more. The correct graphs are integrated with the right analyses. Because *JMP* is task-oriented, not method-oriented, you don't need to be a professional statistician to use it. You only need to know what questions you wish answered.

You might wonder: why not use a spreadsheet program like *Excel*? Spreadsheet software is, at best, like a Swiss army knife. It can do many of things okay—but not very well. Spreadsheet software was designed to manage lists and tables of values and to perform bookkeeping-like calculations—not statistical computations and graphics. Spreadsheet software is not nearly as easy to use as *JMP*; it can't manipulate data nearly as well as *JMP*, and the underlying statistical and mathematical algorithms in spreadsheets are not as well tested and reliable as those in *JMP*. The American Statistical Association, a 175-plus-year-old organization of professional statisticians, strongly states: "Efficient computing tools are essential for statistical research, consulting, and teaching. Generic packages such as *Excel* are not sufficient for teaching of statistics ...."

*JMP* runs under the Windows, Macintosh, and Linux operating systems. Academic researchers and industry use the professional version of *JMP* extensively. An inexpensive student version, *JMP 6 Student Edition*, which is available from W. H. Freeman, is intended for introductory statistics courses. Data table creation is limited to 1000 rows.

Special thanks goes to the authors for crafting an engaging introductory statistics textbook, and to John Sall and the *JMP* Software team at SAS Institute for creating a software system that allows users to focus on understanding the data. I would also like to thank Xander Kasten and Erika Klein for their work on the problem statements and solutions.

The Student CD-ROM contains *JMP* data tables and text files with data from most of the examples, exercises, and tables from the textbook. *JMP* data sets are also available at the book companion Web site. We have adopted the following filename convention throughout this manual.

Textbook Source	File Name
Example 1.6	eg01_006.JMP
Exercise 1.38	ex01_038.JMP
Table 1.1	ta01_001.JMP
Figure 1.1	fg01_001.JMP





# Chapter 0

## Introduction to *JMP* Statistical Software

Statistics is best learned by practicing with real data. The purpose of this manual is to provide instructions on how to use *JMP* statistical software to automate the statistical calculations and graphics presented in an introductory statistics textbook like *Introduction to The Practice of Statistics*, Sixth Edition, (*IPS*) by David S. Moore and George P. McCabe, or *Practice of Business Statistics*, Second Edition, (*PBS*) by David S. Moore, et al. I hope that this allows you to concentrate on the meaning and purpose of the calculations.

This manual parallels the above textbooks. Chapters 1 and following of this manual correspond to separate chapters in the textbooks. These chapters guide you in the use of *JMP* for the calculations and graphics of the corresponding textbook chapters. This chapter, Chapter 0, serves as an introduction to *JMP* statistical software and discusses:

- Getting acquainted with *JMP* and the *JMP* data table
- Entering and saving data
- Working with variables and individuals
- Customizing your *JMP* environment

If you read this chapter carefully, using *JMP* will quickly become second nature. You will be able to produce appropriate graphs and calculations at the click of a mouse and the touch of a button.

## 0.1 Getting Acquainted with *JMP* and the *JMP* Data Table

### 0.1.1 Getting Started and Quitting

*JMP* is started like any other application on your operating system. Either:

- double-click a *JMP* data table or script, or
- double-click the *JMP* icon.

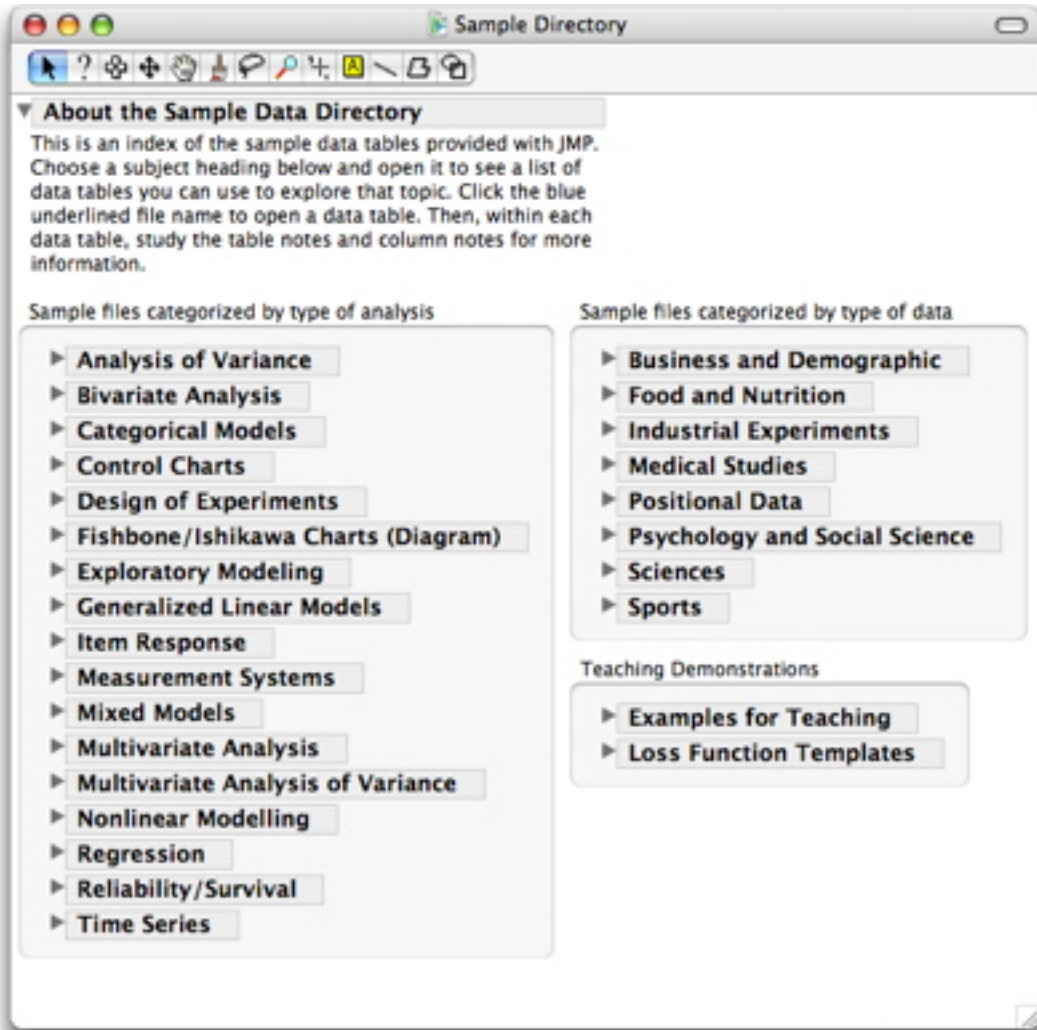
To quit *JMP*:

- select **File** ⇒ **Exit** or press **ctrl-Q** in Windows, or
- select **JMP** ⇒ **Quit** or press **command-Q** on the Macintosh.

### 0.1.2 The *JMP* Data Table

Data to be processed in *JMP* must be in a *JMP data table*. A *data table* is similar to a spreadsheet but the rows and columns have a special purpose. Start *JMP* and open the *JMP* data table **Big Class** in the Sample Data Directory installed with the *JMP* application on your computer.

1. Select **Help** ⇒ **Sample Data Directory** from the menu bar.



2. Press the gray triangle next to **Examples for Teaching**.
3. Select the first file **Big Class.jmp**.

Inspect the data table.

	name	age	sex	height	weight
1	KATIE	12	F	59	95
2	LOUISE	12	F	61	123
3	JANE	12	F	55	74
4	JACLYN	12	F	66	145
5	LILLIE	12	F	52	64
6	TIM	12	M	60	84
7	JAMES	12	M	61	128
8	ROBERT	12	M	51	79
9	BARBARA	13	F	60	112
10	ALICE	13	F	61	107
11	SUSAN	13	F	56	67
12	JOHN	13	M	65	98
13	JOE	13	M	63	105
14	MICHAEL	13	M	58	95
15	DAVID	13	M	59	79
16	JUDY	14	F	61	81
17	ELIZABETH	14	F	62	91

The *data table* looks like a spreadsheet with some enhancements. In the upper left-hand corner, you can see that the data table has 40 rows and 5 columns. Look more closely and notice that each of the columns—**name**, **age**, **sex**, **height**, and **weight**—contains the values of a *variable* and each of the rows is an *individual*. Therefore, **Big Class** contains 5 variables and 40 individuals. In general, the columns of a data table contain *variables* and the rows contain *individuals*.

### 0.1.3 Menu Headings

*JMP* provides a menu bar and an icon bar of commands. The two pull-down menus at either end of the menu bar should look familiar. Let's examine the items on the menu bar.



<b>File</b>	performs most routine file functions, such as opening, closing, printing, and saving.
<b>Edit</b>	performs most editing functions, such as cutting and pasting.
<b>Tables</b>	performs table functions, such as sorting, subsetting, and merging.
<b>Rows</b>	performs row operations (i.e., operations on individuals).
<b>Cols</b>	performs column operations (i.e., operations on variables).
<b>DOE</b>	performs tasks associated with designing statistical experiments.
<b>Analyze</b>	performs most statistical analyses.

<b>Graph</b>	generates a variety of graphs.
<b>Tools</b>	displays a special palette of tools that determine the effect of a mouse action.
<b>View</b>	manages the tool bars and displays the Status Bar (Windows only).
<b>Window</b>	selects among, organizes, and performs routine window operations on opened windows.
<b>Help</b>	accesses the main help feature in <i>JMP</i> .

The menu bar in Mac OS X is the same as above with the addition of a **JMP** menu. The **JMP** menu contains the standard items in an application menu. You can change preferences and quit *JMP* from it.



## Remark

- Instructions in this manual will focus on accessing commands through the menus. As you become more familiar with *JMP*, you may wish to explore the icon alternatives.

### 0.1.4 Column Attributes

Inspect the data table **Big Class**. Notice that each column/variable has a *name*. Also, note that some of the columns are left-aligned and some are right-aligned. Alignment is determined by *data type*. The *data type* of a column, or variable, determines how its values are formatted in the data table, how they are stored internally, and whether they can be used in calculations. The two *data types* of interest to us are:

- Numeric* for columns/variables with numeric values that can be used in calculations. These data are right-aligned. **Age**, **height**, and **weight** are *numeric* variables.
- Character* for columns/variables with numeric and/or character values that can be used to describe different levels of the variable. These data are left-aligned. **Name** and **sex** are *character* variables.

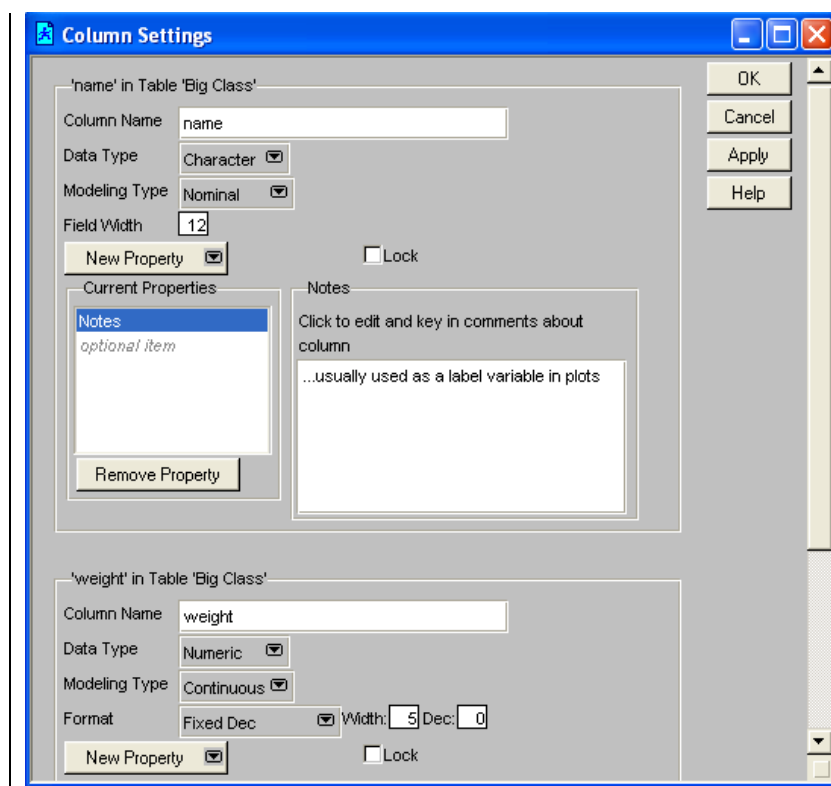
*Name* and *data type* are two attributes of a column/variable. To see other attributes, use the **Column Info** command found in the **Cols** menu.

#### Example 0.1 Obtaining information about a column

Obtain column information on the variables **name** and **weight**.

1. Highlight the columns **name** and **weight** by clicking first on the top of the column **name** and then ctrl-clicking (command-clicking on a Macintosh) on the top of the column **weight**.
2. Select **Cols** ⇒ **Column Info**.

Inspect the dialog.



Notice that:

- the variable **name** has *data type* character and **weight** has *data type* numeric.
- the *format* for **weight** is *Fixed Dec* with a width of 5 and no decimal places.
- both variables have *notes* attached to them.

The *modeling type* of a variable is very important. It is not just a descriptive tag but rather it tells *JMP* how to analyze and graph the data. For example, the **Distribution** command displays histograms and boxplots for “Continuous” variables and stacked bar charts for “Nominal” variables.

**Note:** The default *modeling type* of numeric data is “Continuous” and the default *modeling type* of character data is “Nominal.”

## 0.2 Entering and Saving Data

To process data in *JMP*, it must be in a *JMP data table*. You build a *JMP data table* either by:

- **Creating a new table** with the **New** command in the **File** menu and filling it with values by typing or pasting values into the data grid, constructing a formula, or using an external measuring instrument, or
- **Importing data** from a text file or from another application with the **Open** or the **Database** command in the **File** menu.



## 0.2.1 Creating a New *JMP* Data Table

The **New** command in the **File** menu displays an empty data table with no rows and one column, named **Column 1**.

To add variables:

- use the **New Column** or **Add Multiple Columns** command in the **Cols** menu.

To add individuals:

- use the **Add Rows** command from the **Rows** menu, or
- simply type in a cell anywhere beyond the last row of the table.

You can use the usual editing commands, such as cut and paste, to enter data values. You can also use drag and drop to copy or rearrange columns.

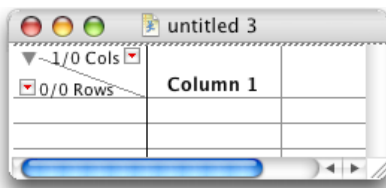
### Example 0.2 Binge drinking and gender

A survey of 17,096 students in U.S. four-year colleges collected information on drinking behavior. One question of interest was the relationship between binge drinking and gender. Here are the data summarized by gender and frequent binge drinking.

Frequent Binge Drinker	Gender	Count
Yes	Female	1684
	Male	1630
No	Female	8232
	Male	5550

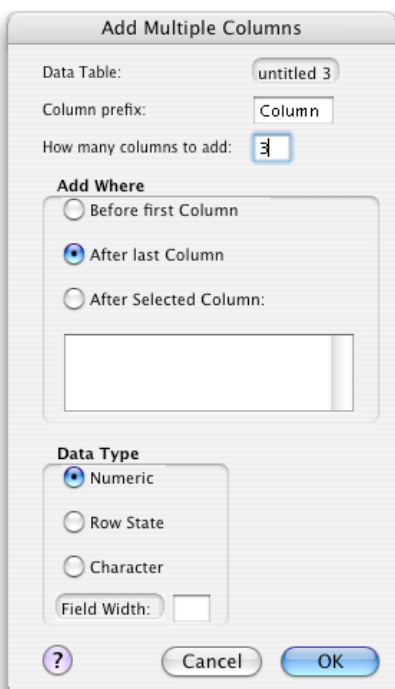
Let's create a *JMP* data table for the data.

1. Select **File** ⇒ **New** from the menu bar.



### Adding Columns

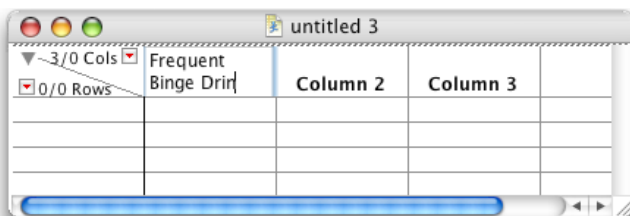
2. Select **Cols** ⇒ **Add Multiple Columns** to accommodate the three variables **Frequent Binge Drinker**, **Gender**, and **Count**.



3. Enter 3 after **How many columns to add** and press OK.

Now let's change the name of the first column to **Frequent Binge Drinker**.

4. Select the first column of the data grid to highlight that column.
  - a. Click on the name **Column 1** to highlight the column name.
  - b. Type **Frequent Binge Drinker**.

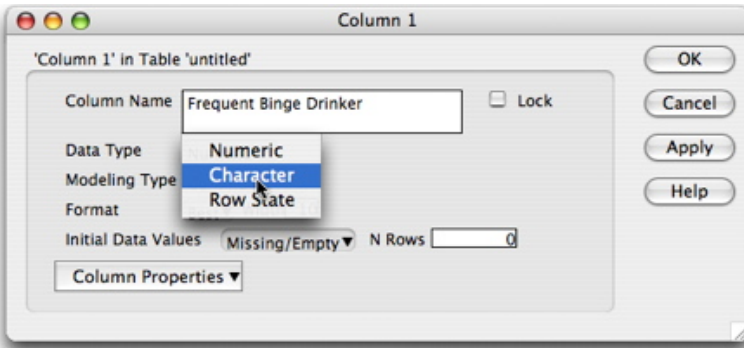


5. Repeat Step 4 on the other columns to change the names to **Gender** and **Count**.

### Setting the Data Type of a Column

By default, columns contain numeric data. However, **Frequent Binge Drinker** and **Gender** have character values. Use the **Column Info** command to change the data type of the variable **Frequent Binge Drinker**.

6. Select the first column, **Frequent Binge Drinker**.
  - a. Select Cols ⇒ Column Info.
  - b. Select Character from the pop-up menu for **Data Type**.
  - c. Press OK.

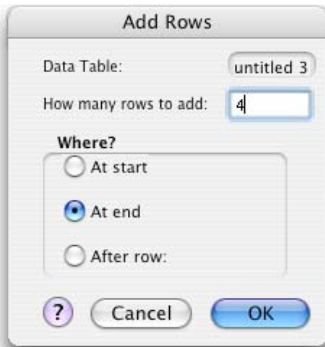


7. Repeat Step 6 for the second column, **Gender**, to change its data type to character.

## Adding Rows

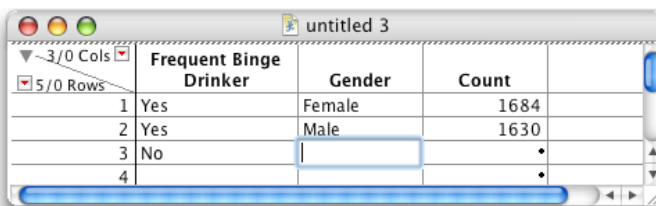
Adding rows is easy.

8. Select **Rows** ⇒ **Add Rows** from the menu bar and enter 4 and press **OK**.



## Entering Data

Entering data into the data table is similar to entering data into a spreadsheet.



9. Select the first cell in the first row and enter **Yes**.
  - a. Press the **Tab** key, enter **Female**, press **Tab** again, and enter **1684**.
  - b. Press **Return** and enter **Yes**.
  - c. Press the **Tab** key, enter **Male**, press **Tab** again, and enter **1630**.
  - d. Continue until you have finished entering the data.

In the next section, you will learn how to save the data table for later use.

## 0.2.2 Saving and Naming a Data Table

The **Save** command in the **File** menu writes the current *JMP* data table to a file.

**Note:** *JMP* analysis windows are not saved with the data table. However, you can use *JMP* tools to copy reports to other applications or you can save the JSL script that produced the analysis.

### Example 0.2 Binge drinking and gender (cont'd.)

---

To save the data table for analysis in conjunction with Chapter 2,

1. Select **File** ⇒ **Save**.
2. Type **Binge Drinking.jmp** in the **Name** field and press **Save**.

## 0.2.3 Importing Data

The **Open** command in the **File** menu directly reads existing *JMP* data tables, text files with any column delimiter, *Excel* files, *JMP* journal and script files, and *SAS* data sets into *JMP* data tables. Also, under **Windows**, the **Database** command can access any database on your system that has an ODBC driver. We illustrate importing *text* files for Windows and Macintosh OS X computers here.

### Windows Text Import

*JMP* offers two choices for importing text files under MS Windows. From the **File of type** drop-down list, you may choose:

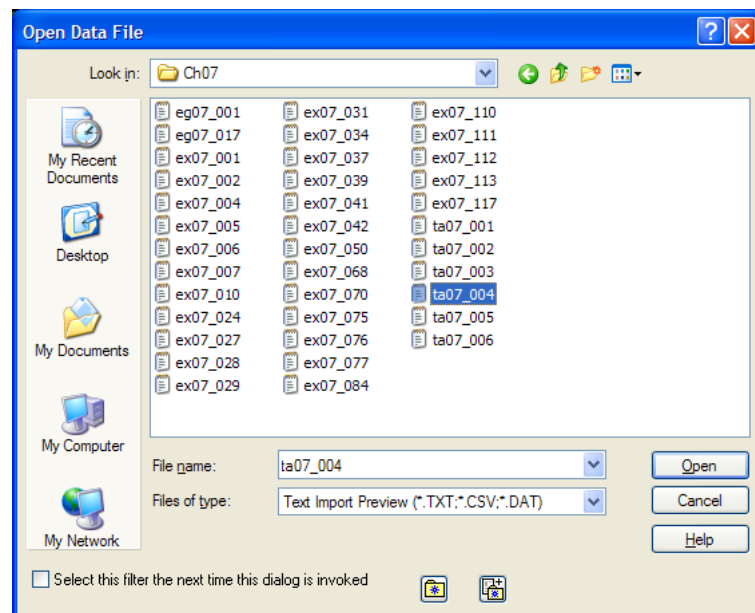
- **Text Import Files**, which opens the file and creates a *JMP* data table using either the default rules (set in the preferences panel) to interpret end-of-field and end-of-line delimiters or the best guess heuristics of *JMP*. These are sufficient for a rectangular text file with no missing fields, a consistent delimiter, and an end-of-line delimiter or with fixed width fields.
- **Text Import Preview**, which allows you to modify the default field and column specifications and displays default variable names, variable data types, and data values for the first two individuals.

### Example 0.3 Is a new teaching method effective?

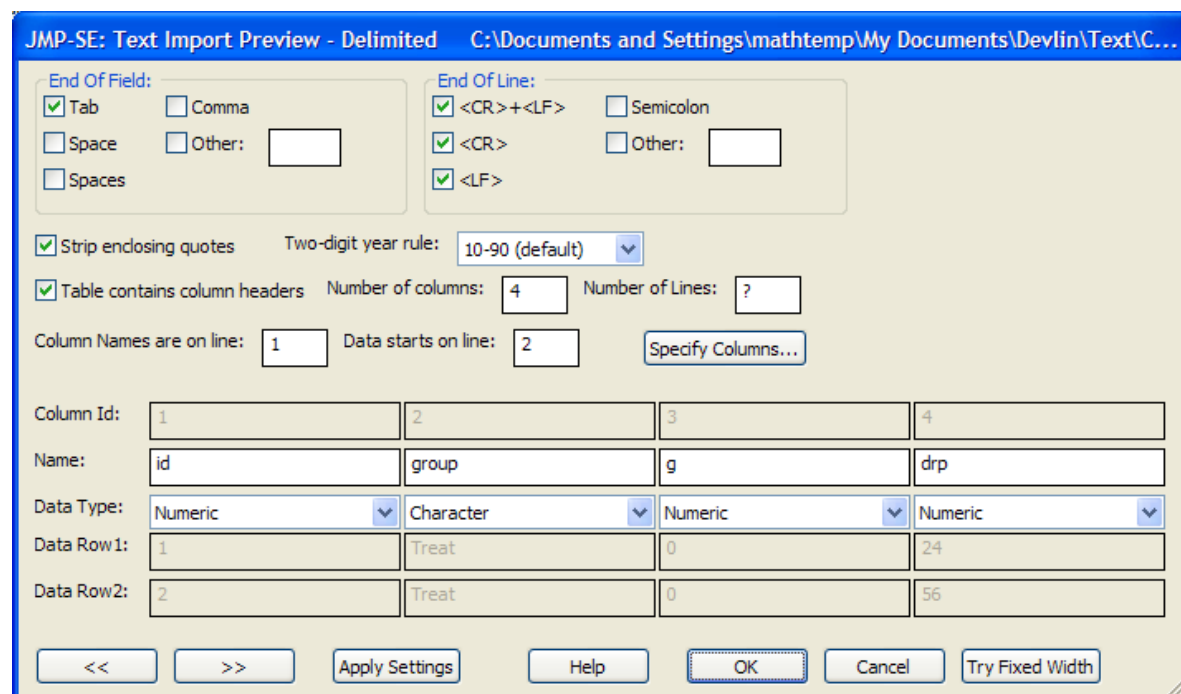
---

An educator believes that new directed reading activities will help elementary school pupils improve some aspects of their reading ability. A class of 21 third-grade students took part in the new directed reading activities and another third-grade class of 23 followed the same curriculum without these activities. Suppose that the Degree of Reading (DRP) test scores for the 44 students are found in the text file **ta07\_004.txt**. We will import the information in that file into a *JMP* data table.

1. Select **File** ⇒ **Open** from the menu bar.



2. Identify the text file in the panel that opens.
  - a. Select **Text Import Preview** from the **Files of type** menu.
  - b. Select the folder that holds the file.
  - c. Select the file **ta07\_004.txt**.
  - d. Click **Open**.



3. a. Change the four column names to **Student**, **Group**, **DRP**, and **Score**, respectively.
- b. Notice that *JMP* has chosen data types for these columns. You may change those choices.
- c. Click **OK**.

The following *JMP* data table is created.

	Student	Group	DRP	Score
16	16	Treat	0	46
17	17	Treat	0	67
18	18	Treat	0	43
19	19	Treat	0	49
20	20	Treat	0	57
21	21	Treat	0	53
22	22	Control	1	42
23	23	Control	1	46
24	24	Control	1	43
25	25	Control	1	10
26	26	Control	1	55
27	27	Control	1	17
28	28	Control	1	26
29	29	Control	1	60

Notice that the data for the second group of subjects is placed directly below the data for the first group rather than adjacent to it as one often sees in textbooks. Recall that, in a *JMP* data table, the rows are the *individuals* and the columns are the *variables*.

## Macintosh Text Import

To import text data on the Macintosh, first select **All Text Documents** from the **Enable** field. The **Open As** menu then appears and gives four choices:

- **Text** opens the file in a text editing window without creating a *JMP* data table.
- **Data (Best Guess)** opens the file and creates a *JMP* data table using the best guess heuristics of *JMP* to arrange the data.
- **Data (Using Preferences)** opens the file and creates a *JMP* data table using the default rules (set in the preferences panel) to interpret end-of-field and end-of-line delimiters.
- **Data (Using Preview)** opens the file, creates a *JMP* data table using delimiters that you designate, and displays default variable names and data types and the data values of the first two individuals.

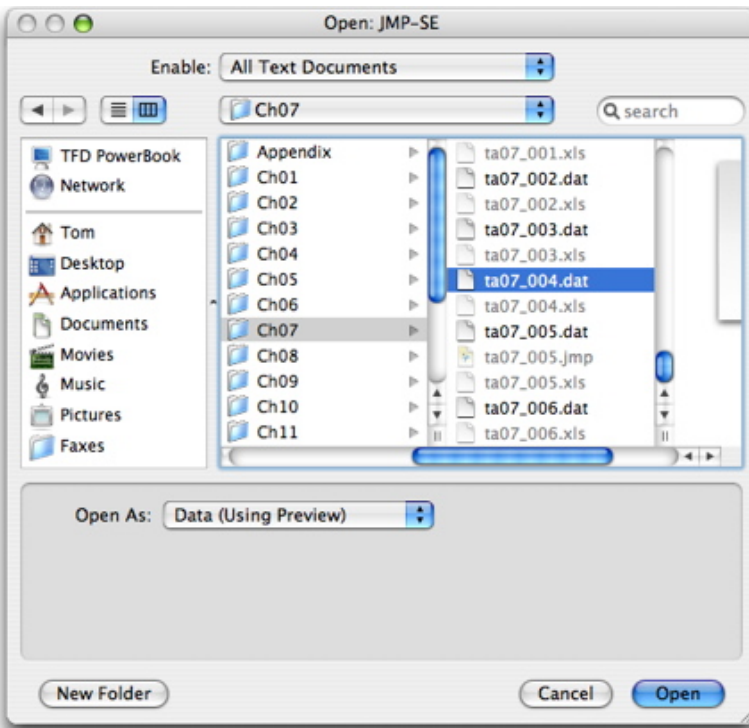
We illustrate the last choice.

### Example 0.3 Is a new teaching method effective?

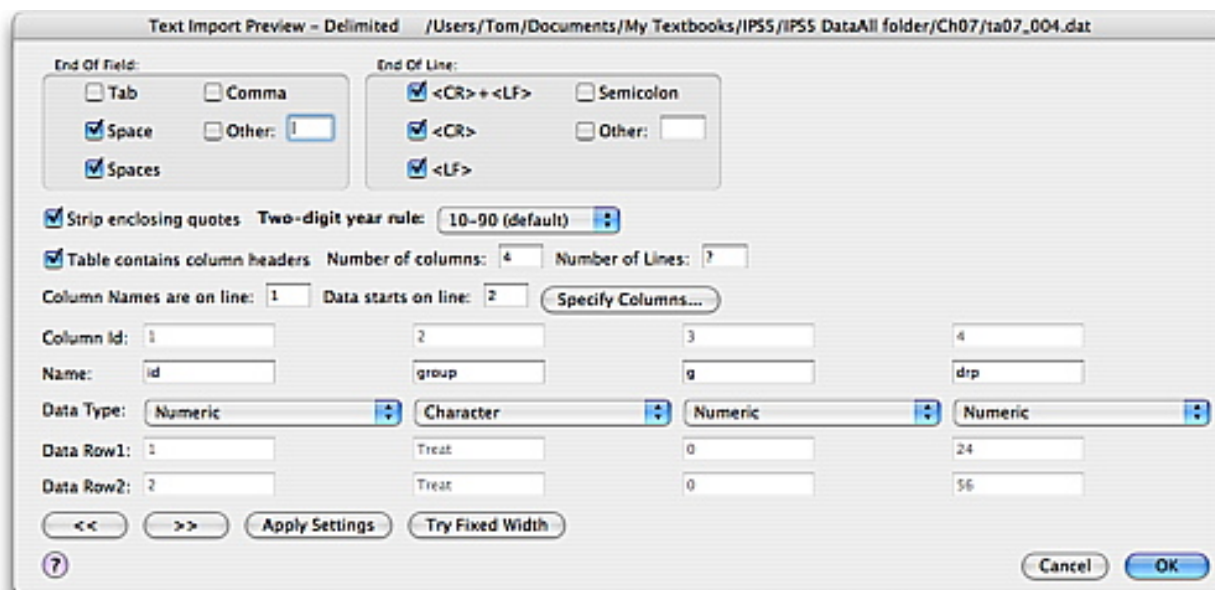
An educator believes that new “directed reading activities” will help elementary school pupils improve some aspects of their reading ability. A class of 21 third-grade students took part in the new directed reading activities and another third-grade class of 23 followed the same curriculum without these

activities. Suppose that data for the 44 students are found in a text file **ta07\_004.dat**. Let's import the information in that file into a *JMP* data table.

1. Select **File** ⇒ **Open** from the menu bar.



2. Identify the text file in the panel that opens.
  - a. Select **All Text documents** from the **Enable** menu.
  - b. Select the folder that holds the file.
  - c. Select the file **ta07\_004.dat**.
  - d. Select **Data (Using Preview)** from the **Open As** menu.
  - e. Click **Open**.



3. a. Notice that *JMP* has used the character values in the first row as the *column names*. You may change these choices here.
- b. Notice that *JMP* has chosen *data types* for these columns. You may change those choices also.
- c. Click **OK**.

The *JMP* data table on the next page is created.

	id	group	g	drp
1	1	Treat	0	24
2	2	Treat	0	56
3	3	Treat	0	43
4	4	Treat	0	59
5	5	Treat	0	58
6	6	Treat	0	52
7	7	Treat	0	71
8	8	Treat	0	62
9	9	Treat	0	43
10	10	Treat	0	54
11	11	Treat	0	49
12	12	Treat	0	57
13	13	Treat	0	61
14	14	Treat	0	33
15	15	Treat	0	44
16	16	Treat	0	46
17	17	Treat	0	67
18	18	Treat	0	43
19	19	Treat	0	49
20	20	Treat	0	57
21	21	Treat	0	53
22	22	Control	1	42
23	23	Control	1	46
24	24	Control	1	43
25	25	Control	1	40



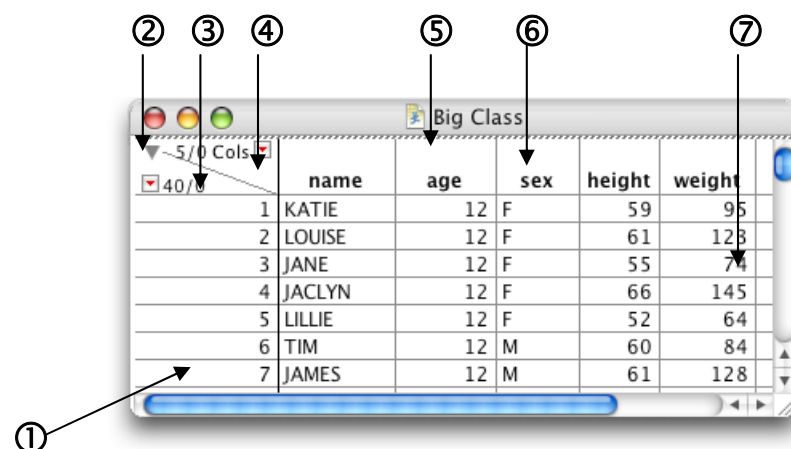
Notice that the data for the second group of subjects is placed directly below the data for the first group rather than adjacent to it as one often sees in a textbook. Recall that, in a *JMP* data table, the rows are the *individuals* and the columns are the *variables*.

## Remarks

- *JMP* automatically detected that **Group** should be given a data type of character.
- Default choices for **End Of Field** and **End Of Line** *delimiters* and other import settings can be changed in the preferences panel. See Section 0.4.2 for details.

## 0.3 Working with Variables and Individuals

### 0.3.1 Selecting Individuals and Columns



Select the area marked:

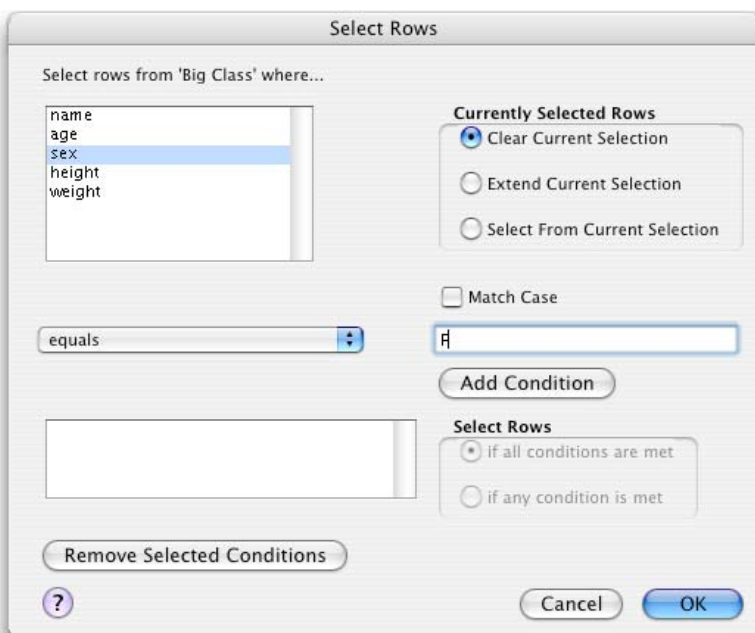
- ① to select a row/individual
- ② to open the left side panels
- ③ to deselect all rows/individuals
- ④ to deselect all columns/variables
- ⑤ to select a column/variable
- ⑥ to change the column/variable name
- ⑦ to edit a cell

After selecting an individual (or a variable), press the:

- **Shift** key to select a block of adjacent individuals (or variables).
- **Ctrl** (command on the Macintosh) key to select nonadjacent individuals (or variables).

You can also use the **Row Selection** command in the **Rows** menu and the **Select Where...** option to select rows/individuals that meet a criterion. Suppose that you wish to select the female students in the *JMP* data table **Big Class** that is stored in the Sample Data folder located with the *JMP* application.

1. Select **File** ⇒ **Open** ⇒ **Big Class**.
2. Select **Rows** ⇒ **Row Selection** ⇒ **Select Where...**



- a. Select **sex** in the list of columns.
- b. Select **equals** (the default) in the comparison menu.
- c. Type **F** as the value.
- d. Select **OK**.

	name	age	sex	height	weight
1	KATIE	12	F	59	95
2	LOUISE	12	F	61	123
3	JANE	12	F	55	74
4	JACLYN	12	F	66	145
5	LILLIE	12	F	52	64
6	TIM	12	M	60	84
7	JAMES	12	M	61	128
8	ROBERT	12	M	51	79
9	BARBARA	13	F	60	112
10	ALICE	13	F	61	107
11	SUSAN	13	F	56	67
12	JOHN	13	M	65	98
13	JOE	13	M	63	105
14	MICHAEL	13	M	58	95
15	DAVID	13	M	59	79
16	JUDY	14	F	61	81

### 0.3.2 Changing the State of an Individual

There are times when we wish to:

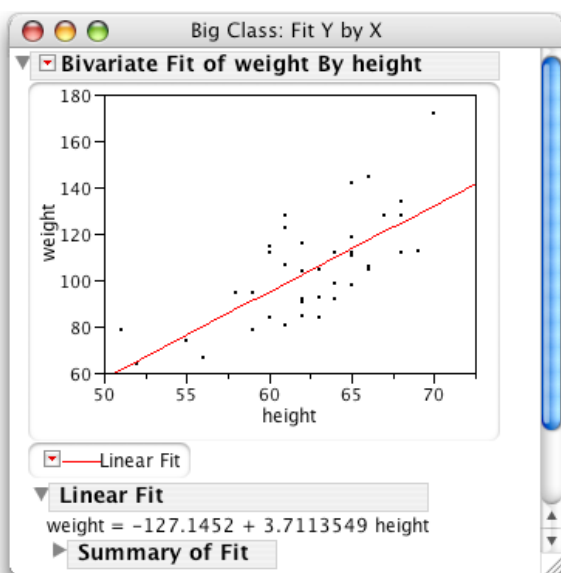
- *exclude* one or more individuals from analysis.
- *hide* one or more individuals in a plot.
- *color* points representing one or more individuals in a plot.

These tasks are easily accomplished in *JMP* by changing the *state* of an individual. The operations involve individuals so commands to perform them are found on the **Rows** menu.



To illustrate these features, we will first produce an analysis with all individuals in their default states.

1. Deselect all individuals in **Big Class** by selecting **Rows** ⇒ **Clear Row States**.
2. Select **Analyze** ⇒ **Fit Y by X**.
  - a. Select the column **weight** and click **Y, Response**.
  - b. Select the column **height** and click **X, Factor**.
  - c. Press **OK**.
3. Click on the red triangle next to **Bivariate Fit of ...** and select **Fit Line** from the menu that opens.



Now change the state of some individuals. Let's color the female students red and change their plotting symbol.

4. a. Use the **Rows Selection** command in the **Rows** menu to select the female students as we did above in Step 2.
- b. Select **Rows** ⇒ **Colors** ⇒ **red**; then select **Rows** ⇒ **Markers** ⇒ **6**.

Let's hide all 12-year-olds in the scatterplot.

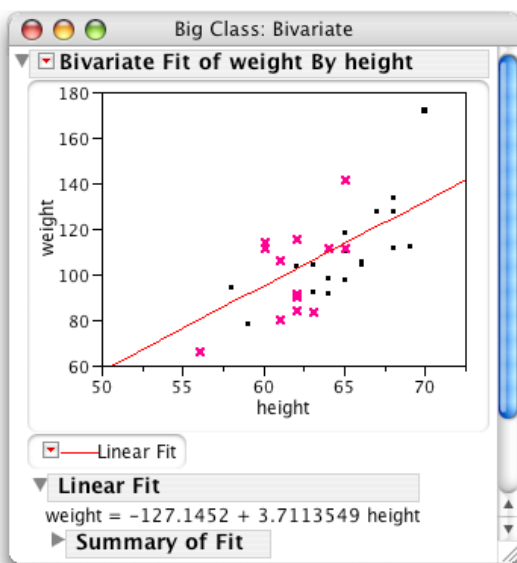
5. Select rows where the age is 12 by using the **Rows Selection** command again.
  - a. Select **Rows** ⇒ **Row Selection** ⇒ **Select Where...**
  - b. Select **age** in the list of columns.
  - c. Type **12** as the value and press **OK**.
6. Select **Rows** ⇒ **Hide/Unhide**.

Finally, we will set individual 40, Lawrence, to be excluded from future analyses. Lawrence is the tallest and heaviest student.

7. Select row **40** and then **Rows** ⇒ **Exclude/Unexclude**.

The data grid updates to indicate all of the new row states.

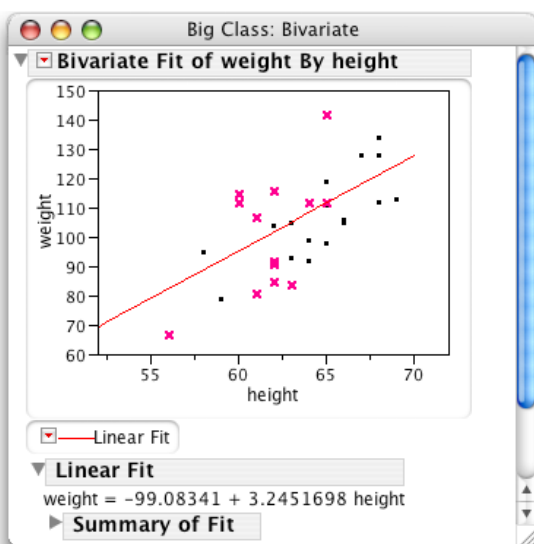




Move the cursor to the point in the upper right-hand corner. Notice that Lawrence is still there. To see the effects of changing the state of Lawrence to exclude, we must redo the analysis with Lawrence in that state. Let's use a shortcut to do that.

9. Click on the red triangle next to **Bivariate Fit....**
  - a. Select **Script** ⇒ **Redo Analysis** on the menu that opens.

Notice that Lawrence is no longer in the scatterplot and the equation of the line is quite different.



### 0.3.3 Creating a New Variable Using a Formula

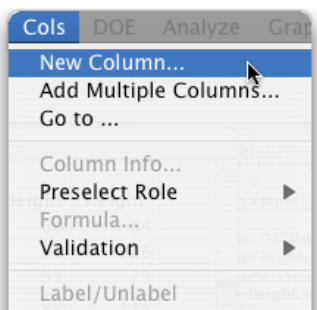
Sometimes we may wish to add or subtract the values of several variables for each individual. For example, we may want to look at the differences between before and after values in a study. Often, a

statistician needs to re-express a variable. For example, the square root of the variable amount might follow a more recognizable pattern than the amounts themselves. Or the relationship of the logarithm of the dose of a new medication and the clinical response may be simpler to understand and describe than the dose-response relationship. At other times, we may wish to randomly generate data from a particular distribution. In each case, we need to construct a new variable from one or more existing variables or from a mathematical or statistical function. The Formula Editor in *JMP* is a powerful and easy-to-use tool for doing these tasks and more. We illustrate its use with a simple example.

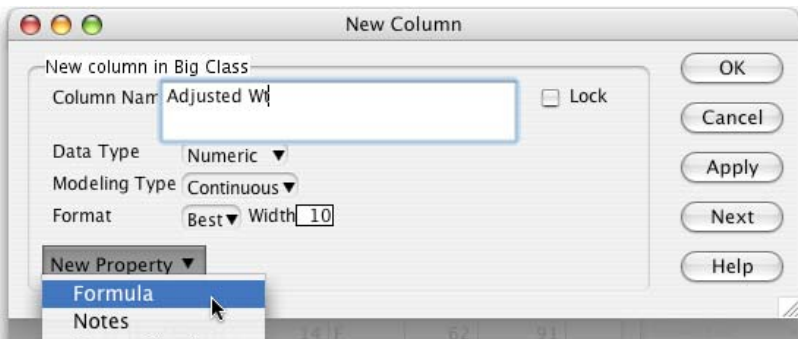
### Example 0.4 Finding the ratio of the weight to height of a child

The *JMP* data table **Big Class** contains the heights and weights of 40 teens and preteens. It is likely that their heights and weights are related. We might wonder, then, if the ratio of a child's weight to his or her height is relatively constant. To investigate this, we decide to construct a new variable—**weight** divided by **height**.

1. Select **File** ⇒ **Open** and select **Big Class.jmp** from the Sample Data folder that comes with the *JMP* application.
2. Select **Cols** ⇒ **New Column**.



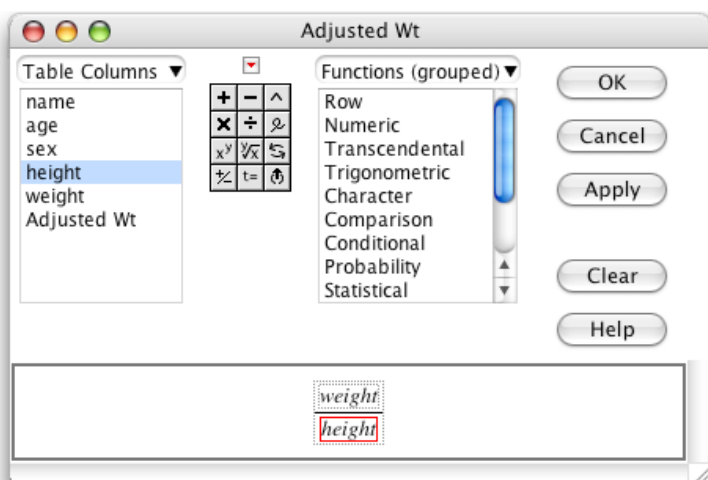
3. Name the column **Adjusted Wt** and select a format of **Fixed Dec** with **2** decimal places.



4. Select **Column Properties** ⇒ **Formula** to open the Formula Editor window.

We now build the formula to calculate the variable **Adjusted Wt**.

- a. Press  $\div$  on the keypad and select **weight** from the list of columns.
- b. Select the denominator of the ratio and then select **height** from the list of columns.
- c. Press **Apply**. Look at the data table.



The values of the ratio **Adjusted Wt** range from about 1.20 to 2.50 lbs/inch.

	name	age	sex	height	weight	Adjusted Wt
1	KATIE	12	F	59	95	1.61016949
2	LOUISE	12	F	61	123	2.01639344
3	JANE	12	F	55	74	1.34545455
4	JACLYN	12	F	66	145	2.1969697
5	LILLIE	12	F	52	64	1.23076923
6	TIM	12	M	60	84	1.4
7	JAMES	12	M	61	128	2.09836066
8	ROBERT	12	M	51	79	1.54901961
9	BARBARA	13	F	60	112	1.86666667
10	ALICE	13	F	61	107	1.75409836

## Remark

Look at the rich list of functions. Some are used in later chapters.

1. Select **Transcendental** from the **Functions (grouped)** List.
2. Also, scroll down the **Functions (grouped)** list and select **Random**.





Transcendental Functions



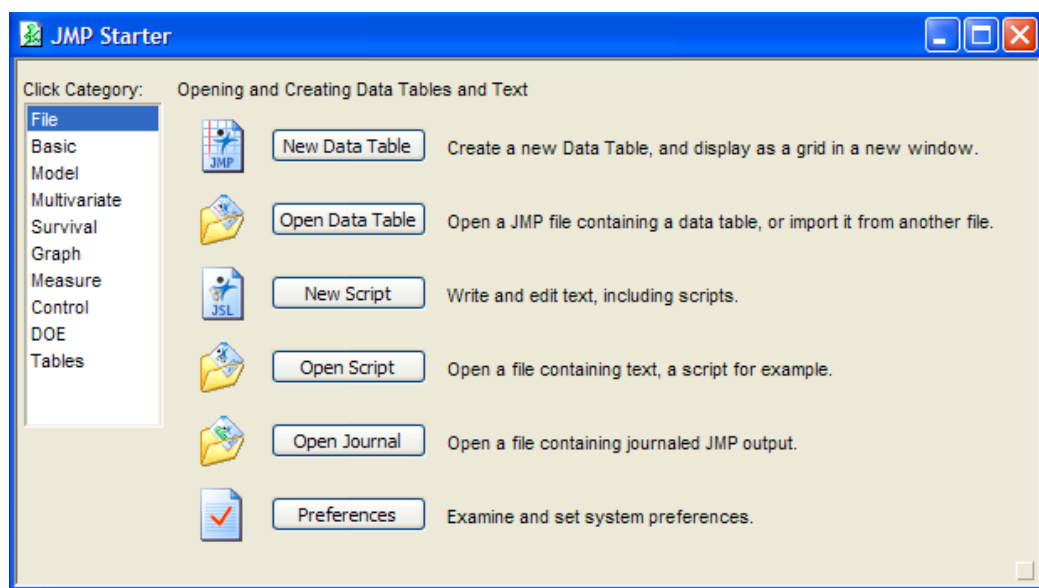
Random Functions

## 0.4 Customizing Your *JMP* Environment

This section discusses the *JMP* Starter window, table information panels, and setting preferences. We make specific recommendations for and show you how to customize your session environment.

### 0.4.1 The *JMP* Starter Window

If a *JMP* data table is not selected before starting, *JMP* may begin by opening a special navigation window—the *JMP Starter*.



The *JMP* Starter window presents an alternate way to access *JMP* commands that we will not use. All these commands are accessible through the menu bar. We recommend that the Starter window be closed at startup. See “Setting Preferences” below for details.

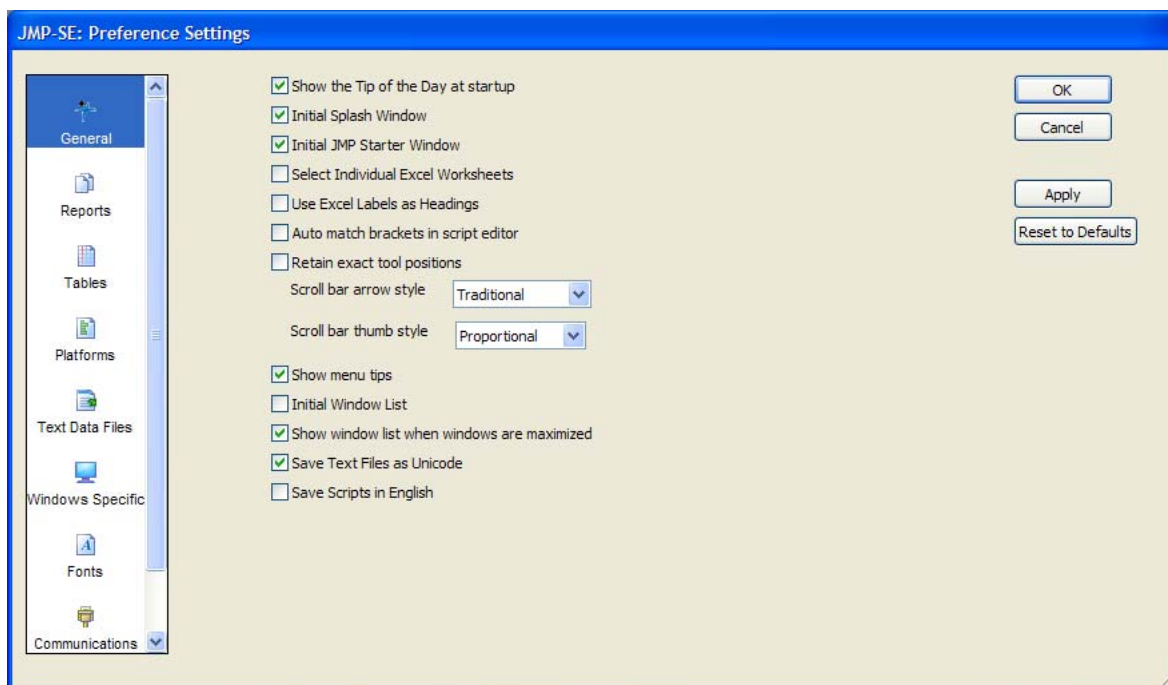
## 0.4.2 Setting Preferences

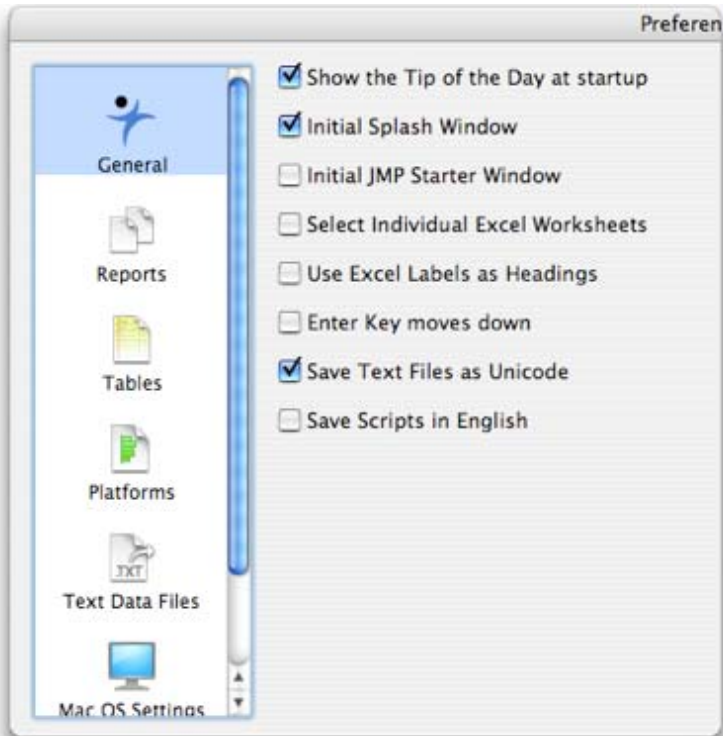
You may customize your session environment using the **Preferences** command. The **Preferences** command displays a panel with tab pages. The following displays the preferences panel for the Windows environment. Most preferences are also available on the Macintosh.

### Example 0.5 Setting preferences

Let's set preferences to close the Starter Window and the table information panels at startup.

1. To open the preferences window, select **File** ⇒ **Preferences** under Windows, or **JMP** ⇒ **Preferences** in Mac OS X.





2. Select the **General** icon.
  - a. Deselect **Initial JMP Starter Window** and press **OK**.

Open the preferences panel again and look at the other icons. Most options are either off or on. Check the items that you want or select from a menu of items. You can press **Apply** to see the results without closing the **Preferences** window.

# Chapter 1

## Looking at Data: Exploring Distributions

This chapter examines the distribution of a variable. Appropriate graphical and semigraphical methods for *displaying a distribution* are presented; numbers that describe the *location and spread of a distribution* are discussed; and *models (density curves) for a distribution* are presented along with methods of assessing the quality of their fit.

### 1.1 Displaying Distributions with Graphs

Almost all graphs and statistical computations for this chapter are performed in the **Distribution** platform of the **Analyze** menu. There is no need to tell *JMP* that you want a bar chart or a histogram. *JMP* automatically produces the appropriate graph and numeric summaries depending on the type of variable. For a categorical or nominal variable, *JMP* produces bar charts and a frequency table. For a continuous, quantitative variable, it produces a histogram and calculates the five-number summary, the mean, and standard deviation.

#### 1.1.1 Categorical Variables: Bar Graphs and Pie Charts

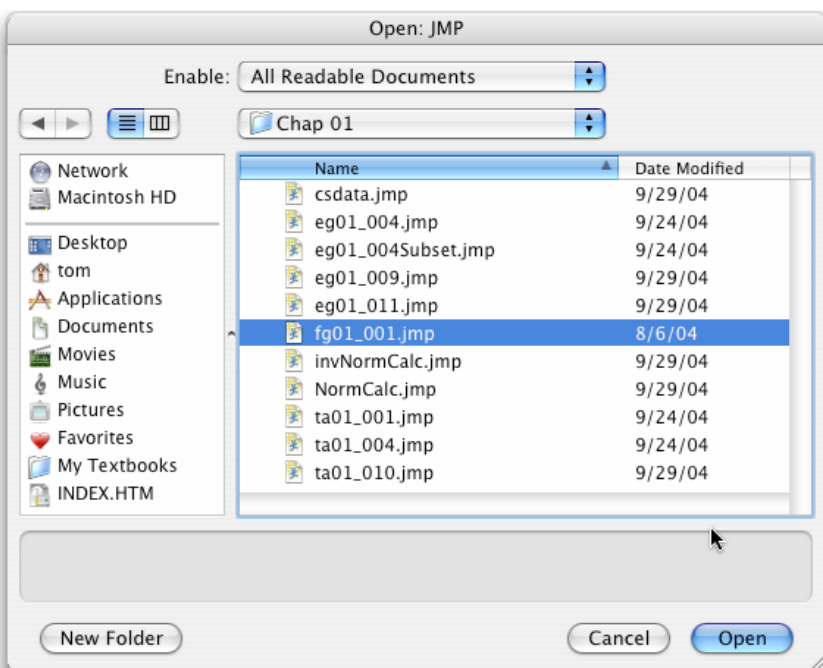
---

**Example 1.1 Educational level of 30-something adults**

---

Suppose that the *JMP* data table **fg01\_001.jmp** contains the highest educational level data of adults aged 25 to 34 years. Create a bar graph to display the distribution of educational levels.

1. To open the data table, select **File** ⇒ **Open** in the menu bar.



2. Select the data table **fg01\_001.jmp** and press **Open**.

fg01\_001.jmp

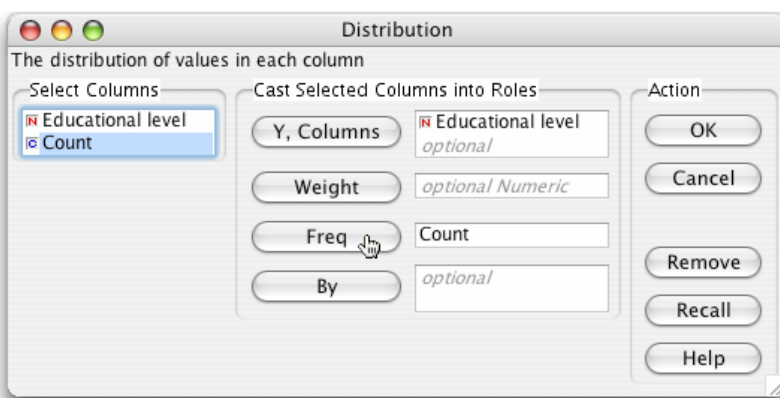
2/0 Cols

6/0 Rows

	Educational level	Count
1	Not HS Grad	4600
2	HS grad	11600
3	Some college	7400
4	Associate	3300
5	Bachelor's	8600
6	Advanced	2500

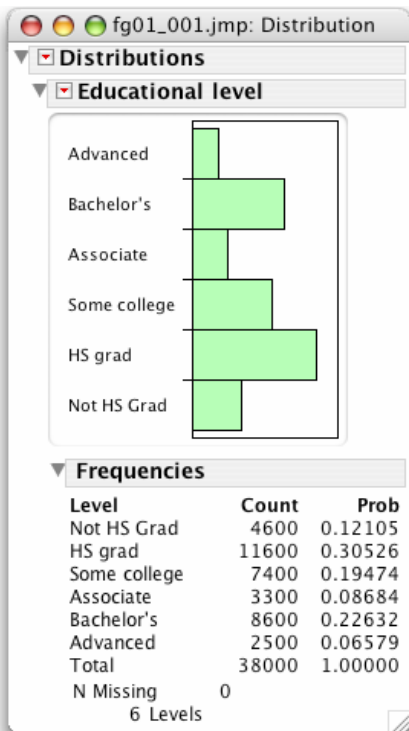
To display the distribution of educational levels, use the **Distribution** analysis platform.

3. Select **Analyze** ⇒ **Distribution** from the menu bar.



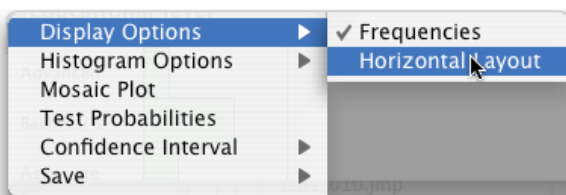
- a. Select the column **Educational level** and press the **Y, Columns** button.
- b. Select the column **Count** and press the **Freq** button.
- c. Select **OK**.

*JMP* displays a frequency table, a regular bar graph, and a stacked bar graph to quickly compare the proportion of adults in each of the six educational levels. (The ordering of the categories has been changed for clarity. See the third remark of Section 2.5 of Chapter 2 for details on changing the ordering.)

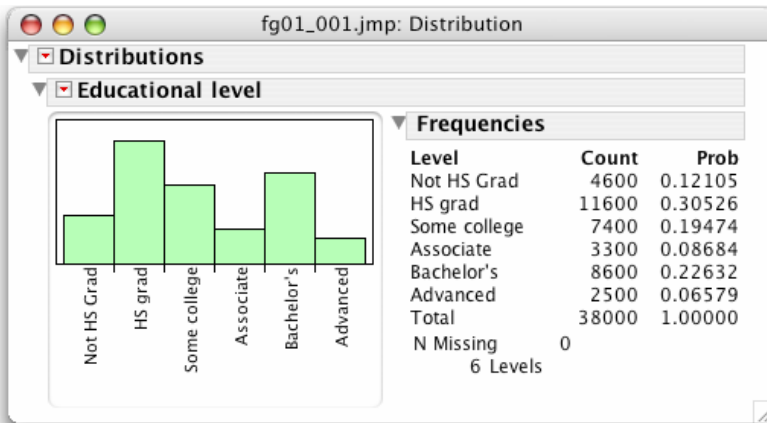


To display the bar graph horizontally as in examples in your textbook:

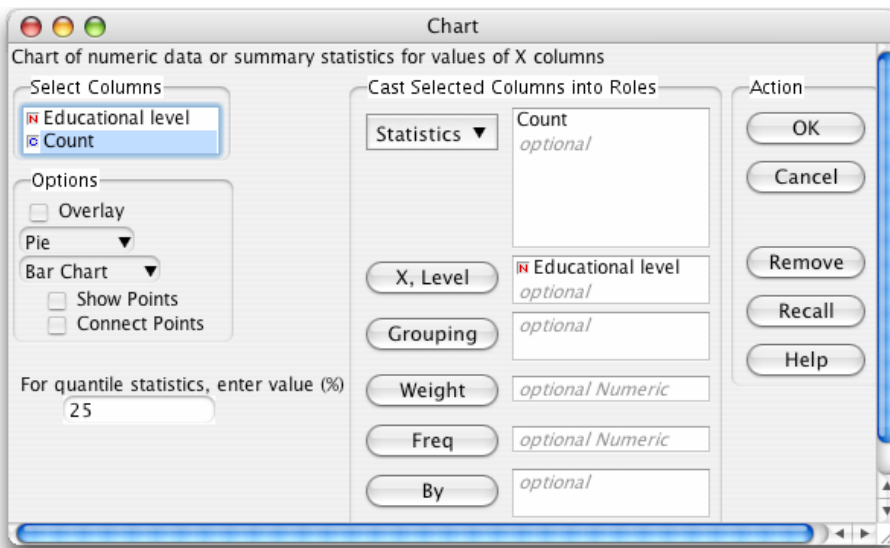
4. Click on the red triangle in the title of the **Educational level** report.
  - a. Select **Display Options** ⇒ **Horizontal Layout** from the menu that opens.



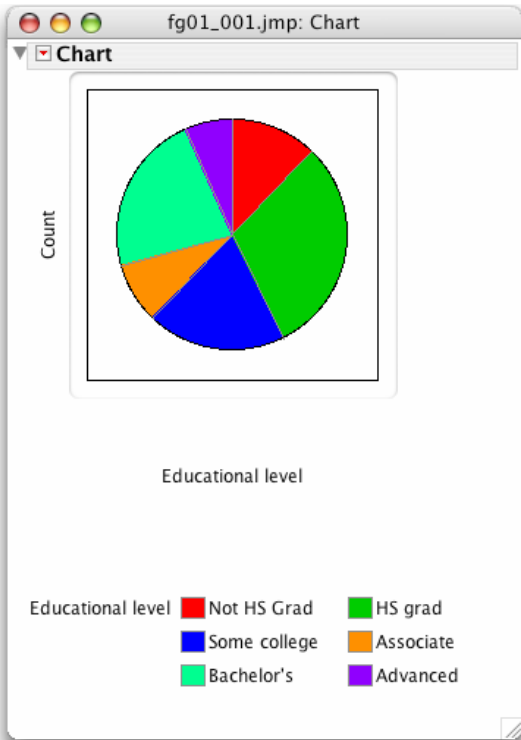
Notice that the levels of education are easier to read in the vertical layout. This is why it is the default layout.



To obtain a pie chart of the levels of education, use pie option on either the **Chart** or **Pareto Plot** commands on the **Graph** menu. Let's use the **Chart** command.



5. Select **Graph** ⇒ **Chart**.
  - a. Select the column **Educational level** and press the **X, Level** button.
  - b. Select the column **Count** and press the **Statistics** ⇒ **Data** button.
  - c. Select **Pie** from the pull-down menu titled **Vertical** under **Options**.
  - d. Select **OK**.



### 1.1.2 Quantitative Variables: Histograms and Stemplots

For quantitative variables, the values of the variable must be grouped together to see the distribution of the values. A *histogram* and a *stemplot* are used to display the distribution of values for a quantitative variable. Both are extremely easy to produce in *JMP*.

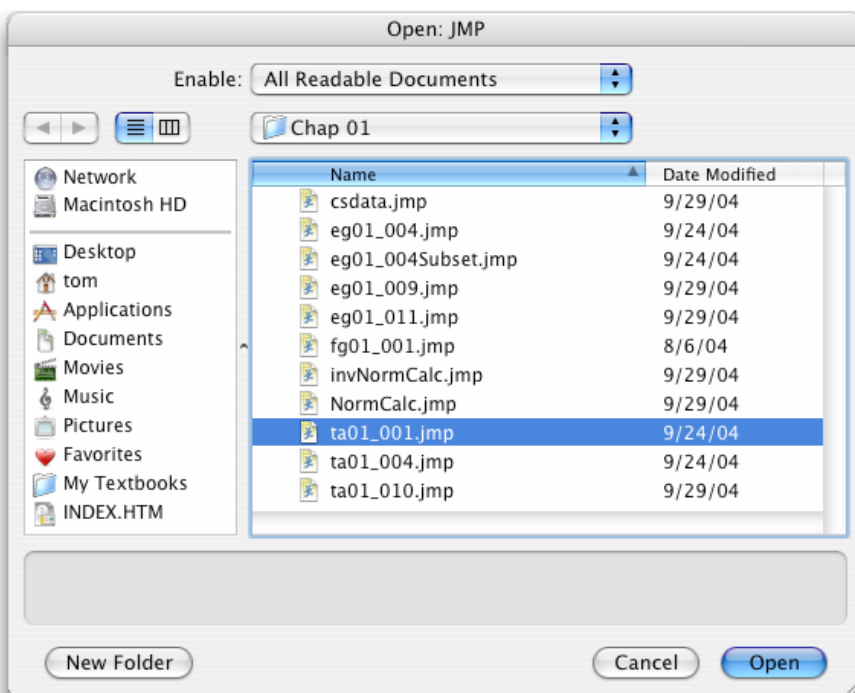
#### Histograms

##### Example 1.2 Service times (seconds)

Suppose that Table 1.1 of your textbook presents the lengths of the first 80 calls to a customer service center of a small bank in a month. The *JMP* data table **ta01\_001.jmp** available on the textbook CD will contain the data values. Let's create a histogram to examine the distribution of the length of the calls.

1. Select **File** ⇒ **Open** in the menu bar to open the data table.





2. Select the file **ta01\_001.jmp** and press **Open**.

	length
1	77
2	289
3	128
4	59
5	19
6	148
7	157
8	203
9	126
10	118

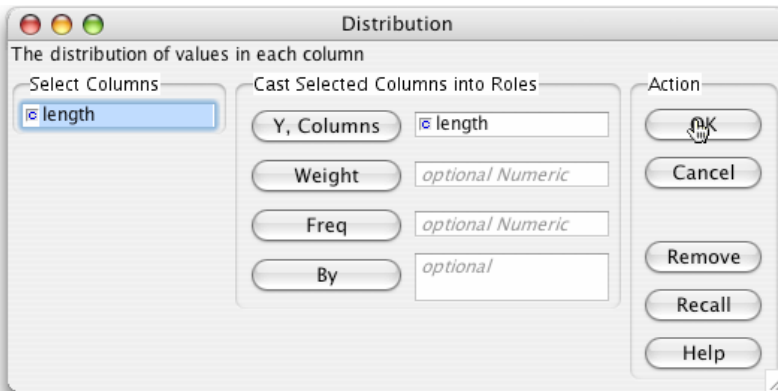
Suppose that we do not wish to include calls longer than 1200 seconds (20 min.) in our histograms. Then, we need to exclude from analysis call 29, the only one in our sample with length over 1200 seconds. To do this, we change the state of row 29. (See Section 0.3 of Chapter 0.)

3. a. Select row **29** in the data table.  
 b. From the menu bar, select **Rows** ⇒ **Exclude/Unexclude**.

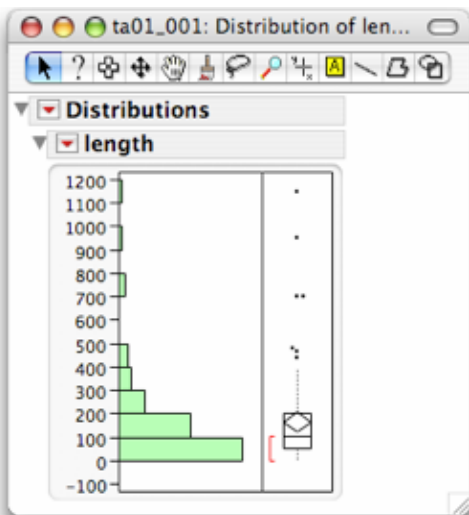
	length	
25	179	
26	1	
27	68	
28	386	
29	2631	
30	90	
31	30	
32	57	
33	89	
34	116	

Now, let's display the distribution of the length of the calls. We use the first analysis platform, **Distribution**.

4. Select **Analyze** ⇒ **Distribution**.



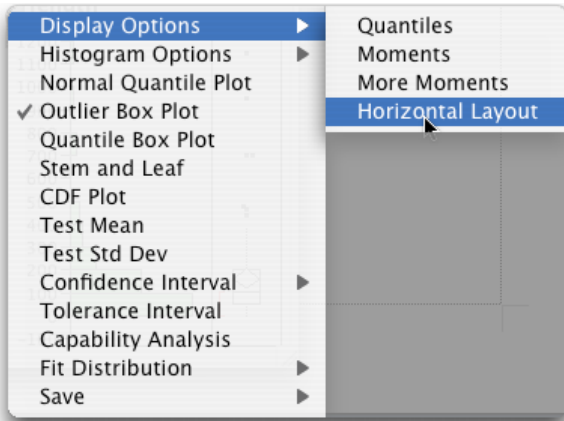
a. Select the column **length** and press the buttons **Y, Columns** and **OK**.



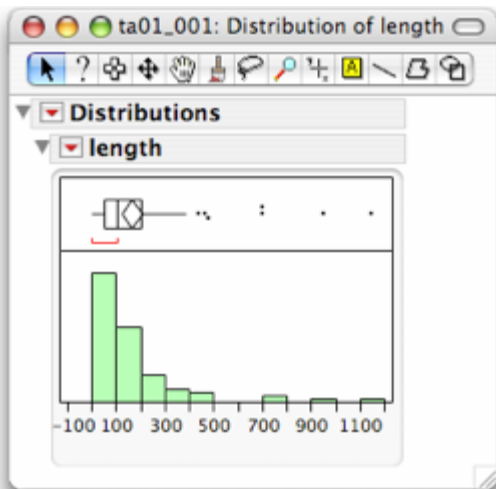
Notice that the lengths of the calls are skewed toward higher values.

By default, *JMP* displays the histogram values on a vertical scale. You may wish to see the more familiar horizontal layout.

5. Click on the red triangle in the title of the **length** report.

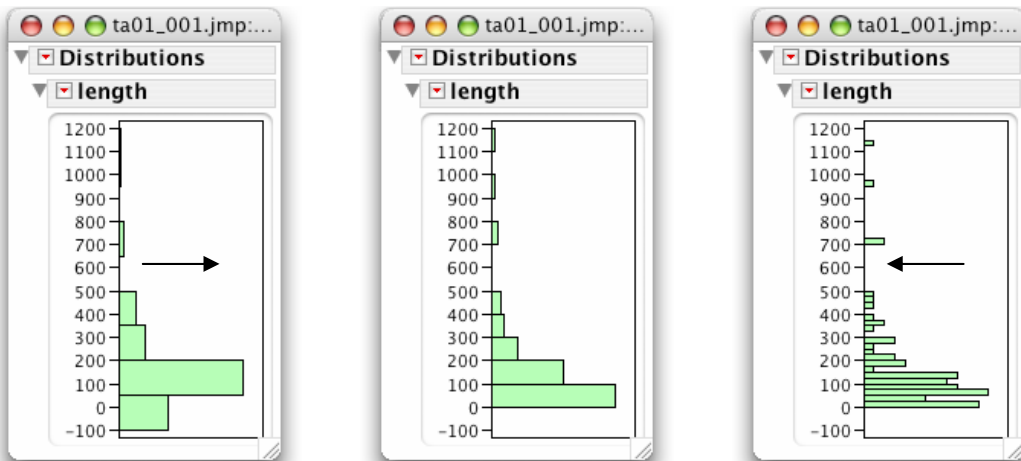


- a. Select **Display Options** ⇒ **Horizontal Layout**.



**Adjusting the Number of Bars.** Usually, we investigate whether another choice of interval width and starting point provides more insight into the overall pattern or into the deviations from the pattern. That is very easy to do in *JMP* using the *grabber* (hand) tool. First, switch back to the default vertical layout.

1. a. Click on the red triangle in the title of the **length** report.  
b. Deselect **Display Options** ⇒ **Horizontal Layout** to uncheck the horizontal layout option.
2. Now select the grabber (hand) tool from the **Tools** menu or from the Toolbar.
3. Move the mouse in the direction of the frequency scale (side-to-side for the default vertical layout). Notice that the number and width of the bars, or intervals, changes.



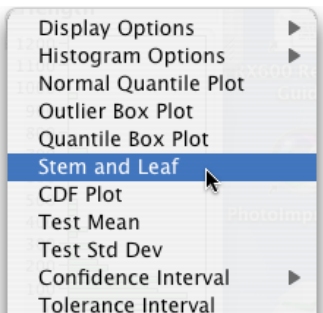
4. Move the mouse up and down along the values axis. Notice that the starting point of the intervals changes.
5. Experiment with this. Notice that the histogram on the right reveals the large number of short calls.

## Stemplots

### Example 1.2 Services times (cont'd.)

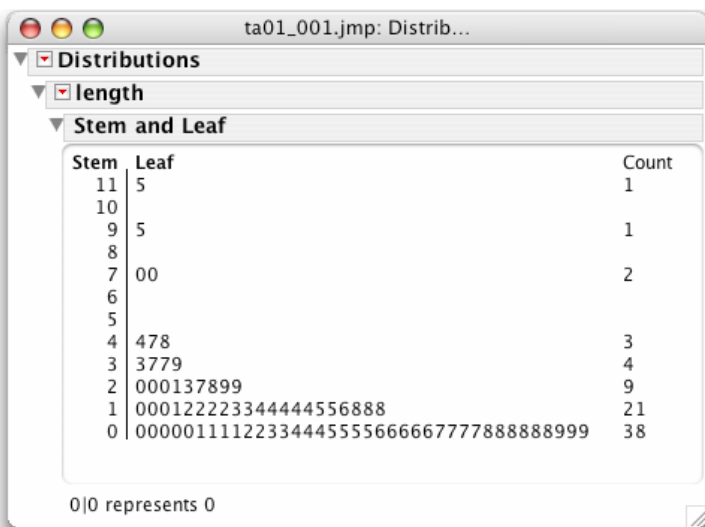
To display a stemplot, select the **Stem and Leaf** command from the menu on the report title bar.

1. Click on the red triangle in the title of the **length** report.



2. Select **Stem and Leaf** from the pop-up menu.

You may need to scroll down to the bottom of the report to see the stemplot. Or you can click on the blue icons (closure/disclosure buttons) for **Quantiles** and **Moments** to close those reports.



Stem	Leaf	Count
11	5	1
10		
9	5	1
8		
7	00	2
6		
5		
4	478	3
3	3779	4
2	000137899	9
1	0001222334444556888	21
0	0000011112233444555666667777888888999	38

0|0 represents 0

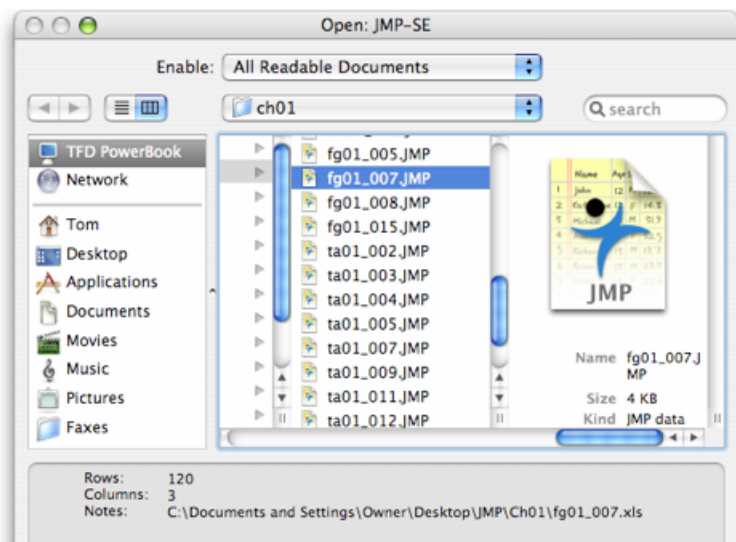
### 1.1.3 Time Plots

To create a time plot, use the **Time Series** command in the **Analyze** menu.

#### Example 1.3 The price of oranges

The average price of fresh oranges is collected each month as part of the government's reporting of retail prices. Suppose that a *JMP* data table **fg01\_007.jmp** contains these prices over the decade from 1991 to 2000, recorded as an index number relative to the average price of oranges in the years 1982 to 1984.

1. Select **File** ⇒ **Open** in the menu bar to open the data table.



2. Select the file **fg01\_007.jmp** and press **Open**.

	Year	Month	Price of Oranges
7	1991	Jul	285.7
8	1991	Aug	298.8
9	1991	Sep	316.9
10	1991	Oct	272.1
11	1991	Nov	206.4
12	1991	Dec	186.8
13	1992	Jan	187.6
14	1992	Feb	178.7
15	1992	Mar	171.6
16	1992	Apr	166.4
17	1992	May	178
18	1992	Jun	188.8
19	1992	Jul	178.6
20	1992	Aug	180.6

3. Select **Analyze** ⇒ **Modeling** ⇒ **Time Series**.
  - a. Select **Price of Oranges** from the list of columns and press **Y, Time Series**.
  - b. Click **OK**.

Report: Time Series - Autocorrelations

Modeling a variable by its lagged values over time

Select Columns: Year, Month, Price of Oranges

Cast Selected Columns into Roles:

- Y, Time Series: Price of Oranges (optional Numeric)
- X, Time ID: (optional Numeric)
- By: (optional)

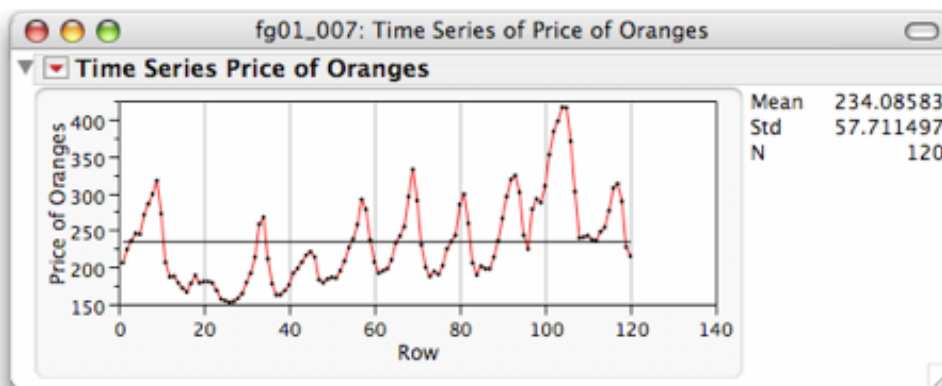
Autocorrelation Lags: 25

Forecast Periods: 25

Action: OK, Cancel, Remove, Recall

Data must be sorted by time, evenly spaced

*JMP* has automatically supplied a time variable with values of 1 through the number of rows, 120 in this case. Notice the gradual increase in prices over time (trend) and the repeating pattern (seasonal variation) in the plot.



## Remarks

- If the data is not listed in order of time but there is a time variable, then select the time variable and press **X, Time ID** after Step 3a above in the time series dialog window.
- See Chapter 18 to use *JMP* to identify the trend and seasonal variation in time series.

## 1.2 Describing Distributions with Numbers

All of the numeric summaries presented in *IPS* are found in the text reports of the **Distribution** platform of *JMP*.

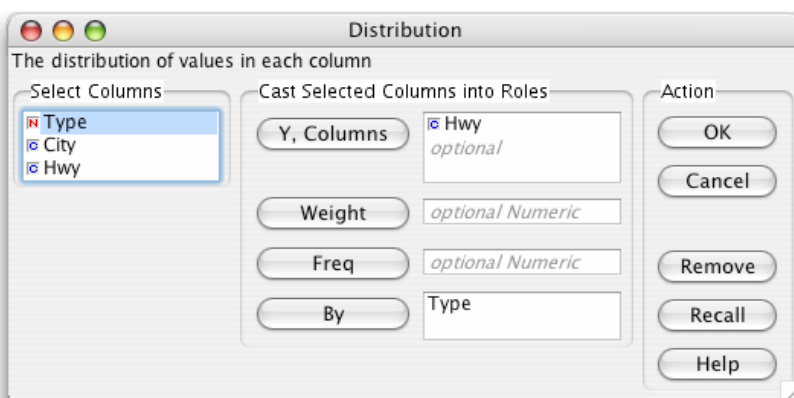
### Example 1.4 Highway mileage for sporty cars

Suppose that we have a *JMP* data table **ta01\_010.jmp** with the city and highway mileage for “two-seater” and “minicompact” cars. Each row of the *JMP* data table **ta01\_010.jmp** contains data for one car—the type, **Type**, city mileage, **City**, and highway mileage, **Hwy**. Let’s use the **Distribution** platform of *JMP* to calculate numerical measure of center and spread for the highway mileage of the two-seaters in the table.

1. Open the *JMP* data table **ta01\_010.jmp**.
  - a. Select **File** ⇒ **Open** in the menu bar.
  - b. Select the file **ta01\_010.jmp**.

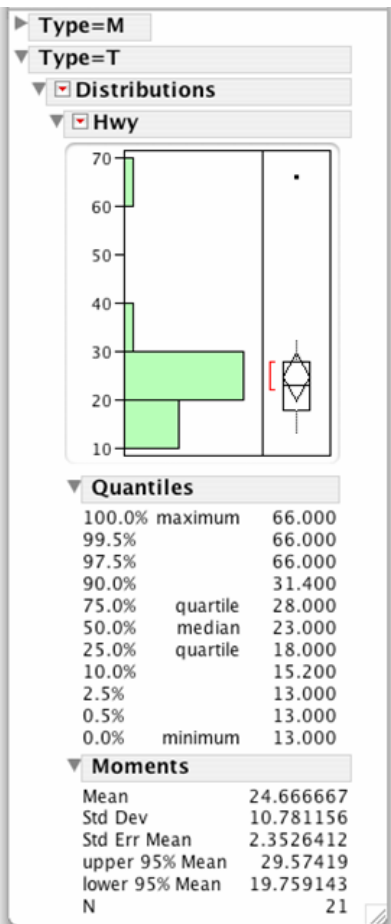
Since we later wish to compare the mileage for the two types of cars, we will use the **By** command in the **Distribution** dialog to get numerical summaries for each type of car.

2. Select **Analyze** ⇒ **Distribution**.
  - a. Select the column **Hwy** and press **Y, Columns**.
  - b. Select the column **Type** and press **By**.
  - c. Press **OK**.

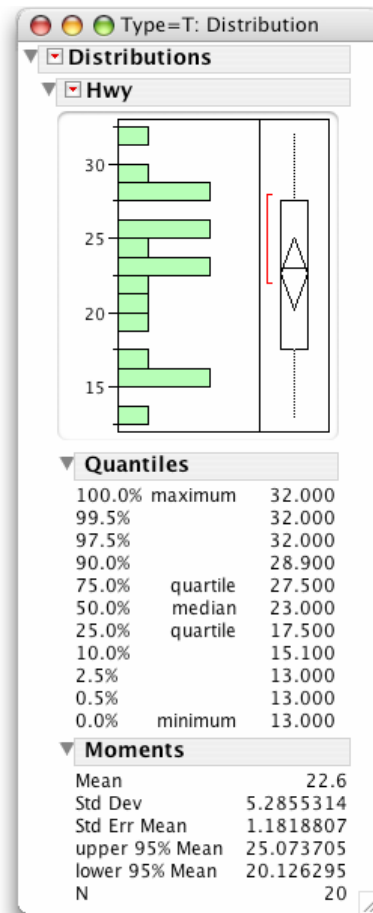


For now, we consider only the two-seaters.

3. Click on the disclosure button for the **Type=M** (minicompacts) button to hide the report for the minicompact cars.



All 23 two-seaters



Without the Honda Insight

The five-number summary is found in the **Quantiles** report, while the mean and standard deviation are found in the **Moments** report. The mean highway mileage for the 21 two-seaters is 24.67 miles per gallon while the median highway mileage is 23 mpg.

There is an outlier: The Honda Insight is a gas-electric hybrid car, while the others are gasoline-powered cars. To determine the effect of the Honda Insight on the mean and standard deviation, redo the analysis without the very high value 66 mpg.

4. Select **Window** ⇒ **ta01\_010.jmp** to bring the *JMP* data table forward.
  - a. Select **Row 10** with a **Hwy** mileage of 66.
  - b. Select **Rows** ⇒ **Exclude/Unexclude**.
5. Go the **ta01\_010.jmp: Distribution** report.
  - a. Press the red triangle on the **Distributions** title bar.
  - b. Select **Script** ⇒ **Redo Analysis** from the menu that opens.

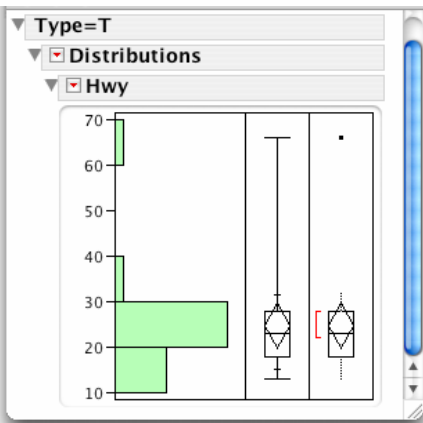
The single outlier reduces the mean by over 2 mpg while the median is unchanged.



## Boxplots

Notice that *JMP* has automatically displayed a modified boxplot of **Hwy mileage** in each of the preceding **Distribution** reports and placed it to the right of the histogram. These boxplots are an enhanced version of the type described in the textbook and are designed to help identify outliers. The whiskers in a modified boxplot are drawn to the last point within 1.5 times the interquartile range of the quartiles. Points beyond the distance are plotted individually. To display the regular boxplot for the report with the Honda Insight:

6. a. Click the red triangle in the **Hwy** report title.
- b. Select **Quantile Box Plot** from the menu that opens.

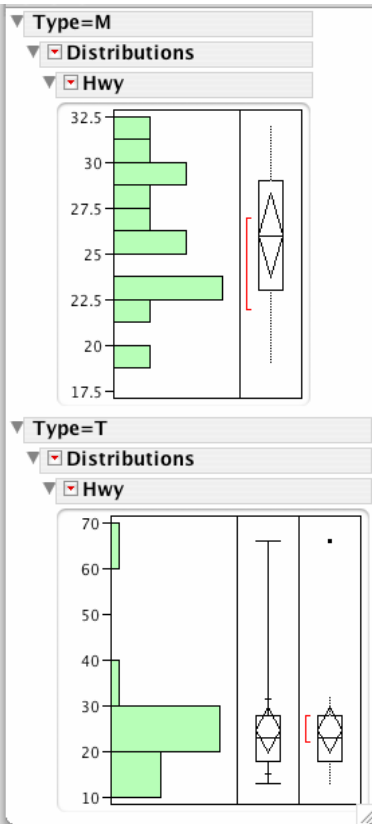


**Comparing Distributions.** Boxplots are very useful for comparing distributions. We can use the **By** command on the **Distribution** platform, or a different analysis platform entirely.

### Example 1.5 Highway mileage for two-seater and minicompact sporty cars

The **By** command computed boxplots for both types of cars. Open the **Type=M** report.

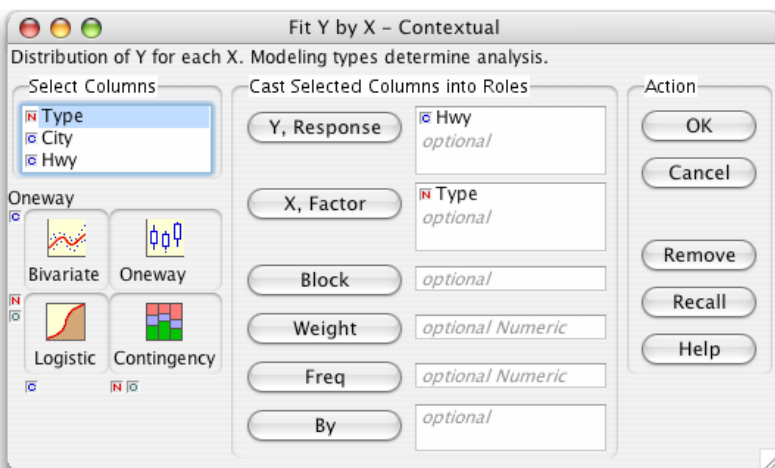
1. Click on the disclosure button for the **Type=M** report.

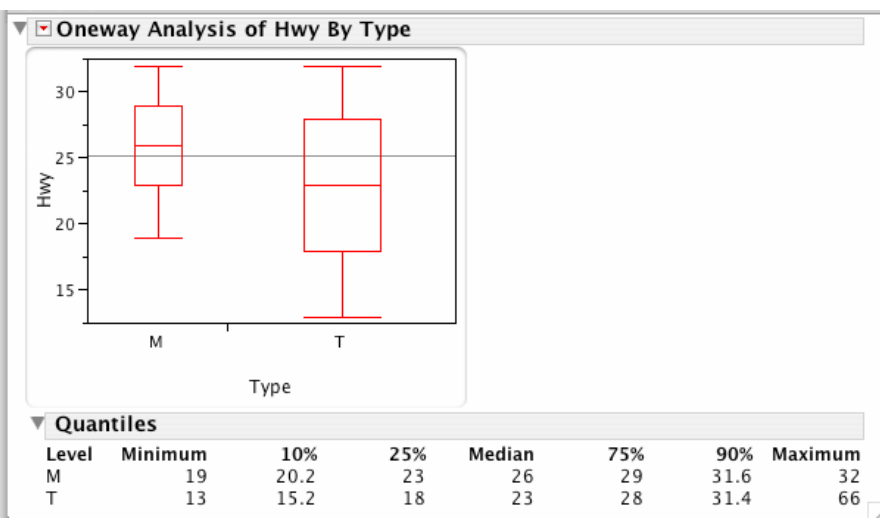


To put the two boxplots of highway mileage side-by-side, we use a different platform, **Fit Y by X**, which is explained more fully in Chapter 2.

2. Select **Analyze** ⇒ **Fit Y by X**.

- Select the column **Hwy** and click **Y, Response**.
- Select the column **Type** and click **X, Factor**, and **OK**.
- Press the red triangle in the **Oneway Analysis of Hwy by Type** report.
- Select **Quantiles** from the menu that opens.
- Deselect **Display Options** ⇒ **Points** to remove the points.





## Remark

- Some of the other commands on the red triangle menu in the **Distribution** analysis platform will be useful in later chapters on statistical inference.

## 1.3 Models for Distributions: Normal Distributions

### 1.3.1 Normal Distribution Calculations

You can use *JMP* to calculate areas under normal distributions and find percentiles of a normal distribution. Just put the values in one column and create another column of areas or percentiles using the appropriate normal function.

#### Example 1.6 Gas mileage

The average miles per gallon ratings for 2001 model vehicles vary according to an approximately Normal distribution with mean  $\mu = 21.22$  mpg and standard deviation  $s = 5.36$  mpg. What percent of 2001 model vehicles had mileage ratings between 30 and 35 mpg?

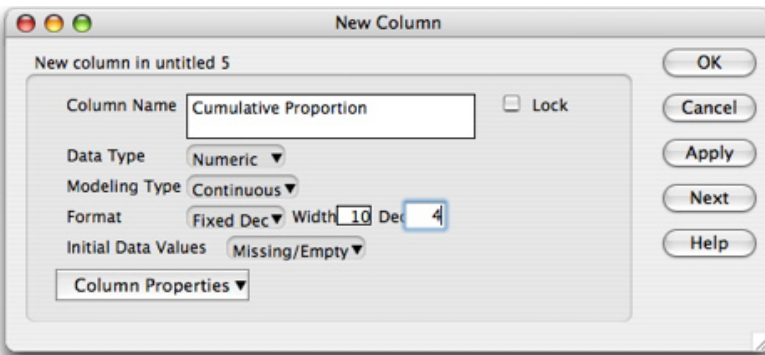
The Solution: In the following steps, we will create a new *JMP* data table with two columns—**Mileage** and **Cumulative Proportion**. Then, you will enter 35 and 30 in the first two rows of the first column. Finally, you will use the formula editor to create the corresponding areas under the normal density curve to the left of these values.

- Select **File**  $\Rightarrow$  **New**.
  - Name the first column **Mileage**. (See Section 0.2.1 in Chapter 0 for more detail.)
  - Select **Rows**  $\Rightarrow$  **Add Rows**.
  - Type 2 as the number of rows to add and press **OK**.
  - Type 35 in the first row and 30 in the second.

1	2
35	30
30	30

3. Select **Cols** ⇒ **New Column...**

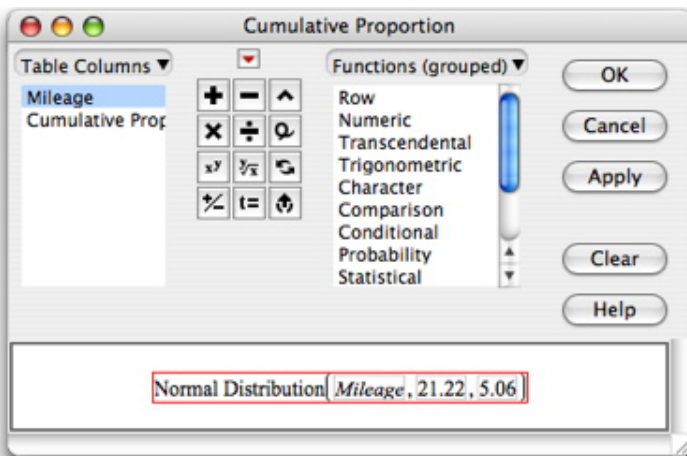
- a. Name the column **Cumulative Proportion** and select a format of **Fixed Dec** with **4** decimal places.



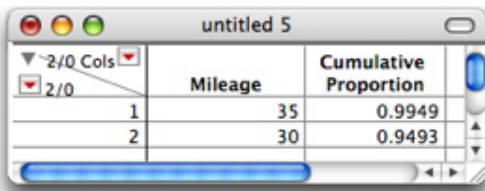
Now we let *JMP* calculate the cumulative proportion.

4. Select **Column Properties** ⇒ **Formula**.

- In the formula editor, select **Probability** ⇒ **Normal Distribution** from the list of functions.
- Select **Mileage** from the list of columns.
- Press the insert key ^ on the formula keypad, type **21.22**, press the insert key ^ again, and then type **5.36** to specify the normal distribution with mean 21.22 and standard deviation 5.36.
- Press **OK** and **OK** again in the **New Column** dialog.



From the resulting data table, we see that the proportion of 2001 model vehicles with miles per gallon ratings between 30 and 35 is then  $0.9949 - 0.9493 = 0.0456$ .



	Mileage	Cumulative Proportion
1	35	0.9949
2	30	0.9493

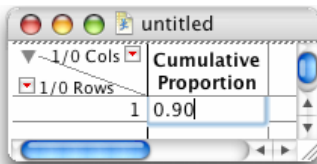
Suppose that you want to do the reverse. Find the observed value corresponding to a given relative frequency. This requires the use of the inverse function *normal quantile*.

### Example 1.7 Gas mileage: “backward” Normal calculations

Miles per gallon ratings of compact cars (2001 models) follow approximately the  $N(25.7, 5.88)$  distribution. How many miles per gallon must a vehicle get in order to place in the top 10% of all 2001 model compact cars?

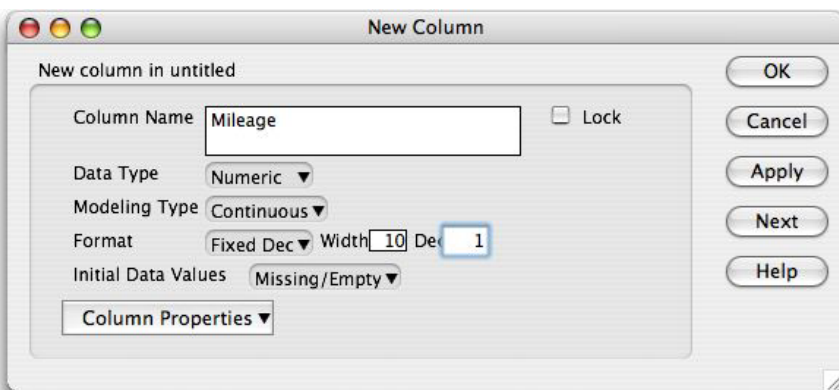
The Solution: You reverse the process this time. Create a column with the cumulative proportion below the mileage value and use a formula to calculate a column with the corresponding mileage value.

1. Select **File** ⇒ **New**.
2. Name the first column **Cumulative Proportion**.
  - a. Enter **.90** in the first row and first column of the data table.



Cumulative Proportion
0.90

3. Select **Cols** ⇒ **New Column...**
  - a. Name the column **Mileage** and select a format of **Fixed Dec** with **1** decimal place.



New Column

New column in untitled

Column Name:  ☐ Lock

Data Type:

Modeling Type:

Format:  Width  Dec

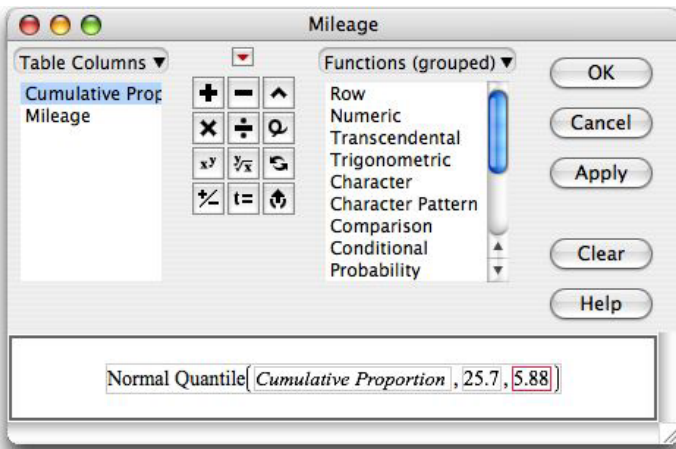
Initial Data Values:

Column Properties ▼

OK Cancel Apply Next Help

6. Select **Column Properties** ⇒ **Formula**.
  - a. In the formula editor, select **Probability** ⇒ **Normal Quantile** from the list of functions.

- b. Select **Cumulative Proportion** from the list of columns.
- c. Press the insert key ^ on the formula keypad, type **25.7**, press the insert key ^ again, and then type **5.88** to specify the normal distribution with mean 25.7 and standard deviation 5.88.
- d. Press **OK** and **OK** again in the **New Column** dialog.



	Cumulative Proportion	Mileage
1	0.9	33.2

From the table, we see that a compact car must get at least 33.2 mpg to place in the top 10%.

### 1.3.2 Assessing Normality: Normal Quantile Plots

*Normal quantile plots* are easily constructed in *JMP* using the **Normal Quantile Plot** command in the **Distribution** analysis platform.

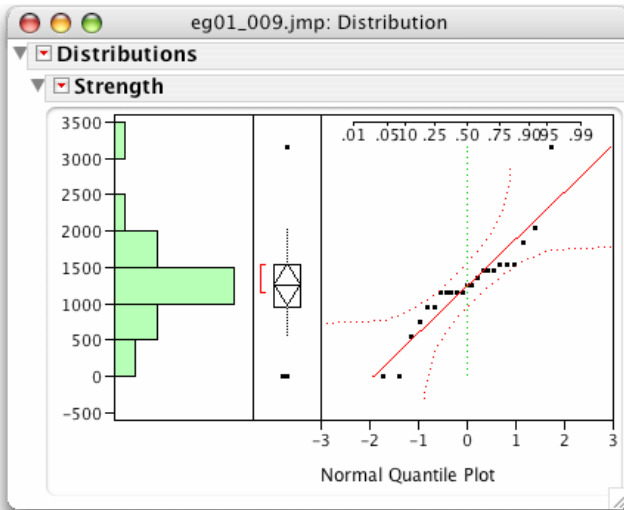
#### Example 1.8 Interpreting Normal quantile plots

Manufacturing an electronic component requires attaching very fine wires to a semiconductor wafer. If the strength of the bond is weak, the component may fail. The measurements on the breaking strength (in pounds) of 23 connections can be found in a file, say **eg01\_009.jmp**. Let's create a Normal quantile plot for the breaking strengths of the connections in the electronic components.

1. Open the *JMP* data table **eg01\_009.jmp**.
2. Select **Analyze** ⇒ **Distribution** from the menu bar.
  - a. Select **Strength** and press **Y, Columns** and **OK** in the dialog window.

Three measurements seem to be far from the others. To get a Normal quantile plot for this data,

- b. Press the red triangle in the title bar and select **Normal Quantile Plot**.



*JMP* enhances the Normal quantile plot with reference bands and a reference line. In general, if points fall outside the reference bands, it's not a good idea to use the Normal model to describe the distribution of a variable. Notice that most of the points in the Normal quantile plot lie close to the line. The high outlier at 3150 pounds deviates from that line substantially.

## 1.4 Summary

With one exception, all statistical graphs and computations in this chapter are performed using the **Distribution** command in the **Analyze** menu. The exception is a time plot. Use the **Time Series** command in the **Analyze** platform to display a time plot of the data.

# Chapter 2

## Looking at Data: Exploring Relationships

This chapter studies relationships between variables. Following an approach similar to Chapter 1, relationships are displayed with graphs; the strength of a linear relationship is described by a number; and, for two quantitative variables, straight lines are used as models for relationships.

All graphs and statistical computations in this chapter can be performed in the second platform **Fit Y by X** of the **Analyze** menu except for the calculation of correlations, which use the **Multivariate** platform.

### 2.1 Displaying Relationships with Graphs

A *scatterplot* displays the relationship between two quantitative variables. *Side-by-side boxplots* and *side-by-side means diamonds* display the relationship between a categorical explanatory variable and a quantitative response variable. In *JMP*, you specify the role (response or explanatory) and modeling type (continuous or nominal) of each variable and *JMP* automatically performs the appropriate methodology.

#### 2.1.1 Two Quantitative Variables: Scatterplots

*Scatterplots* are created whenever the **Fit Y by X** platform is called and both variables are quantitative.

##### Example 2.1 Retail shop sales

---

The owners of Duck Worth Wearing, a shop selling high-quality secondhand children's clothing, toys, and furniture, would like to know what they should expect gross sales to be when a particular number of items have been sold. The *JMP* data table **eg02\_01.jmp** contains retail sales data for April 2000.

To create a scatterplot:

1. Select **File** ⇒ **Open** and the file **eg02\_01.jmp** from the appropriate location.

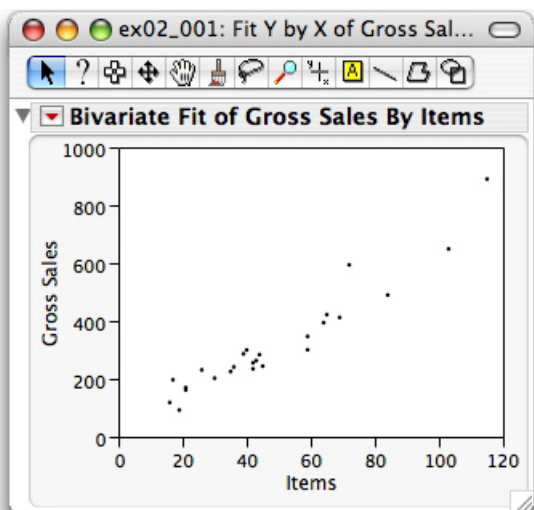


	Date	Gross Sales	Items	Gross Case	Cash Items	Gross Check	Check Items	Gross Credit	Credit Card Items	Day of Week
1	04/01/2000	890.5	115	348.2	55	394.3	56	148	4	Saturday
2	04/03/2000	197	17	42	8	44.5	3	110.5	6	Weekday
3	04/04/2000	231	26	61	9	108.5	10	61.5	7	Weekday
4	04/05/2000	170	21	94	16	76	5	0	0	Weekday
5	04/06/2000	202.5	30	59.5	11	104	14	39	5	Weekday
6	04/07/2000	225.5	35	164.5	26	54	8	7	1	Weekday
7	04/08/2000	489.7	84	125.7	27	220.6	31	143.4	26	Saturday
8	04/10/2000	234.8	42	110.8	19	97.5	18	26.5	5	Weekday
9	04/11/2000	161.5	21	26	5	122	14	13.5	2	Weekday
10	04/12/2000	284	44	109	18	104	14	71	12	Weekday

2. Select **Analyze** ⇒ **Fit Y by X**.

Because **Gross Sales** is the *response variable* and **Items** is the *explanatory variable*,

- Select the column **Gross Sales** and click **Y, Response**.
- Select the column **Item** and click **X, Factor**.
- Press **OK**.

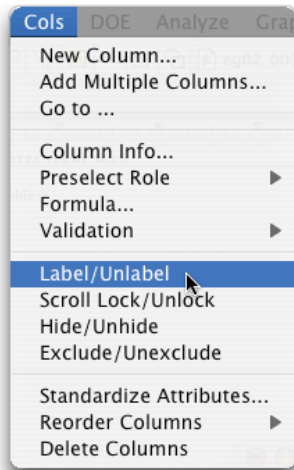


### Identifying Individuals on the Scatterplot

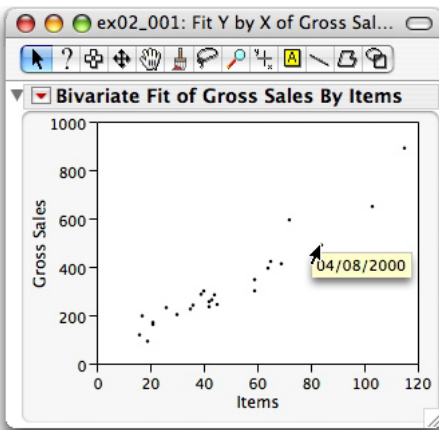
If you hover the cursor over a data point, the row number of the state that the point represents is displayed. We might prefer to display the name of the state instead. The **Label/Unlabel** command in the **Cols** menu tells *JMP* to use a column's values to identify points in plots.

#### Example 2.1 Retail shop sales (cont'd.)


- Select the column **Date** in the data table.
- Select **Cols** ⇒ **Label/Unlabel**.



Return to the **ex02\_01: Fit Y by X** window directly or by using the **Window** menu. Move the cursor over the data points. *JMP* now displays the value of the variable **Date**.



## Remark

- The scatterplot can be enhanced in several ways:
  - Increase (or decrease) the size of the plot by selecting a corner of the plot and dragging.
  - Scroll either axis by moving the hand tool over the numbers.
  - Modify tick marks and the increment between numbers by double-clicking on a scale data value.
  - Modify or enhance an axis name by double-clicking on the axis name.
  - Create editable notes to be displayed and stored with the plot using , the annotate tool.

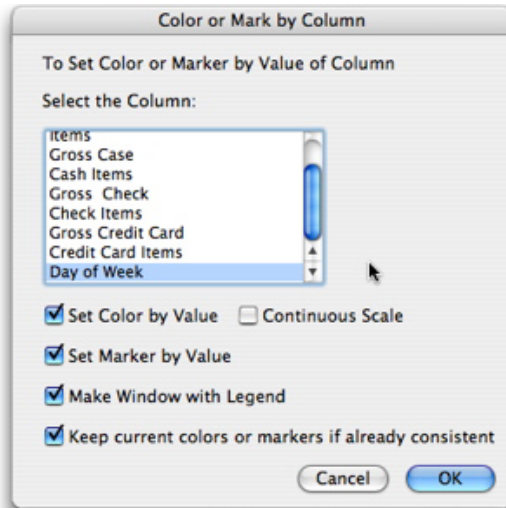
## Adding Categorical Variables to Scatterplots

### Example 2.2 Are Saturdays different?

Duck Worth Wearing is open Monday through Saturday. To compare weekdays with weekends, we might assign the points in the scatterplot associated with Saturdays different colors and symbols than weekdays. To do this, we change the *state* of the rows/individuals that are represented by points in the

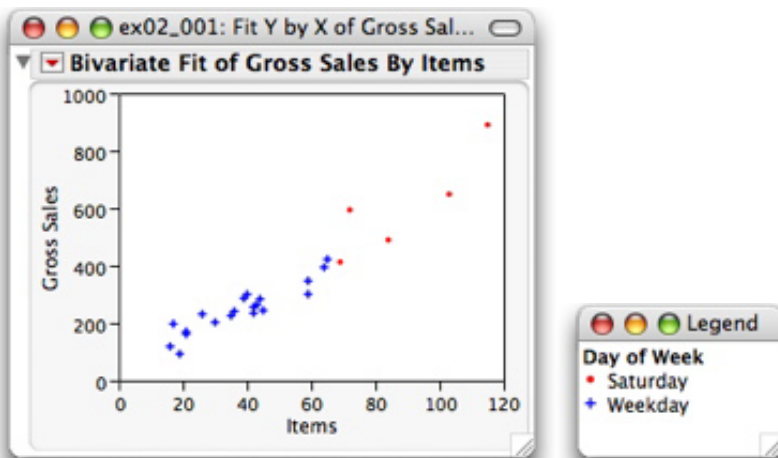
scatterplot. (See Section 0.3.2 in Chapter 0 for more details on row states.) We wish to color and mark the points on the graph differently depending on the values of a variable, or column—in this case, **Day of Week**. The **Color or Mark by Column** command will do this.

1. Select **Rows** ⇒ **Color or Mark by Column**.



- a. Select the column **Day of Week**.
- b. Check **Set Marker by Value** and **Make Window with Legend**.
- c. Press **OK**.

Return to the **ex02\_01: Fit Y by X** window and notice that the day of the week provides an explanation for part of the pattern that we observed.



## 2.1.2 A Categorical Explanatory Variable and a Quantitative Response Variable

To display a relationship between a categorical explanatory variable and a quantitative response variable, we make a side-by-side comparison of the distributions of the response for each category. Some graphs that *JMP* provides for such comparisons are:

- side-by-side boxplots.
- side-by-side point plots.
- side-by-side means diamonds.

The **Fit Y by X** platform in *JMP* automatically creates these displays.

### Example 2.2 Fuel economy for model year 2004 sporty cars

Suppose that you are interested in buying a sporty car but are concerned that it may use too much gas. Most sporty cars are listed by the EPA in two categories, “two-seater” and “minicompact.” You wish to compare the highway mileage of the two groups of vehicles. To put it another way, you wish to see if a car’s highway mileage is related, or depends, on the type of sporty car.

Suppose that the *JMP* data table **ta01\_01.jmp** contains the EPA gas mileage data for these cars.

1. Select **File** ⇒ **Open** in the menu bar to open the data table.
  - a. Select the file **ta01\_01.jmp** located on the CD-ROM.

	Type	City	Hwy
17	T	13	19
18	T	20	26
19	T	20	29
20	T	15	23
21	T	26	32
22	M	12	19
23	M	21	29
24	M	19	27

There are 34 cars and 3 variables describing each car.

### Remark

- Notice that the highway mileages for both groups of cars are listed in one column, **Hwy**, of the *JMP* data table and not two, one for each type of car. This is because each row of a *JMP* data table represents an individual—in this case, a car. Since each column in a *JMP* data table represents a variable, the highway mileages must be put into one variable, **Hwy**, and another variable, **Type**, must be used to identify the type of car. This is *very important to remember* since all statistical computations and graphs assume that the individuals are the rows of a data table and the variables are columns.

We restrict our analysis to gas-powered cars and exclude the hybrid Honda Insight.

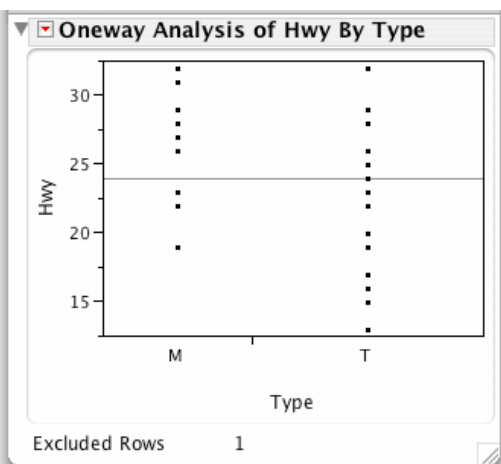
2. a. In the data table, select row **10**.
- b. From the menu bar, select **Rows** ⇒ **Exclude/Unexclude**.

To make a side-by-side comparison of the distributions of the highway mileage for each type of car, we use the **Fit Y by X** platform.

3. Select **Analyze** ⇒ **Fit Y by X**.

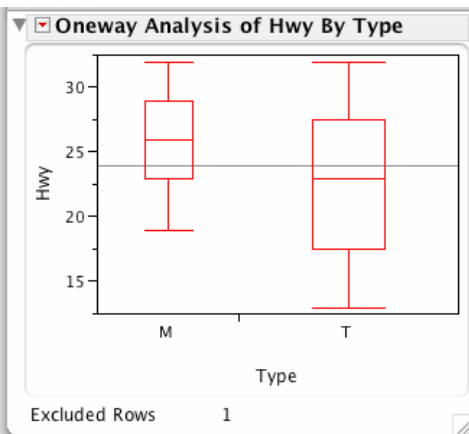
Since **Hwy** is the *response variable* and **Type** is the *explanatory variable*,

- a. Select the column **Hwy** and click **Y, Response**.
- b. Select the column **Type** and click **X, Factor**.
- c. Press **OK**.



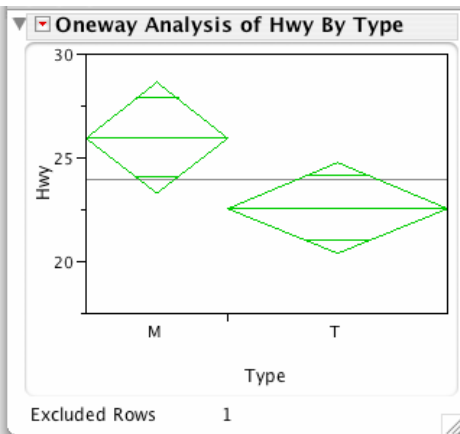
JMP presents *side-by-side point plots* by default. The horizontal line at 23.9 mpg is the overall mean highway mileage for these gas-powered cars. To get *side-by-side boxplots*,

3. Press the red triangle in the **Oneway Analysis of Hwy By Type** report.
  - a. Select **Display Options** ⇒ **BoxPlots** from the menu that opens.
  - b. Deselect **Display Options** ⇒ **Points** to remove the points.



When the mean and standard deviation are used to identify the center and spread of a distribution, *side-by-side means diamonds* are a better tool for comparing the distributions of a response variable among the categories of another variable. To display them:


4. From the red triangle menu in the title **Oneway Analysis of Hwy By Type**:
  - a. Deselect **Display Options** ⇒ **BoxPlots**
  - b. Select **Display Options** ⇒ **Mean Diamonds**



5. Double-click on a value on the **Hwy** (vertical) scale and change the scale increment and range so that your graph looks like the above.

In both displays, it can be seen that the minicompacts get better highway mileage on average than two-seaters.

### Remark

- We will use *side-by-side means diamonds* again in conjunction with the analyses discussed in Chapters 7, 12, and 13. To learn more about *means diamonds*, select the  tool and click on a diamond.

## 2.2 Describing Relationships with Numbers: Correlation

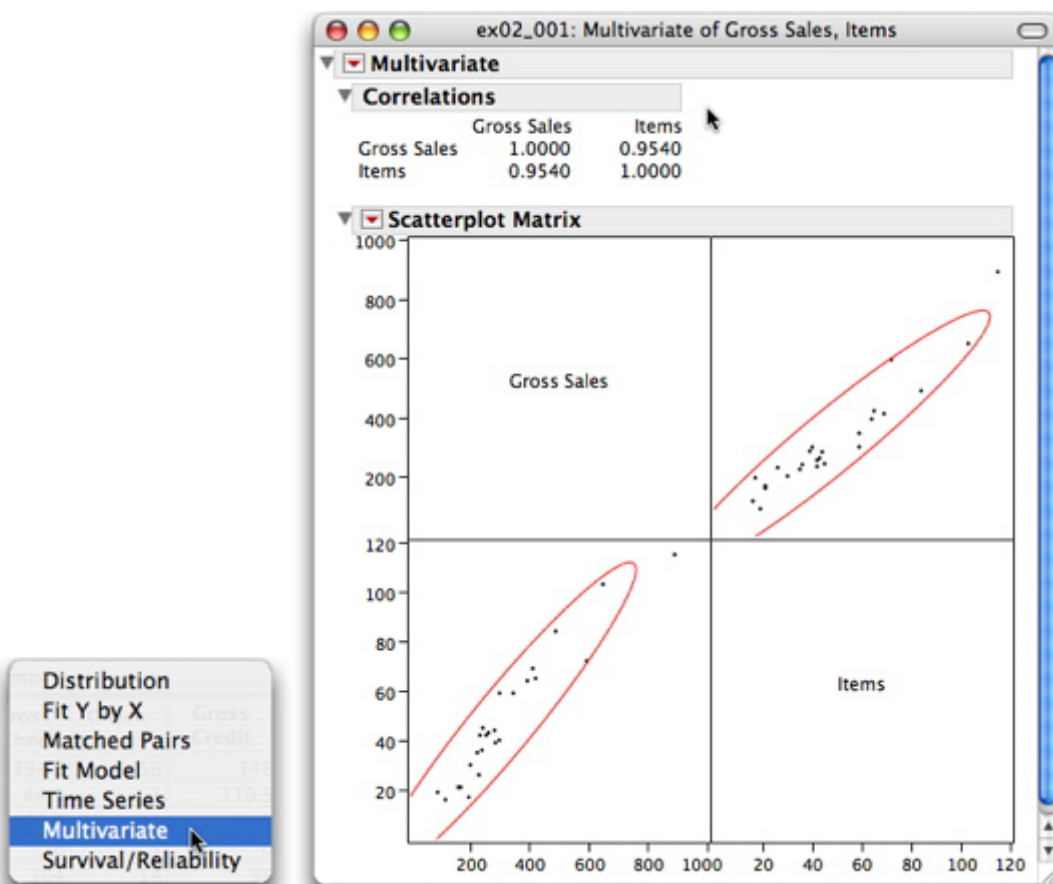
To find the *correlation* between two quantitative variables, use the **Multivariate** platform in the **Analyze** menu of *JMP*.

### Example 2.3 Retail shop sales revisited - correlation

The scatterplot for the number of items sold per day and the gross sales per day shows a somewhat strong positive linear relationship between the variables. Let's calculate the *correlation*.

1. Open the *JMP* data table **eg02\_01.jmp** again if it is closed.
2. Select **Analyze** ⇒ **Multivariate**.

- Select the columns **Gross Sales** and **Items** and press **Y, Columns**.
- Press **OK**.



The correlation between the number of items sold per day and gross sales per day is  $r = 0.9540$ .

**Note:** In the professional version of *JMP*, the **Multivariate** platform is found under the submenu **Multivariate Methods**.

## 2.3 Least-Squares Regression

To fit the least-squares regression line, use the **Fit Line** command in the red triangle menu for scatterplots.

### Example 2.4 Does fidgeting keep you slim?

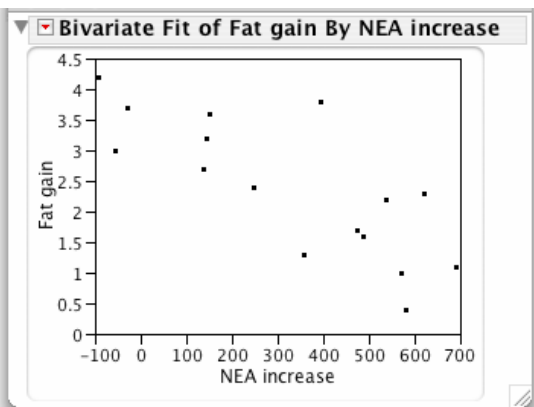
Some people don't gain weight even when they overeat. Perhaps fidgeting and other "nonexercise activity" (NEA) explain why. A scatterplot of fat gain and the increase in NEA of 16 healthy adults in an 8-week study shows a moderately strong negative linear association with no outliers. A straight line might serve as a good model for describing the relationship and for predicting the fat gain for a specific increase in NEA. We can do this easily in *JMP*.

We first create a scatterplot to examine the relationship between the two variables. The NEA increase is the explanatory variable so we wish to plot it on the  $x$ -axis. Suppose the *JMP* data table **eg02\_04.jmp** contains the study data.

1. Open the data table **eg02\_04.jmp**.

Display a scatterplot of **Fat gain** by **NEA increase**.

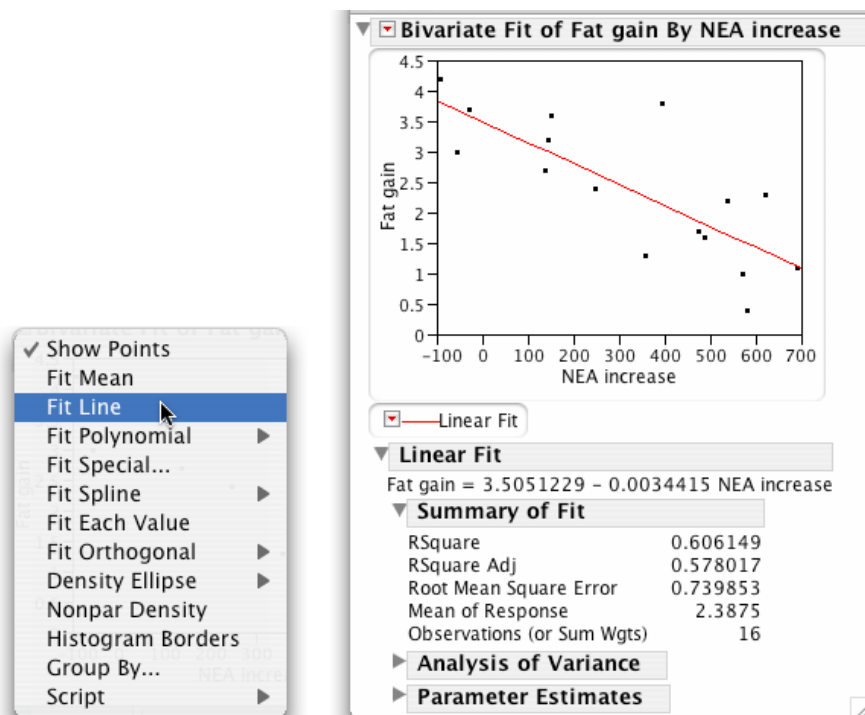
2. Select **Analyze** ⇒ **Fit Y by X**.
  - a. Select **Fat gain** and **Y, Response**.
  - b. Select **NEA increase** and press **X, Factor** and **OK**.



The plot shows a moderately strong negative linear relationship with no outliers. Now, let's fit a straight line to the data.

3. Press the red triangle and select **Fit Line** from the menu that opens.

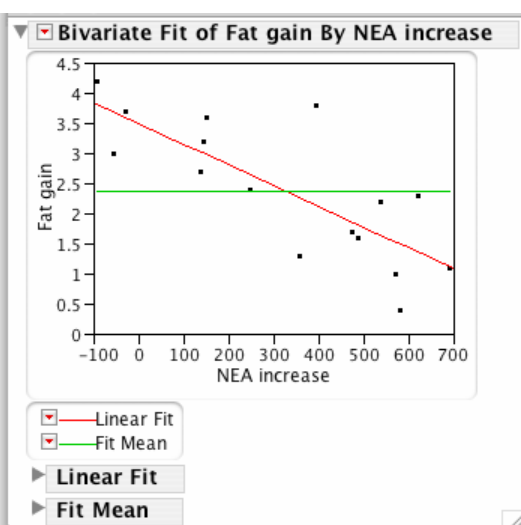




## RSquare, $r^2$

The least-squares equation can be found directly under the **Linear Fit** title bar,  $\text{Fat gain}^{\wedge} = 3.505 - 0.00344 \text{ NEA increase}$ . *RSquare* ( $r^2$ ) can be found directly under the **Summary of Fit** title bar, **RSquare** =  $r^2$  = **0.606149**. Recall that  $r^2$  is the proportion of the variability in **Fat gain** that is explained by the least-squares regression of **Fat gain** on **NEA increase**. To see the variability in **Fat gain** better:

4. Select **Fit Mean** from the red triangle menu on the **Bivariate Fit of Fat gain By NEA increase** title bar.




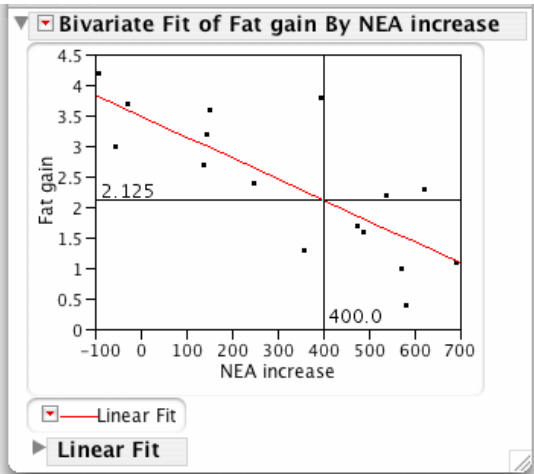
Compare the variability of the points about the horizontal green line with that about the tilted red line. The vertical distances of the points from the red (regression) line are considerably less than from the horizontal green line.

## Prediction

We can use *JMP* to *predict* the response for a specific value of the explanatory variable  $x$ . For example, we might want to predict the fat gain for an individual whose NEA increases by 400 calories when she overeats.

### Example 2.5 Predicting fat gain

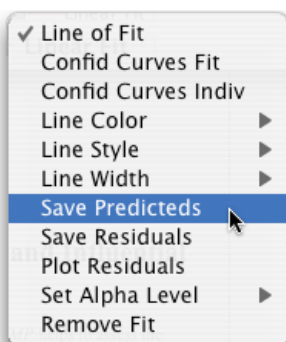
1. Select the **crosshair** tool from the **Tools** palette. 
  - a. Place the cursor, which now resembles a crosshair, on the least-squares regression line directly above 400 and press.



The closest value to 400 calories that is displayed is 399.6. The corresponding predicted value of **Fat gain** for an individual whose NEA increases by 400 calories when she overeats is 2.125, which is the same (to two decimal places) as that calculated in the textbook.

For precise predicted values, *JMP* can calculate and store the *predicted values* of each of the individuals in the data table.

2. Press the red triangle that is directly below the scatterplot and next to the **Linear Fit** title bar. Select **Save Predicteds**.
  - a. Select the data table window and notice that a new column, **Predicted Fat gain**, was created to hold the predicted value for each observation.



Since no individual has an NEA increase of 400, initially there is no row with the corresponding predicted fat gain. To obtain the value of predicted fat gain for an individual with an NEA increase of 400 calories:

- b. Type 400 into the **NEA increase** column of a new row in the data table.

	NEA increase	Fat gain	Predicted Fat gain
14	580	0.4	1.50906043
15	620	2.3	1.37140095
16	690	1.1	1.13049686
17	400	•	2.1285281

## 2.4 Assessing the Fit: Residuals, Outliers, and Influential Observations

Besides fitting models that describe the overall pattern of a relationship, *JMP* helps to assess the appropriateness of a fitted model and to identify striking deviations from that model. The professional statistician does these tasks first—before examining  $r^2$ , the equation of the line, or predicting the response for a specific value of the explanatory variable.

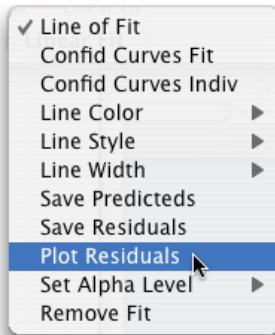
### Residuals

*Residuals* are the vertical deviations of the observed data points from the corresponding predicted values on the least-squares regression line. As such, they represent deviations of the regression model from the data points and a plot of the residuals can help you assess the appropriateness of a regression line as a model for the data. With one command, you can obtain a residual plot and, with another command, you can tell *JMP* to calculate all the residuals and store them in the original data table for later use.

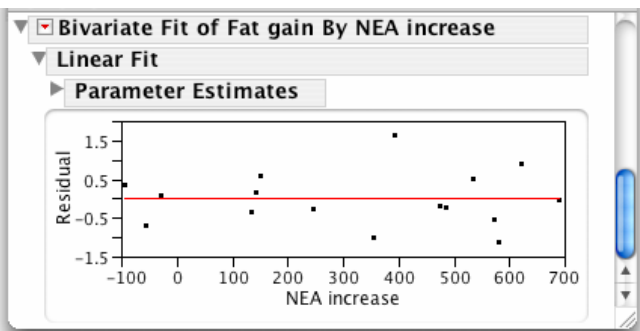
#### Example 2.6 Does fidgeting keep you slim? (cont'd.)

1. Bring the report window **eg02\_04: Fit Y by X** forward. (If you no longer have the window available, repeat the first three steps in Section 2.3.)

- To plot the residuals against the explanatory variable for the linear fit, select **Plot Residuals** from the red triangle menu located directly below the scatterplot next to **Linear Fit** (not the one next to **Fit Mean**, if you are using the report from Section 2.3).



Since the plot is a random band of points centered at zero, the least-squares model,  $\text{Fat gain}^{\wedge} = 3.505 - 0.00344 \text{ NEA increase}$ , provides an appropriate description of the relationship between amount of fat gained and increased nonexercise activity.



- To save the residuals to the *JMP* data table, select **Save Residuals** from the red triangle menu located directly below the scatterplot next to **Linear Fit**.

	NEA increase	Fat gain	Predicted Fat gain	Residuals Fat gain
1	-94	4.2	3.8286227	0.3713773
2	-57	3	3.70128768	-0.7012877
3	-29	3.7	3.60492604	0.09507396
4	135	2.7	3.04052217	-0.3405222
5	143	3.2	3.01299027	0.18700973

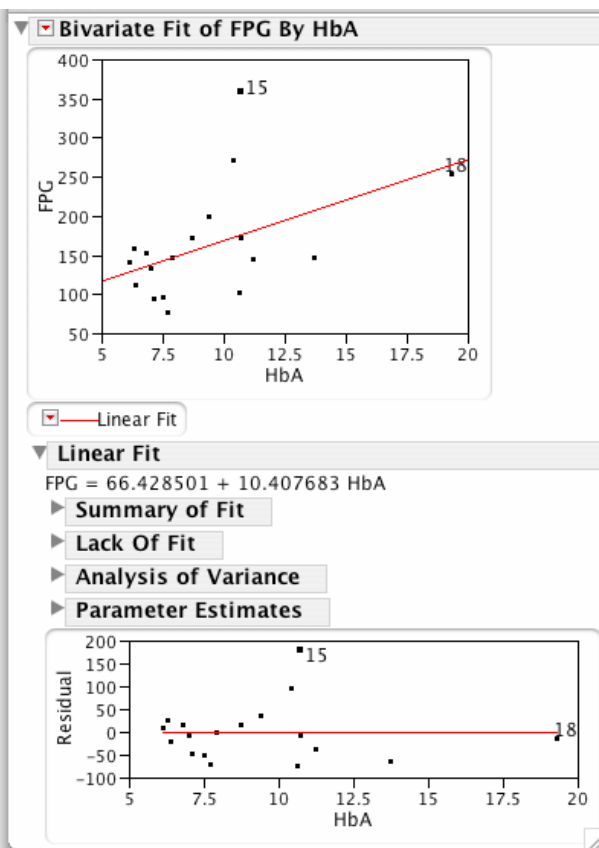
## Outliers and Influential Observations

In addition to judging the appropriateness of a regression line as a model for the data, we need to look for striking individual points, *outliers* and *influential observations*. *Outliers* are points that are outlying in the *y*, or vertical, direction while points that are outlying in the *x*, or horizontal, direction are potentially *influential observations*. Both can be identified using residual plots. To judge the influence of these points, we must fit the regression line with and without the suspect point.

### Example 2.7 Predicting fasting plasma glucose levels

A study of people with diabetes was conducted to investigate the relationship of an easily measurable indicator of blood sugar level, fasting plasma glucose (**FPG**) and the more fundamental indicator called **HbA**. Let's determine if there are any outliers or influential observations in the data. Suppose that the data table **ta02\_01.jmp** contains the data from this study.

1. Open the *JMP* data table **ta02\_01.jmp**.
2. Display a scatterplot of the relationship **FPG** and **HbA**.
  - a. Select **Analyze** ⇒ **Fit Y by X**.
  - b. Select **FPG** and press **Y, Response**.
  - c. Select **HbA**, and press **X, Factor** and **OK**.
3. Fit the least-squares regression line and obtain a residual plot.
  - a. Press the red triangle on the **Bivariate Fit of FPG by HbA** title bar and select **Fit Line**.
  - b. Press the red triangle next to **Linear Fit**, which is directly below the scatterplot, and select **Plot Residuals**.
4. Identify the days associated with the outlying points by hovering the cursor over them.

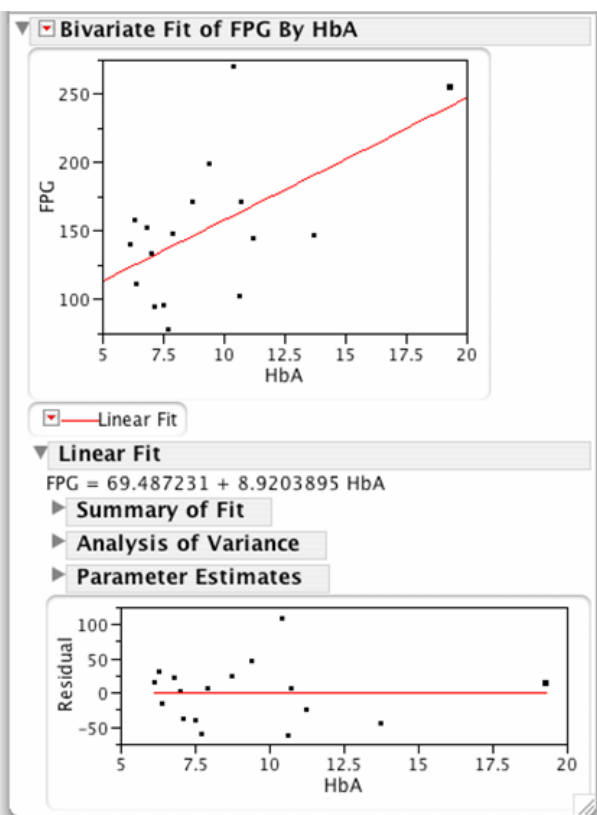


Subject 15 is an outlier. Its vertical deviation from the model is much larger than for other subjects. The model does not fit this person well. Subject 18, which has a very small residual, is outlying in the  $x$

direction and, as such, is potentially influential. Click on these points in the residual plot and notice that the corresponding point in the scatterplot of **FPG** by **HbA** is highlighted. To investigate the influence on the fitted line, you need to exclude a suspicious observation and refit the least-squares regression line. To do this for subject 15, simply change the row state of subject 15 in the data table to **Exclude**. (Row states are discussed in more detail in Section 0.3.2 of Chapter 0.)

5. Select **Window** ⇒ **ta02\_01.jmp** to bring the data table to the front.
  - a. If row 15 (subject 15) is not highlighted, select **Row 15**.
  - b. Select **Rows** ⇒ **Exclude/Include** and notice that the exclusion symbol  $\emptyset$  appears in the row number area next to row 15 at the left of the data grid.
6. To have *JMP* automatically duplicate the analysis without row 15 (subject 15):
  - a. Select **Window** ⇒ **Fit Y by X of FPG by HbA** to bring the report to the front.
  - b. Press the red triangle on the **Bivariate Fit of FPG by HbA** title bar at the top.
  - c. Select **Script** ⇒ **Redo Analysis**.

Compare this scatterplot and the equation of the line with those in the previous report that included subject 15. The slope of the least-squares regression line has been changed notably. Subject 15 is an influential observation.



## 2.5 Data Analysis for Two-Way Tables

*JMP* uses mosaic plots to display the relationship between two categorical variables. Two-way tables, joint distributions, marginal distributions, and conditional distributions describe the relationship and help identify striking deviations from that relationship.

Critical to this data analysis of two categorical variables in *JMP* are the modeling types of the variables and the way in which the data are entered into a *JMP* data table. In the following example, 8235 people (the individuals) are studied. However, we use a *JMP* data table with only 16 rows because there are only  $4 \times 4 = 16$  unique combinations of the values of the two categorical variables.

### Example 2.8 Marital status and job level

Is being married good for your career? The following table summarizes data from a study on the marital status and the job level of 8235 male managers and professionals employed by a large company. Four job grades were assigned. Grade 1 contains jobs with the least value to the company and grade 4 contains jobs with the most value. Both variables are categorical.

Job Grade	Marital Status				Total
	Single	Married	Divorced	Widowed	
1	58	874	15	8	955
2	222	3927	70	20	4239
3	50	2396	34	10	2490
4	7	533	7	4	551
Total	337	7730	126	42	8235

First, we create an appropriate *JMP* data table to summarize the data. (See Section 0.2.2 in Chapter 0.) Here is what we want the data table to look like:

Job Grade	Marital Status	Count
1	Single	58
1	Married	874
1	Divorced	15
1	Widowed	8
2	Single	222
2	Married	3927
2	Divorced	70
2	Widowed	20
3	Single	50
3	Married	2396
3	Divorced	34
3	Widowed	10
4	Single	7
4	Married	533
4	Divorced	7
4	Widowed	4

To better understand how *JMP* data tables for categorical variables are constructed, consider what information you have for each of the 8235 men (the individuals) in the study: the person's job level and marital status. Thus, we need a variable **Job Grade** with four categories and a variable **Marital Status**,

also with four categories, giving a total of  $4 \times 4 = 16$  possible combinations of categories. Instead of entering 8235 rows, one for each person, we can use 16 rows, one for each combination, and include a column, **Count**, of the number of people with each combination of the two categories.

1. Select **File** ⇒ **New** from the menu bar.
2. Select **Cols** ⇒ **Add Multiple Columns** to accommodate the three variables.
  - a. Enter **3** after **How many columns to add** and press **OK**.
  - b. Change the names of the columns to **Job Grade**, **Marital Status**, and **Count**.
  - c. Press **OK**.

By default, columns contain numeric data. Change the data type of the first two variables to “Character.”

3. Select the first two columns.
  - a. Select **Cols** ⇒ **Column Info**.
  - b. Select **Character** from the **Data Type** menu and press **OK**.
4. Select **Rows** ⇒ **Add Rows** from the menu bar and enter **16**.
  - a. Fill in the data grid as above.
5. Select **File** ⇒ **Save** and name the data table **ex02\_08.jmp**.

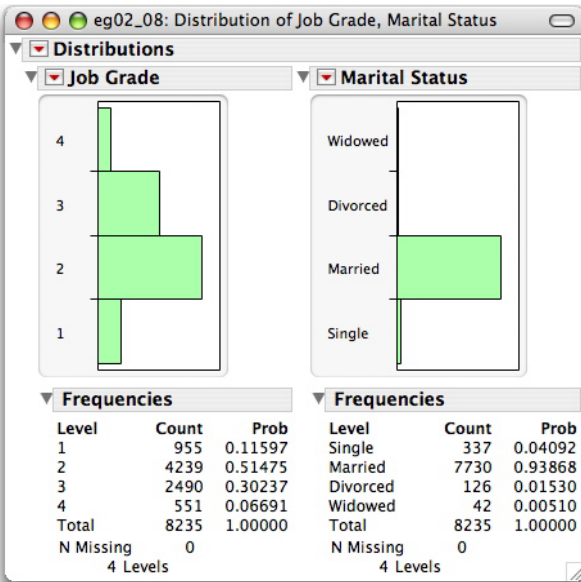
### Marginal Distributions

Now let’s look at the distribution of each variable separately. We do this in the same way that we did in Chapter 1 for a categorical variable.

6. Select **Analyze** ⇒ **Distribution**.
  - a. Select **Job Grade** and **Marital Status** from the list of columns and press **Y, Columns**.
  - b. Select **Count**, and press **Freq**.
  - c. Press **OK**.

Examine the resulting *marginal distributions*.





By default, categories are presented alphabetically. To obtain the above ordering, use the column property, **Value Ordering**. Details can be found in the third remark below.

### Two-Way Table and the Joint Distribution

To examine the relation between the variables, we must look at the variables jointly. A two-way table is a start at doing that. To obtain the *joint distribution* of **Marital Status** and **Job Grade**, proceed as follows:

7. Select **Analyze** ⇒ **Fit Y by X**.
  - a. Select **Marital Status** and press **Y, Response**.
  - b. Select **Job Grade** and press **X, Factor**.
  - c. Select **Count**, and press **Freq** and **OK**.
  - d. Press the red triangle on the **Contingency Table** report and deselect **Row %** and **Col %**.

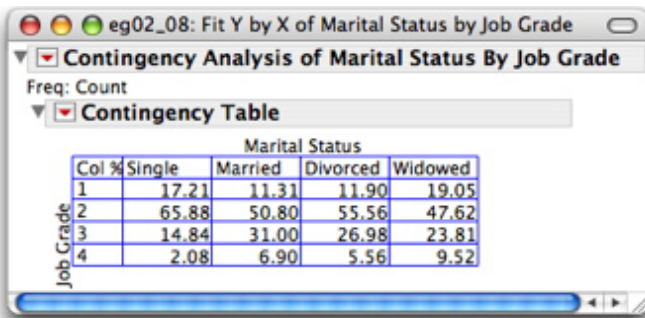
The screenshot shows the Minitab 'Contingency Analysis of Marital Status By Job Grade' window. The frequency is set to 'Count'. The contingency table is displayed below.

		Marital Status				Total %
		Single	Married	Divorced	Widowed	
Job Grade	1	58	874	15	8	955
		0.70	10.61	0.18	0.10	11.60
	2	222	3927	70	20	4239
		2.70	47.69	0.85	0.24	51.48
	3	50	2396	34	10	2490
	0.61	29.10	0.41	0.12	30.24	
	4	7	533	7	4	551
		0.09	6.47	0.09	0.05	6.69
		337	7730	126	42	8235
		4.09	93.87	1.53	0.51	

## Conditional Distributions

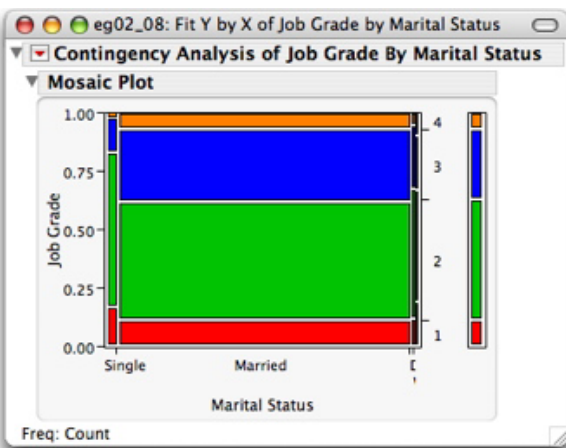
*Conditional distributions* are better for assessing the relationship between two categorical variables. To obtain the conditional distributions of **Job Grade**, we want the row percentages.

8. Press the red triangle on the **Contingency Table** report.
  - a. Select **Row %**.
  - b. Deselect **Total %**.



Each column represents a *conditional distribution* of **Job Grade**. The first column is the conditional distribution of **Job Grade** for single men, the second column contains the conditional distribution for married men, and so on.

A **mosaic plot** can also be used to display the same four *conditional distributions* of **Job Grade**, one for each marital status group, as stacked bar charts.



The stacked bar chart on the right of the *mosaic plot* shows the marginal distribution of **Job Grade** while the proportions along the x-axis of the *mosaic plot* represent the marginal distribution of **Marital Status**. The four other stacked bar charts represent the conditional distributions of Job Grade for the four **Marital Status** categories. The plots for divorced and widowed men are barely visible in this example since there are relatively few of them. From both the **Mosaic Plot** and the **Contingency Table** report, it is clear that job grades 3 and 4 are less common for single men and that job level and marital status are related.

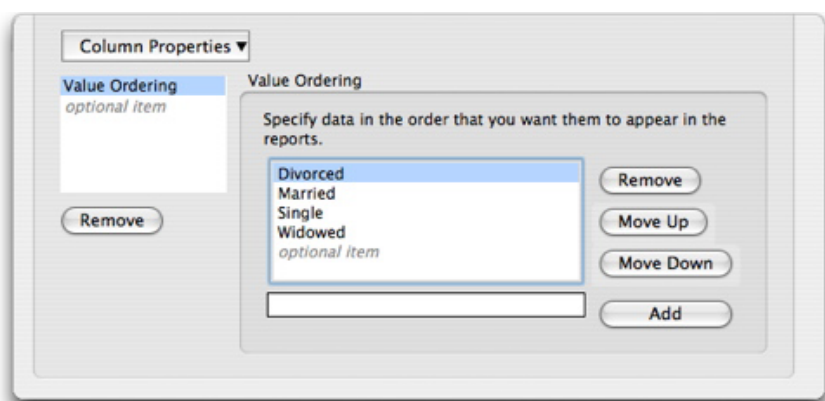
## Remarks

- If the values of a categorical variable are numeric, you must change the modeling type to “Nominal.”
- The **Contingency Table** report and the **Mosaic Plot** have different orientations that are controlled by your choice of **X** and **Y** variables. The **Y, Response** variable is on the vertical axis in the mosaic plot and on the horizontal “axis” in the contingency table report. The above **Mosaic Plot** was obtained by interchanging the **X** and **Y** columns in instruction 7a,b above.
- By default, *JMP* arranges the values of a categorical variable in alphabetical order. To change this ordering, use the column property “Value Ordering.” Here is an example:

### Example 2.9 Rearranging the values of marital status

To rearrange the order of levels of **Marital Status**:

1. Select the column **Marital Status**.
  - a. Select **Cols** ⇒ **Column Info**
  - b. Select **Column Properties** ⇒ **Value Ordering**.



The values of **Marital Status** are displayed in their current order. Select a value and press either **Move Up** or **Move Down**. The new order will be used in mosaic plots and contingency tables.

2. To obtain the order shown in the figures above:
  - a. Select the category **Single**, press **Move Up** twice to place “Single” at the top, and press **OK**.
  - b. Select the category **Divorced** and press **Move Down**.
  - c. Press **OK**.

## Simpson’s Paradox


We can use *JMP* to evaluate the effects of a lurking variable on the relationship between two categorical variables.

### Example 2.10 Which airline is on time?

Suppose that you wish to compare the on-time rates of flights for two airlines. Data on arrival status of flights for one month are available from two western cities. We first look at the relation of delayed

arrival and airline without considering the city from which the flights left. Then, we will examine the relation separately for flights from the different cities.

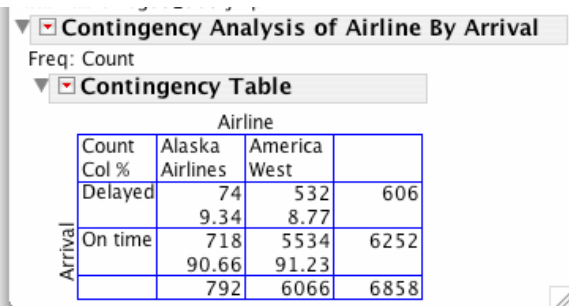
1. Select **File** ⇒ **New** and create four columns named **Departure City**, **Airline**, **Arrival**, and **Count**.
2. Enter the data as follows and save the *JMP* data table.



	Departure City	Airline	Arrival	Count
1	Los Angeles	Alaska Airlines	On time	497
2	Los Angeles	Alaska Airlines	Delayed	62
3	Los Angeles	America West	On time	694
4	Los Angeles	America West	Delayed	117
5	Phoenix	Alaska Airlines	On time	221
6	Phoenix	Alaska Airlines	Delayed	12
7	Phoenix	America West	On time	4840
8	Phoenix	America West	Delayed	415

To display the conditional distributions of **Arrival** for each airline in columns:

3. Select **Analyze** ⇒ **Fit Y by X**.
  - a. Select **Airline** and press **Y, Response**.
  - b. Select **Arrival** and press **X, Factor**.
  - c. Select **Count** and press **Freq** and **OK**.
4. Deselect **Total %** and **Row %** in the **Contingency Table** report.



Contingency Analysis of Airline By Arrival

Freq: Count

☒ Contingency Table

		Airline		
		Alaska Airlines	America West	
Arrival	Delayed	74 9.34	532 8.77	606
	On time	718 90.66	5534 91.23	6252
		792	6066	6858

The table indicates that America West is slightly better—it is delayed less frequently (8.77%) than Alaska Airlines (9.34%). Now let's consider the effect of city of departure on the relation between **Arrival** and **Airline**. To do this in *JMP*, we simply use the **By** option on the **Fit Y by X** platform.

5. Select **Analyze** ⇒ **Fit Y by X**.
  - a. Select **Airline** and press **Y, Response**.
  - b. Select **Arrival** and press **X, Factor**.
  - c. Select **Count** and press **Freq**.
  - d. Select **Departure City** and press **By** and **OK**.
6. Deselect **Total %** and **Row %** in each **Contingency Table** report.

Departure City=Los Angeles				
Contingency Analysis of Airline By Arrival				
Freq: Count				
Contingency Table				
Arrival	Airline			
	Count	Alaska Airlines	America West	
	Col %			
	Delayed	62	117	179
		11.09	14.43	
On time		497	694	1191
		88.91	85.57	
		559	811	1370

Departure City=Phoenix				
Contingency Analysis of Airline By Arrival				
Freq: Count				
Contingency Table				
Arrival	Airline			
	Count	Alaska Airlines	America West	
	Col %			
	Delayed	12	415	427
		5.15	7.90	
On time		221	4840	5061
		94.85	92.10	
		233	5255	5488

Notice that the delay rate is less for Alaska Airlines for both departure cities! Thus, you should choose Alaska Airlines to minimize delays. Clearly, it is not a good idea to ignore a lurking variable when considering the relationship between two variables.

## 2.6 Summary

Except for calculating the correlation, all graphs and statistical computations in this chapter can be performed in the second platform, **Fit Y by X**, of the **Analyze** menu. To calculate correlations, we use the **Multivariate** platform.

Activity	Command
Displaying relationships	<b>Analyze</b> ⇒ <b>Fit Y by X</b>
Scatterplots	
Side-by-side boxplots	... ⇒ <b>Display Options</b> ⇒ <b>BoxPlots</b>
Side-by-side means diamonds	... ⇒ <b>Display Options</b> ⇒ <b>Means/Diamonds</b>
Two-way tables and mosaic plots	<b>Analyze</b> ⇒ <b>Fit Y by X</b> ⇒ ... ⇒ <b>Freq</b>
Correlation	<b>Analyze</b> ⇒ <b>Multivariate</b>
Least-squares regression	<b>Analyze</b> ⇒ <b>Fit Y by X</b> ⇒ <b>Fit Line</b>

# Chapter 3

## Producing Data

### 3.1 Design of Experiments

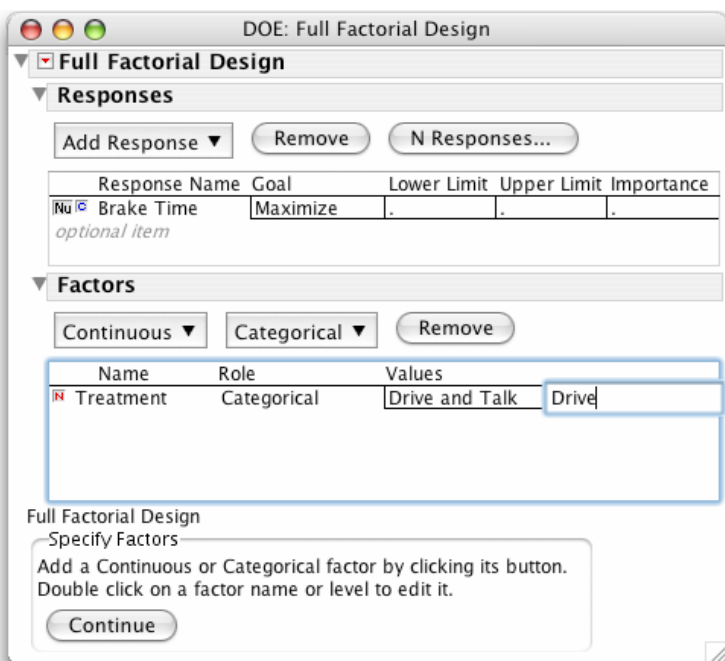
In a completely randomized design, all the experimental units are allocated at random among the treatments. *JMP* has a separate menu (**DOE**) for designing completely randomized designs, randomized block designs, screening designs, response surface designs, and many other complex and useful designs. Random allocation of the experimental units is automated for each.

#### Example 3.1 Does talking on a hands-free cell phone distract drivers?

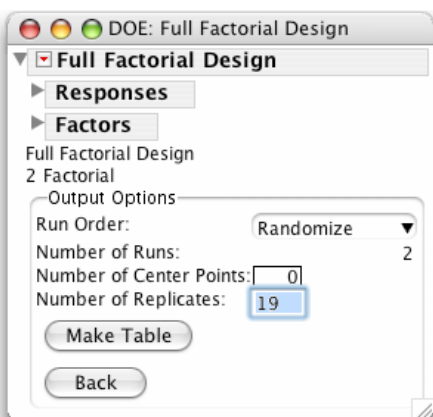
---

In a study of the effects of cell phone usage and driving, undergraduate students “drive” in a special driving simulator equipped with a hands-free cell phone. The car ahead brakes. How quickly does the subject respond? Twenty students (the control group) are to simply drive. Another 20 (the experimental group) are to talk on the cell phone while driving. You are asked to randomly assign 40 students, 20 each, to the two treatments.

1. Select **DOE** ⇒ **Full Factorial Design**.
2. Open the **Responses** panel by pressing the blue disclosure button on the **Responses** title bar.
  - a. The default name for the response is **Y**. Double-click on **Y** and change the name to **Brake Time**.
3. Under factors, select **Categorical** ⇒ **2 Level**.
  - a. Change the default name of the factor by double-clicking on **X1** and typing **Treatment**.
  - b. Enter the treatment names by tabbing to **L1** and **L2** and typing **Drive and Talk** and **Drive**, respectively.
  - c. Select **Continue**.

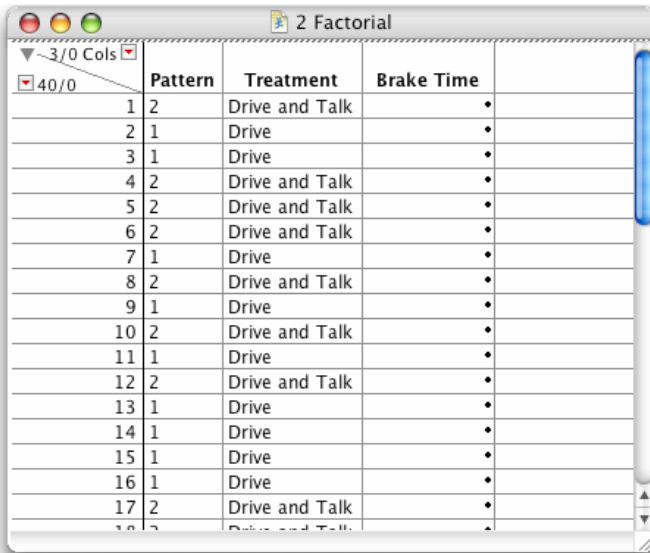


Examine the **Output Options**. A run is an observation. By default, the run order will be randomized. If you have the experimental units in a set order beforehand, this effectively randomizes the assignment of the treatments to the experimental units.



Note that the default number of runs is 2, one for each treatment. To randomly assign the 40 students, you need to change the number of replicates to 19. This results in the original 2 observations plus 38 (19 additional sets of 2 observations each), for a total of 40.

4. Click in the **Number of Replicates** field and change it to 19.
  - a. Select **Make Table**.



	Pattern	Treatment	Brake Time
1	2	Drive and Talk	•
2	1	Drive	•
3	1	Drive	•
4	2	Drive and Talk	•
5	2	Drive and Talk	•
6	2	Drive and Talk	•
7	1	Drive	•
8	2	Drive and Talk	•
9	1	Drive	•
10	2	Drive and Talk	•
11	1	Drive	•
12	2	Drive and Talk	•
13	1	Drive	•
14	1	Drive	•
15	1	Drive	•
16	1	Drive	•
17	2	Drive and Talk	•
18	2	Drive and Talk	•

The resulting *JMP* data table has three columns. The first column is the **Pattern**, which for more than one factor designs would contain the combination of factors used for a particular observation. The second column is the factor for this experiment, **Treatment**. The third column is the response variable, **Brake Time**.

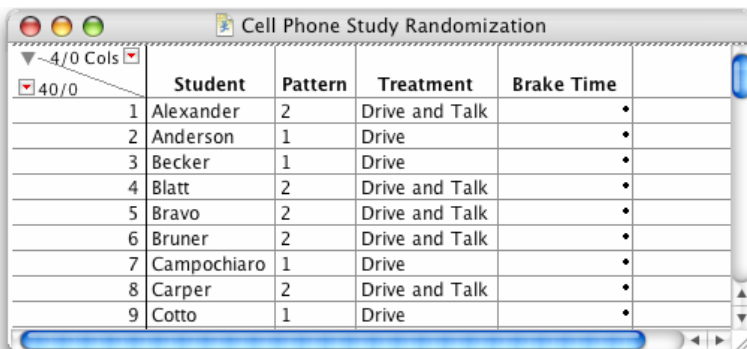
## Remark

Suppose you have a list of the 40 students in a *JMP* data table, **eg03\_006.jmp**. You can include them in the table above using the **Join** command on the **Tables** menu.

### Example 3.1 Cont'd.

6. Open the *JMP* data table **eg03\_006.jmp**.
7. Select **Tables** ⇒ **Join**.
  - a. Select **2 Factorial**, the table above that *JMP* created holding the design.
  - b. Name the output table **Cell Phone Study Randomization**.
  - c. Press **Join**.

The resulting table explicitly matches each student with the treatment he or she is to receive.



	Student	Pattern	Treatment	Brake Time
1	Alexander	2	Drive and Talk	•
2	Anderson	1	Drive	•
3	Becker	1	Drive	•
4	Blatt	2	Drive and Talk	•
5	Bravo	2	Drive and Talk	•
6	Bruner	2	Drive and Talk	•
7	Campochiaro	1	Drive	•
8	Carper	2	Drive and Talk	•
9	Cotto	1	Drive	•



## 3.2 Designing Samples

A simple random sample of  $n$  individuals from a population is a subset of  $n$  individuals chosen in such a way that all subsets of  $n$  individuals have the same chance of being selected. To obtain a simple random sample using *JMP*, you create a *JMP* data table containing all the individuals in the population and then use the **Subset** command on the **Tables** menu.

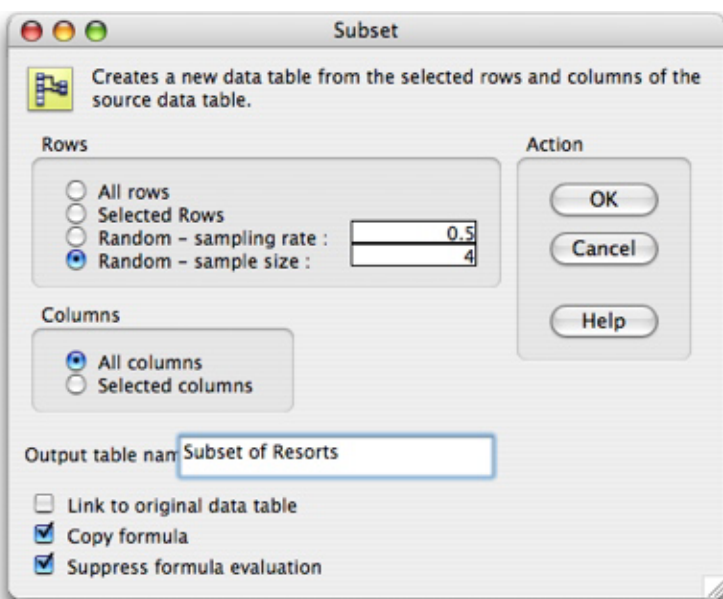
### Example 3.2 How to choose an SRS

A campus newspaper plans a major article on spring break destinations. The authors intend to call a few randomly chosen resorts at each destination to ask about their attitudes toward groups of students as guests. We wish to select an SRS of four resorts from one of the cities for the newspaper. First, place a list of the resorts in the city in a *JMP* data table.

1. Create a *JMP* data table with one variable **Resort** and 28 rows containing the 28 resorts.

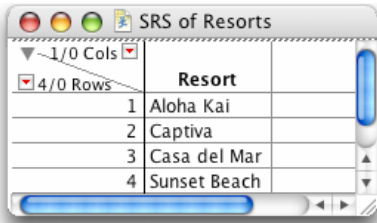


2. Select **Tables** ⇒ **Subset** from the menu bar.



- a. Enter a name for the sample, **SRS of Resorts**.

- b. Press the button next to **Random Sample** and enter a sample size of 4.
- c. Press **OK**.



Save the table and send a copy of the resulting simple random sample to the newspaper.

### 3.3 Summary

#### Activity

Completely randomized designs  
Simple random sampling

#### Command

DOE ⇒ Full Factorial Design  
Tables ⇒ Subset

# Chapter 4

## Probability

### 4.1 Randomness

The *probability* of an outcome of a random phenomenon is the proportion of times the outcome occurs in a very long series of repetitions. Because *JMP* has a built-in *random number generator*, it can imitate repeatedly performing a random phenomenon. We use *JMP* to illustrate the idea of *probability*.

#### Example 4.1 Coin tossing

---

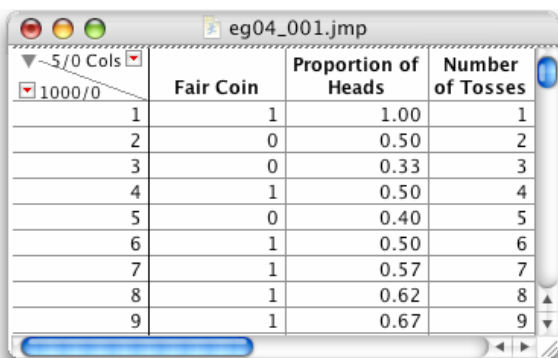
When you toss a fair coin, one of the two sides lands up, heads or tails. There are two outcomes. If the outcomes occur with the same frequency in a large number of tosses of the coin, we say that the *probability* of a head is 0.5 and that the coin is fair. The *JMP* data table **eg04\_001.jmp** available at the textbook web site imitates tossing a coin many times and illustrates the *probability* of getting a head.

1. Open the *JMP* data table **eg04\_001.jmp**.

Each row of the table imitates tossing a coin once. The values of the column **Fair Coin** are either 1 for heads or 0 for tails. Let's imitate tossing this coin 1000 times.

2. Select **Rows** ⇒ **Add Rows**; enter **1000** and press **OK**.

Here is the result of one such simulation. Your 1000 tosses will differ.

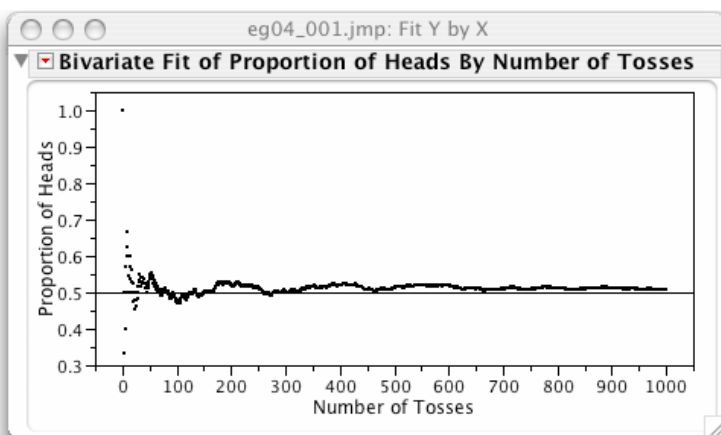


	Fair Coin	Proportion of Heads	Number of Tosses
1	1	1.00	1
2	0	0.50	2
3	0	0.33	3
4	1	0.50	4
5	0	0.40	5
6	1	0.50	6
7	1	0.57	7
8	1	0.62	8
9	1	0.67	9

For each number of tosses, from 1 to 1000, the column **Proportion of Heads** contains the proportion of those tosses that were heads. The first toss results in a 1 = head, so the proportion of heads is 1.00. The second toss results in a 0 = tail, decreasing the proportion of heads to 0.50 after two tosses. The third toss is also a 0 = tail, decreasing the proportion of heads to  $1/3 = 0.33$  after three tosses. The next two tosses are a head followed by a tail, so the proportion of heads after five tosses is  $2/5 = 0.40$ . What happened for your first few tosses?

Let's plot the proportion of heads versus the number of tosses.

3. Select **Analyze** ⇒ **Fit Y by X**.
  - a. Select **Proportion of Heads** and press **Y, Response**.
  - b. Select **Number of Tosses** and press **X, Factor** and **OK**.
4. Add a horizontal reference line at 0.5.
  - a. Click on one of the numbers on the **Proportion of Heads** axis.
  - b. Enter **0.5** in the field to the left of the **Add Ref Line** button, and press **Add Ref Line** and **OK**.
5. Put the cursor on the right edge of the graph. It will change into a two-sided arrow. Drag the edge of the plot to the right to stretch the x-axis.



The proportion of tosses that results in heads varies quite a bit at first but gradually converges to 0.5. Hence, the coin that we simulated tossing can be considered a fair one and the *probability* of a head for it is 0.5.

# Chapter 5

## Sampling Distributions

### 5.1 The Central Limit Theorem

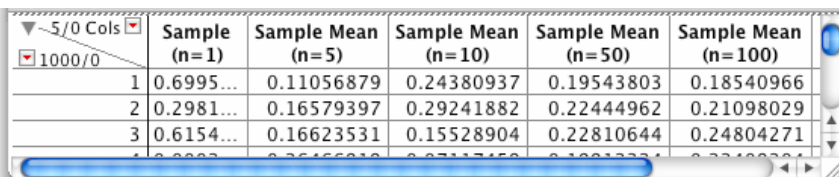
The *central limit theorem* tells us that the distribution of  $\bar{x}$  will be approximately Normal no matter what the shape of the population distribution as long as the sample size is large enough (and the standard deviation is finite).

To help understand this result and reinforce the concept of a sampling distribution, we will simulate the sampling distributions for means of SRSs of different sizes from a very skewed distribution. Included with JMP software is a JMP data table **Central Limit Theorem.jmp**.

#### Example 5.1 Sampling from a non-Normal population

1. Select **Help**  $\Rightarrow$  **Sample Data Directory**.
  - a. Press the gray triangle next to **Examples for Teaching** under **Teaching Demonstrations**.
  - b. Select the JMP data table **Central Limit Theorem.jmp**.
2. Select **Rows**  $\Rightarrow$  **Add Rows**.
3. Type **1000** and press **OK**.

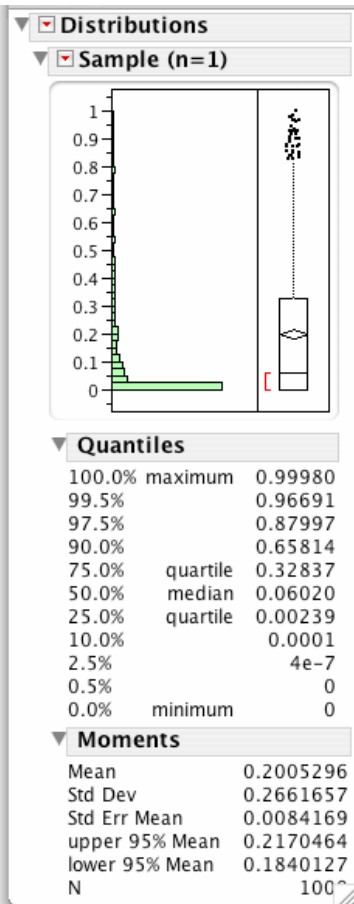
Here is a subset of one possible data table. Since the data values are generated at random, your table will have different values.



	Sample (n=1)	Sample Mean (n=5)	Sample Mean (n=10)	Sample Mean (n=50)	Sample Mean (n=100)
1	0.6995...	0.11056879	0.24380937	0.19543803	0.18540966
2	0.2981...	0.16579397	0.29241882	0.22444962	0.21098029
3	0.6154...	0.16623531	0.15528904	0.22810644	0.24804271

The first column contains 1000 random values from the population distribution and their distribution will approximate the population distribution. Let's use *JMP* to look at this distribution.

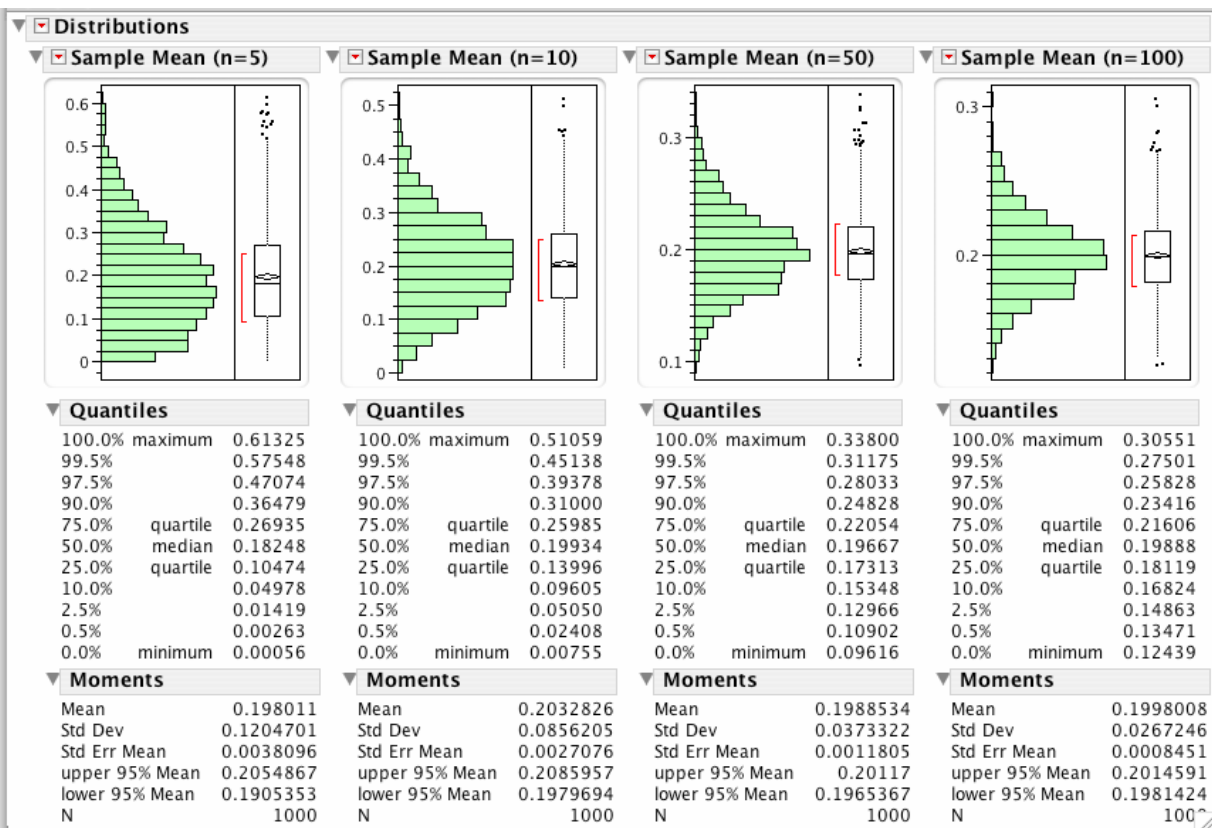
4. Select **Analyze** ⇒ **Distribution**.
5. Select **Sample (n=1)** and press **Y, Columns** and **OK**.



Thus, the mean  $\mu$  and the standard deviation  $\sigma$  of the population distribution are approximately 0.2005296 and 0.2661657, respectively. (Of course, since the data values are generated at random, your results will differ slightly.) You can also see that the distribution is strongly skewed toward larger numbers. Waiting times for service (e.g., a highway tollbooth lane) often follow such a distribution.

The other 4 columns contain the means of samples from this very skewed distribution. Let's look at the distributions of the sample mean for SRSs of sizes 5, 10, 50, and 100.

6. Select **Analyze** ⇒ **Distribution**.
7. Select **Sample Mean (n=5), ..., Sample Mean (n=100)** and press **Y, Columns** and **OK**.



First, look at the means for each distribution and compare them to the mean of the first column, which approximates the population mean. Notice that they are almost equal.

Second, consider the standard deviations of these distributions, 0.1204701, 0.0856205, 0.0373322, and 0.0267246, respectively. Note that they become progressively smaller as expected. According to the formula, these standard deviations should each be approximately  $1/\sqrt{n}$  times the population standard deviation. For the simulation shown here, the population standard deviation is approximately 0.2661657, and so the standard deviations of these sampling distributions should have been approximately 0.1190329, 0.084169, 0.0376415, and 0.0266166, respectively. They are remarkably close to the corresponding empirical standard deviations. How do the empirical standard deviations of your simulation compare with the theoretical results?

Now, look at the shapes of the distributions. Notice that they tend to look more like a Normal curve as the sample size increases although the population distribution is very skewed. That's what the famous *central limit theorem* says!

# Chapter 6

## Introduction to Inference

### 6.1 Estimating with Confidence

Calculations in this section are best done with a calculator. The calculations are not difficult and reinforce their dependence on the sampling distribution of the sample mean. If you are given only the raw data values from a large sample, then calculating a confidence interval estimate with *JMP* is much easier. We illustrate calculating a confidence interval estimate for the population mean first. Then, we illustrate what is meant by a confidence interval estimate using a *JMP* script.

**Calculating a confidence interval estimate for the population mean** If you have the raw data values from a large sample, *JMP* can easily compute a confidence interval estimate for the mean. Since only one variable is involved, we use the **Distribution** platform. The **Confidence Interval** command calculates a confidence interval estimate for the population mean.

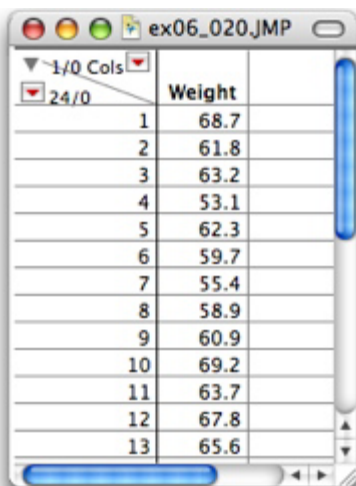
#### Example 6.1 Fuel efficiency

---

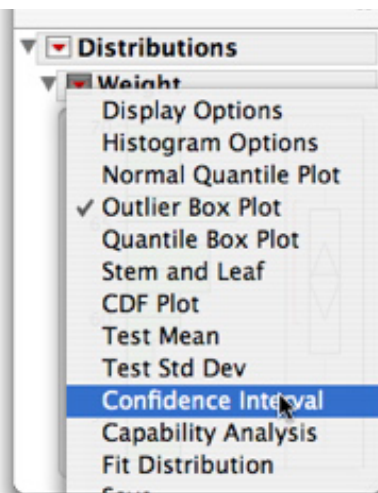
Your company sells exercise clothing and equipment on the Internet. To design the clothing, you collect data on the physical characteristics of your different types of customers. Weights (in kg) for a sample of 24 male runners are stored in a single column of a *JMP* data table. Assume that these runners can be viewed as a random sample of your male customers. Suppose also that the standard deviation of the populations is known to be  $\sigma = 4.5$  kg. Compute a 95% confidence interval for the mean of the population from which this sample is drawn.

1. Open the *JMP* data table.
2. Select **Analyze**  $\Rightarrow$  **Distribution**.
  - a. Select **Weight** and press **Y, Columns** and **OK**.

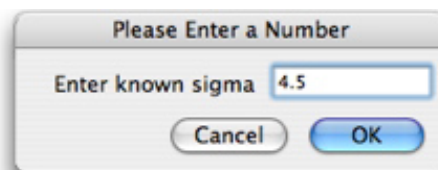
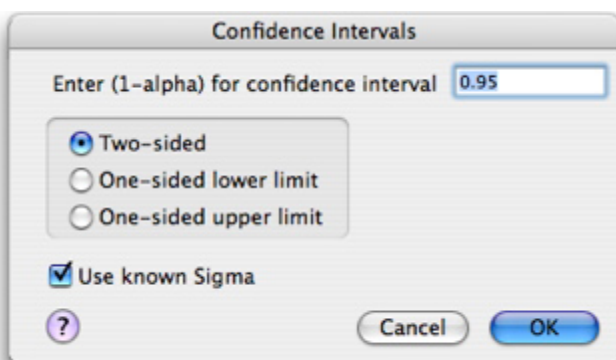




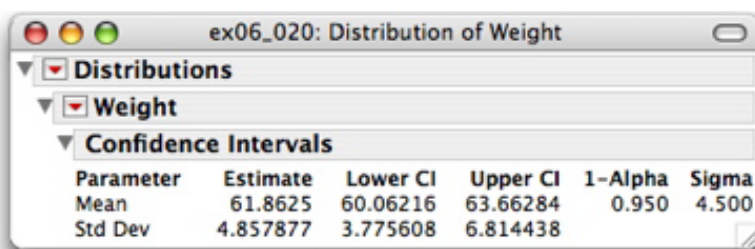
	Weight
1	68.7
2	61.8
3	63.2
4	53.1
5	62.3
6	59.7
7	55.4
8	58.9
9	60.9
10	69.2
11	63.7
12	67.8
13	65.6



3. Click on the red triangle on the title bar of the **Weight** report and select **Confidence Interval** from the menu that opens.



- Check the box for **Use known Sigma** and press **OK**.
- Enter **4.5** in the box that appears and press **OK**.



Parameter	Estimate	Lower CI	Upper CI	1-Alpha	Sigma
Mean	61.8625	60.06216	63.66284	0.950	4.500
Std Dev	4.857877	3.775608	6.814438		

Thus, a 95% confidence interval for the mean weight of your potential male customers is from 60.1 to 63.7 kg.

**Understanding confidence** You have learned that a 95% confidence interval for the mean is arrived at “by a method that gives correct results 95% of the time.” You have also learned that:

- An interval calculated from an SRS of size 100 will be narrower than one based on an SRS of size 20.
- An interval for a population with a standard deviation of 2 will be wider than one with a standard deviation of 1.
- A 99% confidence interval is wider than a 95% confidence interval.

To better understand these notions, we illustrate them using a *JMP* script.

### Example 6.2 Confidence

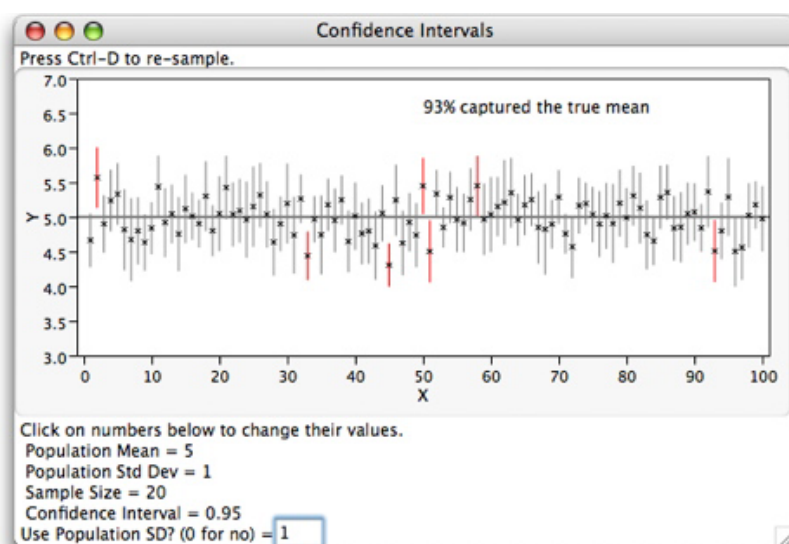
1. Select **Help**  $\Rightarrow$  **Sample Scripts for Students**.
  - a. Click on the script named **confidence.jsl**.

A text window containing the script will open. To run the script:

2. Select **Edit**  $\Rightarrow$  **Run Script**.

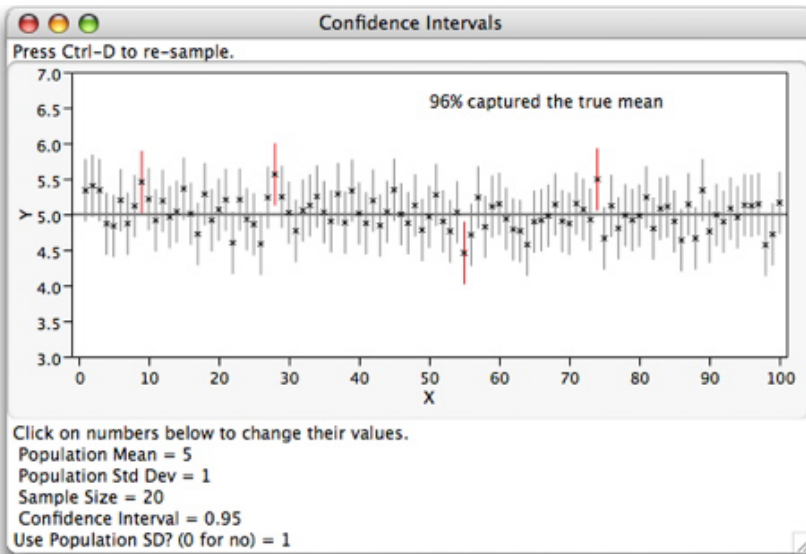
A window appears with a graph containing 100 confidence intervals at the top and with the set of sampling parameters under which the intervals were obtained at the bottom. First, let's simulate taking 100 SRSs of size 20 from a population with known mean  $\mu = 5$  and known standard deviation  $\sigma = 1$  and calculating the 100 associated 95% confidence intervals for these samples. To do that, we only need to change the parameters at the bottom of the window.

3. Select the number **0** that in the **Use Population SD?** field, type **1**, and press **Return**.

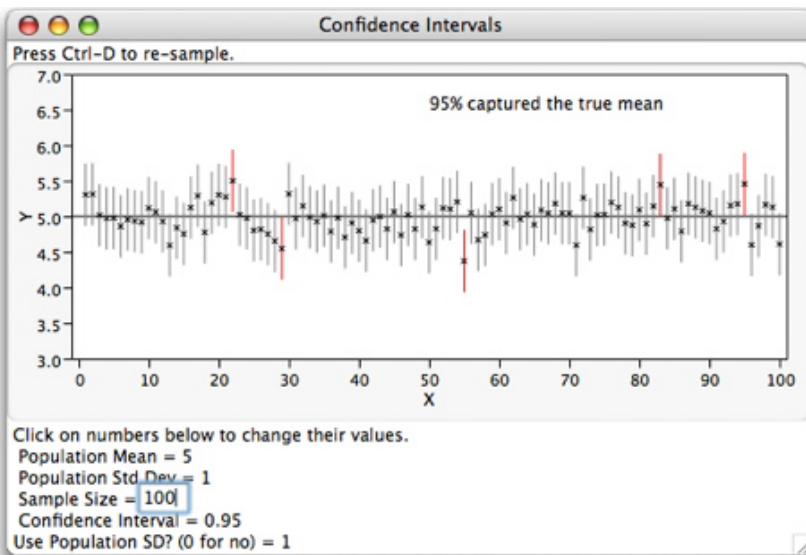


Notice that the graph at the top of the **Confidence Intervals** window changed. A 95% confidence interval for the population mean from the one SRS is calculated and plotted, then a confidence interval from a second SRS, and so on until 100 confidence intervals have been calculated and plotted. The horizontal axis indexes the samples; the values of the variable are on the vertical axis. Red intervals indicate that the interval did not include the population mean. (These correspond to the intervals that are entirely above or below the horizontal line at 5.0.) The percent that capture the true mean is displayed at the top of the graph. We expect that percent to be near 95% and in the long run to converge to 95%. For

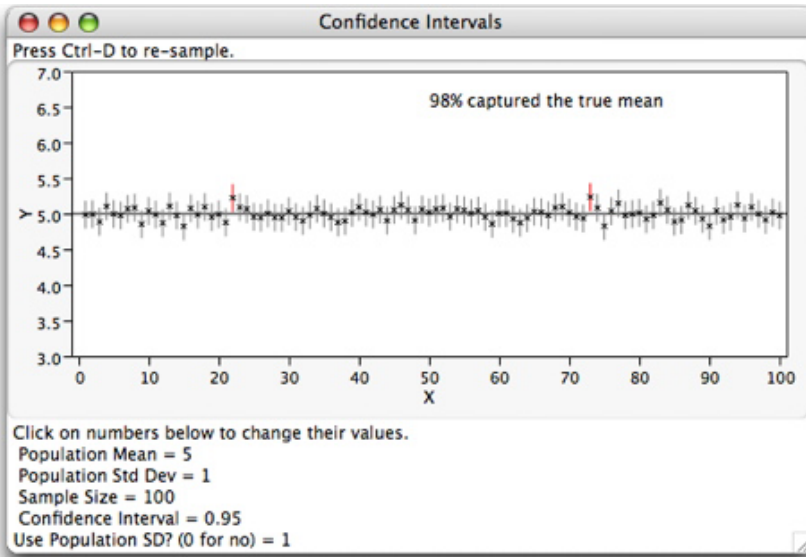
the simulation shown here, four of the samples have intervals that do not include the population mean. Hence, 96% capture the true mean.



Let's see the effect of increasing the sample size from 20 to 100.



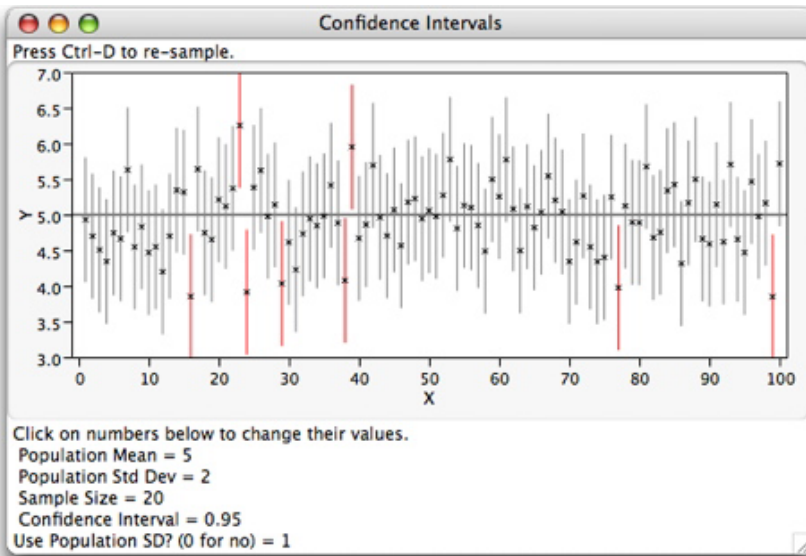
4. Click on the number 20 that specifies **Sample Size**, type 100, and press **Return**.



Notice that the intervals are noticeably narrower, as expected. Next, let's see the effect of increasing the size of the standard deviation.

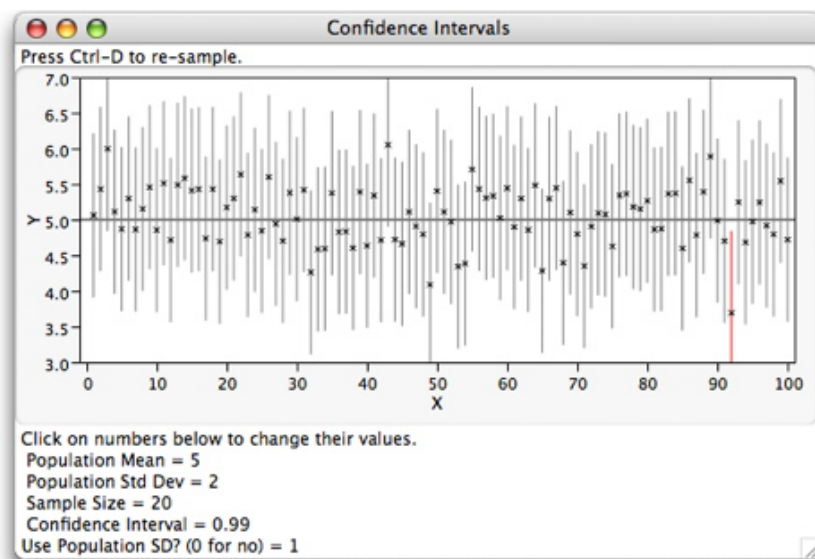
5. a. Return to a sample size of 20 by clicking on the value **100** for **Sample Size** and typing **20**.
- b. Click on the value **1** for **Population Std Dev**, type **2**, and press **Return**.

As expected, these intervals are noticeably wider than the ones with a standard deviation of 1.



Now let's see the effect of increasing the confidence level to 99%.

6. Click on the confidence level of **0.95**, type **0.99**, and press **Return**.



These 99% confidence intervals tend to be wider than the 95% confidence intervals for SRSs of size 20.

**Why aren't they called probability intervals?** In addition, what does it mean to say, “The confidence level shows how *confident* we are that the *procedure* will catch the true population mean”?

On the one hand, when we calculate a confidence interval, we normally get one and only one interval, say 4.2 to 5.1. In addition, that interval either contains the unknown mean or not. Since the population mean is not random, only unknown, and since our interval [4.2, 5.1] is fixed, there is no random experiment. So the probability of our interval containing the mean is either 1 or 0, depending on whether or not the (unknown) population mean is included in the interval. Hence, we **cannot** refer to them as probability intervals.

On the other hand, the method by which a confidence interval is constructed provides that if you repeatedly take SRSs of the same size and calculate 95% confidence intervals, then 95% of those intervals will contain the population mean. To see this, simulate taking many 95% confidence intervals using the **confidence** script.

7. Reset the sample size to 20 and the confidence level to 0.95.

Repeat this by pressing Ctrl-D and calculate the proportion of intervals that capture the population mean after 100 intervals, 200 intervals, 500 intervals, and 1000 intervals. While the percent will vary from one set of 100 intervals to another, it should be close to 95% and the overall percent should approach 95% as more intervals are generated.

That's what we mean when we say, “The confidence level shows how *confident* we are that the *procedure* will capture the true population mean.”

## 6.2 Tests of Significance

Tests of significance are easy to perform in *JMP*. Since only one variable is involved, we use the **Distribution** platform. The **Test Mean** command performs *z* tests.

### Example 6.3 Concentration of active ingredient

The Deely Laboratory has been asked to evaluate the claim that the concentration of the active ingredient in a specimen is 0.86 grams per liter (g/L). Repeated measurements on the same specimen give slightly different results but follow a Normal distribution quite closely. The true concentration is the mean  $\mu$  of the population of repeated analyses. The standard deviation of this population is a property of the analytic procedure and is known to be  $\sigma = 0.0068$  g/L. The laboratory analyzed the specimen three times and found the following concentrations:

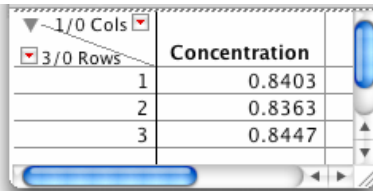
0.8403                  0.8363                  0.8447

Is there convincing evidence to show that the mean of the population of all repeated analyses (the true concentration) differs from 0.86 g/L? To answer this question, we carry out a test of significance on the alternative hypothesis that

$$H_a: \mu \neq 0.86 \text{ g/L}$$

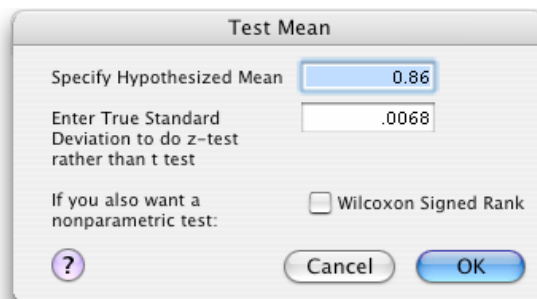
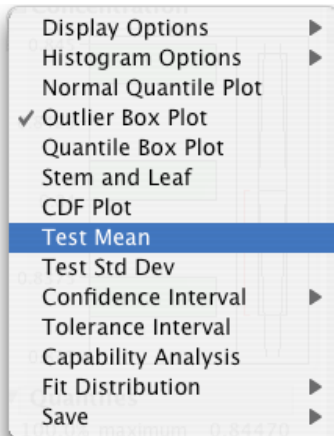
1. Select **File**  $\Rightarrow$  **New**.

Create a *JMP* data table with one variable named **Concentration** and three rows. Save the data table for later use. (See Section 0.2.1 in Chapter 0 for details on creating a new *JMP* data table.)

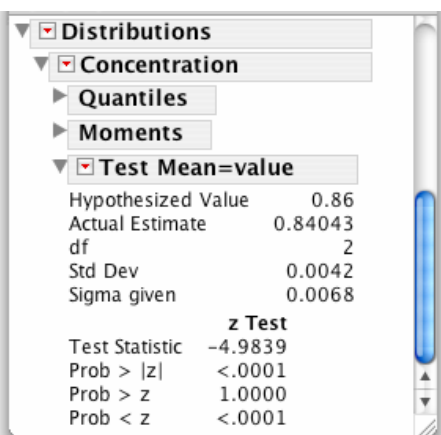


	Concentration
1	0.8403
2	0.8363
3	0.8447

2. Select **Analyze**  $\Rightarrow$  **Distribution** to look at the distribution of **Concentration**.
  - a. Select **Concentration** and press the **Y, Columns** and **OK** buttons just as we did in Chapter 1.
3. Click on the red triangle next to the title bar for the **Concentration** report and select **Test Mean** from the menu that opens.



- Type **0.86** in the **Specify Hypothesized Mean** field.
- Type **3** for the True **Standard deviation**.
- Select **OK**.



Notice that *JMP* calculates the value of the **z test statistic** for the sample  $z = -4.99$ , and provides three *P*-values, one for each of the three possible alternative hypotheses.

<u>Alternative Hypothesis</u>		<u><i>P</i>-value in <i>JMP</i></u>
$\mu \neq 0.86$	Two-tailed	Prob >  z
$\mu > 0.86$	Upper tail	Prob > z
$\mu < 0.86$	Lower tail	Prob < z

Since you wish to test that  $H_a: \mu \neq 0.86$  g/L, use the *P*-value for a two-tailed test,  $< .0001$ . A sample result like this ( $\bar{x} = 0.84043$ ) would happen by chance in less than .01% of samples if the true concentration  $\mu$  were 0.86 g/L. Since the chance of observing a difference as large as this ( $0.01957 = 0.86000 - 0.84043$ ) is virtually impossible if  $\mu = 0.86$ , we conclude that the true concentration of the active ingredient is not 0.86; it is less than 0.86 g/L.

## ***P*-values**

To see how a *P*-value measures the difference between the hypothesized mean and the sample mean, let's look at a graph. Consider the following example.

### **Example 6.4 Fill the bottles**

Bottles of a popular cola drink are supposed to contain 300 milliliters (ml) of cola. There is some variation from bottle to bottle because the filling machinery is not perfectly precise. Assume the standard deviation  $\sigma$  of the filling process is 3 ml. An inspector, who suspects that the bottler is underfilling, measures the contents of a *sample* of six bottles. The results are

299.4   297.7   301.0   298.9   300.2   297.0



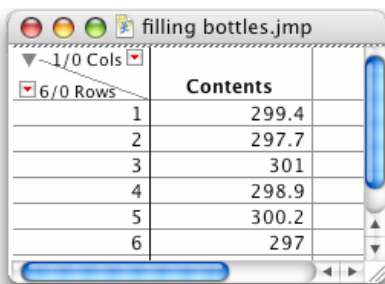
Is this convincing evidence that the mean contents of *all cola bottles* filled by this machinery is less than the advertised 300 ml? To answer the question, we carry out a test of significance on the alternative hypothesis that

$$H_a: \mu < 300 \text{ ml}$$

1. Select **File** ⇒ **New**.

Create a *JMP* data table with one variable named **Contents** and six rows. Save the data table as **filling bottles.jmp** for later use.

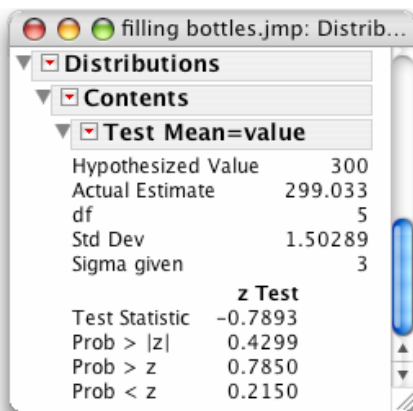
2. Select **Analyze** ⇒ **Distribution** to look at the distribution of **Contents**.
  - a. Select **Content** and press the **Y, Columns** and **OK** buttons.



	Contents
1	299.4
2	297.7
3	301
4	298.9
5	300.2
6	297

Proceed as above to perform a test of significance on the mean **Contents**.

3. Click on the red triangle next to the title bar for the **Contents** report and select **Test Mean** from the menu that opens.
  - a. Type **300** in the Specify Hypothesized Mean field.
  - b. Type **3** for the True Standard deviation.
  - c. Select **OK**.

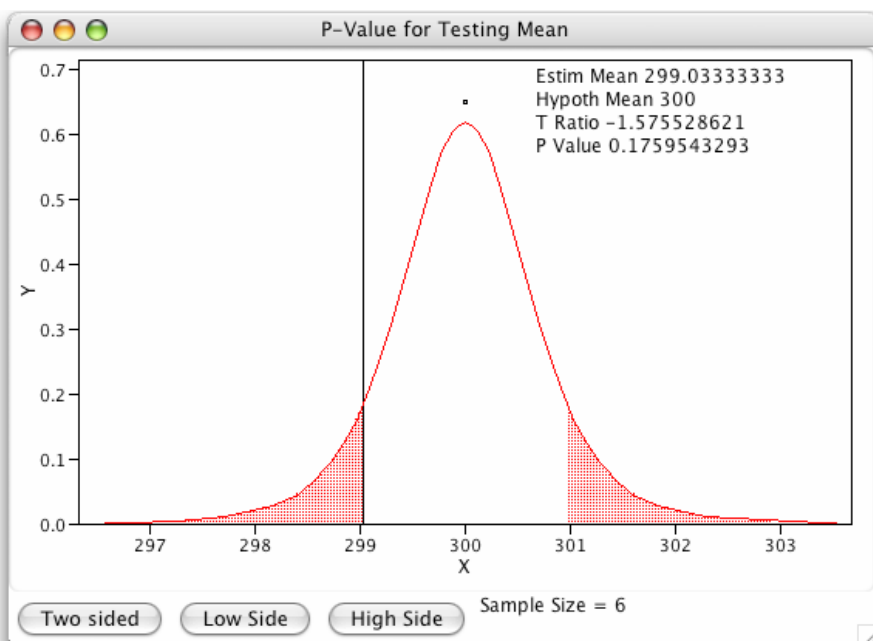


filling bottles.jmp: Distrib...	
<input checked="" type="checkbox"/> Distributions	
<input checked="" type="checkbox"/> Contents	
<input checked="" type="checkbox"/> Test Mean=value	
Hypothesized Value	300
Actual Estimate	299.033
df	5
Std Dev	1.50289
Sigma given	3
<b>z Test</b>	
Test Statistic	-0.7893
Prob >  z	0.4299
Prob > z	0.7850
Prob < z	0.2150

Since you wish to test that  $H_a: \mu < 300$  ml, use the *P*-value for a lower-tailed test, 0.2150. A sample result like this ( $\bar{x} = 299.03$ ) would happen just by chance in 21.5% of samples taken from a machine that was filling properly ( $\mu = 300$  ml). An outcome that could so easily happen just by chance is not good evidence that the machine is underfilling.

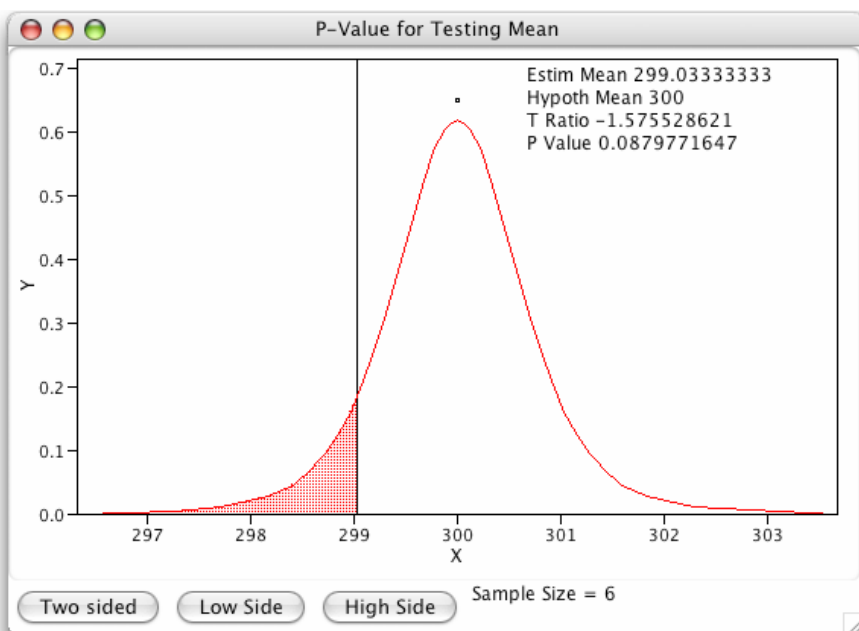


4. Click on the red triangle in the title bar of the **Test Mean=Value** report and select **PValue animation** from the menu that opens.



Examine the graph that opens. Notice that both tails are shaded and that the shaded area is associated with a two-sided alternative hypothesis. Our alternative hypothesis is a one-sided one.

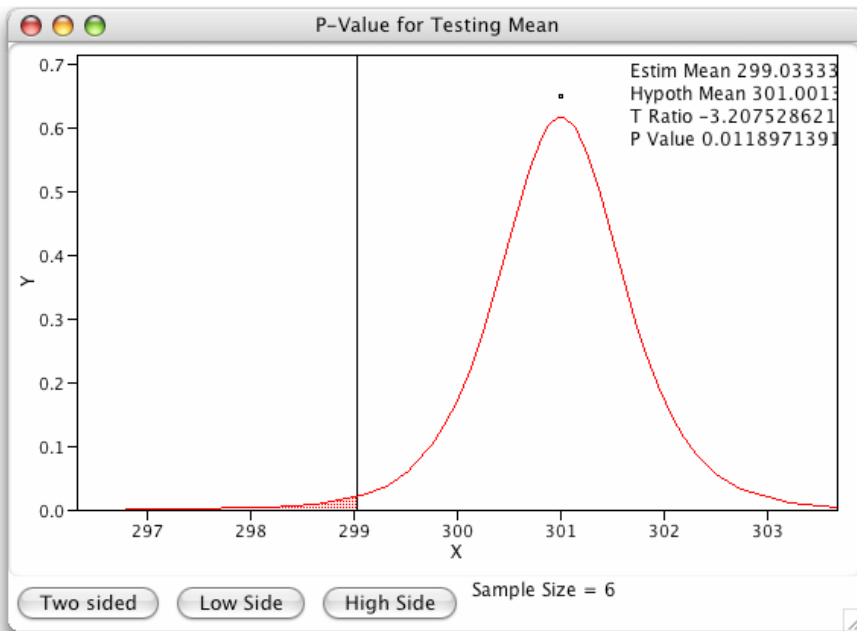
- a. Press the **Low Side** button. Notice that now only the lower tail is shaded.



The  $P$ -values calculated in the animation window do not assume that the population standard deviation is known. Therefore, they will not correspond exactly to the  $P$ -values that we calculate in this chapter. Nonetheless, their order of magnitude will be helpful.

Let's see what happens to the  $P$ -value as we move the hypothesized mean away from the observed (sample) mean.

- b. Press the handle (a small square) directly above the peak of the bell-shaped curve and drag it to the right.



The entire curve moves and the hypothesized mean becomes larger. The  $P$ -value becomes increasingly smaller as the hypothesized mean moves further away from 300. It changes from about 0.09 to 0.01 as the hypothesized mean goes from 300 to 301.

## 6.3 The Power of a Test

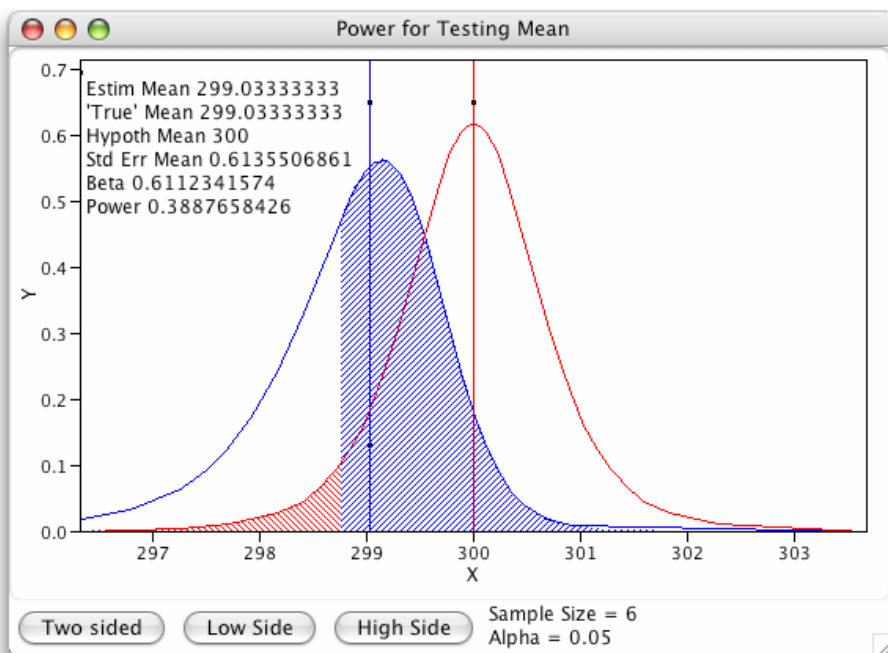
The *power* of a test is the probability that it will determine that the alternative hypothesis is true when it really is. When  $\mu$  is very close to the null mean (300 for the bottle filling example), the test should find it hard to distinguish  $\mu$  from the null mean (300 in this case) and the *power* of the test should be small. When  $\mu$  is far from the null mean, it should be easy to determine that the alternative is true and the *power* of the test should be large.

To help you better understand how the *power* is calculated and to see how the *power* increases as the difference between the null mean and a specific alternative mean increases, let's look at another animation.

**Example 6.4 Fill the bottles (cont'd.)**

1. Click on the red triangle in the title bar of the **Test Mean=Value** report and select **Power animation** from the menu that opens.
  - a. Press the **Low Side** button because the alternative is one-sided and the test is lower-tailed.

Examine the graph that opens. Notice that there are two bell-shaped curves. The one on the right is the sampling distribution of the sample mean when the null hypothesis is true; the one on the left is the sampling distribution of the sample mean when an alternative mean is true—in this instance 299.3, the observed value of the mean.



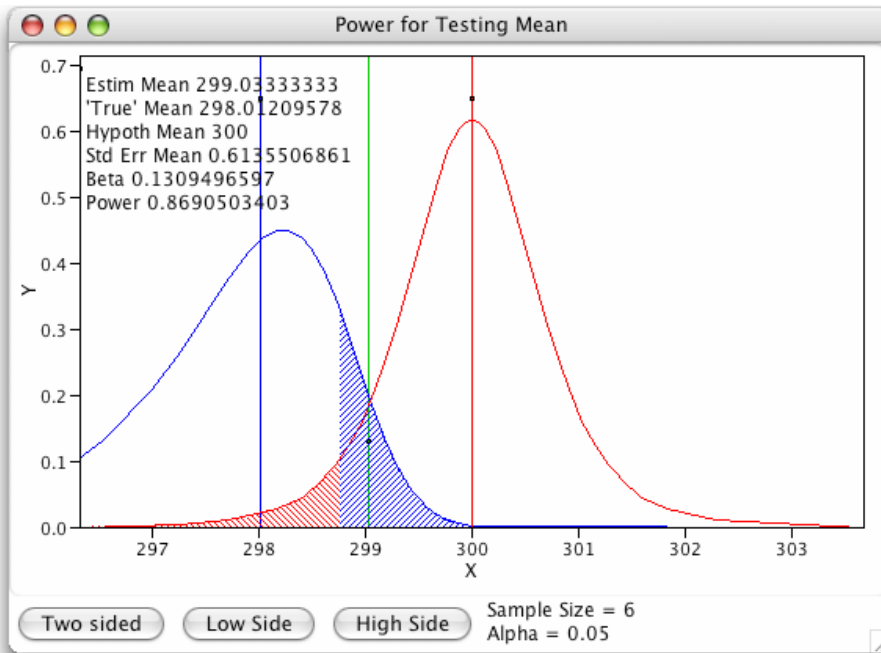
The red-shaded areas of the distribution on the right lie above values of the sample mean for which we would reject the null hypothesis. The blue-shaded area of the distribution on the left represents *1 minus the power*. The size of the blue-shaded area is 0.6212 and hence the power is 0.3888.

The power calculated in the animation window does not assume that the population standard deviation is known. Therefore, it will not correspond exactly to the power calculated in the textbook. Nonetheless, its relationship to the difference between the null mean and the alternative mean can be seen.

Let's see what happens to that shaded area, and hence the power, when we change the specified alternative mean to a value that is further from the null mean.

3. Select the handle (a small square) directly above the peak of the bell-shaped curve on the left and drag its center, the specific alternative hypothesis, to the left, say to about 298.

The size of the blue-shaded area decreases and the power increases to about .87 (0.8691 on the graph following).



What will happen if we change the specified alternative mean to a value that is close to the null mean? Move the center bell-shaped curve on the right closer to the mean by dragging the handle above its peak to the left.

## 6.4 Summary

The focus of this chapter is to introduce you to confidence intervals and significance tests and the reasoning behind them. Although *JMP* software can be useful in some ways, most of the problems can be done by hand or with a calculator.

All graphs and statistical computations in this chapter are performed in the first platform, **Distribution**, of the **Analyze** menu.

### Activity

Confidence Intervals  
Understanding Confidence

Significance tests  
Understanding *P*-values  
Power of test

### Command

**Analyze** ⇒ **Distribution** ⇒ **Confidence Interval**  
**Help** ⇒ **Sample Scripts for Students** ⇒ **Confidence.jsl**  
**Edit** ⇒ **Run Script**  
**Analyze** ⇒ **Distribution** ⇒ **Test Mean**  
**PValue animation**  
**Power animation**

# Chapter 7

## Inference for Distributions

This chapter emphasizes the practice of statistical inference. Population standard deviations are no longer (unrealistically) assumed to be known; we learn about a new family of sampling distributions, the Student  $t$  distributions; and we learn to compare two distributions.

For inference on a single variable, we employ the **Distribution** platform in *JMP*. To compare the means of two populations, we use the **Fit Y by X** platform because this is equivalent to determining whether two variables are related.

### 7.1 Inference for the Mean of a Population

Since we are concerned with only one variable in this section, the **Distribution** analysis platform is used. When the population standard deviation is unknown, as is usually the case, *JMP* calculates the sample standard deviation and uses it in place of the true standard deviation in the test statistic. The  $t$ -statistic and associated  $P$ -value are automatically calculated for tests of significance and the appropriate quantile of the  $t$  distribution is used for confidence intervals. The following examples illustrate this.

#### 7.1.1 The One-Sample $t$ Confidence Interval

---

**Example 7.1 Estimating the level of vitamin C in corn soy blend**

---

Vitamin C is an important nutrient used to fortify corn-soy blend (CSB). The following data are the amounts of vitamin C, measured in milligrams per 100 gm of CSB, for a random sample of size 8 from a production run of CSB:

26    31    23    22    11    22    14    31

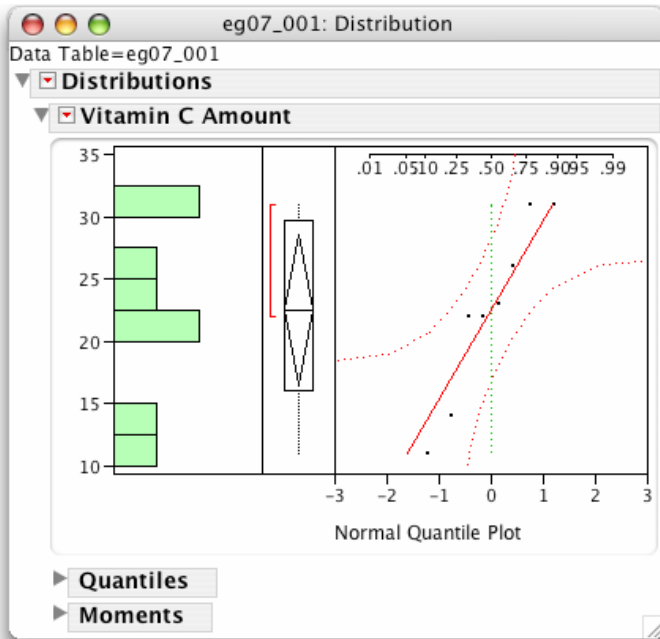
Find a 95% confidence interval for  $\mu$ , the mean vitamin C content of the CSB produced during this run.

1. Select **File**  $\Rightarrow$  **New**.

Create a *JMP* data table with one variable named **Vitamin C Amount** and six rows. Save the data table for later use. (See Section 0.2.1 in Chapter 0 for details on creating a new *JMP* data table.)

2. Select **Analyze** ⇒ **Distribution**.

- Select **Vitamin C Amount** and press the **Y, Columns** and **OK** buttons as we did in Chapter 1.
- Press the red triangle next to **Vitamin C Amount** and select **Normal Quantile Plot**.

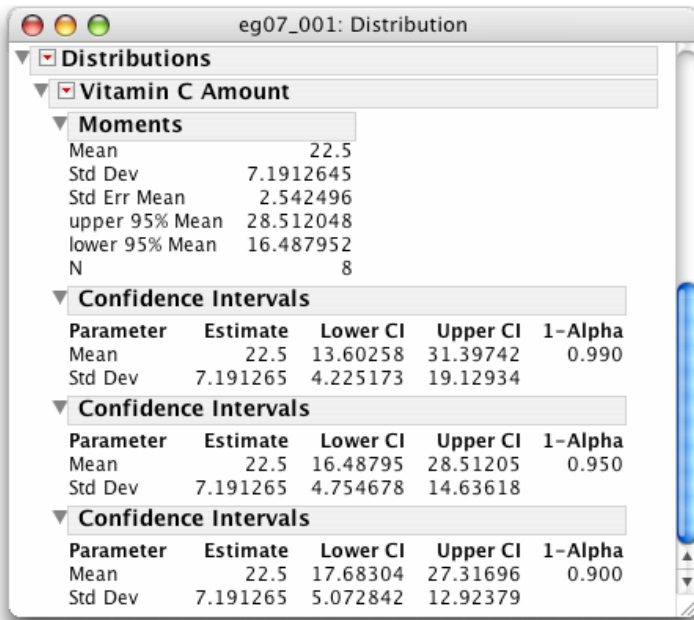


While it is difficult at best to check the conditions for the use of the  $t$  confidence interval with only 8 observations, we can see from the histogram and the Normal quantile plot that there are no striking deviations from Normality. (See Section 1.3 for more on Normal quantile plots.)

Besides the 95% confidence interval, let's also obtain 99% and 90% confidence intervals for the mean  $\mu$  vitamin C content of the CSB produced during this run.

3. Click on the red triangle next to **Vitamin C Amount** and select **Confidence Interval** ⇒ **.99**.

Repeat this for a 95% confidence interval and for a 90% interval to obtain the following reports:



The 95% confidence interval for the mean amount of vitamin C in this production run is (16.49, 28.51). If we are willing to be less confident (90%), the interval will be more precise; i.e., the interval will be narrower and the margin of error smaller (17.68, 27.32). If you want to be more confident (99%), the interval will be less precise; i.e., the interval will be wider and the margin of error larger (13.60, 31.40).

## Remarks

- There was no need to request the 95% confidence interval for the mean. Since 95% is a frequently used confidence level, *JMP* displays the endpoints of the 95% interval in the **Moments** report.
- You can change the number of decimal places displayed for the mean and confidence intervals. Simply double-click on any of the values and a numeric format dialog panel will be displayed.
- Notice the diamond in the boxplot. Its horizontal axis is located at the sample mean (22.5) and the vertical span represents the 95% confidence interval (16.5, 28.5). *JMP* calls this a *means diamond* and displays it when the mean and standard deviation are appropriate descriptors for a distribution. We first encountered means diamonds in Section 2.1.2 in Chapter 2 of this guide. We will use them later in this chapter and again in Chapters 12 and 13 when we compare means of several distributions.

## 7.1.2 The One-Sample $t$ Test

We also use the **Distribution** platform to perform one-sample  $t$  significance tests.

### Example 7.2 Diversify or be sued

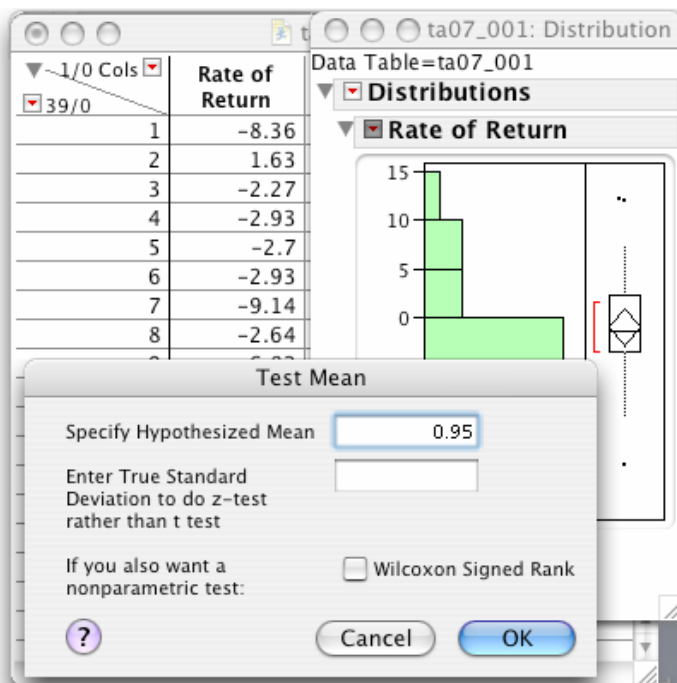
An investor with a stock portfolio worth several hundred thousand dollars sued his broker and brokerage firm because lack of diversification in his portfolio led to poor performance. The arbitration panel compared the rates of return for the 39 months that the account was managed by the broker with the average of the S&P 500 for the same period. The S&P 500 gained a little under 1% (0.95% to be exact)

during that period. Consider the 39 monthly returns as a random sample from the population of monthly returns that the brokerage would generate if it managed the account forever. Thus, the arbitration panel wishes to test the alternative hypothesis

$$H_a: \mu \neq 0.95$$

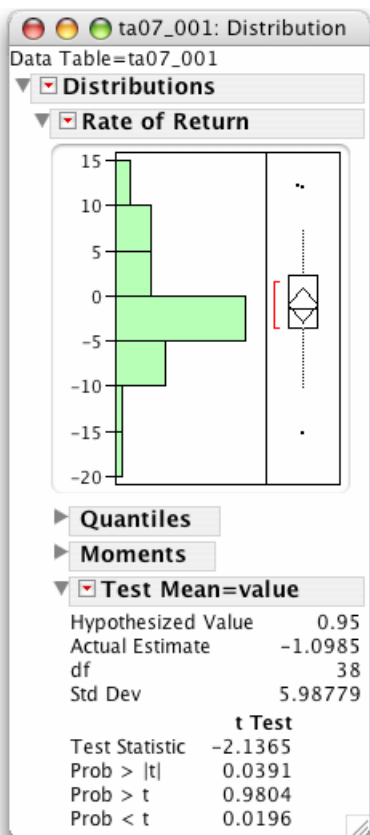
where  $\mu$  is the mean of the population of monthly returns. The *JMP* data table **ta07\_001.jmp** contains the rates of return.

1. Select **File** ⇒ **Open** and the file **ta07\_001.jmp** from the appropriate location.
2. Select **Analyze** ⇒ **Distribution** to display the distribution of **Rate of Return**.
  - a. Select **Rate of Return** and press the **Y, Columns** and **OK** buttons.
  - b. Click on the red triangle next to the title bar for the **Rate of Return** report and select **Test Mean**.



3. Type **0.95** in the **Specify Hypothesized Mean** field and select **OK**.





As was the case for the  $z$  test in Chapter 6, *JMP* calculates the value of the one-sample  $t$  test for the sample and provides three  $P$ -values, one for each of the three possible alternative hypotheses. We choose the one appropriate to the alternative hypothesis that we wish to test.

<u>Alternative Hypothesis</u>		<u><math>P</math>-value in <i>JMP</i></u>
$\mu \neq 0.95$	Two-tailed	Prob >  t
$\mu > 0.95$	Upper tail	Prob > t
$\mu < 0.95$	Lower tail	Prob < t

Since our alternative hypothesis is that  $H_a: \mu \neq 0.95$ , we use the  $P$ -value for a two-sided test. Thus, we conclude that the mean monthly return on investment for this client's account,  $\bar{x} = -1.0985\%$ , is significantly lower than the S&P average for the same period,  $\mu = 0.95\%$  with ( $t = -2.1367$ ,  $df = 38$ , and  $P = 0.0391$ ).

### 7.1.3 Matched Pairs $t$ Procedures

The analysis of a matched pairs design is straightforward. A matched pairs study has two measurements on each individual and the objective is to compare the measurements by examining the difference between the measurements. To do this, we create a new column that contains the difference between the two measurements for each individual. Then, we analyze the difference using one-sample  $t$  procedures as above.

### Example 7.2 The effects of the moon on aggressive behavior

Many people believe that the moon influences the actions of some individuals. A study of dementia patients in nursing homes recorded various types of disruptive behaviors every day for 12 weeks. For each patient the average number of disruptive behaviors was computed for moon days (within 1 day of a full moon) and for all other days. The data for subjects with aggressive behavior are found in the *JMP* data table **ta07\_002.jmp**.

To assess whether there is a difference in aggressive behaviors on moon days versus other days, we test the hypothesis that the mean number of aggressive behaviors on moon days differs from that for other days. Thus, if  $\mu$  represents the mean difference in average number of aggressive behaviors, moon minus other days, we test the alternative hypothesis

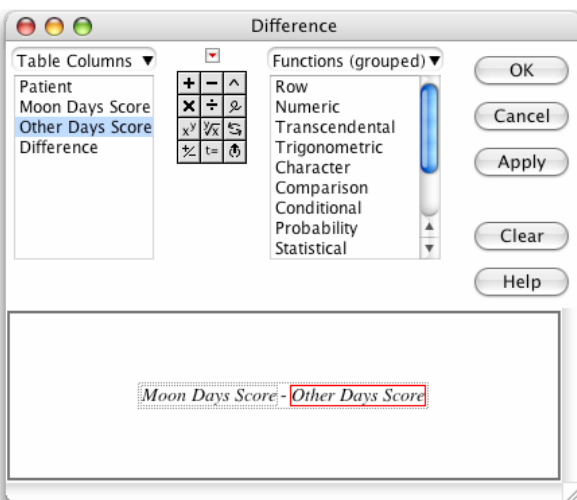
$$H_a: \mu \neq 0$$

This is a matched pairs study. Each patient (individual) has two aggressive behavior scores, one for moon days and another for other days, and we compare the scores by examining the difference between these (aggressive behavior on moon days minus that for other days). We create a new variable called **Difference** that holds the differences. The 20 differences form a single sample. We then use the **Test Mean** command to determine whether the population mean gain is greater than zero.

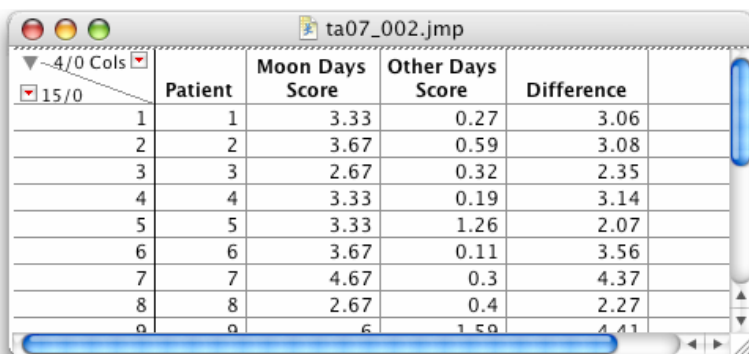
1. Select **File**  $\Rightarrow$  **Open** and the file **ta07\_002.jmp** from the appropriate location.

The JMP data table contains three columns, **Patient**, **Moon Days Score**, and **Other Days Score**, respectively. We need to create a new variable called **Difference**, which is the **Moon Days Score** minus the **Other Days Score** (see Section 0.3.3 in Chapter 0).

2. Select **Cols**  $\Rightarrow$  **New Column** and name the column **Difference**.
3. Select **Column Properties**  $\Rightarrow$  **Formula** and **Edit Formula** to open the Formula Editor window.
  - a. Select **Moon Days Score** from the list of columns.
  - b. Press **-** (minus) on the keypad and select **Other Days Score** from the list of columns.
  - c. Press **OK** and **OK** again.



Check that the new variable **Difference** is indeed the **Moon Days Score** minus the **Other Days Score**.

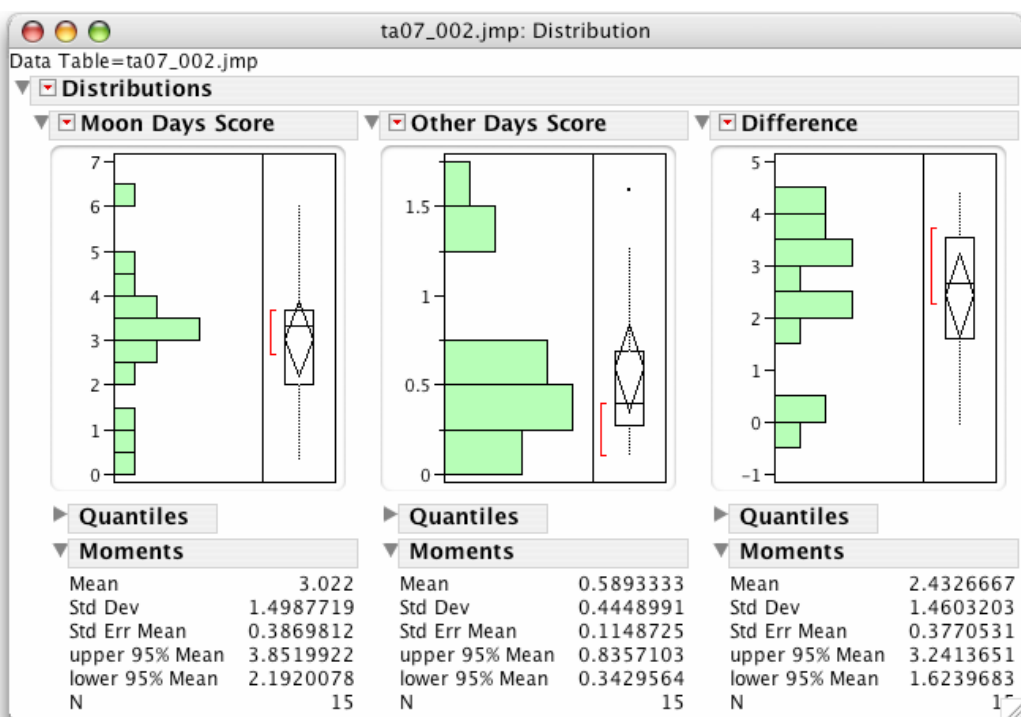


	Patient	Moon Days Score	Other Days Score	Difference
1	1	3.33	0.27	3.06
2	2	3.67	0.59	3.08
3	3	2.67	0.32	2.35
4	4	3.33	0.19	3.14
5	5	3.33	1.26	2.07
6	6	3.67	0.11	3.56
7	7	4.67	0.3	4.37
8	8	2.67	0.4	2.27
9	9	6	1.59	4.41

Let's calculate the mean moon days score, the mean other days score, and the mean difference.

4. Select **Analyze** ⇒ **Distribution**.

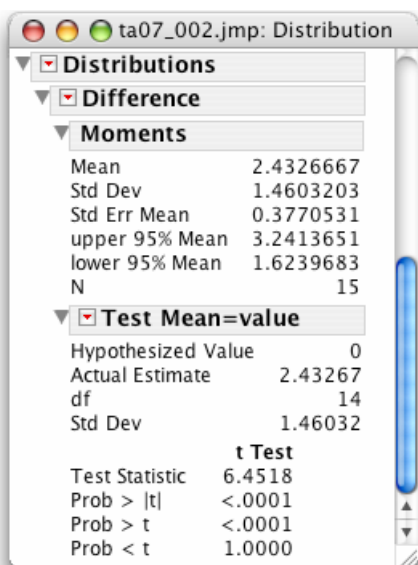
- Select **Moon Days Score**, **Other Days Score**, and **Difference** from the list of columns and then press the **Y, Columns** and **OK** buttons.



The mean **Difference** for the sample of 15 patients was 2.433, which is the difference between the mean **Moon Days Score** 3.022 and the mean **Other Days Score** 0.589. To test whether a mean difference of 2.433 is statistically significant, we use the **Test Mean** command.

5. Press the same red triangle next on the report title **Difference** and select **Test Mean**.

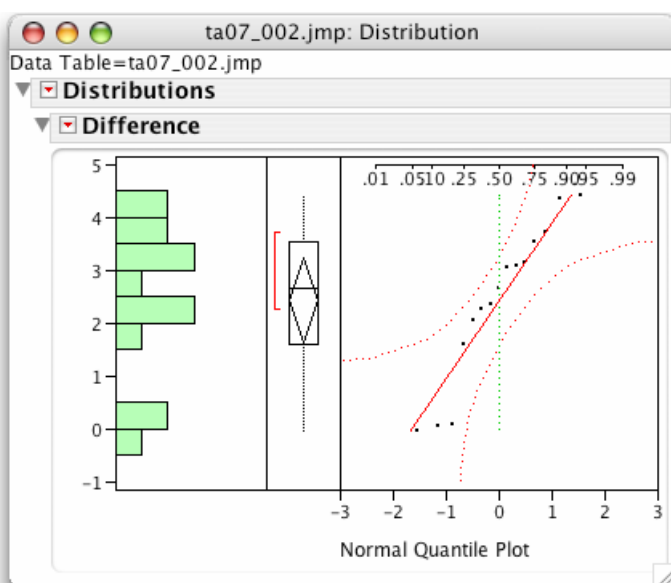
- Type 0 in the **Specify Hypothesized Mean** field and select **OK**.



The mean improvement,  $\bar{x} = 2.433$ , is significantly different from zero ( $t = 6.4518$ ,  $df = 14$ ,  $P < 0.0001$ ). This conclusion is evident upon a close look at the boxplot for **Difference**. The means diamond does not contain zero. A 95% confidence interval for the mean **Difference** is (1.62, 3.24).

## Remarks

- Look again at the *JMP* data table **ta07\_002.jmp**. Each row represents a subject's results. The scores are placed in two columns, not one, because each subject has two scores. All matched pairs studies share a similar data table layout.
- Since a one-sample  $t$  test requires that the data be Normally distributed, before submitting our results (actually, before performing the test), we would examine a Normal quantile plot for the difference. Press the red triangle next to the report on **Difference** and select Normal Quantile Plot.



We notice that there are three patients with fairly small differences. Whether or not these are outliers is a matter of judgment.

- *JMP* offers another approach to the analysis of matched pairs studies—the **Matched Pairs** command found on the Analyze menu. It provides the same output as above together with an interesting plot. To learn more about the command and plot, select **Analyze** ⇒ **Matched Pairs**. Then, press the **Help** button.

## Power of the $t$ Test

The **Power animation** command in the **Test Mean=value** report displays the power of a given  $t$  test to detect alternative values of the population mean. *JMP* offers a powerful method of calculating prospective power and determining sample size, the **Sample size and Power** command in the **DOE** platform.

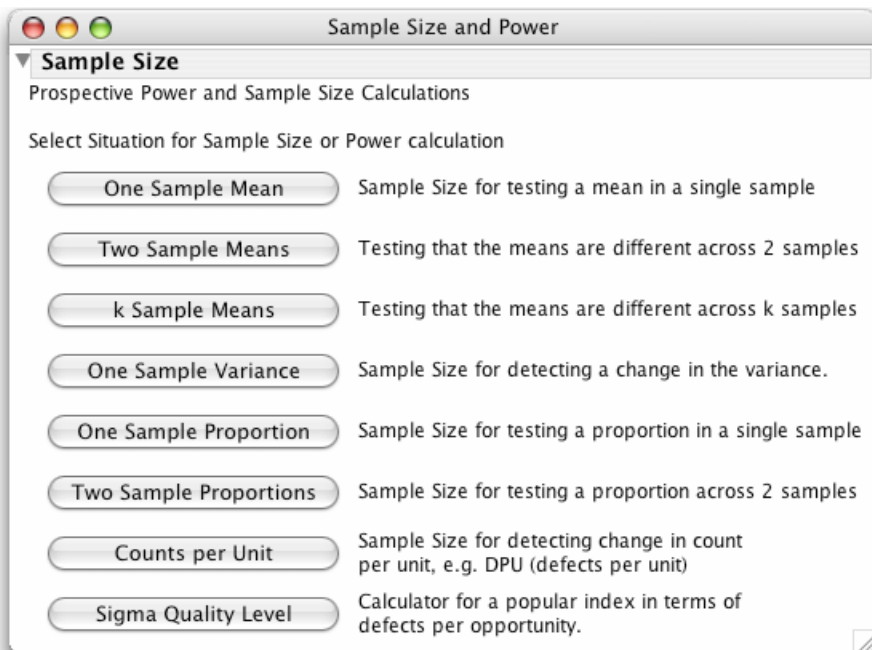
### Example 7.3 The effects of the moon on aggressive behavior

In the previous example, we examined the effect of the moon on the aggressive behavior of dementia patients in nursing homes. Suppose that we want to perform a similar study in a different setting. Determine the power of the  $t$  test for the alternative  $\mu = 1$  when  $\sigma = 1.5$ ,  $\alpha = 0.05$ , and there are 20 subjects. Since we expect the moon days will be associated with an increase in aggressive behavior, the  $t$  test is for

$$H_0: \mu = 0 \text{ versus } H_a: \mu > 0$$

against the alternative  $\mu = 1.0$ .

1. Select **DOE** ⇒ **Sample Size and Power**.



2. Select **One Sample Mean**.

Sample Size and Power

**Sample Size**

One Mean

Testing if one mean is different from the hypothesized value.

Alpha

Error Std Dev

Extra Params

Supply two values to determine the third.  
Enter one value to see a plot of the other two.

Difference to detect

Sample Size

Power

**Continue**

**Back**

**Animation Script**

Since *JMP* assumes a two-sided test is being performed, use  $\alpha = 2 \times .05 = 0.10$ .

- Type 0.10 for **Alpha**.
- Type 1.5 for **Err Std Dev**.
- Type 1 for **Difference to detect**.
- Type 20 for **Sample Size**.
- Select **Continue**.

Sample Size and Power

**Sample Size**

One Mean

Testing if one mean is different from the hypothesized value.

Alpha

Error Std Dev

Extra Params

Supply two values to determine the third.  
Enter one value to see a plot of the other two.

Difference to detect

Sample Size

Power

**Continue**

**Back**

**Animation Script**

The power of this experiment is 0.8902.

## Remark Regarding Inference for Non-Normal Distributions

- Taking logarithms is a common remedy for a lack of Normality due to right skewness. Simply create a new variable **log X** using the Formula Editor to compute the logarithm of the original variable **X**. See Section 0.3.3.

## 7.2 Comparing Two Means

Two-sample problems are among the most common designs encountered in statistical practice. We often select simple random samples from two populations to compare their means or carry out a completely randomized design to compare two groups of individuals to a response variable. Unlike the matched pairs designs, there is no matching of the units in each group or sample. No individual belongs to more than one group and the responses in one group are independent of those in the other group. Hence, the data table for such a design consists of a column for the response variable and a column to identify the groups or treatments. The column that identifies the groups or treatments can be thought of as an explanatory variable or a factor. Then, comparing groups is equivalent to seeing if the response variable depends on the explanatory variable. To put it another way, we are interested in the relationship between the two variables where one, the response variable, is quantitative and the other, the explanatory variable, is categorical.

We used the **Fit Y by X** analysis platform to look at this type of relationship in Section 2.1.2 in Chapter 2. Side-by-side comparisons of the distributions of the response variable for each group were displayed with side-by-side boxplots and side-by-side means diamonds. To extend our conclusions to the populations from which the data are samples, we use the **Means/Anova/t-Test** command. (If you skipped Section 2.1.2 in Chapter 2 or have forgotten the material, now is a good time to review it.)

### Example 7.4 Are directed reading activities effective?

An educator believes that directed reading activities in the classroom would help improve some aspects of the reading ability of third graders. A class of 21 third-grade students took part in the new directed reading activities and another third-grade class of 23 followed the same curriculum without these activities. The Degree of Reading (DRP) test scores for the 44 students are found in a text file named **ta07\_004.txt**. The educator hoped to show that children who receive the new directed reading activities (the treated group) will do better than those who do not (the control group); the alternative hypothesis to be tested is

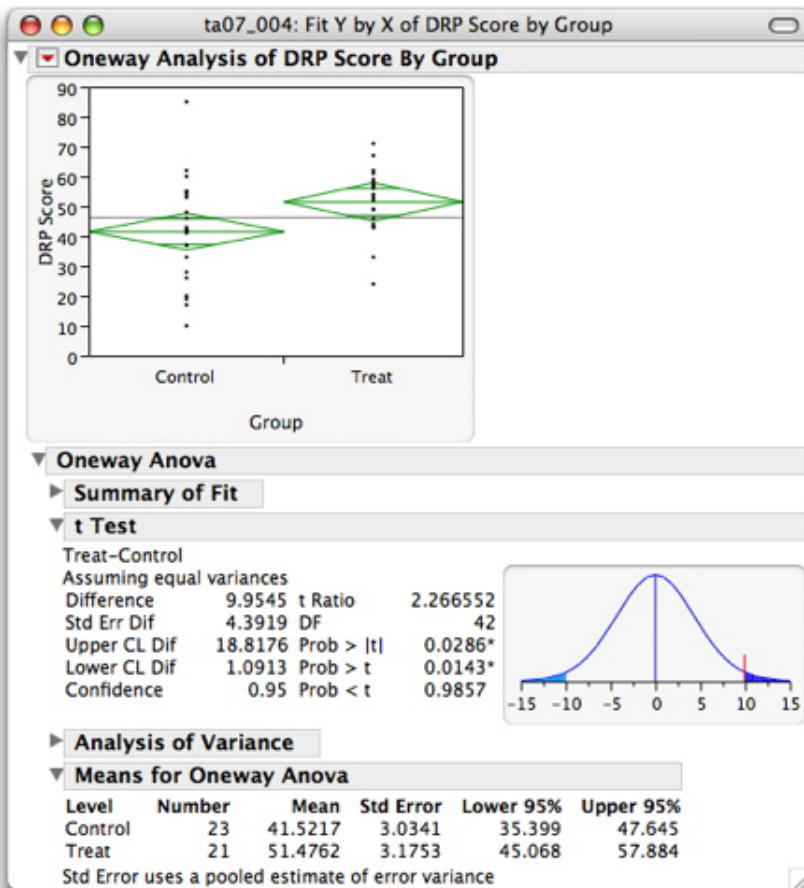
$$H_a: \mu_T > \mu_C$$

To perform the analysis in *JMP*, the data must be placed in a *JMP* data table in a certain way. There must be 44 rows, one for each student. There must be a column containing the DRP scores, the response variable, and a column identifying the group to which the student was assigned.

1. Import the file **ta07\_004.txt** into *JMP* to see how the data is coded. Fortunately, the data file is in the correct format. (See Section 0.2.3 in Chapter 0 for details on importing a text file into *JMP*.)
2. Name the four columns **Student**, **Group**, **Group Code**, and **DRP Score**, respectively.

Proceed as in Section 2.1.2 in Chapter 2 to study the relationship between the DRP scores and the groups.

3. Select **Analyze**  $\Rightarrow$  **Fit Y by X**.
  - a. Select **DRP score**  $\Rightarrow$  **Y, Response**.
  - b. Select **Group**  $\Rightarrow$  **X, Factor** and then select **OK**.
4. Click on the red triangle next to **Oneway Analysis of DRP Score By Group** and select **Means/Anova/Pooled t**.



Inspect the side-by-side means diamonds. The vertical span of each diamond represents the 95% confidence interval estimate for the mean of each group. The horizontal lines above and below each group mean are called overlap marks. For groups with equal sample sizes, overlapping of these marks indicates that the two group means are not significantly different at the .05 significance level (for a two-sided test). Because the diamonds assume that the groups have equal variances, we should only use the means diamonds as rough indicators.

Inspect the **t Test** report. The *pooled estimate for the variance* is used to calculate the *t* statistic and a confidence interval estimate for the difference. (The *t* test that does not assume that the group variances are equal can be found in Section 7.3 following the test for unequal variances.)

**Difference**      difference between the two sample means (9.9545 points in this example)  
**Std. Error**      estimated standard error of the difference between the two sample means using the pooled estimate of the variance (4.3919)

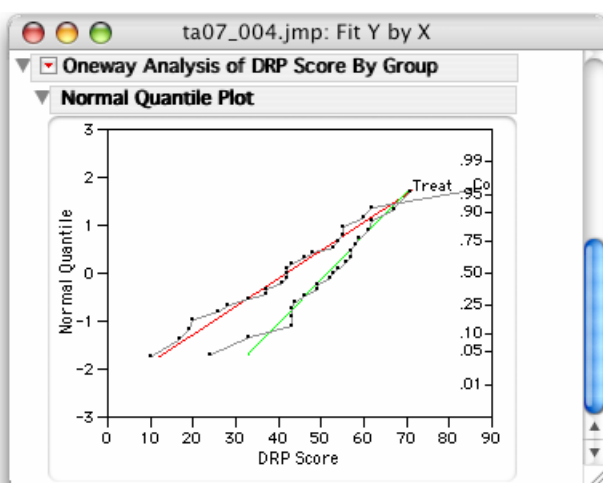


<b>t Ratio</b>	value of the pooled two-sample $t$ statistic (2.2666)
<b>DF</b>	degrees of freedom of the pooled two-sample $t$ statistic (42 here)
<b>Prob &gt;  t </b>	$P$ -value for a two-sided alternative hypothesis, $\mu_T \neq \mu_C$
<b>Prob &gt; t</b>	$P$ -value for an upper-tailed alternative, $\mu_T > \mu_C$
<b>Prob &lt; t</b>	$P$ -value for a lower-tailed alternative, $\mu_T < \mu_C$
<b>Lower 95%</b>	lower bound on the 95% confidence interval for the difference between the means
<b>Upper 95%</b>	upper bound on the 95% confidence interval for the difference between the means

Thus, for our alternative hypothesis, the  $P$ -value is 0.0143. We conclude that the data strongly suggests directed reading activity improves the DRP score ( $t = 2.266$ ,  $df = 42$ ,  $P = 0.0143$ ). The 95% confidence interval estimate for the difference in mean DRP scores  $\mu_T - \mu_C$  between the groups of students is (1.0913, 18.8176).

Our  $P$ -values and hence our conclusions have assumed that the data are Normally distributed. To assess this, we look at Normal quantile plots for the groups.

- Click on the red triangle next to **Oneway Analysis of DRP Score By Group** and select **Normal Quantile Plot**  $\Rightarrow$  **Plot Quantile by Actual** to obtain Normal quantile plots for both groups.



The plots confirm that the DRP scores for both groups are approximately Normally distributed.

## Remark

- Sometimes the treatments or groups are given numeric codes; for example, 0 and 1. To analyze the data, we must ensure that the *modeling type* of the group variable, or factor, is nominal. This is **very important** since *JMP* will display a scatterplot and offer tools for fitting curves when the explanatory variable has a continuous *modeling type* (the default for numeric data type). You can change the *modeling type* of a variable/column with the **Cols Info** command in the **Cols** menu. Alternatively, if the left table information panels (see Section 0.4 in Chapter 0) are open, you may select the icon next to the column name.

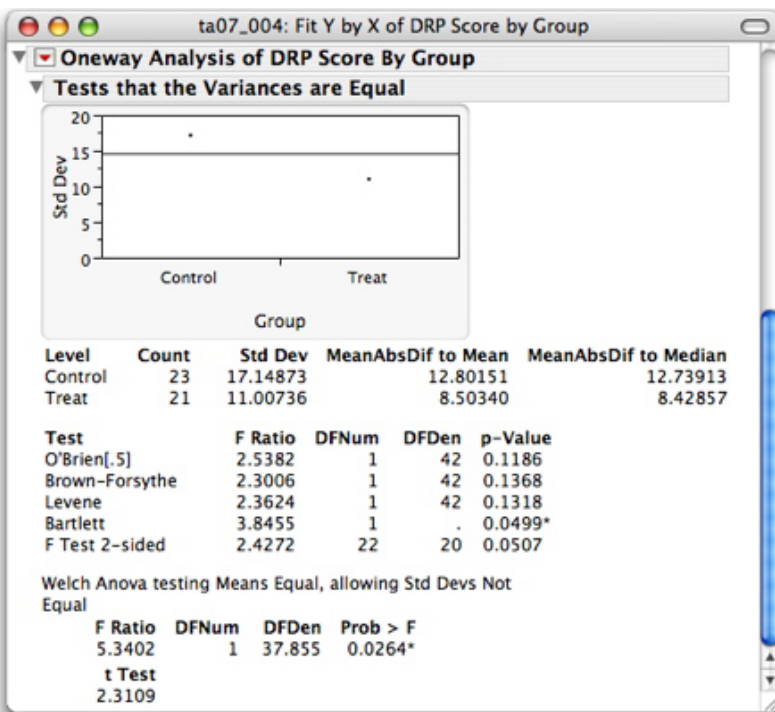
## 7.3 Testing for Unequal Variances


Pooled two-sample  $t$  procedures assume that the variances of both populations are equal. The **UnEqual Variances** command in the **Fit Y by X** platform in *JMP* provides several procedures for verifying that assumption.

### Example 7.5 Are directed reading activities effective? (cont'd.)

To determine whether we can safely use the pooled two-sample  $t$  test, we compare the variances of the two populations.

1. If the *JMP* data table **ta07\_004.jmp** from the last section is closed, open it.
2. Select **Analyze**  $\Rightarrow$  **Fit Y by X**.
  - a. Select **DRP Score**  $\Rightarrow$  **Y, Response**.
  - b. Select **Group**  $\Rightarrow$  **X, Factor** and then select **OK**.
3. Click on the red triangle next to **Oneway Analysis of DRP Score By Group** and select **UnEqual Variances**.



Five different tests for unequal variances  $H_1: \sigma_1^2 \neq \sigma_2^2$  are reported. To see an explanation of each test, select the  tool and click on any of the test names. The **Help** window will open.  $P$ -values for the tests are given in the **Prob > F** column. In practice, if any of the tests are significant, you should **not** use the pooled estimate of the variance and the pooled two-sample  $t$  procedures.

In this case, the  $P$ -value for Bartlett's test is less than .05. The variability in the scores for the control group is significantly larger than that for the treated group. Thus, the pooled estimate of the variance and the pooled two-sample  $t$  test should **not** be used here.

## The Two-Sample $t$ Test (not assuming equal variances)

The two-sample  $t$  significance test, which *does not assume that the unknown population standard deviations are equal*, can be found at the bottom of the Tests that the Variance are Equal report. Items in the "Welsh Anova" of interest to us are:

<b>t Test</b>	two-sample $t$ statistic
<b>DFDen</b>	degrees of freedom of the $t$ distribution (software approximation)
<b>Prob &gt; F</b>	$P$ -value for a two-sided alternative hypothesis. For a one-sided test, the $P$ -value is either half of this value or one minus half of it, depending on the direction of the one-sided alternative.

### Example 7.4 Are directed reading activities effective? (cont'd.)

In our example, we have  $t = 2.3109$ ,  $df = 37.855$ , and  $P = 0.0264$ . We conclude that the data provides sufficient evidence that directed reading activity improves the mean DRP score.

### Remark

- The tests for unequal variances described in this section and the assessment of Normality of the response variable described in the previous section are also used when we compare more than two means in Chapters 12 and 13.

## The Power of the Two-Sample $t$ Test

*JMP* can calculate power and sample size using the **Power and Sample Size** command on the **DOE** platform.

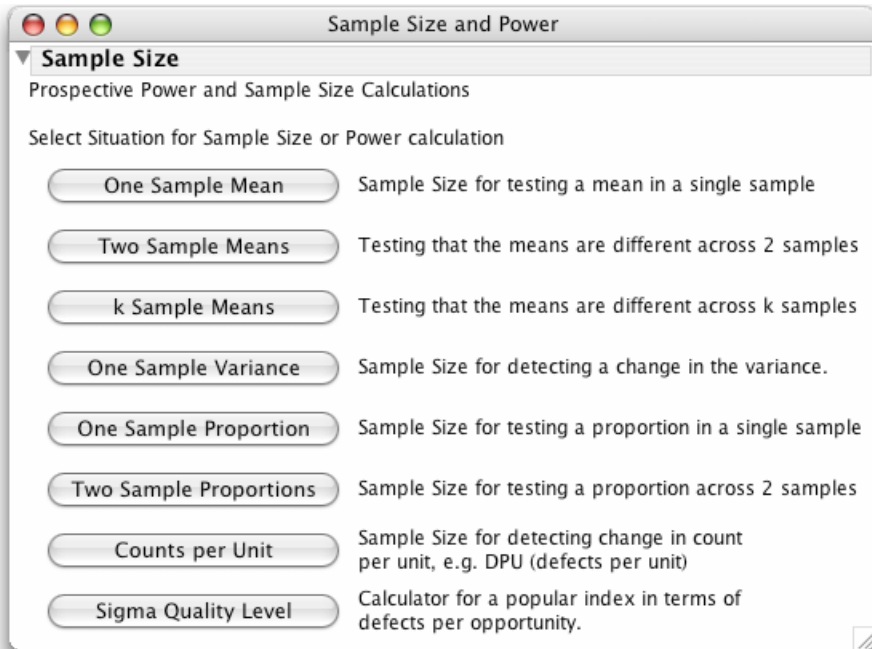
### Example 7.6 The effect of calcium on blood pressure

A study to determine whether calcium lowers blood pressure more than a placebo is planned. It should provide convincing evidence, say at the 0.01 level. There will be 45 subjects in each group (90 overall) and based on previous results, we use  $\sigma = 7.4$  and we chose  $\mu_c - \mu_p = 5$  as an alternative that we would like to be able to detect with  $\alpha = 0.01$ . The hypotheses to be tested are

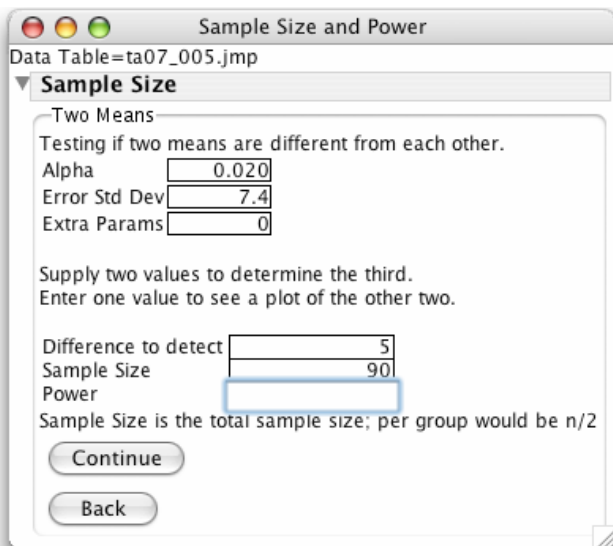
$$H_0: \mu_c = \mu_p \text{ versus } H_a: \mu_c > \mu_p$$

where  $\mu_c$  and  $\mu_p$  are the mean decreases in blood pressure for the calcium group and the placebo group, respectively.

1. Select **DOE**  $\Rightarrow$  **Sample Size and Power**.

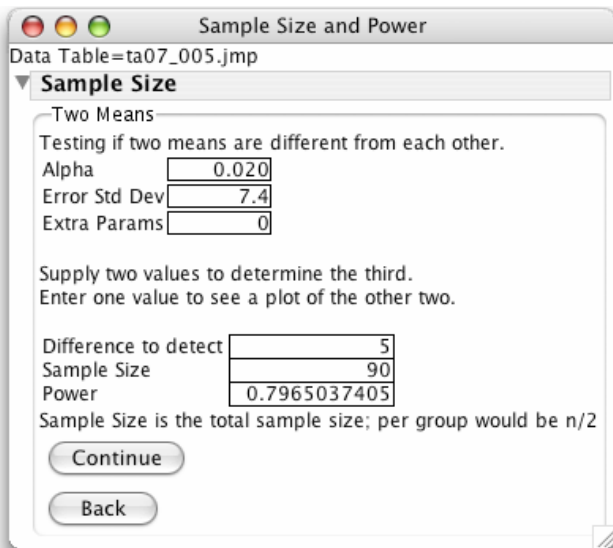


2. Select **Two Sample Means**.



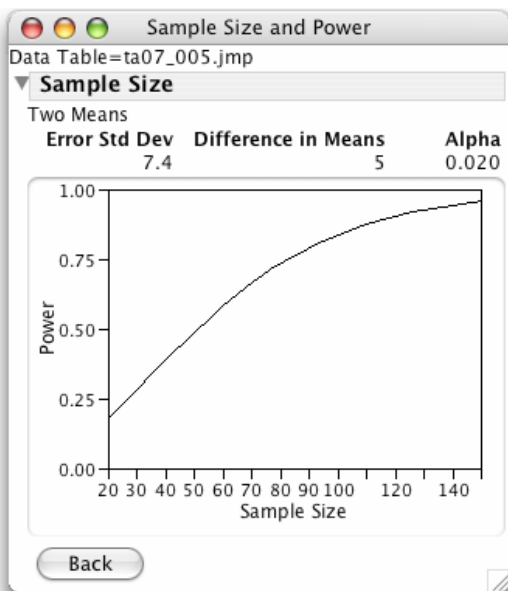
Since *JMP* assumes a two-sided test is being performed, use  $\alpha = 2 \times 0.01 = 0.02$ .

- Type **0.02** for **Alpha**.
- Type **7.4** for **Err Std Dev**.
- Type **5** for **Difference to detect**.
- Type **90** for **Sample Size**.
- Select **Continue**.



The power of this experiment is 0.7983. If the values of sample size and power are missing, *JMP* provides a plot of power versus sample size. Try it.

3. In the **Sample Size and Power** window,
  - a. Delete the value of **Sample Size**.
  - b. Delete the value of **Power** and press Continue.



Examine the graph. What sample size is required for a power of .75? .90?

## 7.4 Summary

All graphs and statistical computations in this chapter are performed in the **Distribution** and **Fit Y by X** platforms of the **Analyze** menu.

<b>Graph/Computation</b>	<b>Command</b>
One-sample procedures	
Confidence intervals	<u>Distribution</u> ⇨ <u>Confidence Intervals</u>
Significance tests	<u>Distribution</u> ⇨ <u>Test Mean</u>
Matched pairs*	<u>Distribution</u> ⇨ <u>Confidence Intervals</u> and <u>Distribution</u> ⇨ <u>Test Mean</u>
Power	<u>Distribution</u> ⇨ <u>Test Mean</u> ⇨ <u>Power Animation</u> <u>DOE</u> ⇨ <u>Sample Size and Power</u> **
Two-sample procedures**	
Pooled two-sample <i>t</i> procedures	<u>Fit Y by X</u> ⇨ <u>Means/Anova/Pooled t</u>
Two-sample significance test	<u>Fit Y by X</u> ⇨ <u>Unequal Variances</u>
Evaluating assumptions	
Normal quantile plots	<u>Fit Y by X</u> ⇨ <u>Normal Quantile Plots</u>
Equal variances	<u>Fit Y by X</u> ⇨ <u>Unequal Variances</u>
Power	<u>DOE</u> ⇨ <u>Sample Size and Power</u>

\* The variable is the *difference* between the paired values.

\*\* The *modeling type* of the explanatory variable must be nominal. The default modeling type of character variables is nominal, but the default modeling type of numeric variables is continuous.

# Chapter 8

## Inference for Proportions

This chapter is the first of two that deal with response variables that are categorical rather than quantitative. In this chapter, inference for one population proportion and inference for comparing proportions in two populations are discussed. In Chapter 9, the case of three or more populations and the more general question of whether two categorical variables are related are presented.

### 8.1 Inference for a Single Proportion

Confidence intervals and tests of significance about a single population proportion are easy to perform in *JMP*. Since there is only one variable involved, we use the **Distribution** platform. The **Confidence Interval** command generates a *score* confidence interval. The **Test Probabilities** command performs a *Pearson chi-squared* test, which is equivalent to a *z* test.

#### Large-sample confidence interval for a single proportion

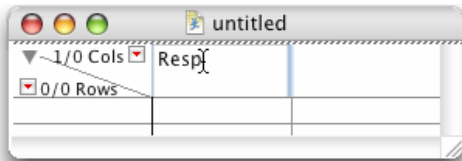
##### Example 8.1 Work stress

---

The human resources manager of a chain of restaurants is concerned that work stress may be affecting the chain's employees. He asks a random sample of 100 employees to respond Yes or No to the question, "Does work stress have a negative impact on your personal life?" Of these, 68 say "Yes." Find a 99% confidence interval estimate for the proportion of all the restaurant chain's employees who feel that work stress is having a negative impact on their personal lives.

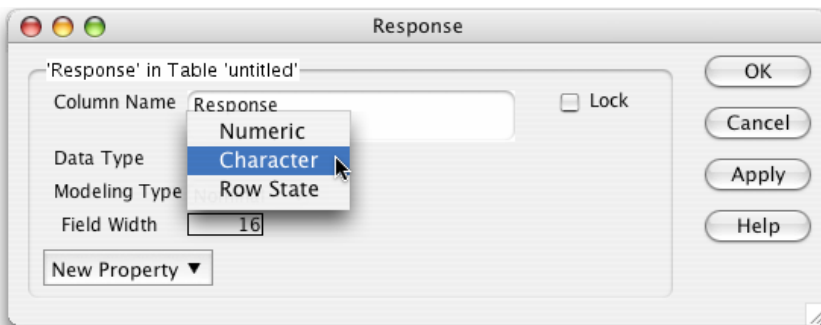
**Entering the data.** For a *JMP* data table that will hold only categorical variables, we usually store the data in a more efficient way than using one row for each individual. Instead we use one row for each category and include a column that holds the frequency of that count among all individuals. The individuals in this study are the employees of the restaurant chain. The variable of interest is an employee's response to the question, "Yes" or "Other." Thus, we can use just two rows (one for each value) and an additional column **Count** in a *JMP* data table to summarize the responses for the 100 employees.

1. **File** ⇒ **New**.
  - a. Select the first column **Column 1**, and type **Response**.



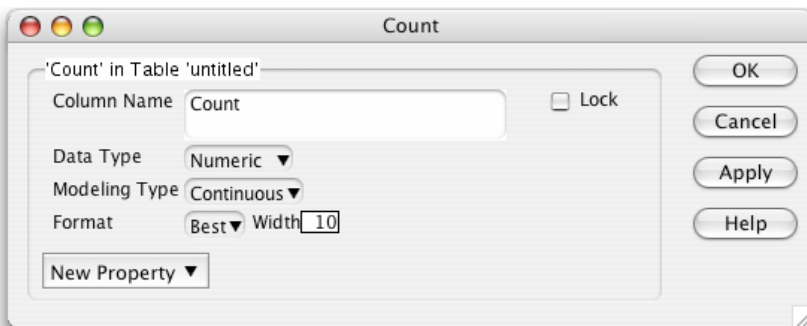
By default, columns contain numeric data. Change the *data type* of **Response** to “Character.” Notice that the *modeling type* automatically changes from “Continuous” to “Nominal.” For a discussion of the modeling type of a variable, see the end of Section 0.1.4 of Chapter 0.

2. Select the column **Response**.
  - a. Select **Cols** ⇒ **Column Info**.
  - b. Select **Character** from the **Data Type** menu and press **OK**.



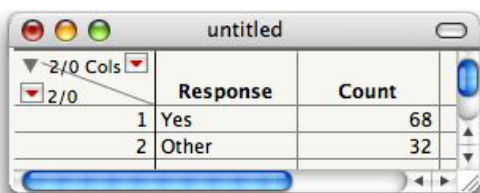
Add the variable **Count**.

3. Select **Cols** ⇒ **New Column...**.
  - a. Type Count in the Column Name field and press **OK**.



4. Select **Rows** ⇒ **Add Rows** from the menu bar and enter **2**.
  - a. Fill in the data grid as in the following figure.



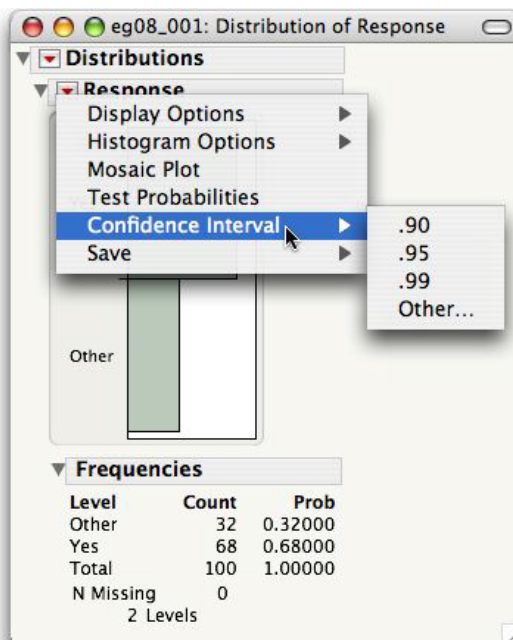
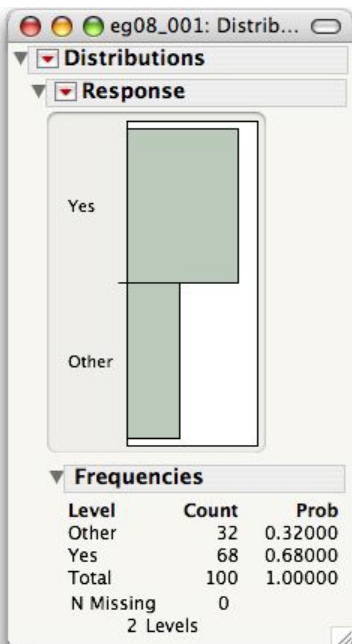


	Response	Count
1	Yes	68
2	Other	32

5. Select **File** ⇒ **Save** and name the data table **eg08\_001.imp**.

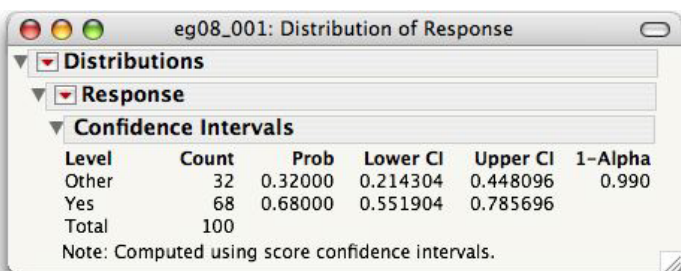
**Finding the confidence interval estimate.** First, let's look at the distribution of responses using the **Distribution** platform. (See Section 1.1.1 of Chapter 1.)

6. Select **Analyze** ⇒ **Distribution**.
- Select **Response** from the list of columns and press **Y, Column**.
  - Select **Count** and press **Freq** and **OK**.



To obtain a 99% confidence interval for the proportion  $p$  of all the restaurant chain's employees who feel that work stress is damaging their personal lives:

7. Press the red triangle on the **Response** report and select **Confidence Interval** ⇒ **.99**.



eg08\_001: Distribution of Response

Distributions					
Response					
Confidence Intervals					
Level	Count	Prob	Lower CI	Upper CI	1-Alpha
Other	32	0.32000	0.214304	0.448096	0.990
Yes	68	0.68000	0.551904	0.785696	
Total	100				

Note: Computed using score confidence intervals.

The interval presented in the *JMP* report is slightly different from that in the textbook. *JMP* calculates a *score* confidence interval, which is similar to a plus four confidence interval. The *score* confidence interval tends to have better coverage probabilities than the large sample confidence intervals, especially with smaller sample sizes (Agresti and Coull, 1998, *The American Statistician*, 52, 119–126).

## Significance test for a single proportion

### Example 8.2 Work stress (cont'd.)

A national survey of restaurant employees found that 75% said that work stress had a negative impact on their personal lives. A random sample of 100 employees of a restaurant chain finds that 68 answer “Yes” when asked, “Does work stress have a negative impact on your personal life?” Is this good reason to think that the proportion of all employees of this chain who would say “Yes” differs from the national proportion  $p = 0.75$ ?

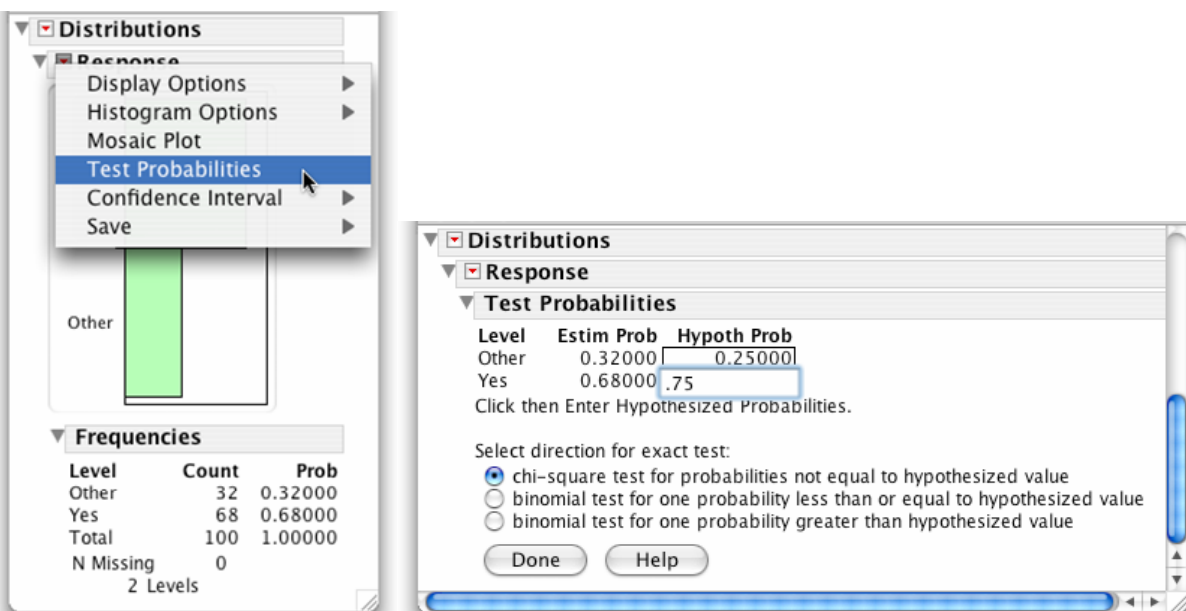
Thus, the alternative hypothesis to be tested is

$$H_a: p \neq 0.75 \text{ where } p \text{ is the proportion of all employees in the chain who would respond “Yes.”}$$

If you have not saved the data table from the previous example, follow steps 1 to 5 above to create a *JMP* data table that summarizes the responses of the 100 employees in the random sample.

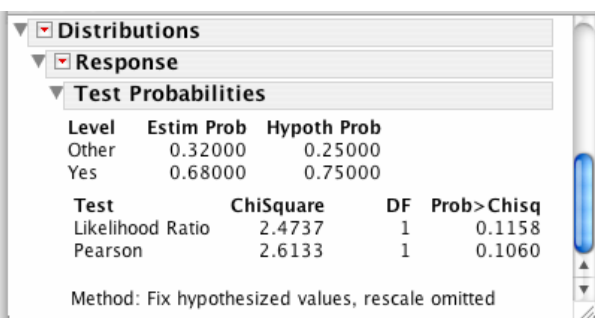
Let’s look at the distribution of responses.

1. Select **Analyze** ⇒ **Distribution**.
  - a. Select **Response** from the list of columns and press **Y, Column**.
  - b. Select **Count** and press **Freq** and **OK**.



To test the hypothesis that the proportion of all employees in the chain who would respond “Yes” is not equal to 0.75, use the **Test Probabilities** command.

2. Press the red triangle on the **Response** report and select **Test Probabilities**.
  - a. Enter .75 in the **Hypoth Prob** field next to **Yes**.
  - b. Enter .25 in the **Hypoth Prob** field next to **Other**.
  - c. Select “**chi-square test for probabilities not equal to hypothesized value.**”
  - d. Press **Done**.



Note that *JMP* does not calculate the value of the *z* statistic. For a two-sided alternative hypothesis, it calculates the related, but more general, *Pearson chi-square test statistic* that gives the same *P*-value. The *Pearson chi-square test statistic* equals the square of the *z* statistic for testing a single proportion. The value of the *Pearson chi-square test statistic* reported by *JMP* is 2.6133, which is the square of  $z = 1.62$ , the *test statistic* given in the textbook. The *P*-value (labeled **Prob>ChiSq**) for a two-sided alternative is 0.1060.

For a one-sided alternative, *JMP* performs an exact one-sided Binomial test. The *P*-value is denoted **p-Value**.

## 8.2 Comparing Two Proportions

Just as comparing two means is equivalent to testing whether a categorical variable and a quantitative variable are related, comparing the proportion of successes in two groups is equivalent to testing whether two categorical variables are related. The groups form a categorical explanatory variable, or factor. The response variable is also categorical in this case, taking the value “success” or “failure” (non-success) for each individual. To compare two proportions, *JMP* uses the same platform **Fit Y by X** that was used for comparing two population means. The test statistic together with an associated *P*-value is presented in the **Tests** report. *JMP* recognizes the different modeling type of the response variable and automatically provides the appropriate test statistic and *P*-value for comparing two proportions.

### Example 8.2 Men, women, and binge drinking

Are men and women college students equally likely to be frequent binge drinkers? A survey of over 17,000 students in U.S. four-year colleges collected information on drinking behavior. Following is the data summary.

Gender	Sample size	Number of frequent binge drinkers	Sample proportion
Men	7,180	1,630	0.227
Women	9,916	1,684	0.170

The sample proportions of frequent binge drinkers (FBDs) are very different. We will perform a significance test to see whether this difference between the two samples is large enough to conclude that the proportion in the population of all male four-year college students and the proportion in the population of all female four-year college students are not equal. Therefore, we test the hypothesis

$$H_0: p_M = p_F \text{ against the two-sided alternative } H_a: p_M \neq p_F$$

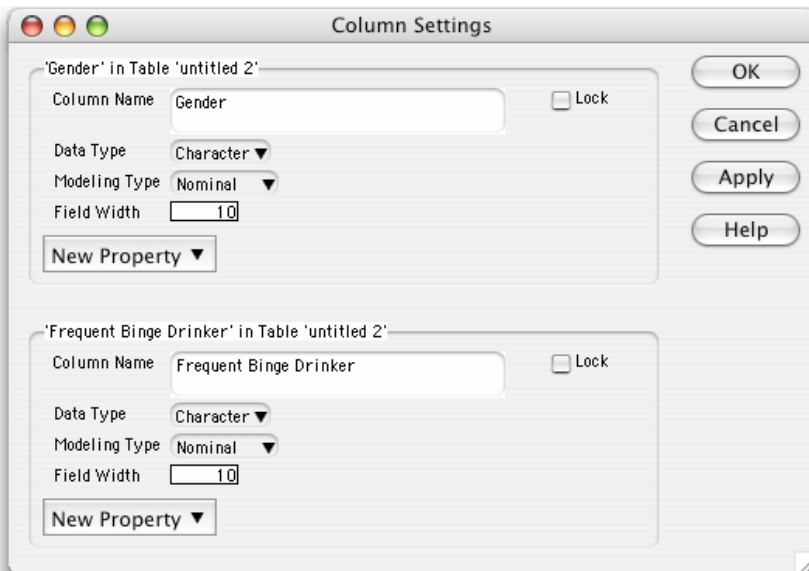
where  $p_M$  and  $p_F$  denote the population proportions of frequent binge drinkers among all male and female students in U.S. four-year colleges, respectively.

To perform the analysis in *JMP*, the data must be first placed in a *JMP* data table. Again, we do so in a certain, but simple, way. Each of the 17,096 students in the study is either male or female and either a frequent binge drinker or not. Thus, we have a variable, **Gender**, with two categories, and a variable, **Frequent Binge Drinker**, with two categories, for a total of  $2 \times 2 = 4$  classes. Each of the 17,096 students in the study falls into one of these four classes. Rather than entering 17,096 rows in the table, we use a column **Count** to summarize the number of individuals in each of the four gender-by-response categories.

1. Select **File**  $\Rightarrow$  **New** from the menu bar to create a new *JMP* data table.
2. Select **Cols**  $\Rightarrow$  **Add Multiple Columns** to accommodate the three variables.
  - a. Enter **3** after **How many columns to add** and press **OK**.
  - b. Change the names of the columns to **Gender**, **Frequent Binge Drinker**, and **Count**.
  - c. Press **OK**.

By default, columns contain numeric data. Change the data type of the first two columns to “Character.” Notice that the *modeling type* automatically changes from “Continuous” to “Nominal.”

3. Select the first two columns.
  - a. Select **Cols** ⇒ **Column Info**.
  - b. Select **Character** from the **Data Type** menu for both **Gender** and **Frequent Binge Drinker**.
  - c. Press **OK**.



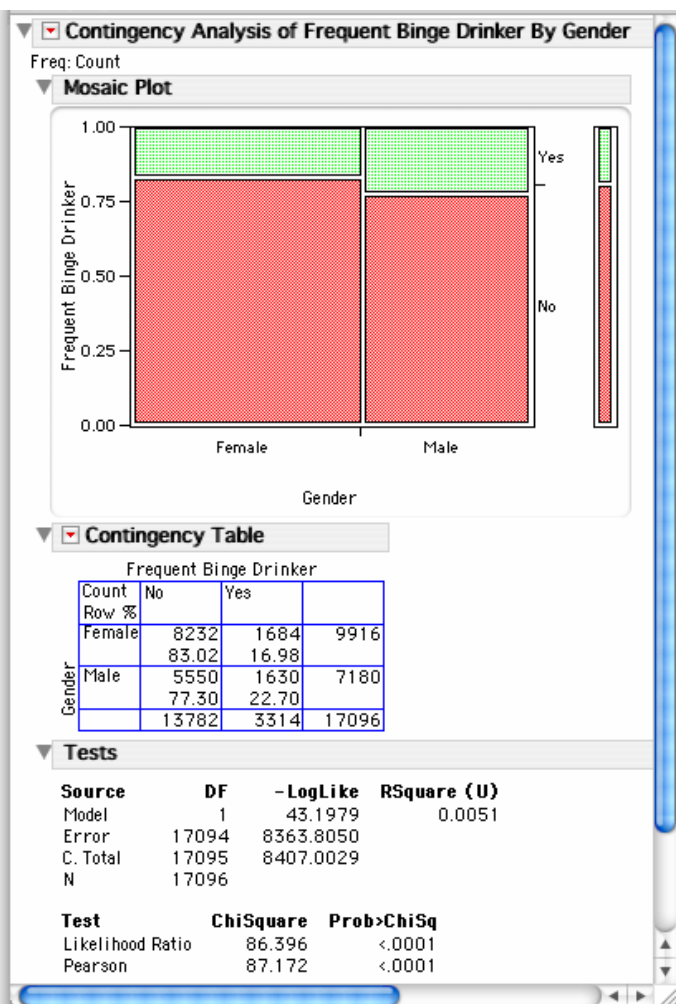
4. Select **Rows** ⇒ **Add Rows** from the menu bar and enter **4**.
  - a. Fill in the data grid as below.

		Gender	Frequent Binge Drinker	Count	
1	Male	Yes	1630		
2	Male	No	5550		
3	Female	Yes	1684		
4	Female	No	8232		

5. Select **File** ⇒ **Save** and name the data table **eg08\_002.jmp**.

Let's examine the sample proportions of each sex who are FBDs.

6. Select **Analyze** ⇒ **Fit Y by X**.
  - a. Select **Frequent Binge Drinker** from the list of columns and press **Y, Response**.
  - b. Select **Gender** and press **X, Factor**.
  - c. Select **Count** and press **Freq** and **OK**.



The **Mosaic Plot** report displays three vertical stacked bar charts for the variable **Frequent Binge Drinker**. (Ignore the width of the bars.) The bar chart on the far right does not consider gender. The two adjacent ones on the left are for Females and Males, respectively. The green bar (Yes) for Males is longer than that for Females. To see the sample proportions more clearly, proceed as follows.

7. Press the red triangle in the **Contingency Table** report and deselect both **Total %** and **Col %**.

16.98% of females are frequent binge drinkers versus 22.70% of males.

To test the hypothesis that the proportion of frequent binge drinkers in the population of all male students is not the same as the proportion in the population of all female students, examine the middle of the **Tests** report. *JMP* does not provide the  $z$  statistic. Instead, it calculates the more general *Pearson chi-square test statistic*; this is discussed in Chapter 9 of the textbook. The *Pearson chi-square test* equals the square of the  $z$  statistic for the case of two proportions and can also be used to compare more than two population proportions. The value of the *Pearson chi-square statistic* reported by *JMP* is 87.172, which is the square of the  $z$  test statistic,  $z = 9.34$ , given in the textbook. The  $P$ -value (labeled **Prob>ChiSq**) for a two-sided alternative is less than 0.0001. The difference is clearly statistically significant.

For a one-sided alternative, the  $P$ -value is either half of the **Prob>ChiSq** value or one minus half of it, depending on the direction of the one-sided alternative.

## 8.3 Summary

All graphs and statistical computations in this chapter are performed in the second platform **Fit Y by X** of the **Analyze** menu.

<b>Activity</b>	<b>Command</b>
A single proportion	
Confidence interval	<u>Analyze</u> ⇒ <u>Distribution</u> ⇒ <u>Confidence Interval</u>
Significance test	<u>Analyze</u> ⇒ <u>Distribution</u> ⇒ <u>Test Probabilities</u>
Comparing two proportions	<u>Analyze</u> ⇒ <u>Fit Y by X</u>

# Chapter 9

## Inference for Two-Way Tables

In Chapter 8, we compared two population proportions. We now study how to compare two or more populations when the response variable has two or more possible values and how to test whether two categorical variables are related. The same statistical test handles both cases. We also learn about a test to evaluate the goodness of fit of a proposed distribution for a categorical variable that also uses the same sampling distribution.

Since analysis of two-way tables is equivalent to examining the relationship between two variables, all graphs and statistical computations in this chapter are performed in the second platform **Fit Y by X** of the **Analyze** menu.

### 9.1 Inference for Two-Way Tables

The same approach and the same test statistic, *Pearson chi-square statistic*, which we used in the previous chapter are used again in this one.

#### Example 9.1 Exclusive territories and the success of new franchise chains

How does exclusive-territory rights relate to the survival of a franchise? A study designed to address this question collected data from a sample of 170 new franchise firms. Two variables were measured for each firm. First, the firm was classified as successful or not based on whether or not it was still franchising as of a certain date. Second, the contract each firm offered to franchises was classified according to whether or not there was an exclusive-territory clause. Here are the data:

Success	Exclusive territory		Total
	Yes	No	
Yes	108	15	123
No	34	13	47
Total	142	28	170



We wish to test:

$H_0$ : There is no association between success and exclusive territory rights versus

$H_a$ : Exclusive territory rights and success are related.

First, we create an appropriate *JMP* data table to summarize the two-way table. Then, we use the **Fit Y by X** platform to look at the relationship between exclusive territory rights and success for franchise firms.

The entries in this two-way table are the numbers of franchises in each success-by-exclusive-territory group. Each of the two columns in the table represents a value of the variable “Exclusive Territory” and each of the two rows represents a value of the variable “Success.” Thus, to enter the data into a *JMP* data table, we need one variable, **Success**, with two categories and one variable, **Exclusive Territory**, with two categories also for a total of  $2 \times 2 = 4$  success-by-exclusive-territory groups. Instead of entering 170 rows, one for each franchise, we use four rows, one for each group, and include a column **Count** of the numbers of franchises in each group.

1. Select **File**  $\Rightarrow$  **New** and create three columns named **Success**, **Exclusive Territory**, and **Count**.
2. Enter the data as follows and save the *JMP* data table.

	Success	Exclusive Territory	Count
1	Yes	Yes	108
2	Yes	No	15
3	No	Yes	34
4	No	No	13

First, let's examine a mosaic plot and the two-way table for this data.

3. Select **Analyze**  $\Rightarrow$  **Fit Y by X**.
  - a. Select **Exclusive Territory** and press **Y, Response**.
  - b. Select **Success** and press **X, Factor**.
  - c. Select **Count** and press **Freq** and **OK**.

		Exclusive Territory		
		No	Yes	
Success	Count	13	34	47
	Total %	7.65	20.00	27.65
	Col %	46.43	23.94	
	Row %	27.66	72.34	
Yes	Count	15	108	123
	Total %	8.82	63.53	72.35
	Col %	53.57	76.06	
	Row %	12.20	87.80	
		28	142	170
		16.47	83.53	

Can you identify the joint distribution, the marginal distributions, and the conditional distributions associated with the two variables? The resulting report contains both counts and percents. The joint proportion of franchises that were a success and that have exclusive territory rights is 108 divided by 170, or 63.53%. The *marginal distribution* of **Success** is found in the right margin of the table. For example, 72.35% of all franchises in the study were a success. The *marginal distribution* of **Exclusive Territory** is in the bottom margin of the table.

To compare firms that have an exclusive territory with those that do not, we look at the *conditional distribution* of **Success** for each column, or value of **Exclusive Territory**. These are included in the **Contingency Table** report. Remove some of the clutter to see them better.

4. Deselect **Count**, **Total %**, and **Row %** in the **Contingency Table** (use the red triangle menu).

**Contingency Analysis of Exclusive Territory By Success**

Freq: Count

**Contingency Table**

		Exclusive Territory		
		No	Yes	
Success	No	13 46.43	34 23.94	47
	Yes	15 53.57	108 76.06	123
		28	142	170

**Tests**

Source	DF	-LogLike	RSquare (U)
Model	1	2.732478	0.0359
Error	168	73.324088	
C. Total	169	76.056566	
N	170		

Test	ChiSquare	Prob>ChiSq
Likelihood Ratio	5.465	0.0194
Pearson	5.911	0.0150

Seventy-six percent of franchises with exclusive territories are successful as opposed to 54% of firms that were not offered exclusivity. The difference is quite large and success and territorial exclusivity are clearly related in the sample data.

To determine whether we can extend our conclusions to the population of all recently established franchises, we perform a test of significance on the hypotheses stated earlier and look at the *Pearson chi-square test* in the **Tests** report. The value of the *Pearson chi-square statistic* is 5.911 and the *P-value* (labeled **Prob>ChiSq**) is 0.0150. Since the *P-value* is small, the data do indeed provide strong evidence that exclusive-territory rights is related to success.

## Remarks

- *JMP* can display the expected counts for all cells in the table. Select **Expected** from the red triangle menu on the **Contingency Table** report. The expected counts are labeled “Expected” and are the last entry in each cell. For example, the expected number of franchises with exclusive territories that are a success is 102.74.

Contingency Analysis of Exclusive Territory By Success

Freq: Count

Contingency Table

		Exclusive Territory		
		No	Yes	
Success	Count	13	34	47
	Expected	7.74118	39.2588	
	No			
	Yes	15	108	123
	Expected	20.2588	102.741	
	Total	28	142	170

- Example 9.1 is about *testing to see if two categorical variables are related*. To test for differences among three or more population proportions, the same test statistic (Pearson's chi-squared) is computed.

## 9.2 Goodness of Fit

The *chi-square statistic* can also be used to test a different kind of hypothesis: that a categorical variable has a specified distribution. We use the **Test Probabilities** command in the **Distribution** platform to construct this test.

### Example 9.2 Cell phones and motor vehicle accidents

A study of 699 drivers who were using a cell phone when they were involved in a collision showed the following distribution across days of the week.

Day	Sun.	Mon.	Tues.	Wed.	Thurs.	Fri.	Sat.	Total
No. of accidents	20	133	126	159	136	113	12	699

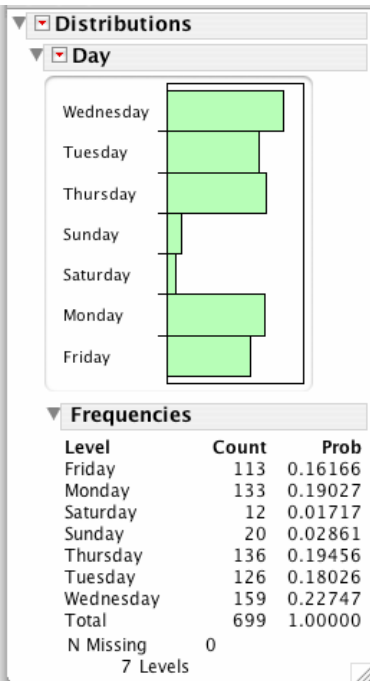
Do these data give significant evidence that the accidents are not equally likely to occur on any day of the week?

1. Create a new *JMP* data table with two variables and seven rows.
2. Name the variables **Day** and **Accidents** and set the data types of **Day** to “character” and **Accidents** as “numeric.”
3. Fill in the data values as below and save the table.

	Day	Accidents
1	Sunday	20
2	Monday	133
3	Tuesday	126
4	Wednesday	159
5	Thursday	136
6	Friday	113
7	Saturday	12

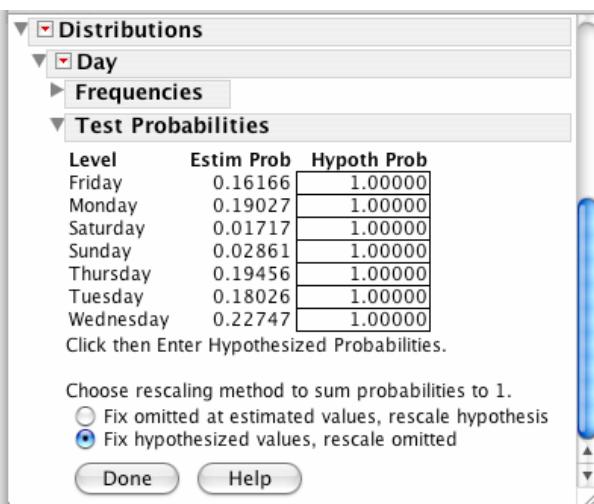
Let's look at the distribution of the categorical variable **Day**, i.e., the distribution of the accidents across days.

4. Select **Analyze** ⇒ **Distribution** as in Chapter 1.
  - a. Select **Day** and press **Y, Columns**.
  - b. Select **Accidents** and press **Freq** and **OK**.



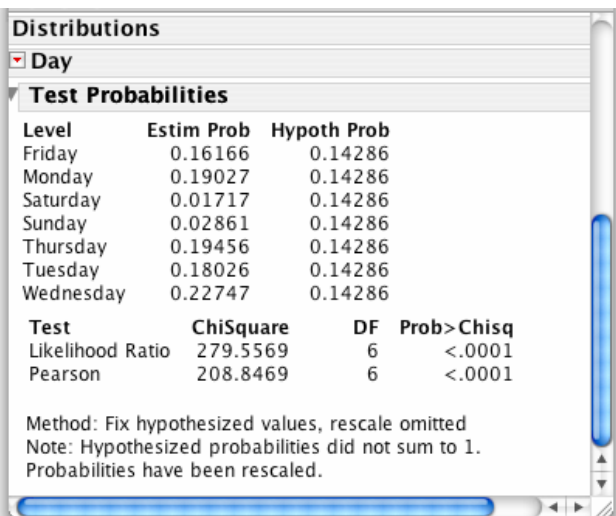
The bars are not all the same size so there is some evidence that the accidents are not equally distributed throughout the week. To test the alternate hypothesis that accidents are not equally likely to occur on all seven days:

5. Press the red triangle on the **Day** report and select **Test Probabilities**.



Because *JMP* scales the hypothesized values that you enter so that the probabilities sum to 1, the easiest way to test that all the probabilities are equal is to enter a 1 in each field.

6. a. Enter a 1 in each of the **Hypoth Prob** fields and press **Done**.



The screenshot shows the 'Distributions' window in JMP. The variable 'Day' is selected. The 'Test Probabilities' table is displayed, showing estimated probabilities and hypothesized probabilities for each day of the week. Below the table, the results of the Chi-Square test are shown, including the Likelihood Ratio and Pearson tests, both with a p-value less than 0.0001. A note at the bottom states: 'Method: Fix hypothesized values, rescale omitted. Note: Hypothesized probabilities did not sum to 1. Probabilities have been rescaled.'

Level	Estim Prob	Hypoth Prob
Friday	0.16166	0.14286
Monday	0.19027	0.14286
Saturday	0.01717	0.14286
Sunday	0.02861	0.14286
Thursday	0.19456	0.14286
Tuesday	0.18026	0.14286
Wednesday	0.22747	0.14286

Test	ChiSquare	DF	Prob>Chisq
Likelihood Ratio	279.5569	6	<.0001
Pearson	208.8469	6	<.0001

Method: Fix hypothesized values, rescale omitted  
 Note: Hypothesized probabilities did not sum to 1.  
 Probabilities have been rescaled.

These data provide very convincing evidence that accidents are not equally likely on all days of the week ( $\chi^2 = 208.85$ ,  $df = 6$ ,  $P\text{-value} < 0.2689$ ). This distribution will be investigated further in one of the exercises.

## 9.3 Summary

All graphs and statistical computations in this chapter are performed in the second platform **Fit Y by X** of the **Analyze** menu. The modeling type of both variables must be “Nominal.” To change the order of the categories of a variable, use the **Column Info** command to select the column property **Value Ordering**.

### Activity

### Command

Inference for two-way tables

**Analyze** ⇒ **Fit Y by X** ⇒ ... ⇒ **Freq**

Goodness of fit tests

**Analyze** ⇒ **Distribution** ⇒ **Test Probabilities**

# Chapter 10

## Inference for Regression

In Section 2.3 of Chapter 2, we modeled the relationship between two quantitative variables as a straight line using least-squares regression. These models allow us to predict a response based on the value of an explanatory variable. Now, we construct confidence interval estimates and perform significance tests for the unknown population regression parameters (the slope and the intercept) and the predicted values.

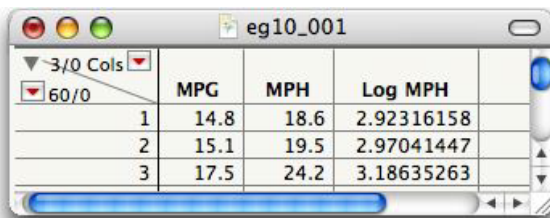
### 10.1 Simple Linear Regression

The first steps in the analysis of the relationship between two *quantitative* variables are to fit the model and to check the assumptions of the model using the residuals. We use the same *JMP* platform—**Fit Y by X**—that we used in Section 2.3 in Chapter 2.

#### 10.1.1 Fitting the model and examining the residuals

##### Example 10.1 Does the speed of a vehicle affect fuel efficiency?

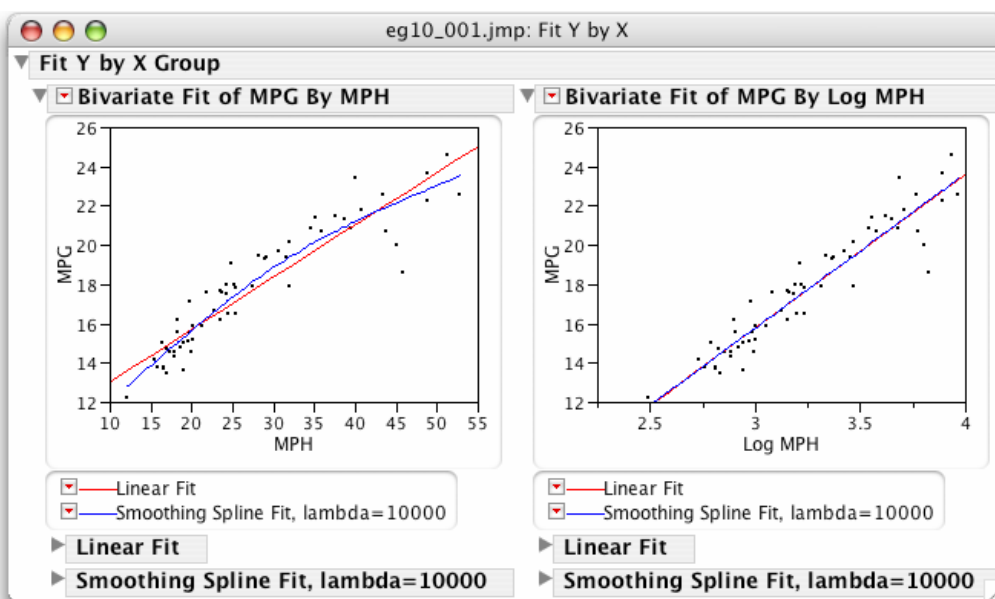
How does the speed at which a particular vehicle is driven affect the fuel efficiency? We have a simple random sample of 60 observations and two measurements—fuel efficiency expressed as miles per gallon and average speed in miles per hour. Suppose that the data are stored in the JMP data table **eg10\_001.jmp**.



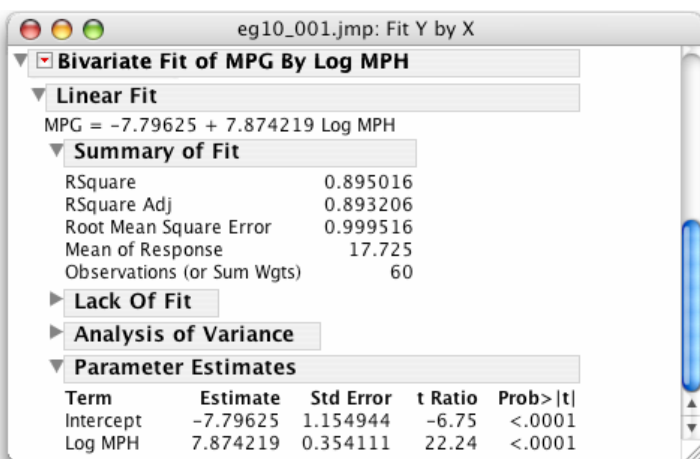
	MPG	MPH	Log MPH
1	14.8	18.6	2.92316158
2	15.1	19.5	2.97041447
3	17.5	24.2	3.18635263

The variable **Log MPH** was created from **MPH** using the formula editor. (See Section 0.3.3 for details.) Open the *JMP* data table and proceed as you did in Chapter 2 to fit the least-squares regression line for predicting the fuel efficiency of the vehicle in terms of average speed.

1. Select **Analyze** ⇒ **Fit Y by X**.
  - a. Select **MPG** from the list of columns and press **Y, Response**.
  - b. Select **MPH** and **Log MPH**, and press **X, Factor** and **OK**.
  - c. Select **Fit Line** from the red triangle menu on each of the **Bivariate Fit** reports.
  - d. Select **Fit Spline** ⇒ **10000** from the red triangle menu on each of the **Bivariate Fit** reports.



It's clear that the relationship between **MPG** and **MPH** is not linear. On the other hand, the relationship of **MPG** and **Log MPH** is approximately linear. Therefore, we study the relationship of fuel economy to speed in terms of **MPG** and **Log MPH**. Inspect the **Linear Fit** report for **MPG** versus **Log MPH**.



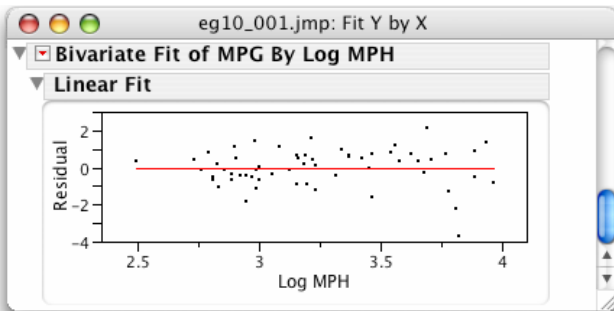
At the top, we find that the least-squares line is  $MPG^{\wedge} = -7.80 + 7.87 \text{ Log MPH}$ , after rounding. In the **Summary of Fit** section, we find that  $r^2$  (labeled **RSquare**) equals 0.8950 and that the *standard error about the line* (labeled **Root Mean Square Error**) equals  $s = 0.9995$ .

Before doing inference, we must check the required regression model assumptions. Checking assumptions involves examining the *residuals*. The *JMP* commands **Plot Residuals** and **Save Residuals** help us to do that.

### Plot the Residuals Versus the Explanatory Variable

To plot the residuals of the 60 observations versus the explanatory variable **Log MPH**, use the **Plot Residuals** command as we did in Section 2.4 in Chapter 2.

2. Select **Plot Residuals** from the red triangle menu next to **Linear Fit**.



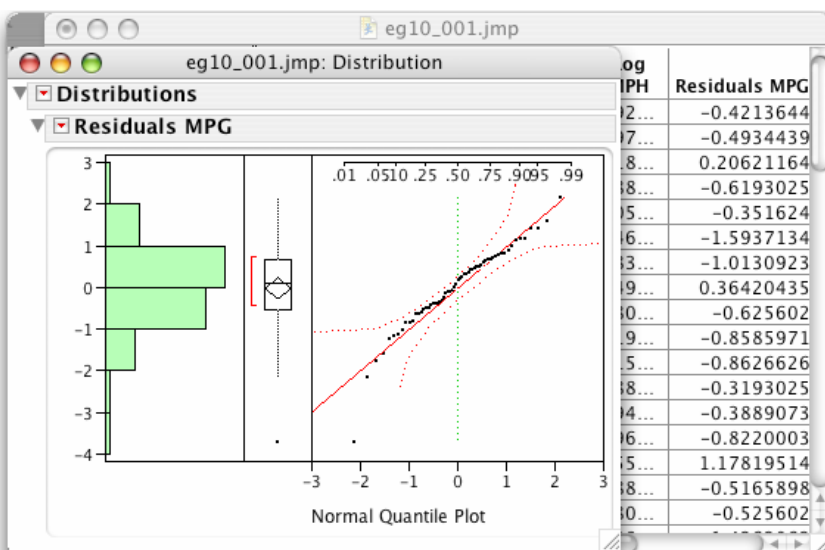
The plot appears at the bottom of the **Linear Fit** report. The residuals are on the vertical axis with a red line at zero. No clear pattern is evident. There is one residual that is somewhat low though.

### Assessing Normality: Normal Quantile Plots

To check for signs of non-Normality, we save the residuals to the *JMP* data table and use the **Normal Quantile Plot** command in the **Distribution** platform (Section 1.3 in Chapter 1).

3. Select **Save Residuals** from the red triangle menu for the **Linear Fit** report.
4. Select **Analyze** ⇒ **Distribution**.
  - a. Select **Residuals MPG** and press **Y, Columns** and **OK**.
  - b. Press the red triangle on the **Residuals MPG** report and select **Normal Quantile Plot**.



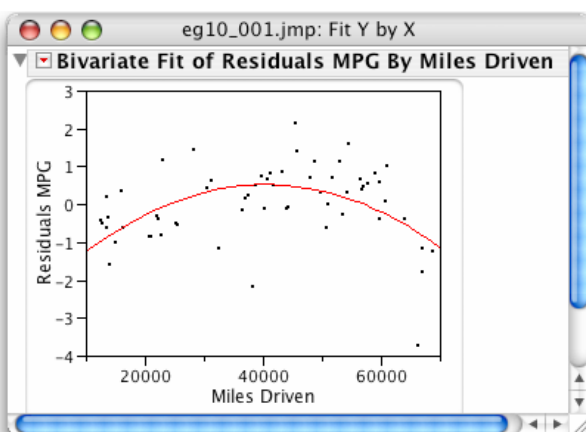


There are no signs of strong non-Normality.

### Plot the Residuals Against Other Variables

We should also plot the residuals against any other known variables, e.g., the individual (or case number) or the time or order in which the observations were taken. In this case, we might wonder whether the total number of miles that the vehicle has been driven also affects fuel economy. If so, we would add a column containing the miles driven, **Miles Driven**, to the data table and plot the residuals against it. Since the residuals have been saved to the *JMP* data table, we can use all of the commands in the **Analyze** menu on the residuals.

5. Select **Analyze** ⇒ **Fit Y by X**.
  - a. Select **Residuals MPG** and press **Y, Columns**.
  - b. Select **Miles Driven** and press **X, Factor** and **OK**.



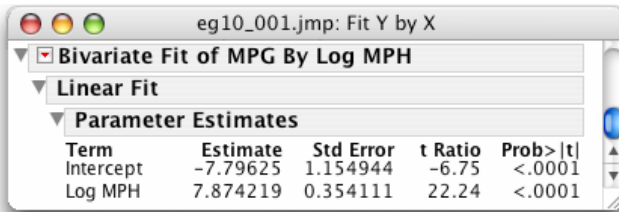
The plot suggests that the residuals increase slightly up to about 50,000 miles and then tend to decrease somewhat. Hence, we may wish to consider a model that includes the age of the vehicle as well as the average speed traveled.

### 10.1.2 Inference for the regression slope and intercept

Having determined that the assumptions of the model hold, we consider extending our conclusions about the relation of fuel economy to average speed beyond the 60 observations in the sample.

#### Example 10.1 Does the speed of a vehicle affect fuel efficiency? (cont'd.)

Use the **Window** menu to bring the **eg10\_001.jmp: Fit Y by X** window with the **Bivariate Fit of MPG By Log MPH** report forward.



Term	Estimate	Std Error	t Ratio	Prob> t
Intercept	-7.79625	1.154944	-6.75	<.0001
Log MPH	7.874219	0.354111	22.24	<.0001

In the **Parameter Estimates** section:

- The **Estimate** column contains the least-squares estimates for the unknown regression parameters. The estimate of the *slope*  $\beta_1$  (7.874219) appears in the “Log MPH” row.
- The **Std Error** column contains the standard errors. In particular,  $SE_{b1} = 0.354111$ .
- To test the hypothesis  $H_0$ : the true *slope* is zero ( $\beta_1 = 0$ ) against  $H_a$ :  $\beta_1 \neq 0$ , *JMP* calculates the *t* statistic (labeled **t Ratio**) and the corresponding *P*-value (labeled **Prob>|t|**) for a two-sided alternative. Hence, to test the hypothesis

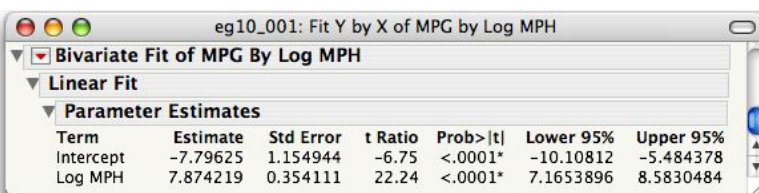
$H_0$ : **Log MPH** is of no value in predicting **MPG** ( $\beta_1 = 0$ ) versus

$H_a$ : **MPG** is linearly related to **Log MPH** ( $\beta_1 \neq 0$ )

the *t* statistic = 22.24 and the *P*-value < 0.0001. Hence, there is very strong evidence that **MPG** is related to **Log MPH**.

To obtain a confidence interval estimate for the true slope  $\beta_1$  (or the intercept  $\beta_0$ ),

1. Context-click (right-mouse click in Windows and control-click on an Apple computer) in the **Parameter Estimates** table.
  - a. Select **Columns**  $\Rightarrow$  **Lower 95%** from the menu that opens.
  - b. Repeat for **Upper 95%**.



Term	Estimate	Std Error	t Ratio	Prob> t	Lower 95%	Upper 95%
Intercept	-7.79625	1.154944	-6.75	<.0001*	-10.10812	-5.484378
Log MPH	7.874219	0.354111	22.24	<.0001*	7.1653896	8.5830484

### 10.1.3 Inference about prediction

*JMP* displays a graph of the *confidence intervals for mean responses* in the form of two curves around the least-squares regression line. To read the confidence interval for the mean of the response  $y$  when  $x$  has the value  $x^*$  from the graph, use the **Crosshair tool**. *Prediction intervals for individual responses* are handled similarly.

#### Example 10.1 Predicting fuel economy from average speed (cont'd.)

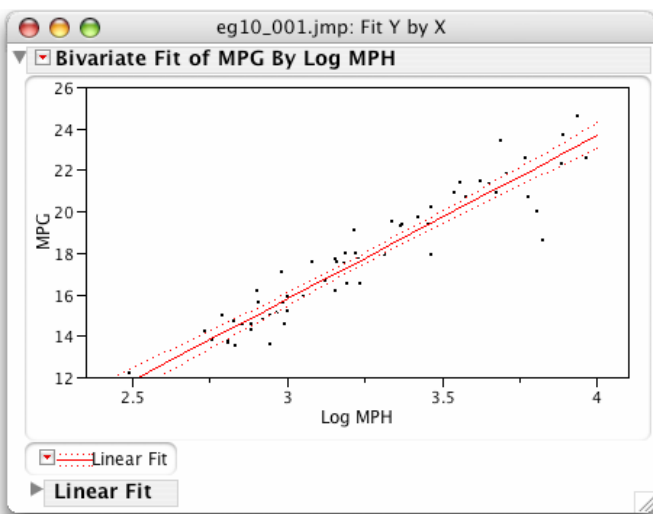
Let's predict the average (mean) fuel economy when traveling at 30 mph.


Use the **Window** menu to bring the **Bivariate Fit of MPG By Log MPH** report window forward.

Since there is good evidence that average speed is useful in predicting fuel economy ( $t = 22.24$ ,  $P$ -value  $< .0001$ ), we use the least-squares line  $MPG^{\wedge} = -7.7925 + 7.8742 \text{ Log MPH}$  to predict mean fuel economy. Since  $\log(30) = 3.4$ , the predicted value of mean fuel economy when averaging 30 mph will be  $-7.7925 + 7.8742(3.4) = 19.0$ . To obtain the 95% *confidence interval for the mean MPG* when **Log MPH** = 3.4:

1. Press the red triangle next to **Linear Fit** and select **Confid Curves Fit** from the menu that opens.

The curves above and below the least-squares line give the 95% *confidence intervals for mean responses*.

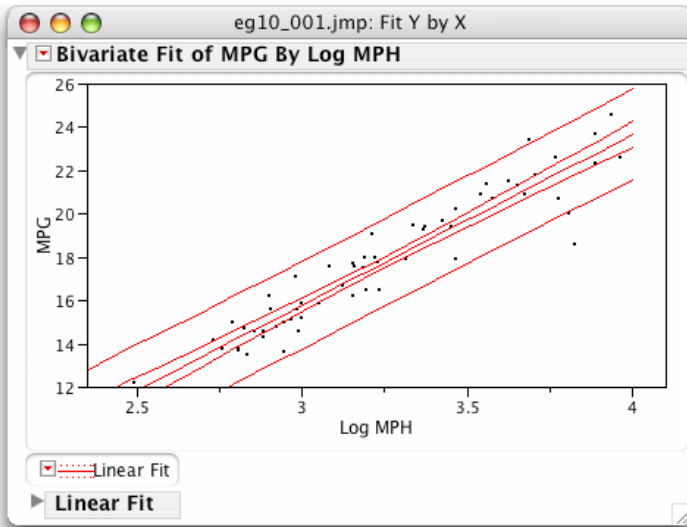


2. Select **Tools**  $\Rightarrow$   (the Crosshair tool).
3. Place the crosshair on the lower curve directly above **Log MPH** = 3.4 (i.e.,  $MPH = 30$ ) to obtain the value of the lower prediction limit. Press and hold the mouse button. The values of the  $x$  and  $y$  axes where the crosshair intersects the plot appear automatically; in this case, 3.4 and 18.69, respectively. You may need to drag the crosshair a bit.
4. Obtain the upper confidence limit by placing the crosshair on the upper curve directly above  $\text{Log MPH} = 3.4$ . A value of 19.31 for **MPG** is displayed.

Hence, the 95% confidence interval for the mean **MPG** is approximately (18.7, 19.3).

The 95% *prediction interval* for a single observation of fuel economy when the vehicle is driven at 30 **MPH** is obtained in a similar way.

5. Select **Confid Curve Indiv** from the red triangle menu next to **Linear Fit**. The curves farther from the least-squares line give the 95% *prediction intervals for individual responses*.

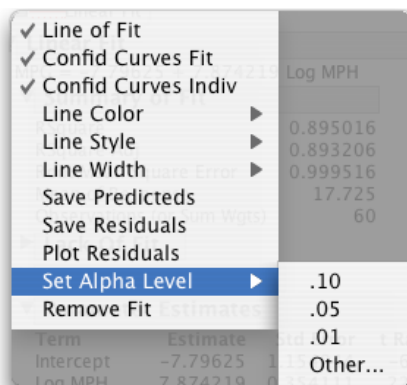


6. Use the crosshair tool to estimate the upper and lower prediction limits when **Log MPH** is 3.4.

The 95% *prediction interval* for the fuel economy **MPG** of an individual observation when driven at 30 **MPH**, (16.96, 21.03), is much wider than the *confidence interval for mean*.

## Remarks

- To obtain prediction intervals and confidence intervals for levels of confidence other than 95%, use the **Set Alpha Level** command on the **Linear Fit** red triangle menu, and choose the value  $1 - C$ , where  $C$  is the level desired.



- Predicted values for the sample data can be saved to the *JMP* data table. The **Save Predicteds** command on the **Linear Fit** red triangle menu creates a new column that contains the sample predicted values. To obtain predicted values for other  $x$  values, simply add rows, containing only the  $x$ -values, to the table.

## 10.2 More Detail of Simple Linear Regression

### 10.2.1 Analysis of variance for regression

*JMP* provides an *analysis of variance table* and an  $F$  statistic as part of the **Linear Fit** report.

#### Example 10.12 ANOVA for fuel economy data

Use the **Window** menu to bring the **eg10\_001.jmp: Fit Y by X** window forward.

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	493.98859	493.989	494.4668
Error	58	57.94391	0.999	Prob > F
C. Total	59	551.93250		<.0001

Note that *JMP* uses the label “C. Total” for the (corrected) total variation in the response variable.

### 10.2.2 Inference for correlation

To obtain significance tests involving the population correlation  $\rho$ , we use the same command that we used to calculate the sample correlation  $r$  in Section 2.2 of Chapter 2, the **Multivariate** command for the **Multivariate Methods** platform

#### Example 10.1 Fuel economy and speed (cont’d.)

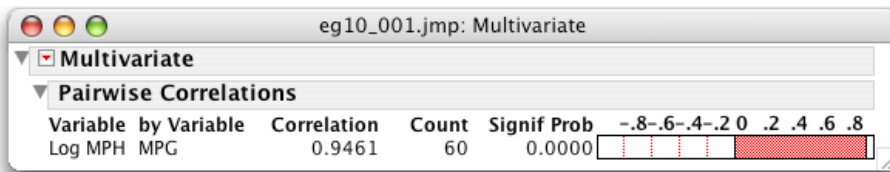
We wish to assess the evidence that, for this particular vehicle, there is a linear relationship between **MPG** and **Log MPH**. Hence, we wish to test the hypothesis:

$$H_0: \rho = 0 \text{ against the alternative hypothesis } H_a: \rho \neq 0$$

Use the **Window** menu to bring the **eg10\_001.jmp: Fit Y by X** window forward.

1. Select **Analyze**  $\Rightarrow$  **Multivariate Methods**  $\Rightarrow$  **Multivariate**.

- Select the columns **MPG** and **Log MPH** and press **Y, Columns**.
- Press **OK**.
- Press the red triangle on the **Multivariate** title bar and select **Pairwise Correlations**.



Notice that the sample correlation between weekly wages and length of service is  $r = 0.9461$ . The number 0.0000 directly below **Signif. Prob** is the  $P$ -value for testing the two-sided alternative hypothesis that the population correlation is not 0. Since the  $P$ -value is so small (zero to four decimal places), the data provide very strong evidence that there is a linear correlation between **MPG** and **Log MPH** for this vehicle.

## 10.3 Summary

All graphs and statistical computations to perform inference for regression use the **Fit Y by X** platform. The residuals can be examined using the **Distribution** platform.

<u>Activity</u>	<u>Command</u>
Inference about the model	<u>Fit Y by X</u> ⇒ <u>Fit Line</u>
Inference about prediction	<u>Fit Y by X</u> ⇒ <u>Fit Line</u> ⇒ ...
To estimate an individual response	<u>Confid Curves Indiv</u>
To estimate the mean response	<u>Confid Curves Fit</u>
Checking assumptions	<u>Fit Y by X</u> ⇒ <u>Fit Line</u> ⇒ <u>Plot Residuals</u>
	<u>Fit Y by X</u> ⇒ <u>Fit Line</u> ⇒ <u>Save Residuals</u>
Inference for correlation	<u>Analyze</u> ⇒ <u>Multivariate Methods</u> ⇒ <u>Multivariate</u> ⇒ ...
	<u>Pairwise</u>

# Chapter 11

## Multiple Regression

The previous chapter discussed inference for a linear relationship between two quantitative variables—a response variable and a *single explanatory* variable. This chapter considers the case when there are *multiple explanatory* variables.

The multiple linear regression model and methods for estimating the unknown population regression parameters are discussed in your textbook. Methods of inference for the regression coefficients and the analysis of variance  $F$  test are also presented. We will use a new analysis platform **Fit Model** in *JMP* to fit and analyze the multiple regression model.

### 11.1 Inference for Multiple Regression

We illustrate fitting and testing a multiple regression model in the context of model building using *JMP*. We begin with some preliminary analyses—examination of the distributions of each variable separately followed by an examination of the relationships between pairs of variables. After this, we fit and test a multiple regression model. Additional models can then be fitted, tested, and compared.

---

#### Example 11.1 Predicting academic success in college

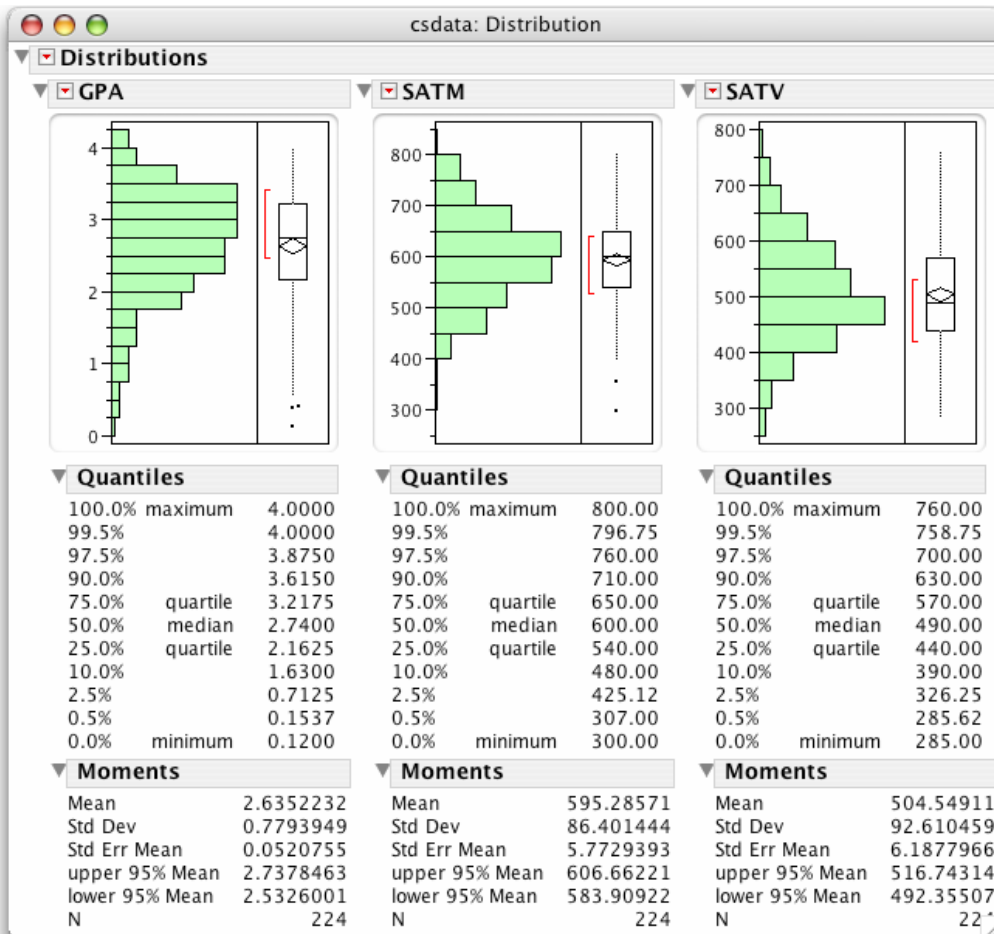
A large university wished to predict success in the early university years for its computer science majors. The response variable is the cumulative GPA after three semesters. The explanatory variables are average high school grades in mathematics (**HSM**), science (**HSS**), and English (**HSE**), as well as the SAT mathematics (**SATM**) and verbal (**SATV**) scores. Suppose that the data for 224 students is contained in a *JMP* data table **csdata.jmp**.

1. Open the data table **csdata.jmp**.

#### Examine the distribution of each variable separately

We use the **Distribution** platform in *JMP* to examine the distributions of the variables of interest.

2. a. Select **Analyze** ⇒ **Distribution**.
- b. Select **GPA**, **SATM**, **SATV**, **HSM**, **HSS**, and **HSE**; press **Y, Columns** and **OK**.

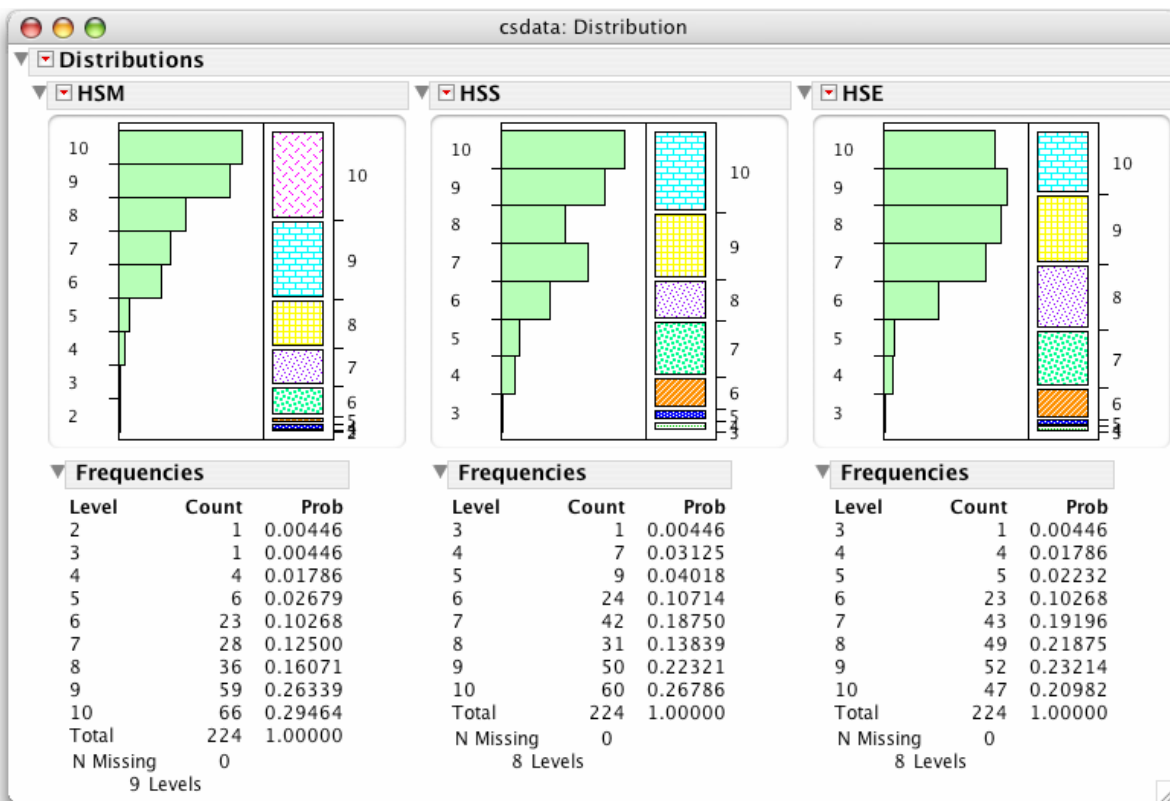


The response variable **GPA** is skewed to smaller numbers with a few students who have probably given up. The five-number summary is 0.12, 2.16, 2.74, 3.21, 4.00. A lack of Normality of the variable **GPA** is not of concern. Only the residuals of a fitted model must be Normally distributed. Examine the distributions of the **SATM** and **SATV** scores to learn something about them alone before attempting to use them in a complex model.

The high school grade variables **HSM**, **HSS**, and **HSE** are discrete variables with ordered values and are best summarized by giving the relative frequencies for each possible value. If we temporarily change their modeling type to “Ordinal,” *JMP* will display a mosaic plot and frequency table for each.

3. Select the columns **HSM**, **HSS**, and **HSE** of **csdata.jmp** in the *JMP* data table window.
  - a. Select **Cols** ⇒ **Col Info**.
  - b. Select **Ordinal** in each of the **Modeling Type** fields and press **OK**.
4. a. Select **Analyze** ⇒ **Distribution**.
- b. Select **HSM**, **HSS**, and **HSE**; press **Y, Columns** and **OK**.



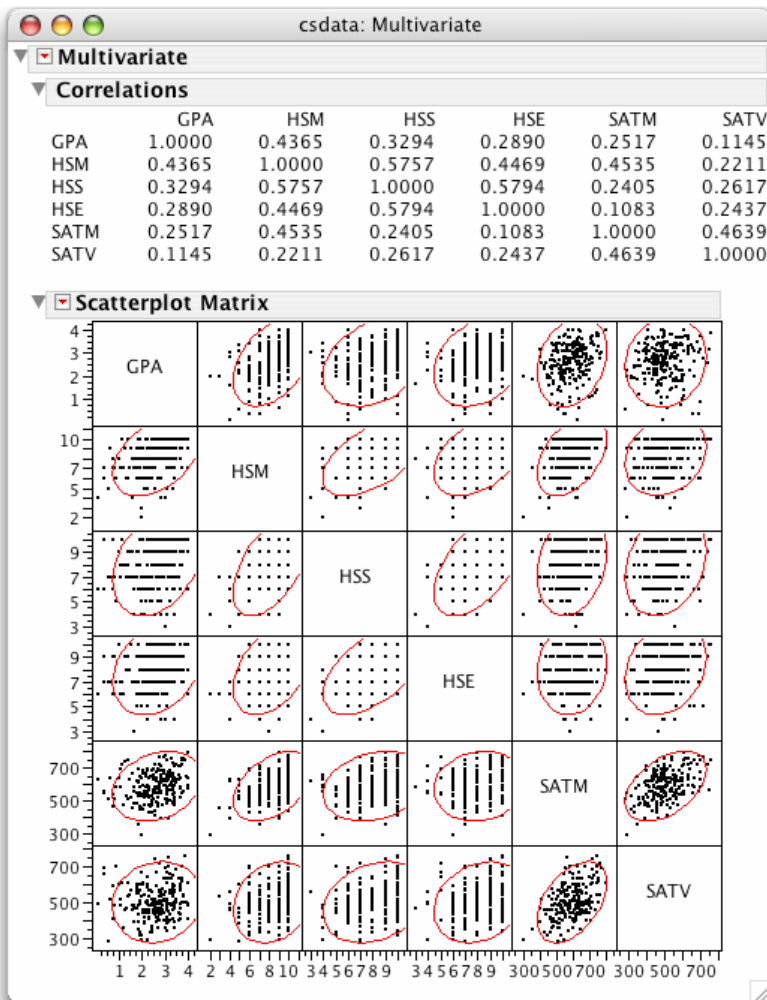


As might be expected, students admitted to the university did well in high school and the high school grade variables are skewed toward the lower grades.

## Examine relationships between pairs of variables

Recall from Chapter 2 that the **Multivariate** command in the **Multivariate Methods** platform provides a succinct summary of the pairwise relations among variables.

4. a. Select **Analyze** ⇒ **Multivariate Methods** ⇒ **Multivariate**.
- b. Select **GPA**, **SATM**, **SATV**, **HSM**, **HSS**, and **HSE**; press **Y, Columns** and **OK**.

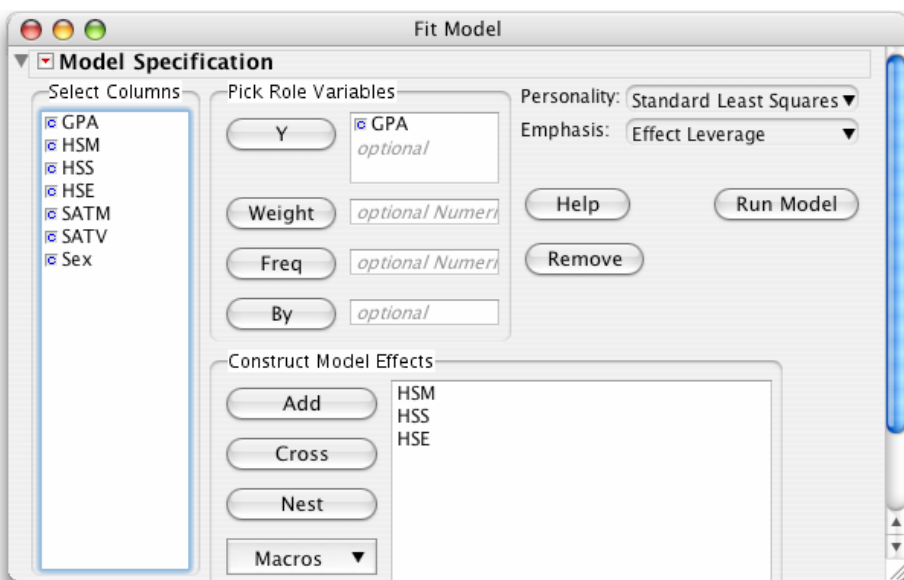


Examine the pairwise correlations and scatterplots. The high school grades are more correlated with GPA than with the SAT scores. The high school grades are correlated with one another, and so are the two SAT scores. The SAT mathematics score correlates well with HSM. Other pairs have correlations less than .27.

## Fitting a Multiple Regression Model and Examining Residuals: Regression on High School Grades

We illustrate fitting a multiple regression model using only the three high school grades to predict GPA. We use a new *JMP* analysis platform (**Fit Model**).

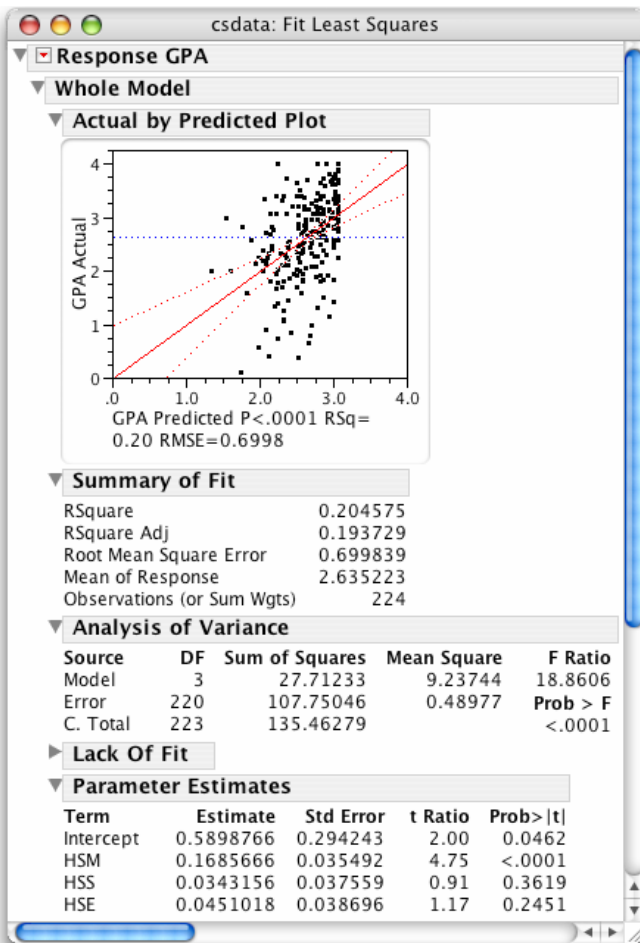
1. Select **Analyze** ⇒ **Fit Model**.
  - a. Select **GPA** from the list of columns and press **Y**.
  - b. Select **HSM**, **HSS**, and **HSE**; press **Add** under **Construct Model Effects**.
  - c. Press **Run Model**.



Examine the **Whole Model** report in the window that opens. The degrees of freedom in the **Analysis of Variance (ANOVA) table** show, as expected, 220 degrees of freedom for Error. The ANOVA *F statistic* (**F Ratio** in the report) is 18.8606 with a *P*-value (**Prob > F**) less than 0.0001. Hence, there is strong evidence to support the alternative hypothesis

$H_a$ : at least one of the  $\beta_j$  is not equal to zero.

We can conclude that at least one of the three population regression coefficients for the high school grades is different from zero, and thus at least one of the high school grades should be used as a predictor for **GPA**.



Investigate the **Summary of Fit** report. The value of  $s$ , the *estimate of the parameter*  $\sigma$ , is 0.6998 and called **Root Mean Square Error**. The squared multiple correlation  $R^2$  (labeled **RSquare**) is 0.204575. Although significant, the model only explains about 20% of the variability in the **GPA** scores.

From the **Parameter Estimates** table, we obtain the *fitted regression equation*:

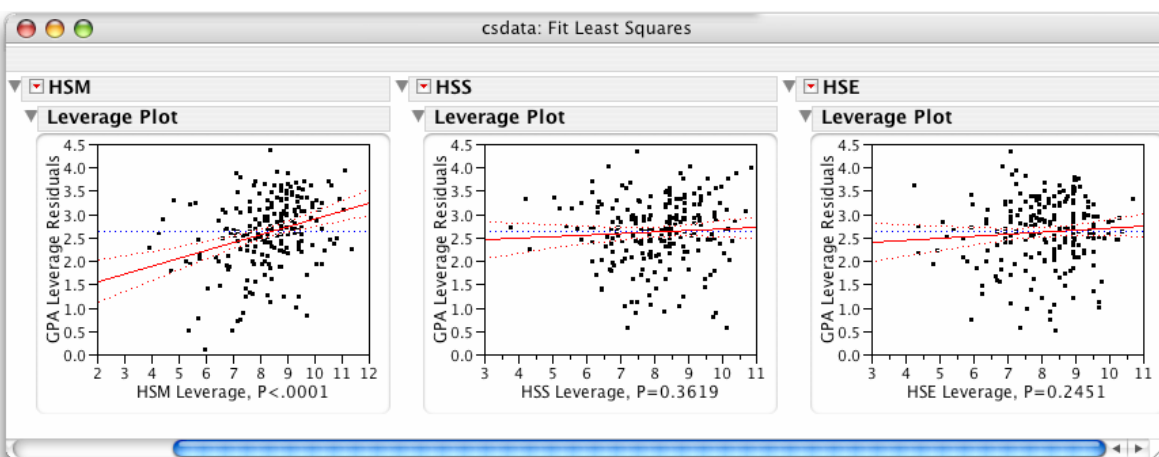
$$\widehat{GPA} = 0.590 + 0.169 HSM + 0.034 HSS + 0.045 HSE.$$

The  $t$  statistics for testing the regression coefficients and their associated  $P$ -values appear in the last two columns labeled **t Ratio** and **Prob>|t|**, respectively. Only the coefficient of **HSM** (high school math grade) is significantly different from zero.

These results are supported by the associated graphics. The plot shown at the top of the **Whole Model** report shows actual **GPA** plotted versus **GPA** predicted by the model. A diagonal 45° reference line and 95% confidence intervals for the model fit are also plotted. Because the confidence curves do not contain the blue line that represents the no effect model, you conclude that the whole model is statistically significant. This confirms the *ANOVA F test* result.

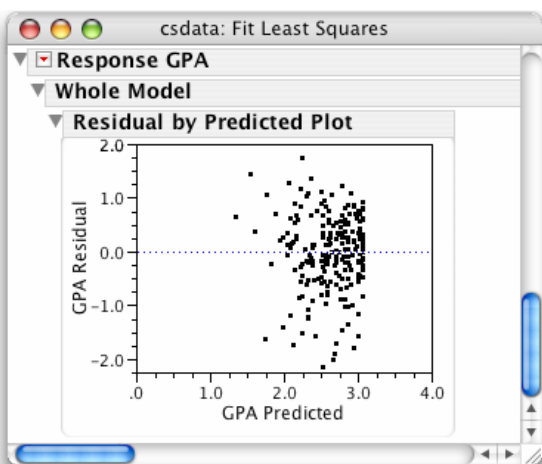
*JMP* also produces *leverage plots*, or *added variable plots*, for each explanatory variable. They confirm the results of the  $t$  statistics. The plots enable you to see the residual of the full model and the residual of the model without the variable of interest. A graphical test of whether a variable is important is obtained

by comparing the confidence curves around the red line to the horizontal blue line just as in the Whole Model test. Only the confidence interval in the leverage plot for **HSM** does not contain the blue line. Therefore, only the coefficient of **HSM** is significantly different from zero ( $\alpha = 0.05$ ).



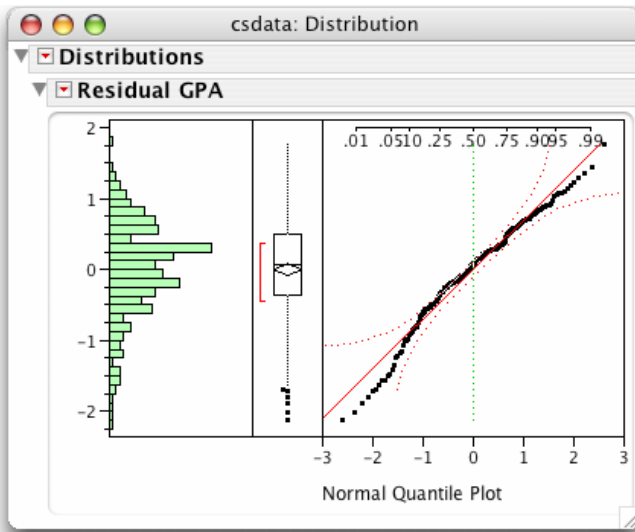
## Residuals

A plot of the residuals versus the predicted values can be found at the bottom of the report. In general, this plot will be much more informative than individual plots of the residuals versus each of the explanatory variables. The plot appears to be random noise.



To obtain a Normal quantile plot of the residuals, we save the residuals to the data table and then use the **Normal Quantile Plot** command in the **Distribution** platform.

2. Press the red triangle next to **Response GPA**.
  - a. Select **Save Columns** ⇒ **Residuals** from the menu that opens.
3. Select **Analyze** ⇒ **Distribution**.
  - a. Select **Residual GPA**, and press **Y, Columns** and **OK**.
  - b. Select **Normal Quantile Plot** from the red triangle menu of the **Residual GPA** report.



While there appears to be some modest departures from Normality in the center of the distribution, they are not likely to affect the  $P$ -values in view of the large sample size.

## 11.2 Summary

All graphs and statistical computations for multiple regression models use the **Fit Model** platform. The residuals can be examined using the **Distribution** platform.

### Activity

Inference about the model  
Saving residuals

### Command

Analyze ⇒ Fit Model  
Fit Model ⇒ Save Columns ⇒ Residuals

# Chapter 12

## One-Way Analysis of Variance

Chapter 7 provided tools for comparing the means of two groups or treatments. In this chapter, we compare any number of means using techniques that generalize the tools of Chapter 7. The same *JMP* commands for comparing two means, **Fit Y by X**  $\Rightarrow$  **Means/Anova**, will allow us to compare more than two means.

### 12.1 Inference for Comparing Three or More Means

The *JMP* data table layout is the same as in Section 7.2 in Chapter 7. Each row contains only one experimental unit or individual. There is a column to identify the groups or treatments and a column for the response variable. The column that identifies the groups or treatments should be thought of as an explanatory variable, or factor. Then, comparing groups is equivalent to evaluating the relationship between the two variables and determining whether the response variable depends on the explanatory variable. Thus, we use the **Fit Y by X** platform in *JMP*.

#### Examples 12.1 Workplace safety

---

In a study of workplace safety, workers were asked to rate the safety of all of their work environments and a composite score called the Safety Climate Index (SCI) was calculated. The workers were classified as unskilled, skilled, and supervisor. One question of interest in the study is whether or not there is sufficient evidence in the data to conclude that the means of the populations corresponding to the three worker groups are not all equal. Suppose that the file **eg12\_003.jmp** contains the data from this study.

Let  $\mu_{un}$ ,  $\mu_{sk}$ , and  $\mu_{su}$  denote the population mean SCI score for all unskilled workers, skilled workers, and supervisors, respectively. Then, we wish to test the hypothesis

$$H_0: \mu_{SK} = \mu_{su} = \mu_{sk} \text{ versus}$$

$$H_a: \text{not all of the } \mu_i \text{ are equal; i.e., at least two of the } \mu_i \text{ differ.}$$

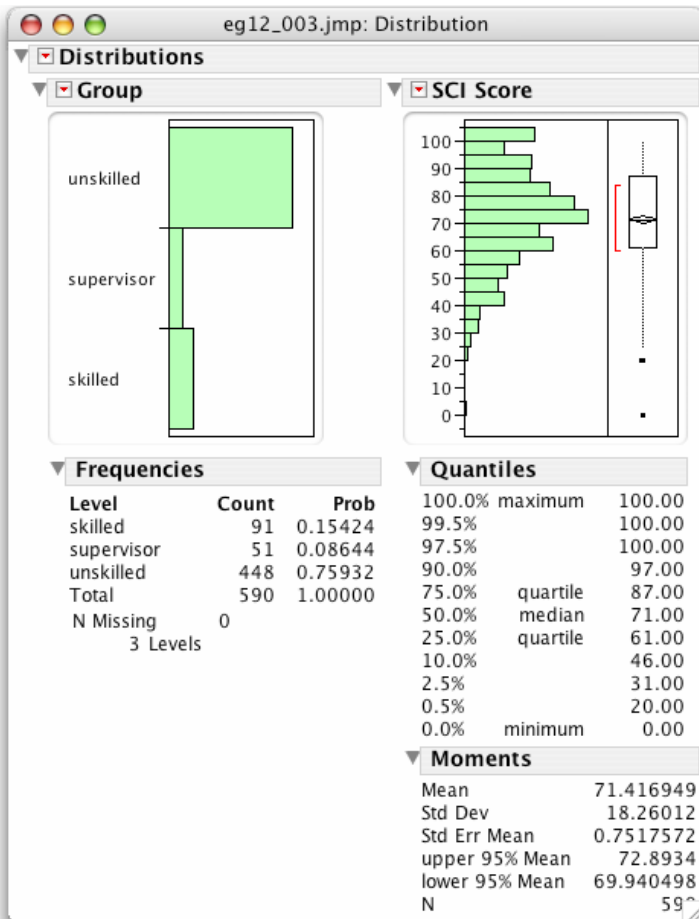
#### Inspect the Data

We first look at the variables separately.

1. Open the *JMP* data table **eg12\_003.jmp** and notice that there are three columns, named **Group**, **Group Code**, and **SCI Score**.

	Group	Group Code	SCI Score
447	unskill	1	61
448	unskill	1	87
449	skilled	2	78
450	skilled	2	31

2. Select **Analyze** ⇒ **Distribution**.
  - a. Select **Group** and **SCI Score** and press **Y, Columns**.
  - b. Press **OK**.

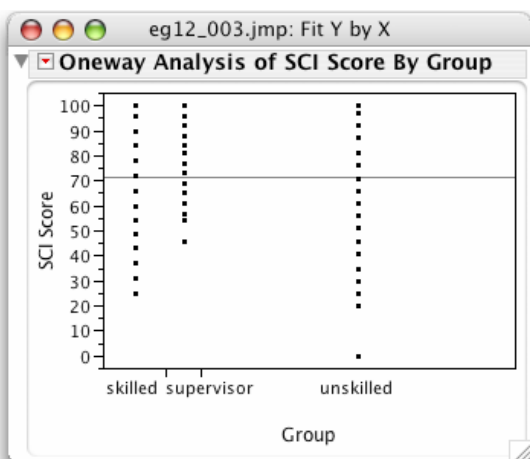


As expected, there are considerably more unskilled workers than skilled workers and supervisors. The overall mean safety score is 71.4 with a standard deviation of 18.26. We now look at the variables together.

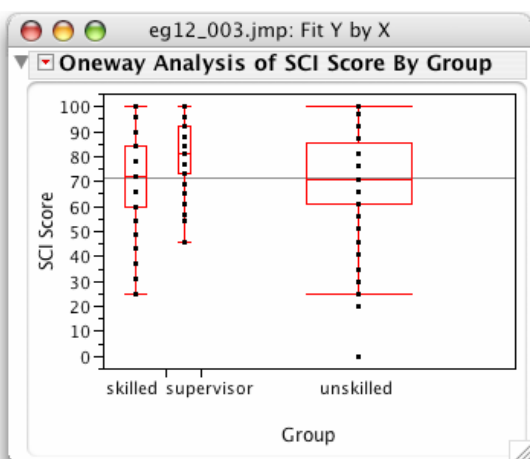
3. Select **Analyze** ⇒ **Fit Y by X**.
  - a. Select **SCI Score** and press **Y, Response**.
  - b. Select **Group** and press **X, Factor** and **OK**.



Side-by-side point plots provide some insight into the effect of group on safety score. However, we gain more by looking at side-by-side boxplots and side-by-side means diamonds.



4. Select **Display Options** ⇒ **Box Plots** from the red triangle menu to display boxplots.

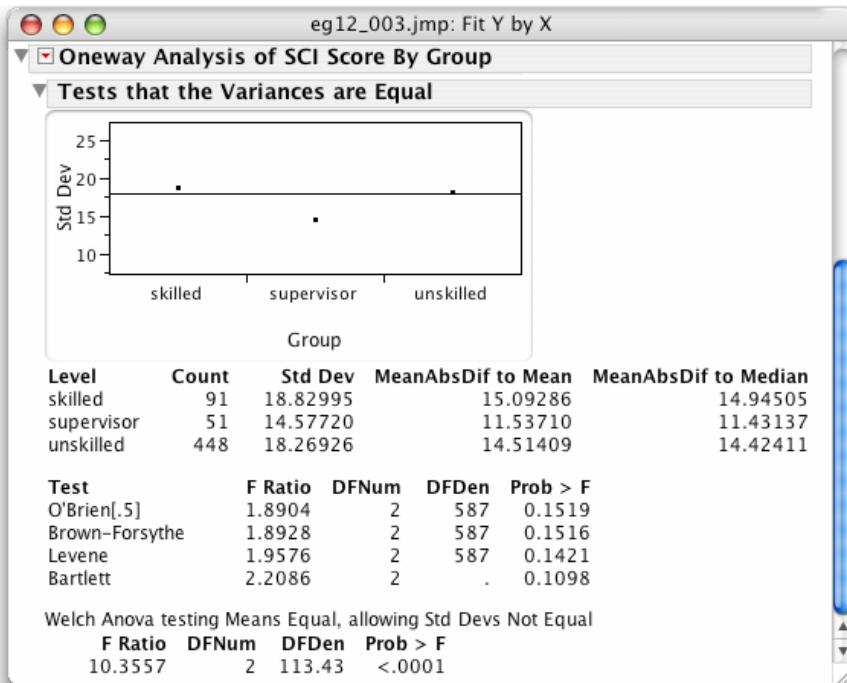


Note that the widths of the boxplots have been constructed proportional to the sample sizes of the groups. The average scores for the unskilled workers and the skilled workers in the samples do not differ while the supervisors have a higher average.

### Check Assumptions

In order to extend our results to the means of three populations of worker types, certain assumptions must be satisfied. As we showed in Section 7.3 of Chapter 7, *JMP* allows us to test for departures from the assumption of equal variances. We illustrate this again.

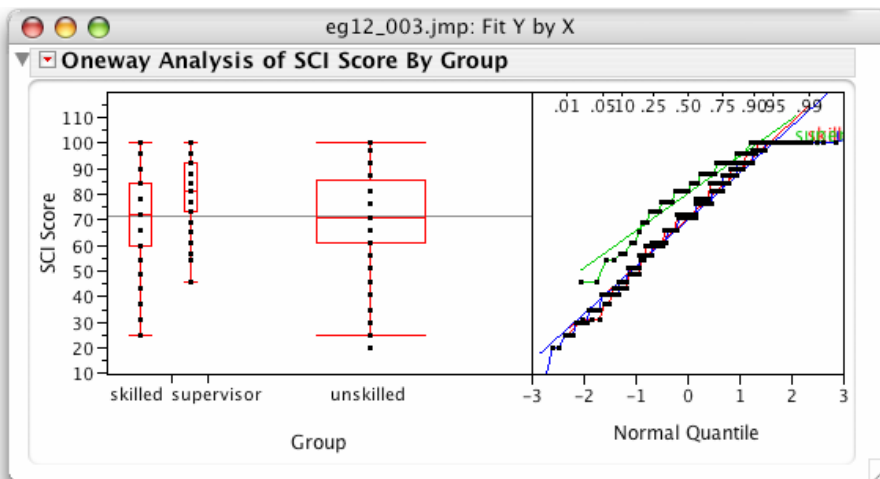
5. Click on the red triangle and select **UnEqual Variances**.



None of the four tests for unequal variances is significant at the 0.10 level.

To check for non-Normality, *JMP* also provides the **Normal Quantile Plot** command.

6. Press the red triangle and select **Normal Quantile Plot** ⇒ **Plot Actual by Quantile**.

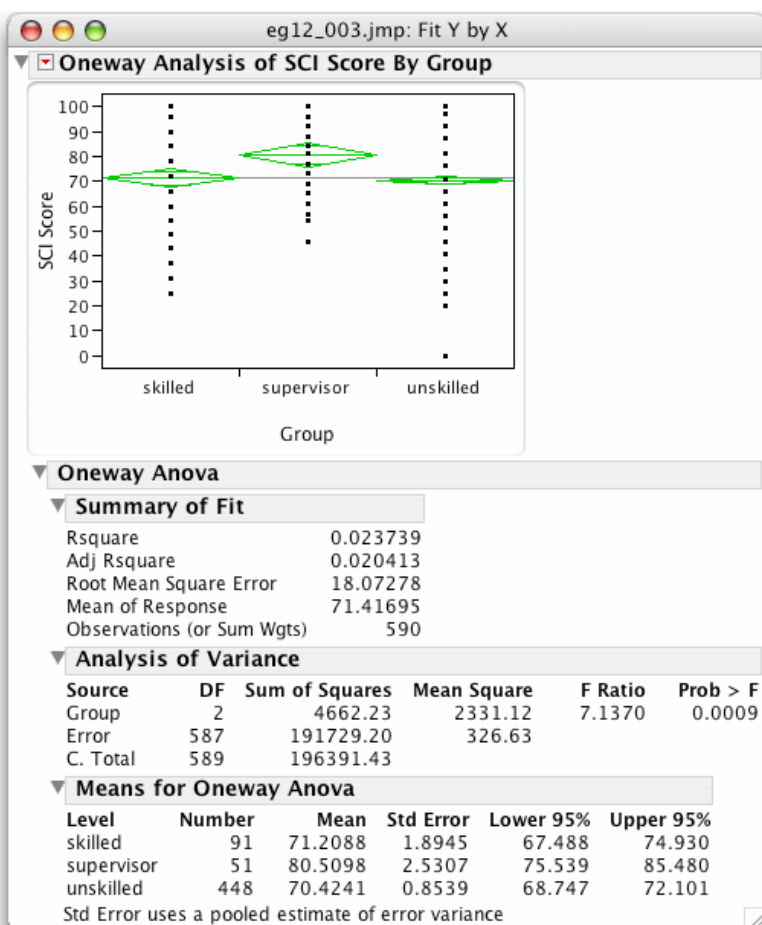


Three Normal quantile plots are displayed next to the side-by-side plots, one for each group. The points for our distributions do not deviate from the lines by much and there are no outliers.

Because the data look reasonably Normal and the standard deviations are about the same, we can compare the three population means using one-way analysis of variance.

## The ANOVA and the $F$ Test

7. Click on the red triangle and select **Means/Anova**. (In the display below, the Normal quantile plots as well as the boxplots have been deselected.)



The **Means for Oneway Anova** report lists the three sample means, 71.2088, 80.5098, and 70.4241. The *root mean square error* (18.07278) and  $R^2$  (labeled **RSquare** and equal to 2.4%) are found in the **Summary of Fit** report.

From the **Analysis of Variance** report, we see that the  $F$  test statistic (labeled **F Ratio**) is 7.1370 and has 2 and 587 degrees of freedom (**DF**). The  $P$ -value (labeled **Prob>F**) for the test statistic is 0.0009. Therefore, there is strong evidence to conclude that the population means differ, that is, the alternative hypothesis is true.

## Remark

- The modeling type of the column that identifies the groups must be “Nominal” (see Section 7.2 in Chapter 7 for details). If you were to perform steps 4 and 5 above using the column **Group Code** as the factor, the commands **Box Plots** and **Unequal Variances** would not be available. This is because *JMP* assumes that you wish to perform least-squares regression since **Group Code** is a numeric variable with modeling type “Continuous.”

## 12.2 Comparing the Means

The *Analysis of Variance F test* is an overall test to determine if there is good evidence of *any* differences among the means that we wish to compare. If the *F* test is statistically significant, then, without further analysis, all that we can safely conclude is that the groups with the largest and smallest means are significantly different. More detailed follow-up is required to decide which of the other group (population) means differ significantly.

Sometimes, specific questions can be formulated about the population means beforehand. In that case, *contrasts* can be used to answer the questions. Otherwise, we use *multiple comparison* procedures.

### Contrasts

Specific questions formulated before examination of the data can be expressed as *contrasts*. Significance tests provide answers to these questions. We use a different platform in *JMP*, **Fit Model**, to construct and test *contrasts*.

#### Example 12.1 Workplace safety (cont'd.)

Experts on safety in workplaces would suggest that supervisors face a very different safety environment than the other types of workers. Therefore, a reasonable question to ask is whether or not the supervisors are different from the others. Also of interest is whether the two non-supervisor groups have different mean SCI scores. Consequently, prior to analyzing the data, the researchers formulated the following hypotheses to be tested.

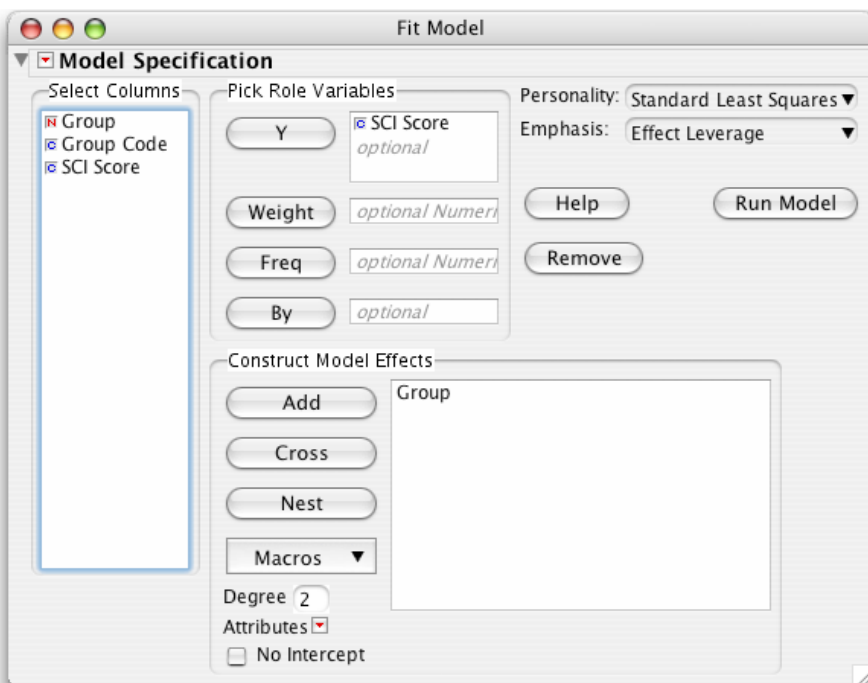
- (1)  $H_{01}: .5(\mu_{UN} + \mu_{SK}) = \mu_{SU}$  versus the alternative  $H_{a1}: .5(\mu_{US} + \mu_{SK}) > \mu_{SU}$ , and  
 (2)  $H_{02}: \mu_{UN} = \mu_{SK}$  versus the alternative  $H_{a2}: \mu_{UN} \neq \mu_{SK}$

Each of these gives rise to a *contrast* to be tested:

- (1)  $\psi_1 = \mu_{SK} - .5\mu_D - .5\mu_S$  and  
 (2)  $\psi_2 = \mu_{UN} - \mu_{SK}$

Let's construct and test these contrasts.

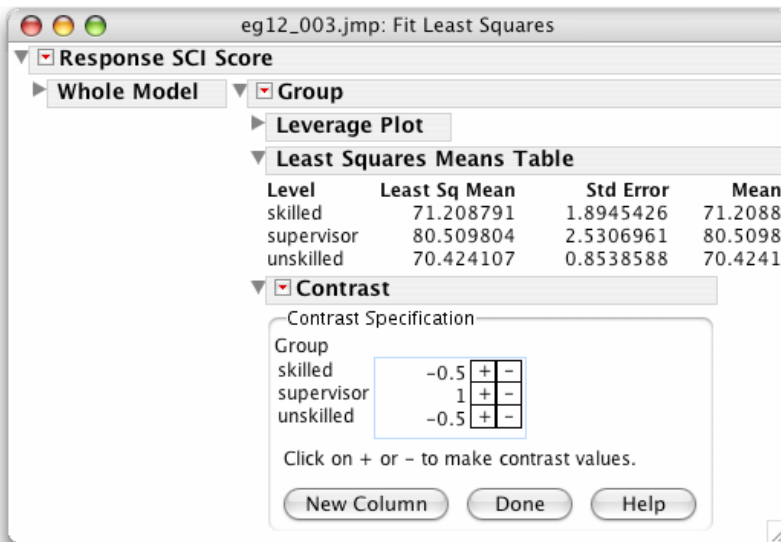
1. Select **Analyze** ⇒ **Fit Model**.
  - a. Select **SCI Score** and press **Y**.
  - b. Select **Group** and press **Add** under **Construct Model Effects**.
  - c. Press **Run Model**.



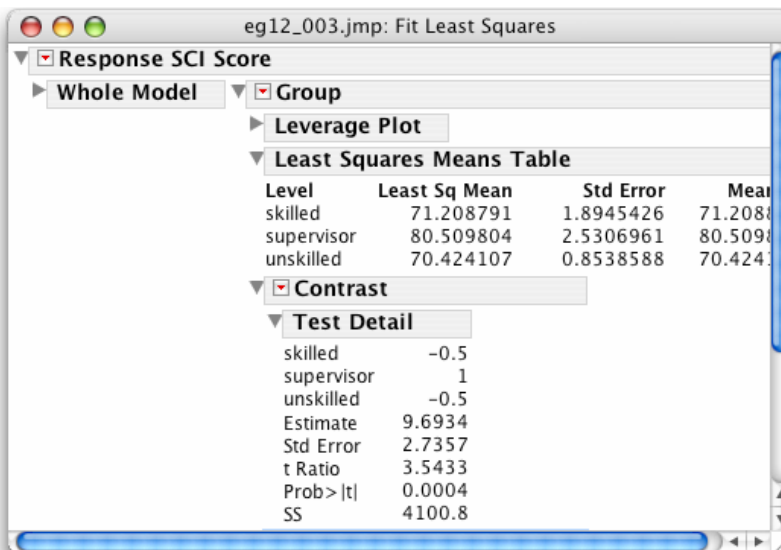
In the resulting window that opens, we are interested in the **Group** report. Close the **Whole Model** report and the **Leverage Plot** for **Group** to obtain the following.

Level	Least Sq Mean	Std Error	Mean
skilled	71.208791	1.8945426	71.2088
supervisor	80.509804	2.5306961	80.5098
unskilled	70.424107	0.8538588	70.4241

2. Press the red triangle next to the **Group** report title and select **LSMeans Contrast...**.
3. To specify the first contrast,
  - a. Press the + (plus) sign next to **supervisor** group,
  - b. Press the - (minus) sign next to **unskilled** group, and
  - c. Press the - (minus) sign next to **skilled** group.

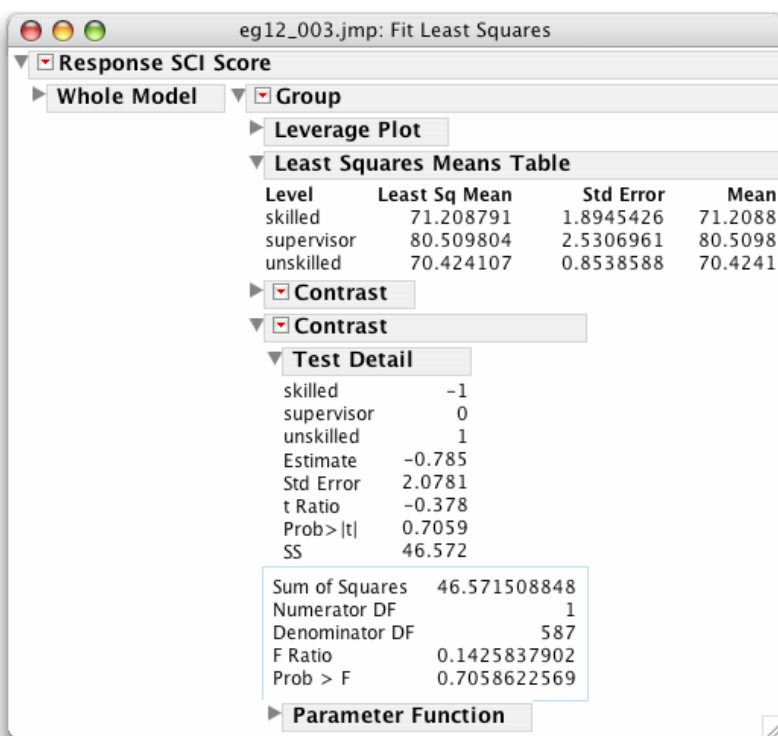


d. Press **Done** and open the resulting **Test Detail** report by clicking the disclosure diamond.



Notice the coefficients of **skilled**, **supervisor**, and **unskilled** are  $-0.5$ ,  $+1$ , and  $-0.5$ , respectively; the estimate of the first contrast is 9.69 and its standard error is 2.7357. The  $t$  statistic is 3.54 and its two-sided  $P$ -value (labeled **Prob>|t|**) is 0.0004. Since the researchers were interested in showing only that the supervisors' SCI scores were better than the workers' scores, they were interested in a one-sided alternative and the  $P$ -value for the first hypothesis is 0.0002 (equal to one-half of 0.0004). The data strongly support the hypothesis that the supervisors have higher mean SCI scores than the non-supervisory groups, in general. To obtain the test for the second contrast, simply:

4. Select **LSMeans Contrast...** from the red triangle menu on the **Group** report title bar again.
  - a. Press the **+** (plus) sign next to **unskilled** and the **-** (minus) sign next to **skilled**.
  - b. Select **Done** and open the **Test Detail** report.



The estimate of the second contrast, which compares the two non-supervisor groups, is  $-0.785$  with a standard error of  $2.078$ . This gives a  $t$  statistic equal to  $-0.378$  and a two-sided  $P$ -value of  $0.7059$ . There is not sufficient evidence to show a difference in population mean SCI scores between unskilled workers and skilled workers.

## Remark

- The **Fit Model** platform for use with analysis of variance is discussed in more detail in the next chapter.

## Multiple Comparisons

Frequently, specific questions cannot be formulated in advance of the analysis. *Multiple comparison* methods are then used to compare pairs of population means. *JMP* provides several multiple comparison methods. We illustrate two: the *least-significant differences*, or *LSD*, method and the *Tukey-Kramer honestly-significant differences*, or *HSD*, method. We use the **Fit Y by X** platform to calculate these. The *JMP* command **Each Pair, Student's t** gives the *least-significant differences* method while the command **All Pairs, Tukey HSD** gives the *Tukey-Kramer honestly-significant differences*, or *HSD*, method. We recommend using the latter method.

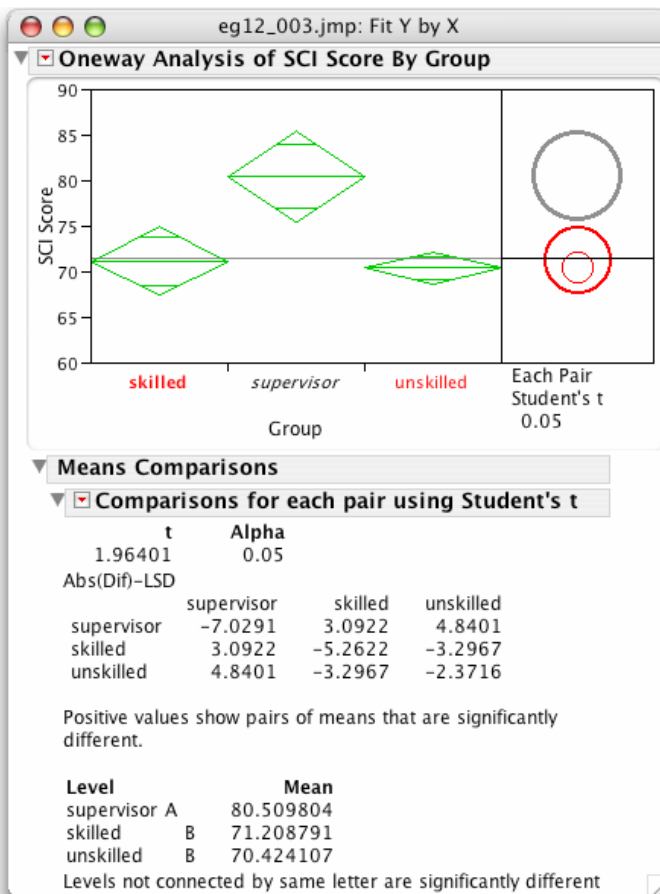
### Example 12.1 Workplace safety (cont'd.): Which means differ?

We start with the *least-significant differences*, or repeated two-sample  $t$  test, procedure.

If the *JMP* data table **eg12\_003.jmp** from the last example is closed, open it and repeat steps 3 and 7 of Section 12.1:

1. Select **Analyze** ⇒ **Fit Y by X**.
  - a. Select **SCI Score** ⇒ **Y, Response**.
  - b. Select **Group** ⇒ **X, Factor** and then select **OK**.
  - c. On the resulting report, press the red triangle menu and select **Means/Anova**.
2. Press the red triangle and select **Compare Means** ⇒ **Each Pair, Student's t**.

Each multiple comparison procedure begins with a *comparison circles* plot, which is a visual representation of group mean comparisons. You can compare each pair of group means visually by clicking on any comparison circle to highlight it. The highlighted circle appears with a thick (red) solid line. Circles representing groups with means that are *not* significantly different from the selected group appear as thin (red) lines. Circles representing means that are significantly different from the selected circle are displayed with a thicker (gray) color. A table of mean comparisons follows:



- a. Click on the *comparison circle* of a group (**skilled** was selected here) to see which group means are significantly different from that group.

The selected circle appears with a thick red line. Since only the circle for supervisors is a thick gray color, only the population mean **SCI score** for supervisors differs from that for the skilled workers.

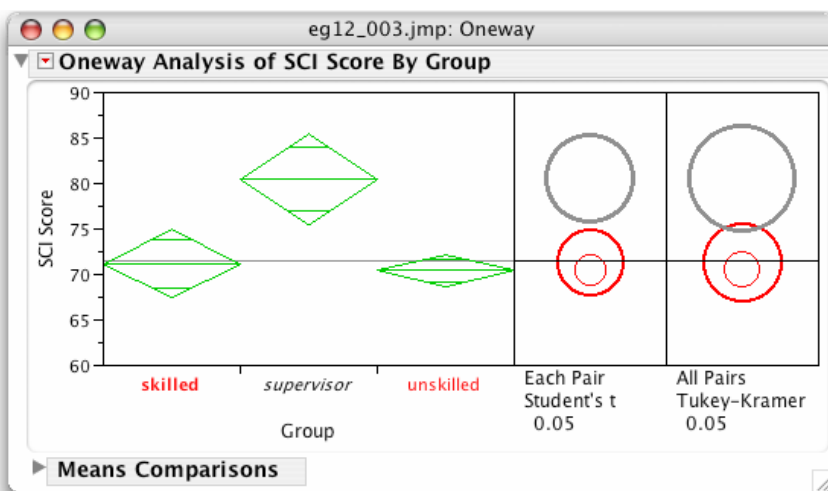
- b. Click on the *comparison circle* of another group, say, **unskilled**.



The unskilled workers are shown to be significantly different from supervisors but not from skilled workers.

As the textbook states, the method of *least-significant differences*, or *LSD*, has some undesirable properties. In general, it should not be used when comparing more than two groups because it does not take into consideration all the comparisons that you are making. The *Tukey-Kramer honestly significant difference*, or *HSD*, method, unlike the *LSD*, does control the overall probability of some false rejection among all pairs. It is less conservative than the *Bonferroni* method and offers just as much protection. We recommend that it be used in place of the Bonferroni method for problems in the textbook. To display the Tukey-Kramer HSD procedure:

3. Press the red triangle and select **Compare Means** ⇒ **All Pairs, Tukey HSD**.



Notice that the circles are a bit larger for this method. For this example, though, we reach the same conclusions with both methods. When the conclusions differ, the Tukey-Kramer HSD method is usually the appropriate method.

## Remarks

- The default level of significance is 0.05. This can be changed using the **Set Alpha Level** command on the red triangle pop-up menu.
- The abbreviation *LSD* in the **Means Comparisons** report refers to the smallest difference between two means that is statistically significant.

## 12.3 The Power of the ANOVA Test

To calculate the power of an ANOVA test in *JMP*, we use the **Power and Sample Size** command on the **DOE** platform that was used in Section 3 of Chapter 7. This platform does not require that data already be produced and so can be used prospectively in planning studies in which ANOVA will be used for analysis. We illustrate its use.

**Examples 12.2 Power and sample size calculations**

Suppose that a study on reading comprehension for three different teaching methods is being planned. A previous study performed in a different setting found sample means of 41, 47, and 44, and the pooled standard deviation was 7. To decide on an appropriate total sample size for the study, we will use this information to create a graph of the power for the alternative  $\mu_1 = 41$ ,  $\mu_2 = 47$ ,  $\mu_3 = 44$ , with  $\sigma = 7$  versus different numbers of subjects. We will use an equal number of subjects for each group and assume that the ANOVA test will be performed at the 5% level of significance.

1. Select **DOE**  $\Rightarrow$  **Sample Size and Power**
2. Select **k Sample Means**.

Sample Size and Power

**Sample Size**

k Means

Testing if there are differences among k means.

Alpha

Error Std Dev

Extra Params

Enter up to 10 Prospective Means showing separation across groups


Enter Power or Sample Size to get the other.  
Enter neither to get a plot of Power vs. Sample Size

Sample Size

Power

Sample Size is the total sample size; per group would be n/k


**Continue**

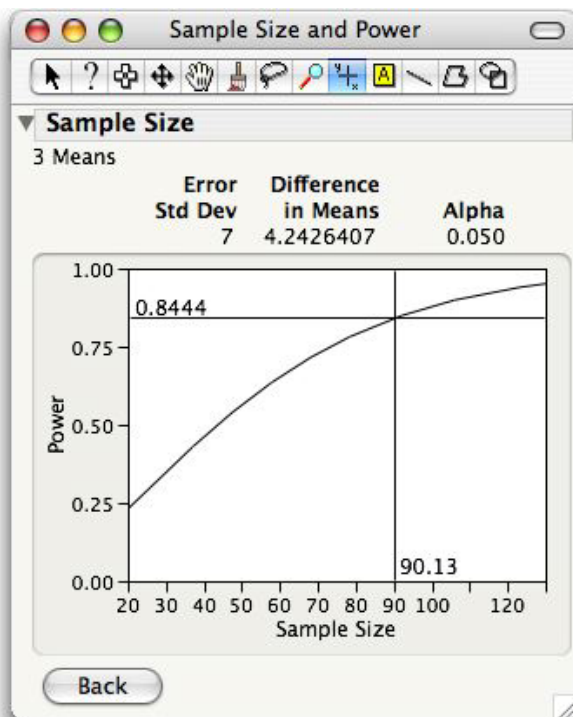
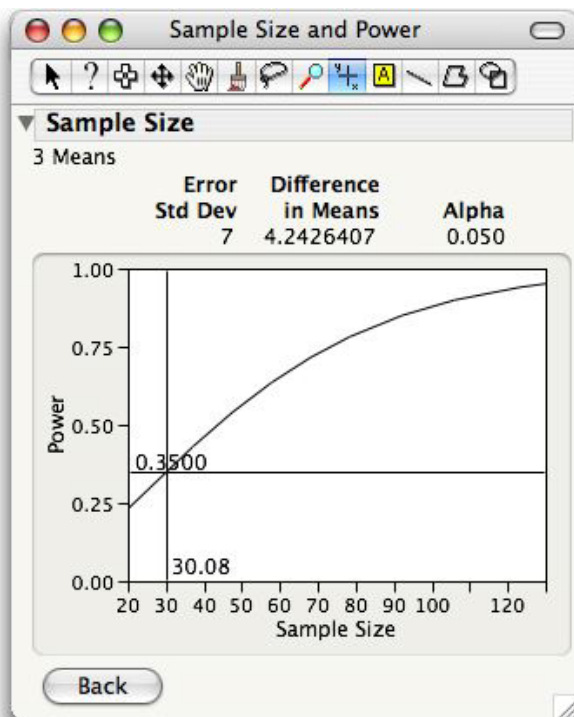
**Back**

Fill in the required information. Leave both the **Sample Size** and **Power** fields blank to get a graph of power versus sample size.

3. a. Enter 0.05 for **Alpha**, the level of significance of the ANOVA test.
- b. Enter 7 for the **Err Std Dev**.
- c. Enter 41, 47, and 44 for the 3 **Prospective Means** of the groups.
- d. Press **Continue**.

To read the power for the ANOVA test when the total sample size is 30 (10 per group) or 90 (30 per group), use the **Crosshair** tool.

4. Select the **Crosshair tool**  from the toolbar.
- Place the crosshair on the curve with the total sample size = 30.
  - Place the crosshair on the curve with the total sample size = 90.



From the plots, we see that the power for these sample sizes is about 35% and 84%, respectively.

## 12.4 Summary

### Graph/Computation

ANOVA  $F$  test  
 Evaluating assumptions  
   Normal quantile plots  
   Equal variances  
 Contrasts  
 Multiple comparisons  
 Power

### Command

Fit Y by X  $\Rightarrow$  Means/Anova  
Fit Y by X  $\Rightarrow$  Normal Quantile Plots  
Fit Y by X  $\Rightarrow$  Unequal Variances  
Fit Model  $\Rightarrow$  LSMeans Contrast  
Fit Y by X  $\Rightarrow$  Compare Means  
DOE  $\Rightarrow$  Sample Size and Power

# Chapter 13

## Two-Way Analysis of Variance

In Chapter 12, inference for the relationship between a response variable  $y$  and a *single factor*, or *categorical explanatory* variable  $x$ , was discussed. Now, we consider the case when there is more than one factor or explanatory variable—specifically, when there are *two factors*.

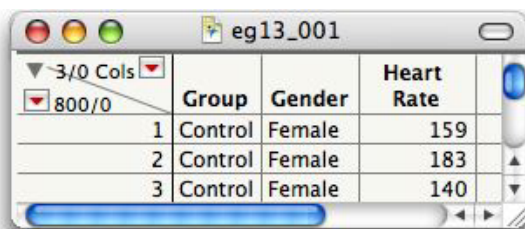
### 13.1 Inference for Two-Way ANOVA

Since there is more than one factor, or explanatory variable, we use the *JMP Fit Model* platform to analyze the data in this chapter.

#### Example 13.1 Lifestyles and cardiac fitness

A study of cardiovascular risk factors compared runners who averaged at least 15 miles per week with a control group described as “generally sedentary.” Both men and women were included in the study. One variable of interest in the study was the heart rate after 6 minutes of exercise on a treadmill.

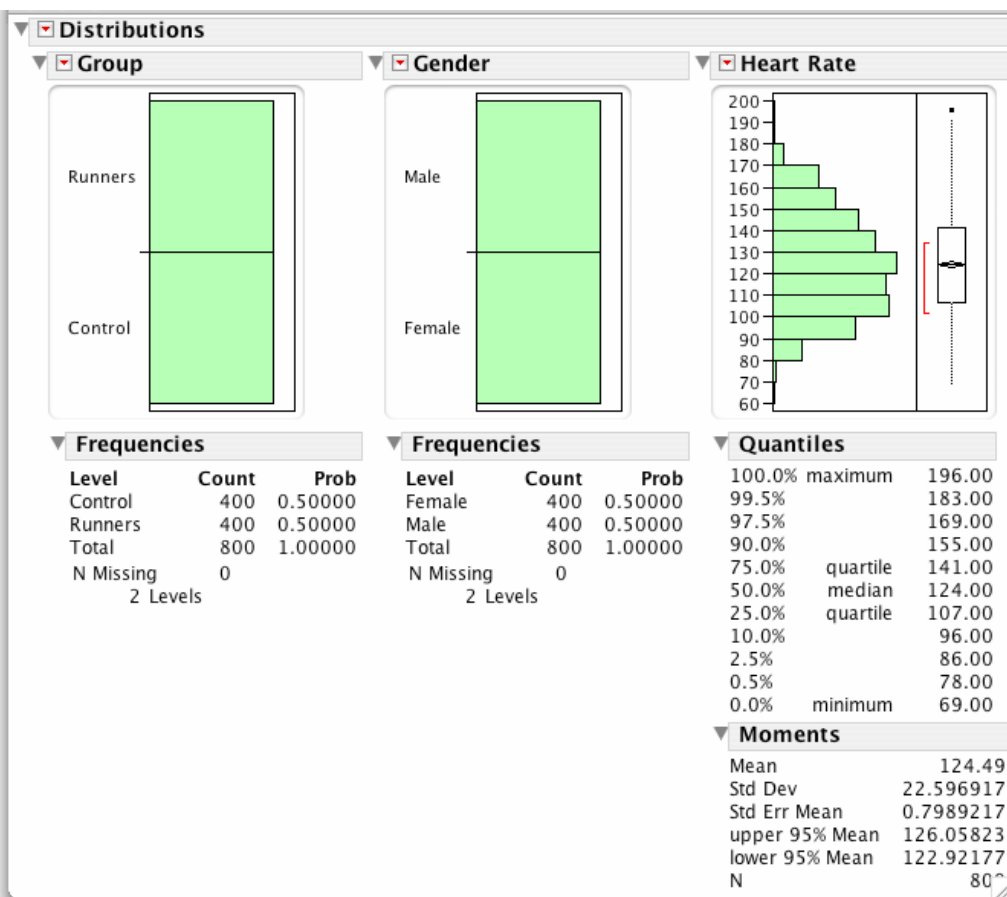
1. Open the *JMP* data table containing the data. There are 800 individuals and four variables, **Subject**, **Group**, **Gender**, and **Heart Rate**.



	Group	Gender	Heart Rate
1	Control	Female	159
2	Control	Female	183
3	Control	Female	140

Use the **Distribution** platform in *JMP* to first look at the distributions of the three variables.

2. Select **Analyze** ⇒ **Distribution**.
  - a. Select **Group**, **Gender**, and **Heart Rate**; press **Y, Columns** and **OK**.

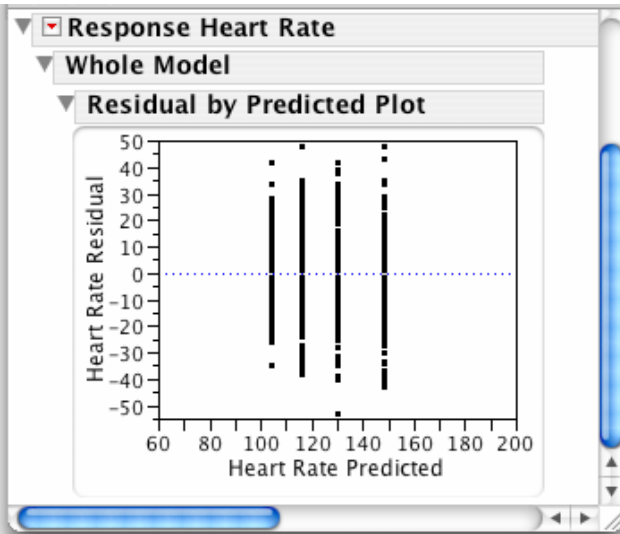


This was a carefully designed study with an equal number of males and females and an equal number of runners and nonrunners. Further examination (**Fit Y by X** with **Group** and **Gender**) would show that there are exactly 200 people in each of the four group-by-gender categories. The response variable **Heart Rate** is symmetrically distributed with a mean of 124.49 beats per minute (bpm) and a standard deviation of 22.60 bpm. Subject 154 is somewhat outlying.

It is possible to examine the relationships of **Heart Rate** to **Group** and to **Gender** separately (using **Fit Y by X**); however, this fails to detect a possible interaction of **Group** and **Gender**. We proceed directly to a model that includes both **Gender** and **Group** as factors.

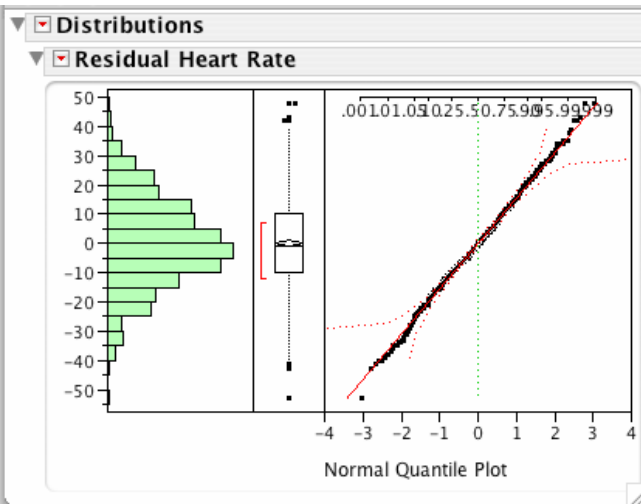
3. Select **Analyze** ⇒ **Fit Model**.
  - a. Select **Heart Rate** and press **Y**.
  - b. Select **Group** and **Gender** and press **Add**.
  - c. Select **Group** and **Gender** and press **Cross** to include an interaction effect.
  - d. Press **Run Model**.

Examine the residuals. First, scroll down to the **Residual by Predicted Plot**. The spread of the four groups appears to be roughly the same.

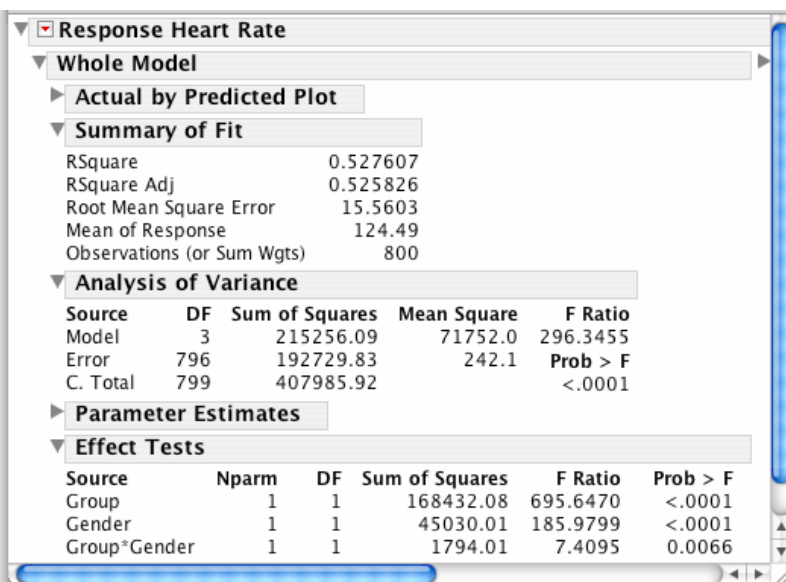


To check the residuals for Normality, save them to the data table and use the **Normal Quantile Plot** command.

4. Press the red triangle next to **Response Heart Rate** and select **Save Columns** ⇒ **Residuals**.
5. Select **Analyze** ⇒ **Distribution**.
  - a. Select **Residuals Heart Rate** and press **Y, Columns** and **OK**.
  - b. Select **Normal Quantile Plot** from the red triangle menu.



The data appear to be reasonably Normal. Now inspect the **Whole Model** report.



**Response Heart Rate**

▼ **Whole Model**

▶ **Actual by Predicted Plot**

▼ **Summary of Fit**

RSquare	0.527607
RSquare Adj	0.525826
Root Mean Square Error	15.5603
Mean of Response	124.49
Observations (or Sum Wgts)	800

▼ **Analysis of Variance**

Source	DF	Sum of Squares	Mean Square	F Ratio	Prob > F
Model	3	215256.09	71752.0	296.3455	
Error	796	192729.83	242.1		
C. Total	799	407985.92			<.0001

▶ **Parameter Estimates**

▼ **Effect Tests**

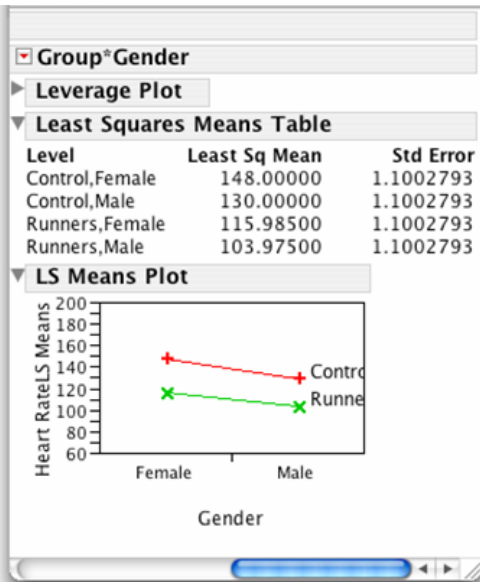
Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
Group	1	1	168432.08	695.6470	<.0001
Gender	1	1	45030.01	185.9799	<.0001
Group*Gender	1	1	1794.01	7.4095	0.0066

In the **Summary of Fit** report, the root mean square error is 15.6 bpm and the *coefficient of determination*,  $R^2$ , equals .53. Thus, **Gender** and **Group** explain about 53% of the variability in the heart rates.

The **Analysis of Variance** report is, in effect, a one-way ANOVA table with four treatments: female control, female runner, male control, and male runner. The  $F$  test statistic (296.3) and its associated  $P$ -value (< .0001) indicate that the population means of the four treatments are not the same.

The **Effect Tests** report contains the sum of squares, degrees of freedom, and  $F$  statistics for the **Group** and **Gender** main effects and the **Group-by-Gender** interaction. These sum of squares and degrees of freedom sum to the sum of squares and degrees of freedom for the Model in the **Analysis of Variance** report. Examine the  $P$ -values (labeled as **Prob > F**). The interaction is strongly significant. To interpret the results, we examine a plot of the cell means.

6. Click on the red triangle next to the **Group\*Gender** report and select **LSMeans Plot**.



The two lines in the plot are not parallel. There is interaction. The difference between controls and runners in mean heart rates is greater for women than for men. The interaction, while significant, is not large. Return to the **Effect Tests** report and notice that the main effects for group and gender are both very significant. Runners, on average, have lower heart rates than the controls for both males and females by about 29 bpm ( $= 139 - 109.98$ ). Women, on average, have higher heart rates than men do by about 15 bpm.

## Remark

- *JMP* provides a shortcut for adding the two main effects and the interaction effect to the model in the **Fit Model** dialog. For the previous example, simply select **Group** and **Gender**, and press **Macros** ⇒ **Full Factorial**.

## 13.2 Summary

All graphs and statistical computations for multiple regression models use the **Fit Model** platform. The residuals can be examined using the **Distribution** platform.

### Activity

Inference about the model  
 Saving residuals  
 Plot of the cell means

### Command

Analyze ⇒ Fit Model  
Fit Model ⇒ Save Columns ⇒ Residuals  
Fit Model ⇒ LSMeans Plot



# Chapter 14

## Bootstrap Methods and Permutation Tests

The methods of this chapter provide alternatives to the methods of earlier chapters for finding standard errors and confidence intervals and for performing significance tests. They can be used to perform inference in settings for which there is no traditional method. They represent a new frontier in statistics resulting from the continuing revolution in computing.

Bootstrap resampling and bootstrap distributions provide an alternative to sampling distributions and interval estimation of parameters.

### 14.1 Bootstrap Methods

Bootstrap resamples are drawn with replacement from the original random. The distribution of the values of a statistic based on repeated resampling make up the bootstrap distribution of that resampled statistic. The bootstrap distribution of the statistic mimics the shape, spread, and bias of sampling distributions and so can be used to establish confidence interval estimates for the associated parameter.

Bootstrap resampling can be performed in *JMP* but, as is currently the case with most commercial statistical software, there is no high-level platform that can be used. The bootstrap distribution of a desired statistic must be generated by a lower-level language. The *JMP* scripting language provides that capability. We illustrate one such script.

#### 14.1.1 The Bootstrap Distribution

Here is an example to generate and describe the bootstrap distribution for the mean from a population that is not Normal.

---

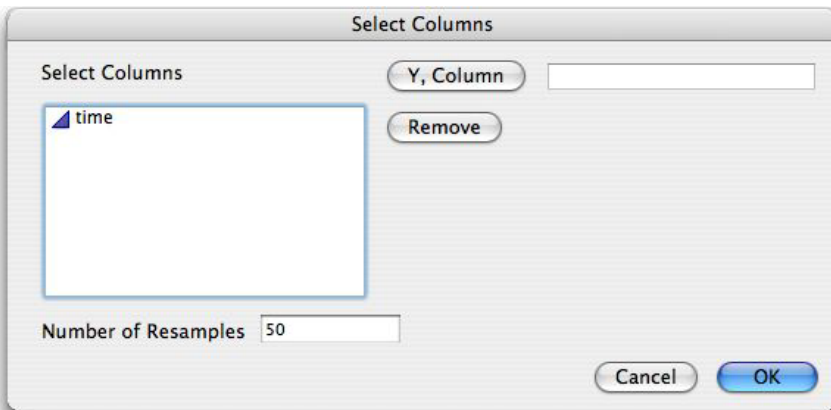
**Example 14.1 Telecommunication repair times**

---

Verizon is a local telephone company that provides repair services for itself and other telephone companies. The *JMP* data table **eg14\_001.jmp** contains data on the repair times of a random sample of 1664 repair calls from its own customers. Create 1000 bootstrap resamples and store them in a *JMP* data

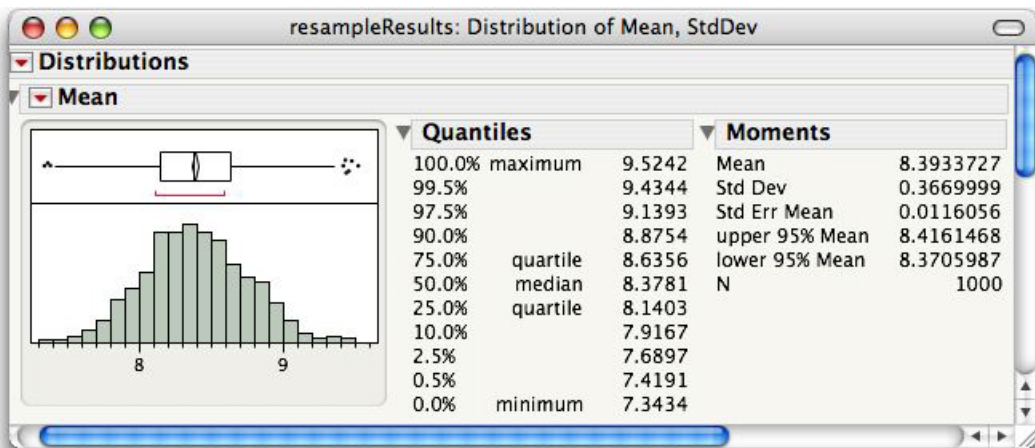
table for further analysis. To do this, we will use a *JMP* script named **BootstrappingTheMean.jsl** found at the textbook Web site.

1. Select **File** ⇒ **Open** and the file **eg14\_001.jmp**.
2. Select **File** ⇒ **Open** and the file **BootstrappingTheMean.jsl**.



- a. Select the column **time**, and press **Y, Column**.
- b. Enter **1000** for the **Number of Resamples** and press **OK**.

The bootstrap resamples are constructed and the distribution of their means is displayed.



The distribution is approximately bell-shaped with a mean of 8.393 and the bootstrap standard error is 0.367. Since the mean of the original sample is 8.412, we see that the bias of the bootstrap mean is small, 0.019.

## 14.1.2 Bootstrap Confidence Intervals

### Example 14.1 Telecommunication repair times (cont'd.)

The **bootstrap  $t$  confidence interval** is easily calculated because the sample mean time and bootstrap error are known, and the critical value of the  $t$  distribution with 1663 degrees of freedom for .95 confidence is 1.96. Thus, the bootstrap  $t$  confidence interval estimate for the mean repair time for Verizon customers is  $8.412 \pm 1.96 (0.367) = (7.993, 9.131)$  hours.

The **bootstrap percentile confidence interval** for selected levels of confidence (.99, .95, .80, .50) can be obtained directly from the Quantiles report. The .95 percentile confidence interval estimate for the mean repair time is (7.690, 9.139) hours.

## 14.1.3 Bootstrapping Other Parameters

Bootstrapping of other parameters can be performed in *JMP* and requires multiple steps that need to be carefully thought out. Without a script, the computations of each resample and summaries are tedious. However, the code varies considerably for different parameters and different designs as is the case for the various platforms in the **Analyze** menu.

## 14.2 Permutation Tests

Permutation tests are significance tests based on permutation resamples drawn at random from the original data. When they can be used, permutation tests have great advantages. They do not require that the population(s) be Normally distributed and they apply to a variety of statistics, not just ones with a well known distribution under the null hypothesis.

As is the case with bootstrap resampling, there are no high-level platforms for permutation tests in *JMP*. Custom *JMP* scripts specific to the study design are needed to carry out construction of the permutation distribution and computation of the  $P$ -value of the test.

## 14.3 Summary

The statistical computations in this chapter are performed using custom scripts found on the textbook Web site.

Graph/Computation	Command
Bootstrap Methods	
Mean	<a href="#"><u>BootstrappingTheMean.jsl</u></a> ⇒ <a href="#"><u>Run Script</u></a>

# Chapter 15

## Nonparametric Tests

The most common methods for inference about means assume that the variables have Normal distributions in the population or populations from which they are drawn. Bootstrap methods and permutation tests do not require a specific form for the distribution of the population(s). Nonparametric methods also do not require Normality or any other specific form for the distribution of the population(s). *JMP* provides a wide range of nonparametric tests. They are especially useful when the sample(s) are small.

### 15.1 The Wilcoxon Rank Sum Test

The *Wilcoxon rank sum test* compares two distributions to assess whether one has systematically larger values than the other. Use the **Nonparametric  $\Rightarrow$  Wilcoxon Test** command on the **Fit Y by X** platform to display the *Wilcoxon rank sum test*.

#### Example 15.1 Weeds and corn yield

---

Does the presence of a small number of weeds reduce the yield of corn? A researcher planted corn at the same rate in eight small plots of ground, then weeded the corn rows by hand to allow no weeds in four randomly selected plots and exactly three lamb's-quarter plants (a weed) per meter of row in the other four plots. Test the hypotheses that

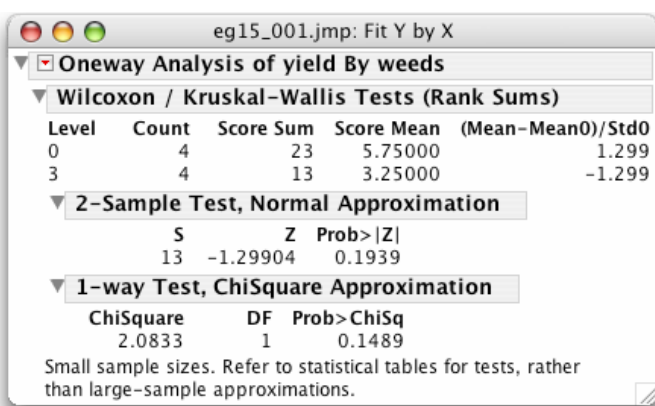
$H_0$ : There is no difference in distribution of yields versus

$H_a$ : Yields are systematically higher in weed-free plots

The methods of Section 7.2 of Chapter 7 assume that the yields are Normally distributed. Because the sample sizes are so small, we cannot rely on the robustness of the two-sample *t* test. We use the *Wilcoxon rank sum test* instead.

Suppose that the data are in the *JMP* data table **eg15\_001.jmp**, which has 2 variables, **weeds** and **yield**, and 8 rows. Suppose also that the variable **weeds** is numeric, with values 0 and 3. For our purposes, it should be treated as categorical.

1. Open the *JMP* data table **eg15\_001.jmp**.
2. Select the column **Weeds** to change its modeling type to “Nominal.”
  - a. Select **Cols** ⇒ **Cols Info**.
  - b. Select **Nominal** from the **Modeling Type** menu and press **OK**.
3. Select **Analyze** ⇒ **Fit Y by X**.
  - a. Select **yield** and press **Y, Response**.
  - b. Select **weeds** and press **X, Factor** and **OK**.
4. Click on the red triangle and select **Nonparametric** ⇒ **Wilcoxon Test**.



Level	Count	Score Sum	Score Mean	(Mean-Mean0)/Std0
0	4	23	5.75000	1.299
3	4	13	3.25000	-1.299

S	Z	Prob> Z
13	-1.29904	0.1939

ChiSquare	DF	Prob>ChiSq
2.0833	1	0.1489

Small sample sizes. Refer to statistical tables for tests, rather than large-sample approximations.

The **2-Sample Test, Normal Approximation** table gives the rank sum statistic 23 (but calls it **Score Sum** rather than *W*) as well as the *z* statistic  $-1.29904$  (using the continuity correction) and the two-sided *P*-value 0.1939. The one-sided *P*-value is one-half of that, or 0.09695.

## 15.2 The Wilcoxon Signed Rank Test

The nonparametric *Wilcoxon signed rank test* is often used in place of the *t* test when the assumption of normality does not hold in the one-sample or matched pairs settings. The *Wilcoxon signed rank test statistic* and associated *P*-values can be requested using the **Test Mean** command in the **Distribution** platform.

### Example 15.2 Storytelling and reading

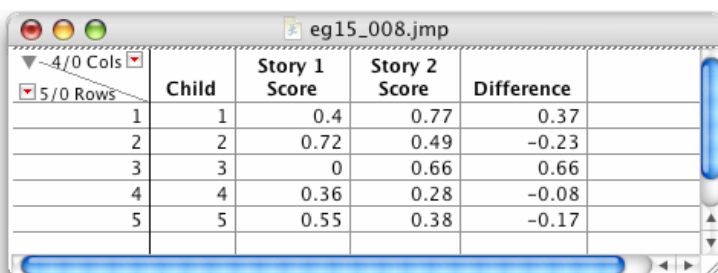
We wonder if illustrations improve how children retell a story. A study of early childhood education asked kindergarten students to tell a fairy tale that had been read to them earlier in the week. Each child told two stories. The first had been read to them, and the second had been read but also illustrated with pictures. An expert listened to a recording of the children and assigned a score for certain uses of language. The hypotheses to be tested are that for the population of “low-progress” readers

$H_0$ : Scores have the same distribution for both stories versus

$H_a$ : Scores are systematically higher for story 2 (which was illustrated)

This is a matched pairs design (see Section 7.1 in Chapter 7). To compare the story-telling scores, we create a column of the differences between the scores for the two stories and test whether the average difference in the scores differs from zero. Suppose that the *JMP* data table **eg15\_008.jmp** contains the data.

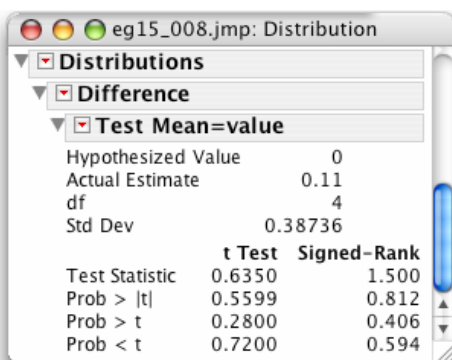
1. Open the *JMP* data table **eg15\_008.jmp**.
2. Select **Cols** ⇒ **New Column**.
  - a. Enter **Difference** in the **Column Name** field.
  - b. Press **New Property** ⇒ **Formula**.
  - c. Select **Story 2 Score** from the list of columns, press – (minus) on the keypad, and then select **Story 1 Score** from the list of columns.
  - d. Press **OK** and **OK**.



	Child	Story 1 Score	Story 2 Score	Difference
1	1	0.4	0.77	0.37
2	2	0.72	0.49	-0.23
3	3	0	0.66	0.66
4	4	0.36	0.28	-0.08
5	5	0.55	0.38	-0.17

To test the alternative hypothesis that the mean difference between the story scores is not zero, we use the **Test Mean** command on the **Distribution** platform.

3. Select **Analyze** ⇒ **Distribution**.
  - a. Select **Difference** and press **Y, Columns** and **OK**.
4. a. Select **Test Mean** from the red triangle menu on the **Difference** title bar.
  - b. Check the box for the **Wilcoxon Signed Rank** nonparametric test and press **OK**.



eg15_008.jmp: Distribution		
Distributions		
Difference		
Test Mean=value		
Hypothesized Value	0	
Actual Estimate	0.11	
df	4	
Std Dev	0.38736	
	t Test	Signed-Rank
Test Statistic	0.6350	1.500
Prob >  t	0.5599	0.812
Prob > t	0.2800	0.406
Prob < t	0.7200	0.594

The value of the *Wilcoxon signed rank test statistic* is 1.50 (*JMP* gives the deviation from the expected value of 7.5) and the *P*-value for an upper-tailed test is 0.406. There is no evidence from this small sample that seeing illustrations improves the story-telling of low-progress readers.

## 15.3 The Kruskal-Wallis Test

The *Kruskal-Wallis test* compares the distributions of several populations based on independent random samples. We use the same command as for the *Wilcoxon signed rank test*, **Nonparametric** on the **Fit Y by X** platform, to display the *Kruskal-Wallis test*.

### Example 15.3 Weeds and corn yield

---

Does the presence of a small number of weeds reduce the yield of corn? A researcher planted corn at the same rate in 16 small plots of ground and then weeded the corn rows by hand to allow a fixed number of lamb's-quarter plants (weeds) to grow in each meter of corn row. These numbers were 0, 1, 3, and 9 in the four groups of plots. All plots received identical treatment except for the weeds. We wish to test the hypotheses that

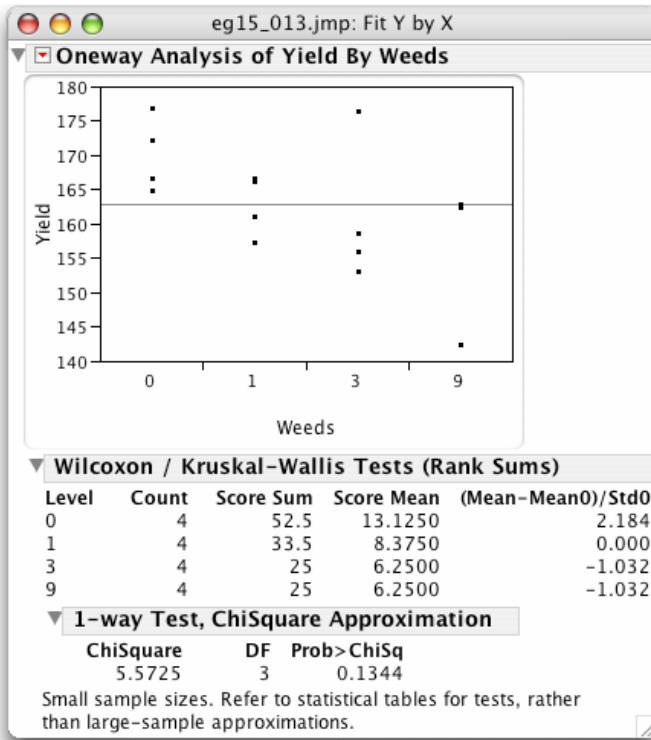
$H_0$ : There is no difference in distribution of yields versus

$H_a$ : Yields are systematically higher in some groups than in others

Because outliers are present, we prefer to compare the medians rather than the means of the distributions. Instead of using the ANOVA *F* test, we must use the *Kruskal-Wallis test*.

1. Open the *JMP* data table **eg15\_013.jmp**.
2. If the column **Weeds** is numeric, as is likely the case, change its modeling type to “Nominal.” (See step 2 in Example 15.1 above for details.)
3. Select **Analyze** ⇒ **Fit Y by X**.
  - a. Select **Yield** and press **Y, Response**.
  - b. Select **Weeds** and press **X, Factor** and **OK**.
  - c. Click on the red triangle and select **Nonparametric** ⇒ **Wilcoxon Test**.

The resulting report is shown on the next page. The ranks sums for each of the four groups are given in the column **Score Sum**. The *Kruskal-Wallis statistic* (labeled **ChiSquare**) equals 5.5725 and the *P*-value equals 0.1344. *JMP* makes a small adjustment for the presence of ties that accounts for the slightly larger value of the statistic. The adjustment makes the associated *P*-value more accurate. Since the *P*-value is not small, we cannot conclude that the presence of the weed, lamb's-quarter, affects corn yield.



## 15.4 Summary

All graphs and statistical computations in this chapter are performed in the **Distribution** and **Fit Y by X** platforms of the **Analyze** menu.

Graph/Computation	Command
One-sample test and matched pairs	
Wilcoxon signed rank test	<u>Distribution</u> ⇒ <u>Test Mean</u>
Two samples	
Wilcoxon rank sum test	<u>Fit Y by X</u> ⇒ <u>Nonparametric</u> ⇒ <u>Wilcoxon Test</u>
Three or more samples	
Kruskal-Wallis Test	<u>Fit Y by X</u> ⇒ <u>Nonparametric</u> ⇒ <u>Wilcoxon Test</u>



# Chapter 16

## Logistic Regression

The linear regression methods studied in earlier chapters modeled the relationship between a *quantitative* response variable and one or more explanatory variables. This chapter discusses similar methods for use when the response variable is *binary*, i.e., has only two possible values.

### 16.1 The Odds Ratio for Two Samples

In Section 2 of Chapter 8, we compared the distribution of a categorical variable in two populations by examining the difference between *proportions* in the two samples. Another way to analyze the data is to examine the ratio of the *odds* in the two samples. In *JMP*, computing the ratio of the odds of two samples is easy. We use the **Odds Ratio** command on the same analysis platform that we used in Chapter 8.2, **Fit Y by X**.

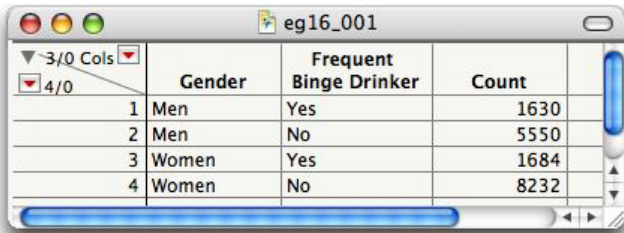
#### Example 16.1 Binge drinking and gender

Are men and women college students equally likely to be frequent binge drinkers? Example 8.2 of this manual discusses a survey of over 17,000 students in U.S. four-year colleges on drinking behavior. A student who reports drinking five or more drinks in a row three or more times in the past two weeks is called a “frequent binge drinker” (FBD). Here is the data on frequent binge drinking by gender:

Gender	Sample size	Number of frequent binge drinkers	Sample proportion
Men	7,180	1,630	0.2270
Women	9,916	1,684	0.1698

Find the odds ratio of being a frequent binge drinker (FBD) for men to women.

1. First create the *JMP* data table **eg16\_001.jmp** below for this data by following steps similar to those in Section 8.2.



	Gender	Frequent Binge Drinker	Count
1	Men	Yes	1630
2	Men	No	5550
3	Women	Yes	1684
4	Women	No	8232

Then, since *JMP* computes the odds of the first level of a categorical value that occurs, change the order of the values of FBD using the column property, **Value Ordering**, so that “Yes” precedes “No.”

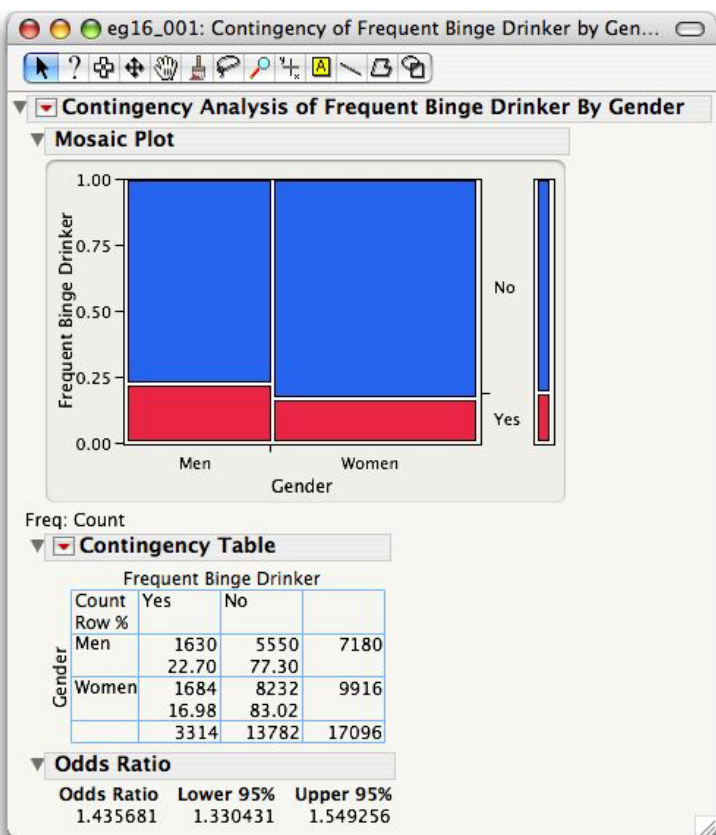
2. Select the column **Frequent Binge Drinker** and press **Cols** ⇒ **Column Info**.
  - a. Press **Column Properties** and select **Value Ordering**.
  - b. Select **Yes** and press **Move Up**.
  - c. Press **OK**.

Now we use the **Fit Y by X** platform to compute the odds ratio of being an FBD for men to women.

3. Select **Analyze** ⇒ **Fit Y by X**.
  - a. Select **Frequent Binge Drinker**, and press **Y, Response**.
  - b. Select **Gender**, and press **X, Factor**.
  - c. Select **Count** and press **Freq**.
  - d. Press **OK**.

The report is on the next page. To display the odds ratio for the contingency table,

4. Press the red triangle menu at the top of the report and select **Odds Ratio**.



Thus, the odds ratio of being an FBD (men to women) is 1.436. So the odds of being an FBD for men are 1.44 times the odds for women.

## 16.2 Inference for Logistic Regression

We use the **Fit Model** platform in *JMP* for statistical inference for logistic regression models.

### Example 16.1 Binge drinking and gender (cont'd.)

To investigate whether frequent binge drinking is related to gender in general, we fit a logistic regression model to the data.

1. Select **Analyze** ⇒ **Fit Model**.
  - a. Select **Frequent Binge Drinker** from the list of columns and press **Y**.
  - b. Select **Gender** and press **Add**.
  - c. Select **Count** and press **Freq**.
  - d. Press **Run Model**.

eg16\_001: Fit Nominal Logistic

**Nominal Logistic Fit for Frequent Binge Drinker**

Iteration History

Freq: Count

**Whole Model Test**

Model	-LogLikelihood	DF	ChiSquare	Prob>ChiSq
Difference	43.1979	1	86.39577	<.0001*
Full	8363.8050			
Reduced	8407.0029			
RSquare (U)	0.0051			
Observations (or Sum Wgts)	17096			

Converged by Gradient

**Parameter Estimates**

Term	Estimate	Std Error	ChiSquare	Prob>ChiSq	Lower 95%	Upper 95%
Intercept	-1.4060375	0.0194228	5240.5	0.0000*	-1.4441055	-1.3679695
Gender[Men]	0.18081958	0.0194228	86.67	<.0001*	0.14275162	0.21888755

For log odds of Yes/No

From the Parameter Estimates report, notice that the parameter estimates are given as  $b_0 = 1.4060$  and  $b_1 = 0.1808$ . These are calculated for the log(odds) of Yes/No (FBD/not an FBD) since we changed the order of the values of the variable **Frequent Binge Drinker** (see step 2 preceding). It's important to notice, though, that *JMP* codes the explanatory variable, or factor, differently than the textbook, using +1 and -1 instead of 1 and 0. Thus, the fitted equation is:

$$\log(\text{odds of being an FBD}) = -1.4060 + 0.1808 X$$

where, in *JMP*,  $X = +1$  for men, and  $X = -1$  for women.

As a result, the intercept is no longer the estimate of the log(odds) for women and the estimate of the slope is no longer the difference between the log(odds) for men and the log(odds) for women. The estimate of the slope is half of that difference; that is, half of the *log* of the odds ratio.

From the equation, we see that the log(odds of being an FBD) for men is  $-1.2252$  ( $= -1.4060 + 0.1808$ ) and for women are  $-1.5868$  ( $= -1.4060 - 0.1808$ ). Thus, the odds of being an FBD are  $e^{-1.2252} = .2937$  for men and  $e^{-1.5868} = .2046$  for women. Thus, the odds ratio (men to woman) is 1.436. So, the odds of being a frequent binge drinker for men are 1.44 times the odds for women. This agrees with the value that we computed in Section 1 of this chapter.

To obtain a 95% confidence interval for the odds ratio, first obtain a 95% confidence interval for the slope,  $\beta_1$ .

2. Press the red triangle, select **Confidence Intervals**, enter 1 minus the confidence level and press **OK**.

The 95% confidence interval for  $\beta_1$ , displayed in the figure above, is (0.14275, 0.21889). Since  $\beta_1$  is half of the log odds ratio, the odds ratio is  $e$  raised to the  $2\beta_1$  power. Therefore, the 95% confidence interval for the odds ratio is from 1.33 ( $= e^{2(0.14275)}$ ) to 1.55 ( $= e^{2(0.21889)}$ ). Compare this to the confidence interval computed in Section 1.

To test the hypothesis  $H_0: \beta_1 = 0$  versus  $H_a: \beta_1 \neq 0$  using the *Wald Chi-Square statistic*,

3. Press the red triangle and select **Wald Tests**.

Source	Nparm	DF	Wald ChiSquare	Prob>ChiSq
Gender	1	1	86.6698132	<.0001*

The value of the *Wald Chi-Square statistic*  $X^2 = 86.67$  and the *P-value* is less than 0.0001. Thus, there is strong evidence that the percentages of men and women who are frequent binge drinkers differ.

## 16.3 Multiple Logistic Regression

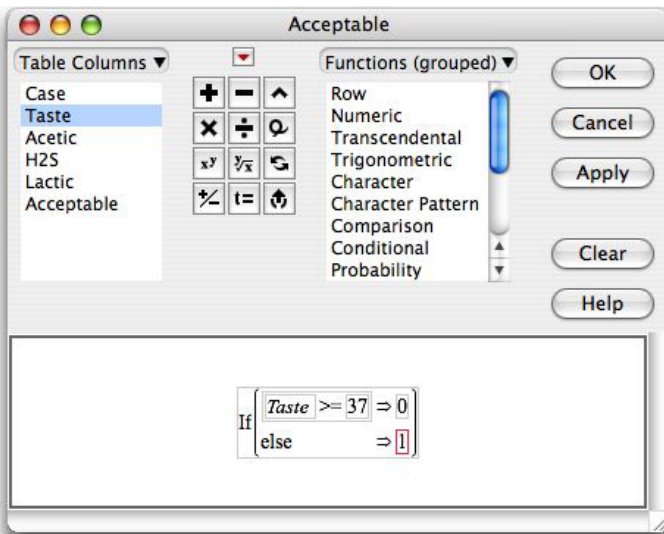
In multiple logistic regression, the response variable again has only two possible values, but there can be several explanatory variables. The **Fit Model** platform is used for analysis. A likelihood ratio Chi-square test and the *Wald Chi-square tests* play roles similar to that of the *F* and *t tests*, respectively, in multiple linear regression.

### Example 16.2 What makes cheddar cheese tasty?

As cheddar cheese matures, a variety of chemical processes take place. The taste of matured cheese is related to the concentration of several chemicals in the final product. In an Australian study, samples of cheddar cheese were analyzed for their chemical composition and were subjected to taste tests. For this example, the cheese is classified as acceptable if the variable **Taste**  $\geq 37$ , and unacceptable otherwise. We wish to predict the odds that the cheese is acceptable using the three variables **Acetic**, **H2S**, and **Lactic**. Data for this study is contained in the *JMP* data table **eg16\_002.jmp**.

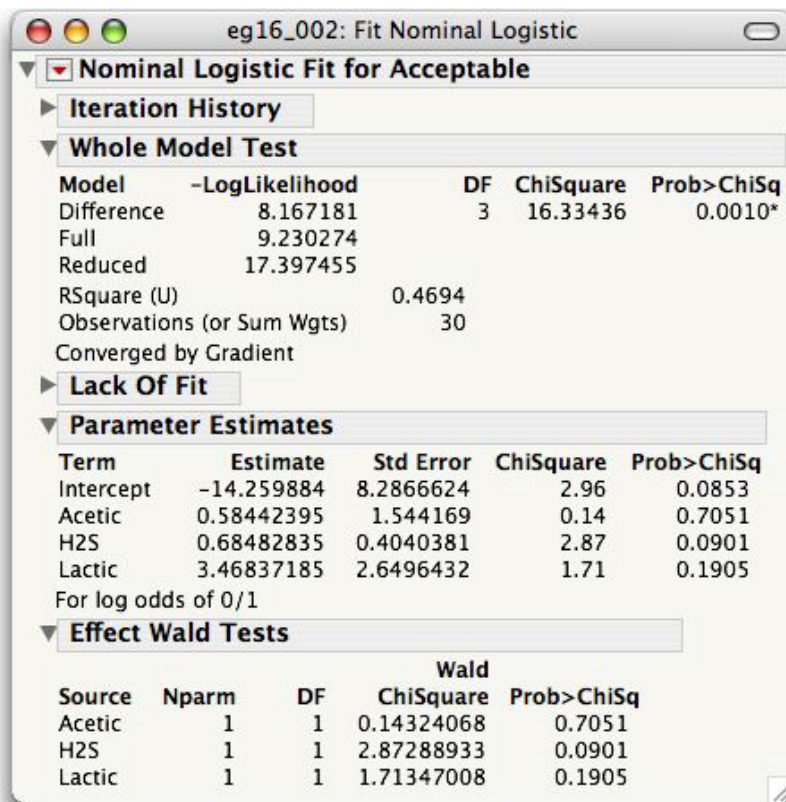
Create the variable **Acceptable** with the values 1 and 0. Since the **Fit Model** platform fits the log odds of the smallest numeric value of the response variable (zero in this case), we assign **Acceptable** = 0 when **Taste**  $\geq 37$ . You can enter the values directly or use the Formula Editor (see Section 0.3.3 in Chapter 0 and following).

1. Open the *JMP* data table **eg16\_002.jmp**.
2. Select **Cols**  $\Rightarrow$  **New Column**.
  - a. Enter **Acceptable** in the **Column Name** field.
  - b. Select **Nominal** from the menu.
  - c. Press **New Property**  $\Rightarrow$  **Formula**.
  - d. Enter the following formula and press **OK**.



To fit the model, we use the **Fit Model** platform and include all the explanatory variables as model effects.

3. Select **Analyze** ⇒ **Fit Model**.
  - a. Select **Acceptable** from the list of columns and press **Y**.
  - b. Select **Acetic**, **H2S**, and **Lactic**, and press **Add**.
  - c. Press **Run Model**.



From the **Parameter Estimates** report, we see that the fitted model is:

$$\log(\text{Odds of acceptable cheese}) = -14.26 + 0.58 \text{ Acetic} + 0.68 \text{ H2S} + 3.47 \text{ Lactic}$$

Examine the **Whole Model Test** report. The hypothesis that at least one of the logistic regression coefficients is not zero is tested by the likelihood ratio statistic (found in the row **Difference** under **ChiSquare** and equal to 16.33). The  $P$ -value (**Prob>ChiSq**) is 0.0010. We can thus conclude that at least one of the chemicals can be used to predict the odds that the cheese is acceptable.

To examine the coefficients of each variable and the tests that each is zero when the other variables are in the model:

4. Press the red triangle and select **Wald Tests**.

The test statistics in the column **Wald ChiSquare** of the **Effect Wald Tests** report are the statistics given in the textbook. The  $P$ -values are 0.7051, 0.0901, and 0.1905. None are statistically significant at the .05 level. Thus, if two of the three variables are in the model, the third is of little to no value. Each of the two variable models should be examined further.

## 16.4 Summary

<b>Activity</b>	<b>Command</b>
Odds ratio for two samples	<u>Analyze</u> ⇒ <u>Fit Y by X</u>
Inference about the model	<u>Analyze</u> ⇒ <u>Fit Model</u>
Odds ratios	<u>Analyze</u> ⇒ <u>Fit Model</u> ⇒ <u>Odds Ratios</u>
Wald Chi-square test	<u>Analyze</u> ⇒ <u>Fit Model</u> ⇒ <u>Wald Tests</u>

# Chapter 17

## Statistics for Quality

Statistical process control uses a combination of graphical and numeric techniques to decide when a process has become unstable and requires intervention. This chapter discusses the use of *JMP* to obtain the most common control charts,  $\bar{x}$  charts,  $s$  charts, and  $p$  charts, and process capability indexes,  $C_p$  and  $C_{pk}$ . Construction of control charts is easily accomplished in *JMP* using the **Control Chart** graph platform. Capability indexes are easily calculated using the **Capability Analysis** command in the **Distribution** platform.

### 17.1 Statistical Process Control

For quantitative variables, the most common control charts are  $\bar{x}$  and  $s$  charts. In *JMP*, control charts are constructed using the **Control Chart** platform on the **Graph** menu. Data must be in the usual form with one column holding the values of the variable under study and another column identifying the samples to which they belong.

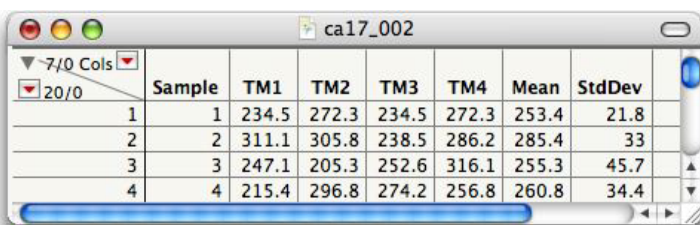
#### Example 17.1 Manufacturing computer monitors

---

A manufacturer of computer monitors must control the tension on the mesh of fine vertical wires that lies behind the surface of the viewing area. The manufacturing process has been stable with mean tension  $\mu = 275$  mV and process standard deviation  $\sigma = 43$  mV. We want to watch the process and maintain its current condition.

The operator measures the tension on a sample of 4 monitors each hour. Suppose that the *JMP* data table **ca17\_002.jmp** contains data for the last 20 samples. We want to construct  $\bar{x}$  and  $s$  control charts to monitor the process. Take a look at the data table **ca17\_002.jmp**.

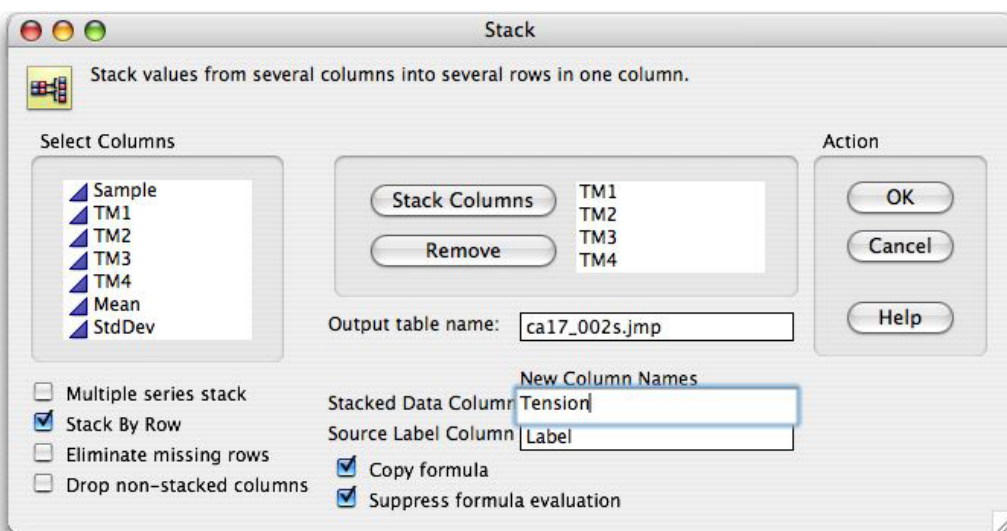




	Sample	TM1	TM2	TM3	TM4	Mean	StdDev
1	1	234.5	272.3	234.5	272.3	253.4	21.8
2	2	311.1	305.8	238.5	286.2	285.4	33
3	3	247.1	205.3	252.6	316.1	255.3	45.7
4	4	215.4	296.8	274.2	256.8	260.8	34.4

The individual values for the tension on the mesh are given in four separate columns, **TM1**, ..., **TM4**. This is a common method of storing small data sets. However, in order to analyze that data we first need to rearrange the data table so that all tension measurements fall in one column (which we will call **Tension**) and each is identified by the sample to which it belongs. We use the **Stack** command in the **Tables** menu to do that.

1. Select **Tables** ⇒ **Stack**.
  - a. Select **TM1**, **TM2**, **TM3**, and **TM4** and press **Stack Columns**.
  - b. Type **ca17\_002s.jmp** for the **Output table name**.
  - c. Type **Tension** for the **Stacked Data Column**.
  - d. Press **Stack**.



The mean and standard deviation of each sample are superfluous as *JMP* will compute them as needed. We can remove them.

2. Select columns **Mean** and **StdDev**.
  - a. Press **Cols** ⇒ **Delete Columns**.

	Sample	Label	Tension
1	1	TM1	234.5
2	1	TM2	272.3
3	1	TM3	234.5
4	1	TM4	272.3
5	2	TM1	311.1
6	2	TM2	305.8
7	2	TM3	238.5
8	2	TM4	286.2
9	3	TM1	247.1
10	3	TM2	205.3

Now, we can construct the control charts:

5. Select **Graph** ⇒ **Control Chart** ⇒ **XBar**.
  - a. Select **Tension** and press **Process**.
  - b. Select **Sample** and press **Sample Label**.
  - c. Under the default **Chart Type**, deselect **R** and select **S**.
  - d. Press the **Specify Stats** button and enter **43** for **Sigma** and **275** for **Mean (measure)**.
  - e. Press **OK**.

**Control Chart**

**XBar Control Chart**

**Select Columns**

- Sample
- Label
- Tension

**Cast Columns into Roles**

Process: Tension (optional numeric)

Sample Label: Sample

Phase: optional

By: optional

**Parameters**

☒ XBar

☐ R

☒ S

**Known Statistics for XBar Chart**

Tension

Sigma: 43

Mean(measure): 275

Mean(range): .

Mean(std dev): .

**Action**

OK

Cancel

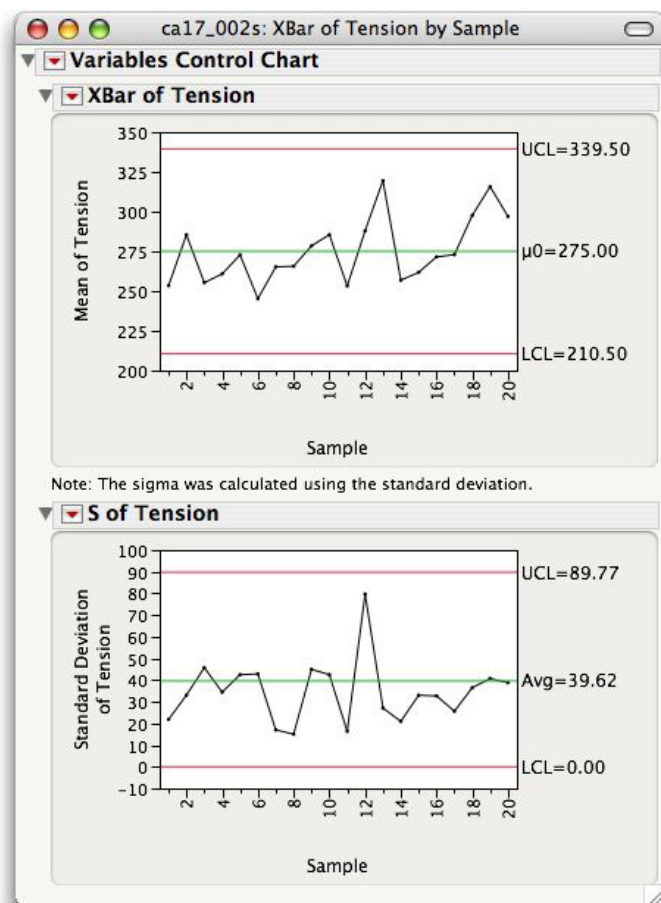
Remove

Recall

Help

Get Limits

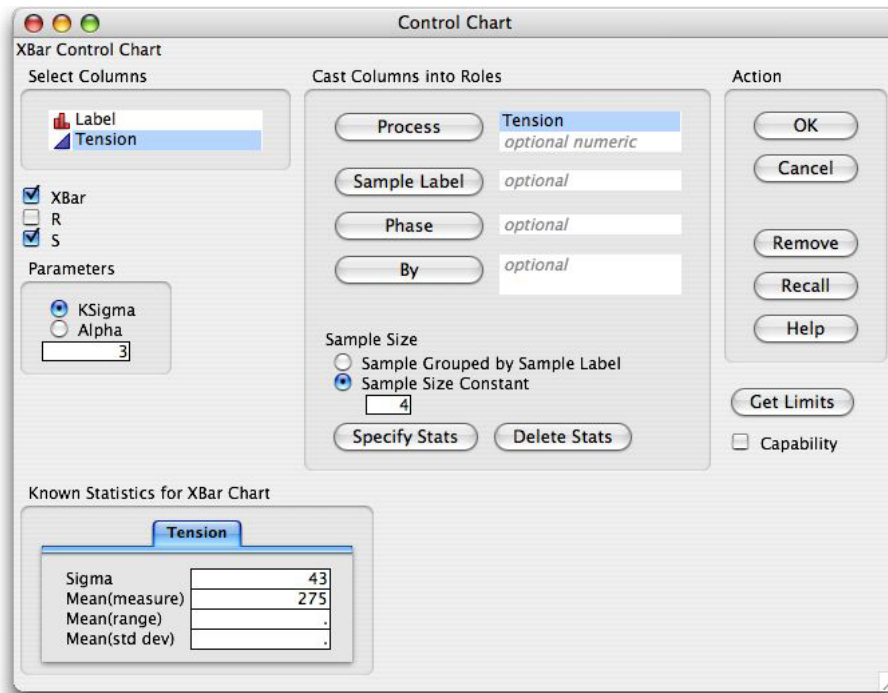
☐ Capability



Both the  $\bar{x}$  chart and the  $s$  chart indicate that the process is in control.

## Remarks

- Often, there is not a variable like **Sample** in the original data table that identifies each sample. If that is the case but each sample contains the same number of data values, we can use the **Sample Size Constant** field at the bottom of the dialog box to specify the size of each sample. For the mesh tension data, which contains 4 values in each sample, the completed **Control Chart** dialog box would look like:



- If the process mean and standard deviation are not known, *JMP* estimates them from the data. Thus, we no longer select “Specify Stats” in the **Control Chart** dialog box.
- *JMP* bases the limits for the  $\bar{x}$  chart on the standard deviation when an *s* chart is chosen in the same report as the  $\bar{x}$  chart. Otherwise, the limits are based on the range.

## 17.2 Process Capability Indexes

Process capability indexes are measures of the ability of a process in control to produce products that fall within certain specified limits. Two important capability indexes are  $C_p$  and  $C_{pk}$ . Capability indexes are easily found using the *JMP* command **Capability Analysis** in the **Distribution** platform.

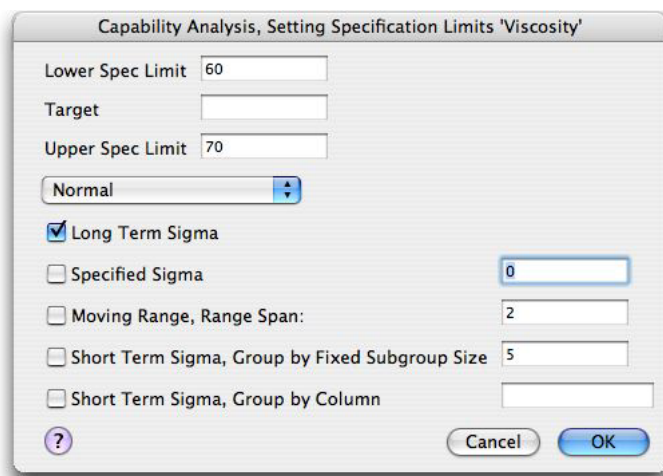
### Exercise 17.2 Elastomer viscosity process capability

Viscosity is a critical characteristic of rubber and rubber-like compounds called elastomers. A specialty chemical company is beginning production of an elastomer that is supposed to have average viscosity of 65 Mooneys (the unit of measurement of viscosity). A major customer wants the specifications for viscosity of this elastomer to be  $LSL = 60$  and  $USL = 70$ . Estimate  $C_p$  and  $C_{pk}$  for this process.

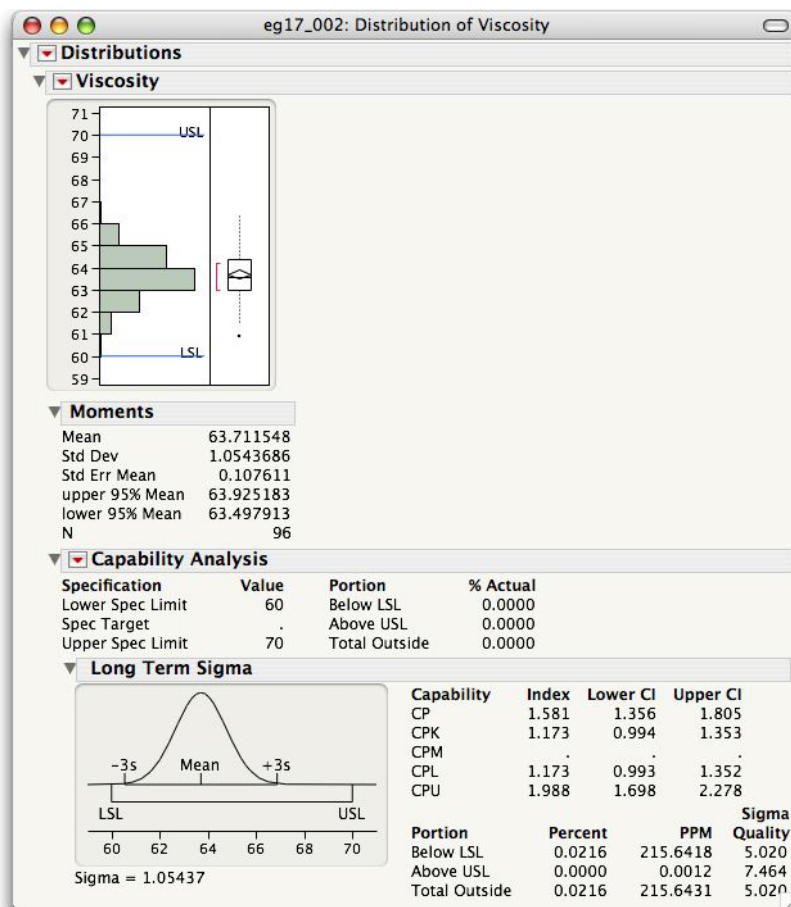
The mean and standard deviation of the distribution of the elastomer’s viscosity are estimated from past data. The *JMP* data table **eg17\_002.jmp** contains data on samples of 4 lots from 24 shifts of production. Open the data table and proceed as follows.

1. Select **Analyze** ⇒ **Distribution**.
  - a. Select the column **Viscosity** and press **Y, Columns**.
  - b. Press **OK**.

2. Press the red triangle on the **Viscosity** title bar and select **Capability Analysis**.
  - a. Type 60 for the **Lower Spec Limit** and 70 for the **Upper Spec Limit**.
  - b. Press **OK**.



*JMP* uses the standard deviation  $s$  of all the data for the **Long Term Sigma**.



$C_p$  is 1.581 while  $C_{pk}$  is 1.173.

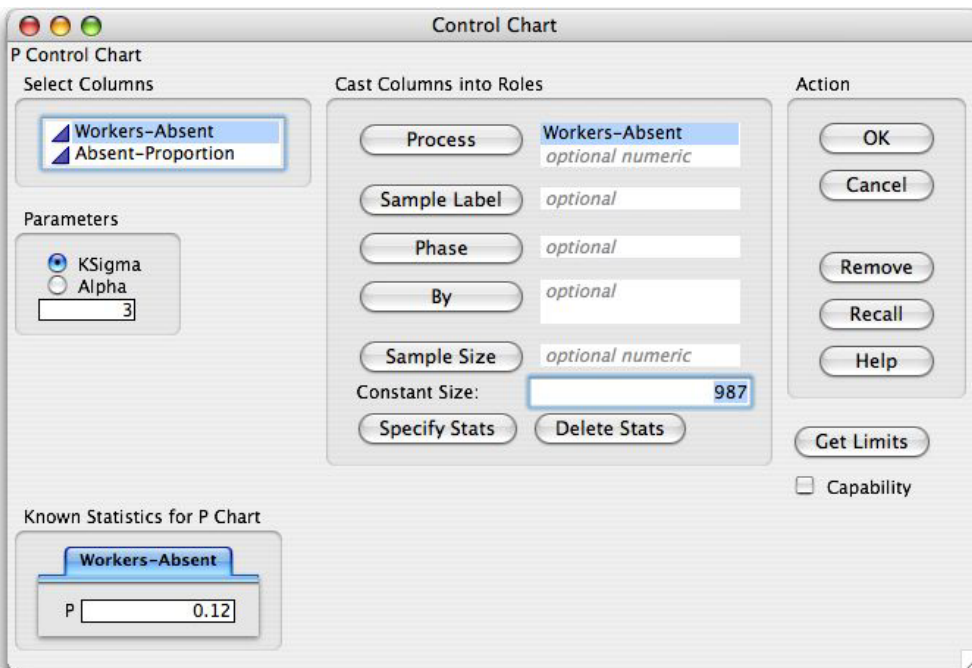
## 17.3 Control Charts for Sample Proportions

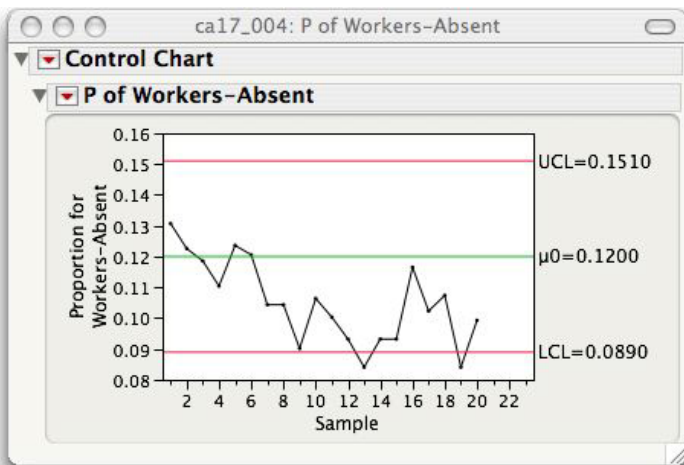
When the variable of interest in a process is categorical, we calculate a sample proportion for each sample taken and plot the sample proportions against the order in which the samples were taken. These charts are called  $p$  charts when control limits are added.

### Example 17.3 Reducing the absenteeism rate

Suppose that you have been asked to improve absenteeism in a production facility where 12% of the workers are now absent on a typical day. You do a thorough background study to identify factors that increase absenteeism and enact some changes. To see whether your actions have been effective you collect data on absenteeism for the next four weeks on 987 production workers. Construct a  $p$  chart of the absenteeism rate. Suppose that the 20 days of data are stored in the *JMP* data table **ca17\_004.jmp**.

1. Select **Graph** ⇒ **Control Chart** ⇒ **P**.
  - a. Select the column **Workers-Absent** and press **Process**.
  - b. Type **987** in the **Constant Size** field to specify the sample size.
  - c. Press the **Specify Stats** button and enter **0.12** for **P**.
  - d. Press **OK**.





## 17.5 Summary

### Activity

$\bar{x}$  and  $s$  charts  
 $p$  charts  
 Capability indexes

### Command

Control Charts  $\Rightarrow$  XBar

Control Charts  $\Rightarrow$  P

Analyze  $\Rightarrow$  Distribution  $\Rightarrow$  ...  $\Rightarrow$  Capability Analysis

# Chapter 18

## Time Series Forecasting

### 18.1 Trends and Seasons

This chapter studies time series. In the first section, regression methods are applied to identify trends and seasonal patterns, seasonally adjusted time series are calculated and plotted, and autocorrelation is discussed and computed. In the second section, models for time series that do not require the often unrealistic independence assumption are discussed. They include the first-order autoregression AR(1) model, moving average models, and the simple exponential smoothing model.

To identify and plot trends and seasonal patterns in time series, we use the **Fit Y by X** analysis platform. For modeling of time series, we use the **Time Series** platform.

#### 18.1.1 Trend

##### Linear Trends

##### Example 18.1 Monthly retail sales

---

The U.S. Census Bureau tracks retail sales using the Monthly Retail Trade Survey. Suppose that data for monthly retail stores of general merchandise stores like Wal-Mart for the period January 1992 through May 2002 is stored in the *JMP* data table **eg18\_001.jmp**. Create a plot of the time series to identify interesting characteristics of retail sales and fit a line to the trend in the data.

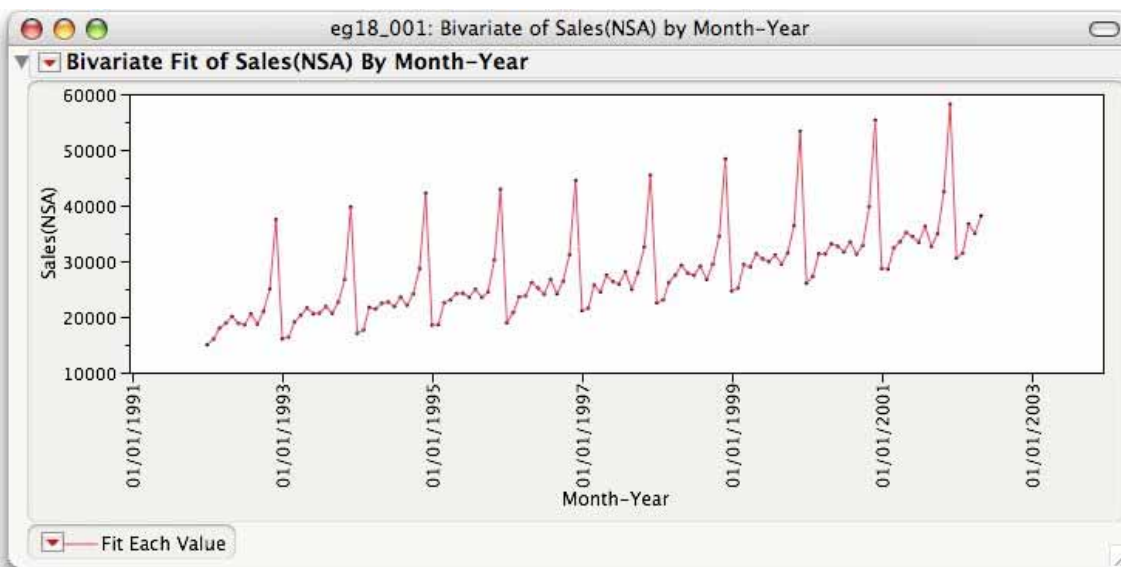
Open the *JMP* data table **eg18\_001.jmp**. We use the **Fit Y by X** platform to plot the data and fit the trend line.

1. Select **Analyze** ⇒ **Fit Y by X**.
  - a. Select **Sales (NSA)** and press **Y, Response**.
  - b. Select **Month-Year**, press **X, Factor** and **OK**.



Expand the time axis by dragging the right edge of the plot to the right. To connect the data points,

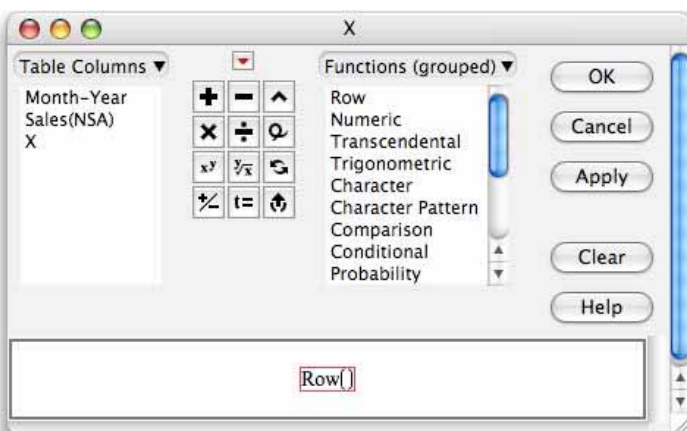
2. Press the red triangle pop-up menu and select **Fit Each Value**.



There is a positive linear trend: overall sales have gradually increased. Also, notice the distinct pattern that repeats itself approximately every 12 months. To model the trend, we use the ideas of linear regression.

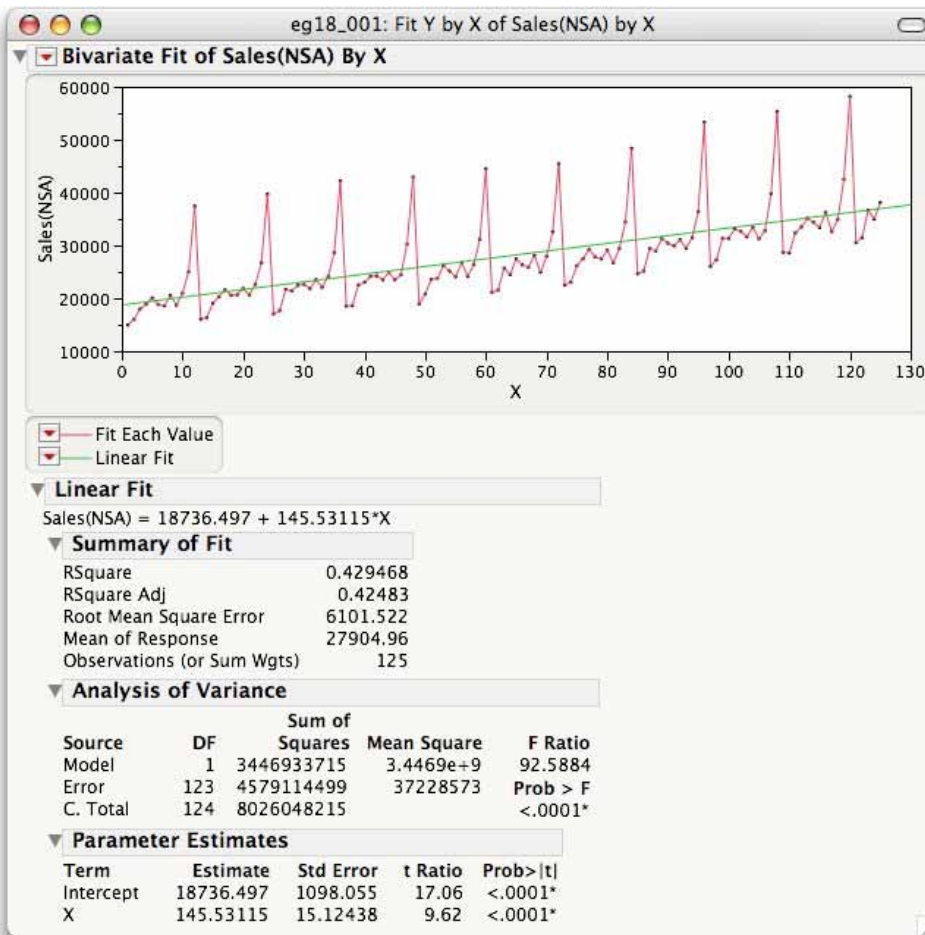
Often the actual dates of a time series are replaced with the number of time points elapsed beginning with the first time point. We create a column  $X$  with values  $x = 1$  corresponding to January 1992,  $x = 2$  corresponding to February 1992, etc.

3. Select **Cols**  $\Rightarrow$  **New Column**.
  - a. Name the column  **$x$** .
  - b. Select **Column Properties**  $\Rightarrow$  **Formula**.
  - c. Select **Row**  $\Rightarrow$  **Row** from the **Functions** list.
  - d. Press **OK** and **OK**.



Now, use the **Fit Y by X** platform as above with the column **X** in place of **Month-Year** and add the command **Fit Line**.

4. Select **Analyze** ⇒ **Fit Y by X**.
  - a. Select **Sales (NSA)** and press **Y, Response**.
  - b. Select **X**, press **X, Factor** and **OK**.
  - c. Press the red triangle pop-up menu and select **Fit Each Value**.
  - d. Press the red triangle next to **Bivariate Fit ...** and select **Fit Line**.



We estimate the linear trend to be  $\text{Sales}^{\wedge} = 18,736.5 + 145.5 X$ .

## Exponential trend

Time series that exhibit exponential trend can be easily plotted and an exponential curve fit using the **Fit Y by X** platform in *JMP*.

To fit an exponential trend model, recall the relationship between the logarithm function and the exponential function. If  $y = ae^{bx}$ , then  $\ln y = \ln a + bx$ .

The latter model is linear with response variable  $\log(y)$ , explanatory variable  $x$ , intercept  $\log(c)$ , and slope  $d$ . In order to fit an exponential model, we use *JMP* to fit a linear model to  $\log(y)$  and  $x$ . The statistics  $d$  and  $c$  of the exponential model are the slope  $b$  of the fitted linear model, and  $a$ , where  $a$  is the intercept of the fitted linear model, respectively. Although this may sound complicated, *JMP* simplifies the process with the **Fit Special** command.

### Example 18.2 DVD player sales

---

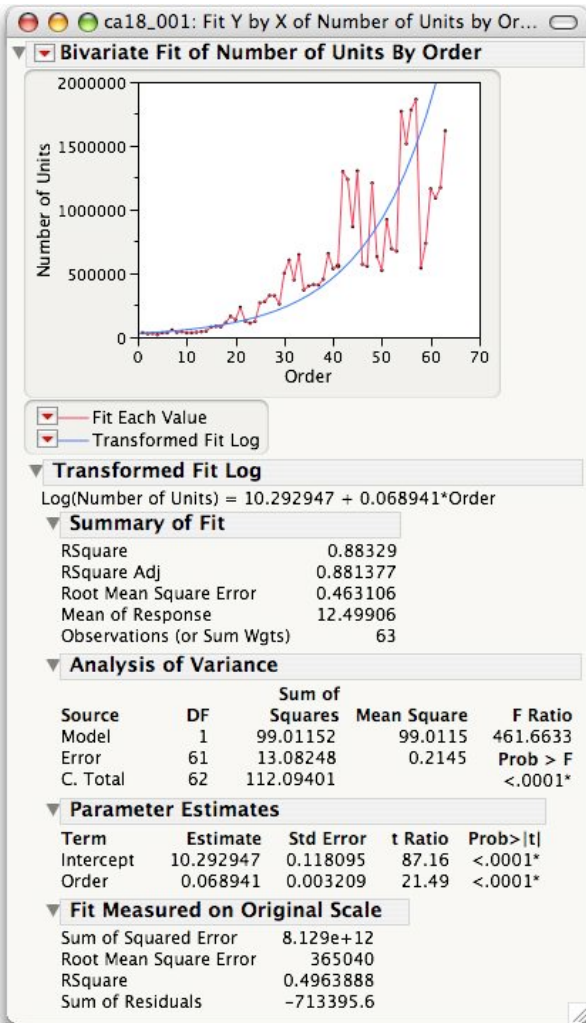
Like many consumer products, sales of DVD players exploded when they were first introduced. The Consumer Electronics Association tracks monthly sales of DVD players. Suppose that the data from April 1997 until June 2002 is stored in a *JMP* data table **ca18\_001.jmp**.

Open the *JMP* data table **ca18\_001.jmp**. We use the **Fit Y by X** platform as we did for a linear trend. For simplicity, assume that a column **Order** indexing the time values is in the data table. As before,

1. Select **Analyze**  $\Rightarrow$  **Fit Y by X**.
  - a. Select **Number of Units** and press **Y, Response**.
  - b. Select **Order**, press **X, Factor** and **OK**.
  - c. Press the red triangle pop-up menu and select **Fit Each Value**.

To fit the *exponential* model, use the **Fit Special** command in place of the **Fit Line** command.

2. Press the red triangle next to **Bivariate Fit ...** and select **Fit Special**.
  - a. Select **Natural logarithm: log (y)**
  - b. Press **OK**.



On the log scale, the relationship is  $\log(\text{Number of Units})^{\wedge} = 10.292947 + 0.068941 \text{ Order}$ . Since  $e^{10.292947} = 29,523.65$ , this means that the equation of the exponential trend on the original scale is:

$$\text{Number of Units}^{\wedge} = 29,532.65e^{0.689417 \text{ Order}}.$$

## 18.1.2 Seasonal Patterns

Time series usually have some form of repeating pattern over time that represents seasonal variability. Seasons are frequently months or quarters. Economists and business leaders need to estimate the size of these seasonal effects to improve the accuracy of their forecasts. Two approaches are frequently used to incorporate seasonality into a trend model—seasonal effects (using indicator variables) and seasonality factors.

### Using indicator variables to obtain seasonal effects

In this approach, we improve our predictions by using indicator variables to add the seasonal pattern in a time series to the trend model. The coefficients of the indicator variables in a multiple regression model estimate the size of the seasonal effects.

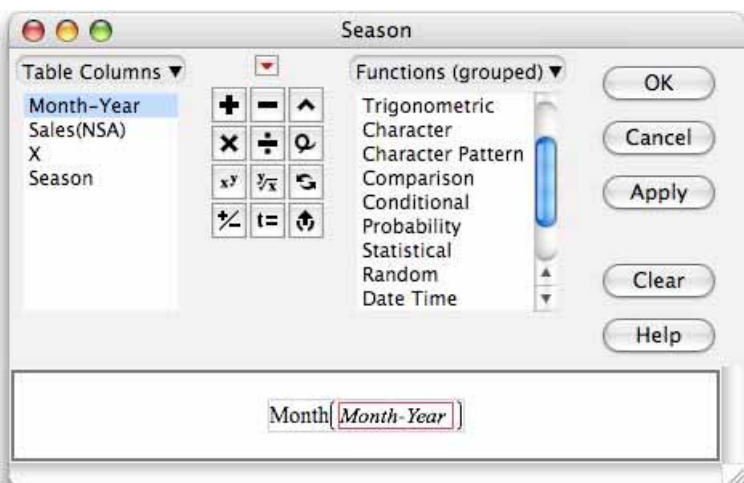
### Example 18.3 Monthly retail sales: seasonal effects

We use the monthly sales data for general merchandise stores discussed earlier to illustrate the use of indicator variables to obtain a *trend-plus-season* model.

Open the *JMP* data table **eg18\_001.jmp**. We use the **Fit Y by X** platform to plot the data and fit the trend line.

First, we create a column **Season** that identifies which month the sales are for. Then, we create 11 indicator variables to indicate the month.

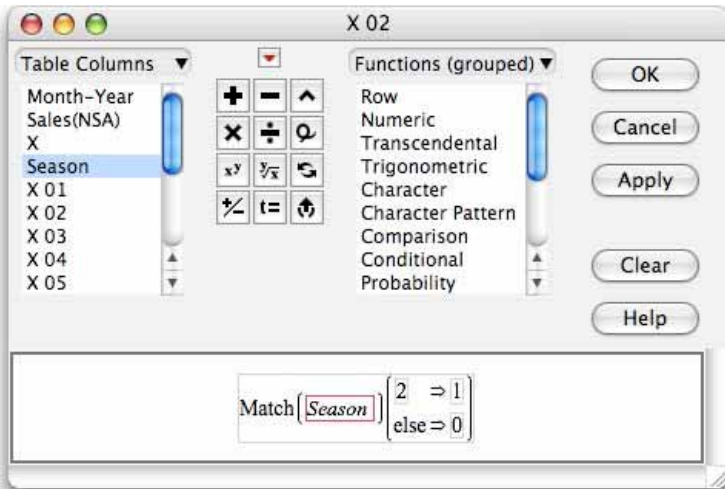
1. Select **Cols** ⇒ **New Column**.
  - a. Name the column **Season**.
  - b. Select **Column Properties** ⇒ **Formula**.
  - c. Select **Date Time** ⇒ **Month** from the **Functions** list.
  - d. Select the column **Month-Year**.
  - e. Press **OK** and **OK**.



2. Select **Cols** ⇒ **Add Multiple Columns**.
  - a. Enter **X** for the Column prefix.
  - b. Enter **11** for the number of columns to add.
  - c. Press **OK**.

Use formulas to define the values of each of the indicator variables. Select all 11 indicator variables and bring up the formula editor. We illustrate the details for the indicator variable for February, **X 02**.

3. Press **Cols** ⇒ **Formula**.
  - a. Select **Conditional** ⇒ **Match** from the list of functions.
  - b. Select the column **Season**.
  - c. Enter **2** for February in the **value** field of the formula and enter **1** in the **then clause** field.
  - d. Press the insert **^** button and enter **0** in the **else clause** field.
  - e. Press **OK**.



Now fit a multiple regression model that includes a linear trend effect and an effect for each month.

4. Select **Analyze** ⇒ **Fit Model**.
  - a. Select **Sales (NSA)** ⇒ **Y**.
  - b. Select the variables **X, X 01 through X 11** ⇒ **Add**.
  - c. Press the **Run Model** button.

eg18\_001s: Fit Least Squares

Response Sales(NSA)

**Summary of Fit**

RSquare	0.987029
RSquare Adj	0.985639
Root Mean Square Error	964.1134
Mean of Response	27904.96
Observations (or Sum Wgts)	125

**Analysis of Variance**

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	12	7921942577	660161881	710.2222
Error	112	104105638	929514.62	Prob > F
C. Total	124	8026048215		<.0001*

**Parameter Estimates**

Term	Estimate	Std Error	t Ratio	Prob> t
Intercept	37472.69	343.3506	109.14	<.0001*
X	140.13045	2.392702	58.57	<.0001*
X 01	-24276.19	421.4213	-57.61	<.0001*
X 02	-23748.69	421.3602	-56.36	<.0001*
X 03	-20271.18	421.3126	-48.11	<.0001*
X 04	-20250.49	421.2786	-48.07	<.0001*
X 05	-18517.99	421.2583	-43.96	<.0001*
X 06	-19574.52	431.4036	-45.37	<.0001*
X 07	-20323.65	431.3306	-47.12	<.0001*
X 08	-18626.88	431.2708	-43.19	<.0001*
X 09	-20878.11	431.2244	-48.42	<.0001*
X 10	-18933.04	431.1912	-43.91	<.0001*
X 11	-13842.17	431.1713	-32.10	<.0001*

This is the trend-plus-season model. Predicted values from it can be saved and plotted along with the original data against X, the time index. One should also examine the residuals to evaluate the effectiveness of this model.

### Using seasonality factors

In the earlier approach to incorporating seasonality into the model, we *added* seasonal effects. Another way to account for seasonality is to calculate adjustment *factors* for each season. To estimate these seasonality factors, we obtain the predicted values from the trend model, divide the observed values by the predicted values, and then, for each season, calculate the mean of these ratios.

Each of these steps is easy to accomplish in *JMP*. We obtain the predicted values from the trend line using the **Fit Y by X** platform, create a column that contains the ratio of the observed to predicted values, and calculate and save the means of these ratios using **Summary** command in the **Tables** menu.

#### Example 18.4 Monthly retail sales: seasonality factors

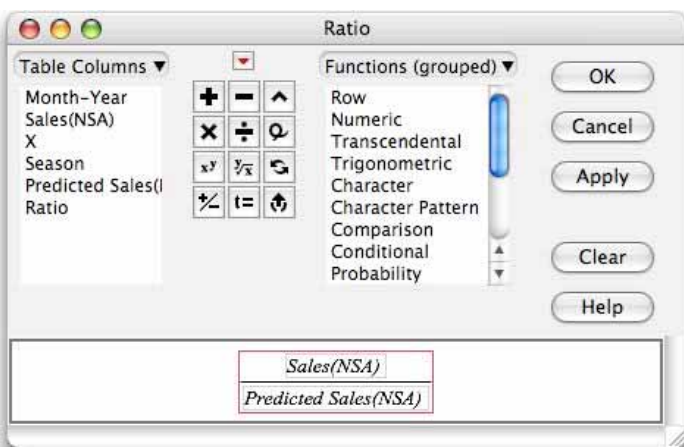
We use the monthly sales data for general merchandise stores discussed earlier to illustrate the use of indicator variables to obtain a *trend-plus-season* model.

Use the *JMP* data table **eg18\_001.jmp** with the variable **Season**, which identifies the month of the year in this example. To obtain the predicted values from the trend-only model, we use the **Fit Y by X** platform.

1. Select **Analyze** ⇒ **Fit Y by X**.
  - a. Select **Sales (NSA)** and press **Y, Response**.
  - b. Select **X**, press **X, Factor** and **OK**.
  - c. Press the red triangle next to **Bivariate Fit ...** and select **Fit Line**.
  - d. Press the red triangle next to **Linear Fit** and select **Save Predicteds**.

Now, we can create a column with the ratios of the observed to predicted values.

2. Select **Cols** ⇒ **New Column**.
  - a. Name the column **Ratio**.
  - b. Select **Column Properties** ⇒ **Formula**.
  - c. Select the column **Sales(NSA)**, press ÷ (divide), and select the column **Predicted Sales(NSA)**.
  - d. Press **OK** and **OK**.

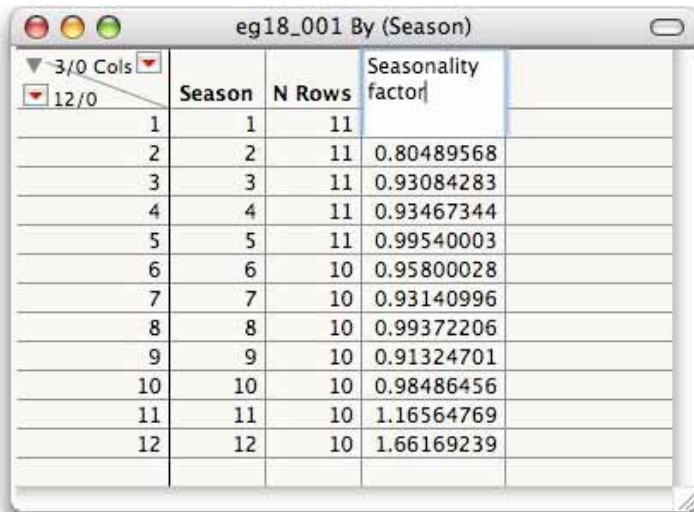


Calculate and save the means of the ratios using the **Summary** command in the **Tables** menu.



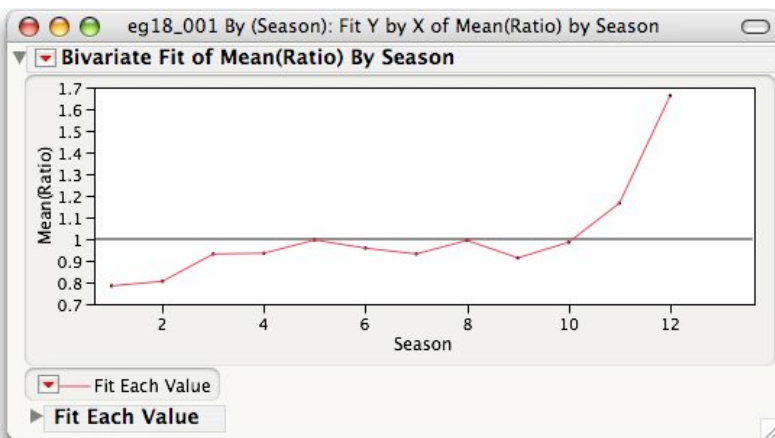
3. Select **Tables** ⇒ **Summary**.
  - a. Select **Ratio** and press **Statistics** ⇒ **Mean**.
  - b. Select **Season** and press **Group** and **OK**.

A JMP data table **eg18\_001 By (Season)** containing the mean ratios for each month is created. These means are the seasonality factors and we rename the **Mean(Ratio)** column by double-clicking on it.



	Season	N Rows	Seasonality factor
1	1	11	
2	2	11	0.80489568
3	3	11	0.93084283
4	4	11	0.93467344
5	5	11	0.99540003
6	6	10	0.95800028
7	7	10	0.93140996
8	8	10	0.99372206
9	9	10	0.91324701
10	10	10	0.98486456
11	11	10	1.16564769
12	12	10	1.66169239

Notice that the seasonality factor for December is 1.66; sales in December are typically 66% above the annual average. We can plot the seasonality factors versus month using **Fit Y by X**.



### Seasonally adjusted data

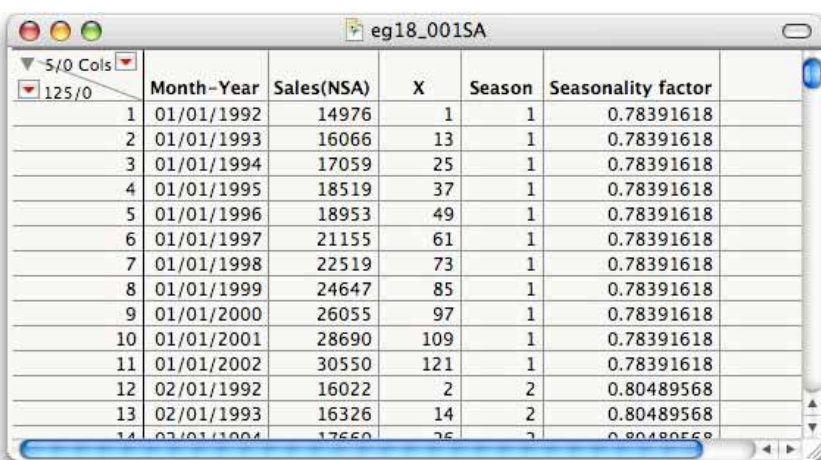
Most economic time series are seasonally adjusted to make the overall trend in the numbers more apparent. A seasonally adjusted time series has had each value divided by the seasonality factor corresponding to the appropriate season.



To calculate seasonally adjusted monthly retail sales, we simply divide each actual sales value by the seasonality factor corresponding to the appropriate month. The first step, then, is to put the seasonality factors in the original data table in such a way that each time value has the appropriate seasonality factor.

With the data tables **eg18\_001.jmp** and **eg18\_001 By (Season)** both open,

1. Select **Tables** ⇒ **Join**.
  - a. Select **eg18\_001 By (Season)** from Join 'eg18\_001' with.
  - b. Select **By Matching Columns** from **Matching Specification**.
    - i. Select **Season** from the **Source Columns** of both tables and press **Match**.
  - c. Select **Select columns for joined table** from **Output Columns**.
    - i. Select **Month-Year, Sales(NSA), X, and Season** in the list for **eg18\_001**.
    - ii. Select **Seasonality factor** in the list for **eg18\_001 By (Season)**.
  - d. Press **Select** and **OK**.

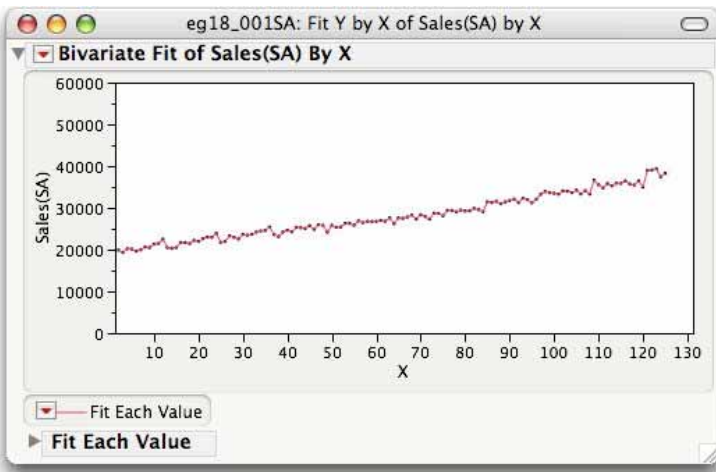


	Month-Year	Sales(NSA)	X	Season	Seasonality factor
1	01/01/1992	14976	1	1	0.78391618
2	01/01/1993	16066	13	1	0.78391618
3	01/01/1994	17059	25	1	0.78391618
4	01/01/1995	18519	37	1	0.78391618
5	01/01/1996	18953	49	1	0.78391618
6	01/01/1997	21155	61	1	0.78391618
7	01/01/1998	22519	73	1	0.78391618
8	01/01/1999	24647	85	1	0.78391618
9	01/01/2000	26055	97	1	0.78391618
10	01/01/2001	28690	109	1	0.78391618
11	01/01/2002	30550	121	1	0.78391618
12	02/01/1992	16022	2	2	0.80489568
13	02/01/1993	16326	14	2	0.80489568
14	02/01/1994	17660	26	2	0.80489568

Now we create a column with the seasonal adjusted series Sales(SA).

2. Select **Cols** ⇒ **New Column**.
  - a. Name the column Sales(SA).
  - b. Select **Column Properties** ⇒ **Formula**.
  - c. Select the column **Sales(NSA)**, press ÷ (divide), and select the column **Seasonality Factor**.
  - d. Press **OK** and **OK**.

Use the **Fit Y by X** platform with the **Fit Each Value** command to obtain a plot of the seasonally adjusted time series.



### 18.1.3 Autocorrelation

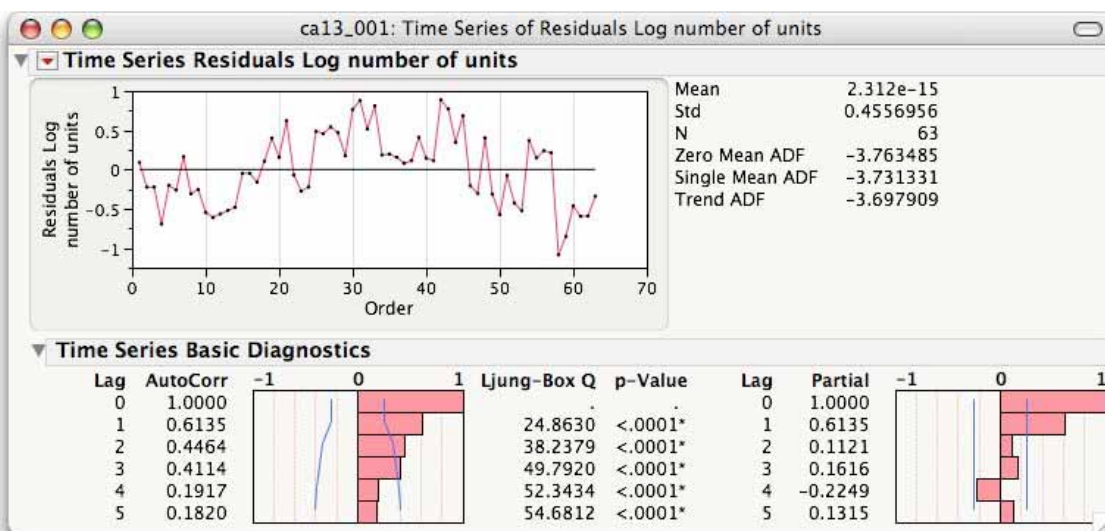
Values in a time series tend to be correlated with one another. It's important to assess the degree of relationships among successive values of a detrended time series. The correlation of a value with the previous one is called first-order autocorrelation or, often simply, autocorrelation. To find autocorrelation using *JMP*, use the **Time Series** analysis platform.

#### Example 18.4 DVD player sales: autocorrelation

Open the *JMP* data table **ca18\_001.jmp** used in Example 18.2 earlier. Assume that an exponential model has been fit and the residuals saved to the data table as **Residuals Log Number of Units**. To obtain the first-order autocorrelation (and more), use the **Time Series** command.

1. Select **Analyze** ⇒ **Modeling** ⇒ **Time Series**.
  - a. Select **Residuals Log Number of Units** and press **Y, Time Series**.
  - b. Select **Order**, press **X, Time ID** and **OK**.

The first order autocorrelation 0.6135 can be found in the **Time Series Basic Diagnostics** report under **AutoCorr** next to **Lag = 1**.



## Remark

- Lagged residual plots can be created using the **Fit Y by X** platform. Simply create a new column that contains the lagged values of the residuals using the **Lag** function which is found in the “Row” group.

## 18.2 Time Series Models

The models of the previous section are not always appropriate. Successive time values often are correlated. In that case, the linear regression model is not appropriate. Instead, we make use of past values of the time series to forecast future values of the series.

### 18.2.1 Autoregressive Models

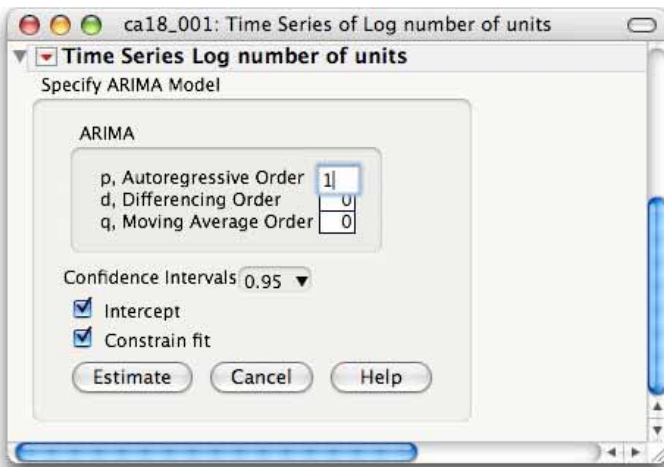
The first-order autoregression model uses only the most immediate previous value to forecast the next. The short-hand for the model is AR(1), and the equation is:

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \epsilon_t$$

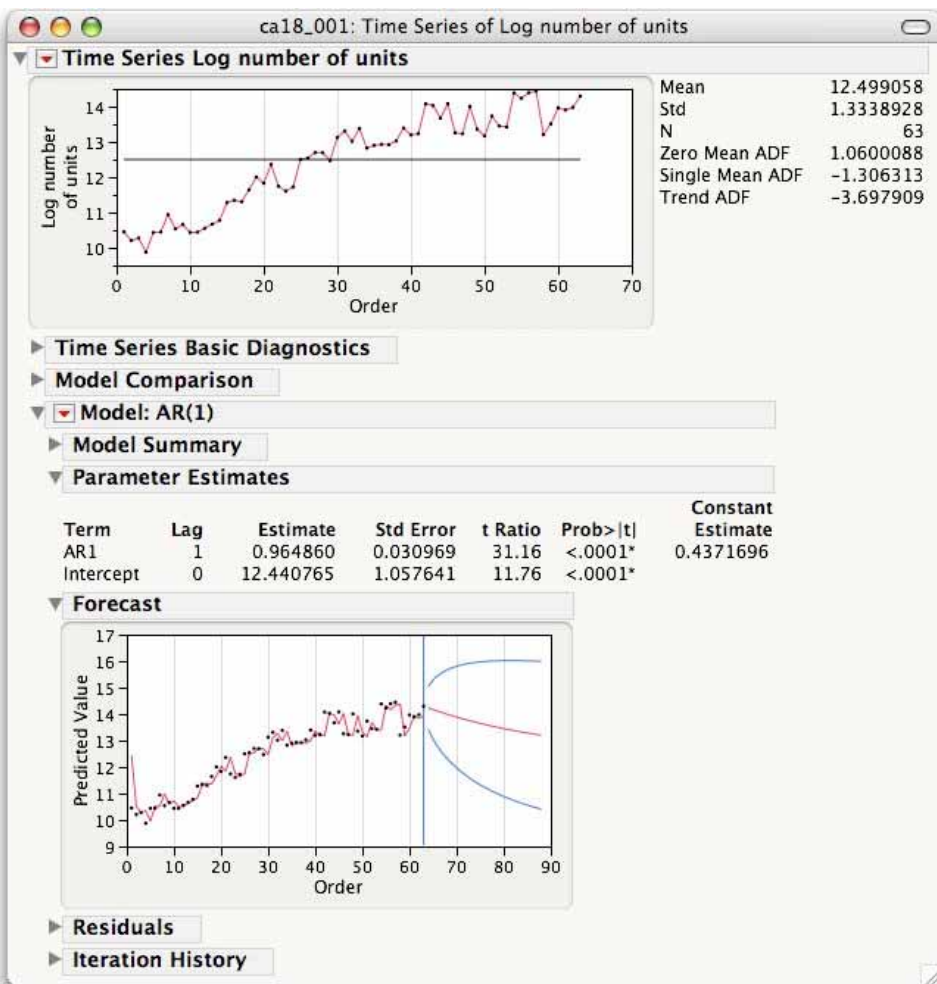
#### Example 18.5 DVD player sales: first-order autoregressive model

In Example 18.4 earlier, we found autocorrelation in the residuals from a regression model of the logarithm of DVD player sales versus time. Let's fit an AR(1) model to this time series. We will use the **Time Series** analysis platform. Open the *JMP* data table **ca18\_001.jmp** again.

- Select **Analyze** ⇒ **Modeling** ⇒ **Time Series**.
  - Select **Log number of units** and press **Y, Times Series**.
  - Select **Order** and press **X, Time ID** and **OK**.
  - Press the red triangle and select **ARIMA**.



2. Enter a 1 for **p**, the **Autoregressive Order**, press **Estimate**.



The **Parameter Estimates** report for the AR(1) model specifies the fitted model. Forecasts of DVD player sales at future time periods are easy to get in *JMP*.

3. Simply select **Save Columns** from the red triangle menu for **Model: AR(1)**.

A *JMP* data table is created that contains forecasts, **Predicted Log number of units**, their standard errors, and 95% prediction interval estimates. These can easily be transformed into forecasts and 95% confidence interval estimates for the raw number of DVD player sales.

	10/0	88/0	Actual Log number of	Predicted Log number of units	Std Err Pred Log number of units	Residual Log number	Upper CL (0.95) Log number of	Lower CL (0.95) Log number of	Predicted Number of	Upper CL (0.95)	Lower CL (0.95)
59			13.5091457	13.1774773	0.40998994	0.3316684	13.9810428	12.3739118	528330	1180021	236549
60			13.9661419	13.4716027	0.40998994	0.49453918	14.2751682	12.6680372	708994	1583532	317438
61			13.9023917	13.91254	0.40998994	-0.0101484	14.7161056	13.1089745	1101893	2461067	493350
62			13.9742086	13.85103	0.40998994	0.1231786	14.6545955	13.0474645	1036158	2314248	463919
63			14.2961437	13.9203233	0.40998994	0.37582048	14.7238888	13.1167578	1110503	2480297	497205
64			*	14.2309456	0.40998994	*	15.0345111	13.4273801	1515030	3383804	678324
65			*	14.1680385	0.56971708	*	15.2846634	13.0514135	1422659	4345552	465754
66			*	14.107342	0.68575413	*	15.4513954	12.7632886	1338877	5133999	349161
67			*	14.0487783	0.77838377	*	15.5743825	12.5231742	1262720	5805884	274628
68			*	13.9922726	0.85565166	*	15.669319	12.3152262	1193347	6384087	223067
69			*	13.9377525	0.92178124	*	15.7444105	12.1310945	1130027	6881936	185553
70			*	13.8851482	0.97933953	*	15.8046184	11.965678	1072120	7309010	157264
71			*	13.8343925	1.03003682	*	15.8532276	11.8155574	1019061	7673072	135342
72			*	13.7854203	1.07508704	*	15.8925522	11.6782884	970358	7980824	117982
73			*	13.738169	1.11539251	*	15.9242982	11.5520399	925574	8238247	103989

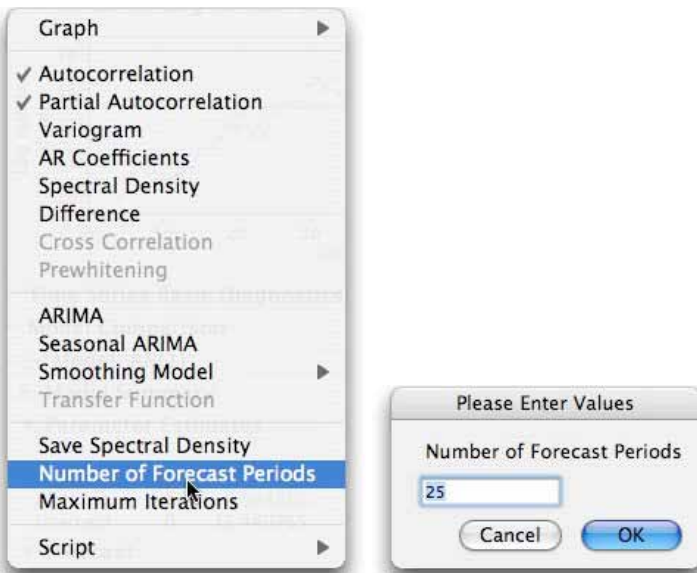
The forecast for July 2002 Log (DVD player sales), the 64th period, is found in the 64 row of the *JMP* data table, 14.2309456, and the forecast for August 2002 is found in the 65th row, 14.1680385. The corresponding forecasts of DVD player sales for July 2002 and August 2002 are 1,515,030 and 1,422,659, respectively. That's roughly 1.5 million and 1.4 million players.

The transcendental function *exp* was used to transform the log data into the raw data. The formula for the **Predicted Number of units** is:

Predicted Number of units
Exp( Predicted Log number of units )

## Remark

- *JMP* saves the forecasts for 25 periods by default. You can modify this with the **Number of Forecast Periods** command on the red triangle menu at the very top of the **Time Series** report.



## 18.2.2 Moving Average Models

*Moving average* models use the average of the last several values of the time series to forecast the next value. To calculate the moving average forecasts, we use the formula editor to create a new column in the original table with the averages.

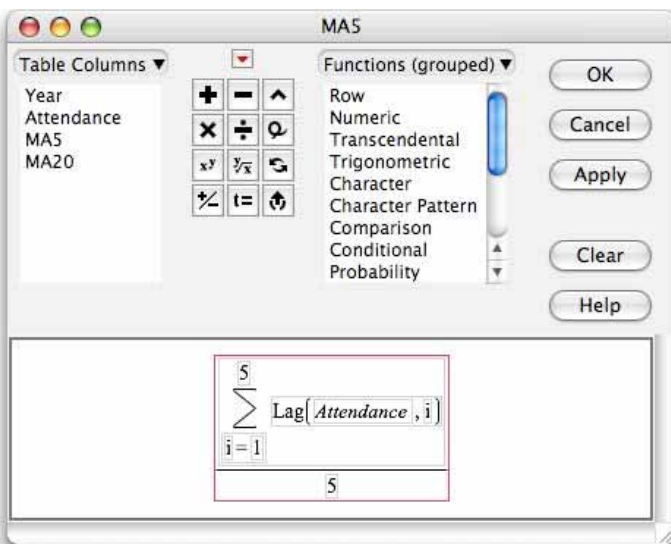
### Example 18.6 Chicago Cubs attendance per game

The Chicago Cubs have been playing their home games at Wrigley Field from 1916 through 2001. Construct the 5-year and 20-year moving average models for attendance per home game at Wrigley Field and use them to forecast attendance in 2002. Suppose that the data are stored in a *JMP* data table **eg18\_006.jmp**.

First, we construct the 5-year moving average series. Open the *JMP* data table **eg18\_006.jmp** and create a new column.

1. Select **Cols** ⇒ **New Column**.
  - a. Name the column **MA5**, and change the format to **Fixed Decimal** with **0 Dec**.
2. Select **New Property** ⇒ **Formula** and enter the formula for the moving average.
  - a. Select **Statistical** ⇒ **Summation** from the **Functions** menu.
  - b. Double-click on **NRow()** and type **5**, the span of this moving average.
  - c. Select the field **Body** and select **Row** ⇒ **Lag** from the **Functions** menu.
  - d. Select **Attendance** from the list of columns.
  - e. Double-click on the number **1** in the formula and type **i** (lower-case letter *i*).
  - f. Select the entire formula, press ÷ (divide) on the **Formula Editor** keypad, and then type **5**.
  - g. Press **OK** and **OK**.





To *forecast* the attendance per home game for 2002 based on the 5-year moving average model, simply add a row to the table.

3. Select **row 87** of the table and enter **2002** in the column **Year**.

	Year	Attendance	MA5
82	1997	27041	28787
83	1998	31990	28944
84	1999	34739	28869
85	2000	34438	29562
86	2001	34314	31121
87	2002	•	32504

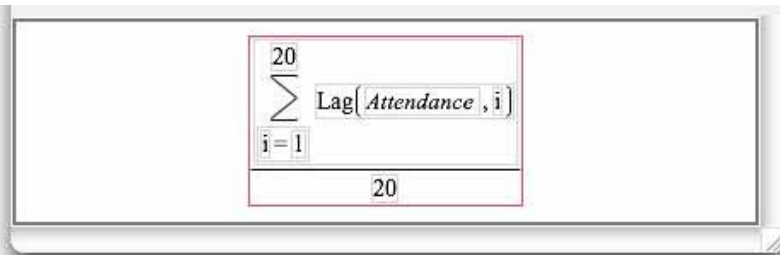
32,504 is the forecast of attendance per game in 2002 based on the 5-year moving average model.

For later time periods, we must first update our time series by replacing the unknown actual attendance values with forecasted ones. Let's obtain the forecast for 2003.

4. a. Select **row 87** of the table and enter 32504 in the **Attendance** column.
- b. Select the next row and enter **2003** in the column **Year**.

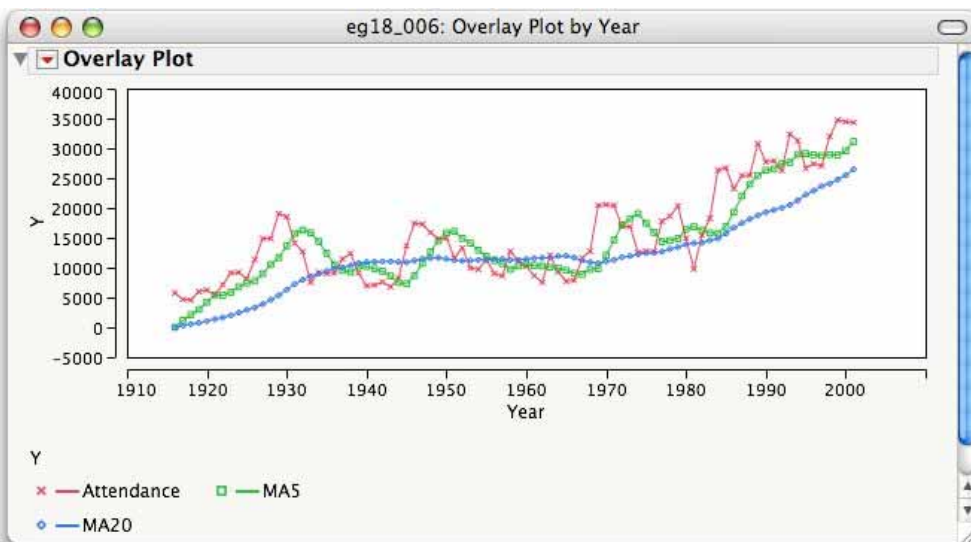
	Year	Attendance	MA5
82	1997	27041	28787
83	1998	31990	28944
84	1999	34739	28869
85	2000	34438	29562
86	2001	34314	31121
87	2002	32504	32504
88	2003	•	33597

Construct the 20-year moving averages by creating a new variable **MA20** in the same way that you did for **MA5**. The formula will look like:



To create a plot containing the original time series and the two moving average series, use the **Overlay Plot** command in the **Graph** platform.

5. Select **Graph** ⇒ **Overlay Plot**.
  - a. Select **Attendance**, **MA5**, and **MA20**, and press **Y**.
  - b. Select **Year** and press **X** and **OK**.
  - c. Select **Y, Options** ⇒ **Connect Points** from the red triangle menu.
  - d. Widen the plot by dragging the right border to the right.



### 18.2.3 Exponential Smoothing Models

The (*simple*) *exponential smoothing* model uses a weighted average of all the values of the time series with most recent time periods receiving the larger weights. The smoothing constant, or weight given to the most recent time period, is a value between 0 and 1. With a smoothing constant close to zero, an exponential smoothing model will follow only major changes in the time series, i.e., the closer to zero the smoothing constant, the smoother the resulting model. The **Time Series** platform in *JMP* can fit a variety of exponential smoothing models.



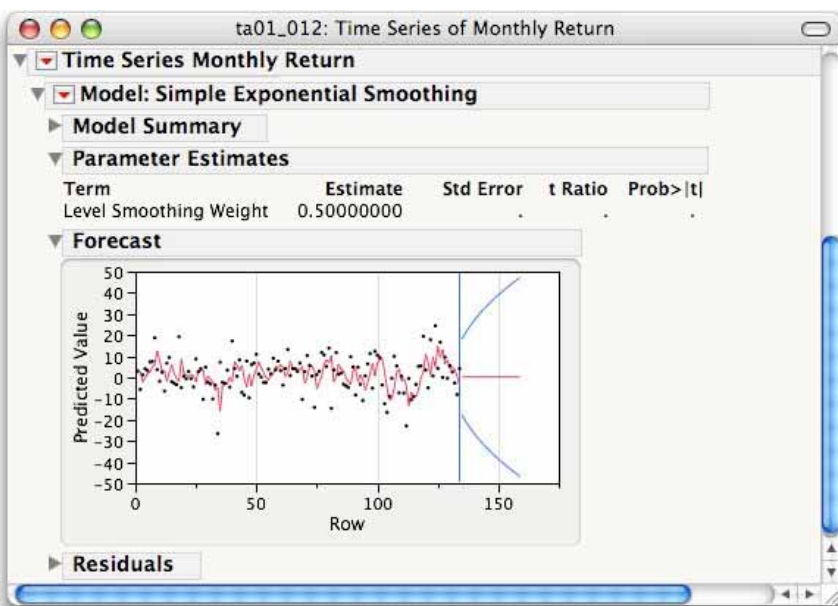
**Example 18.7 Philip Morris returns**

Let's use the exponential smoothing model to forecast the future return on Philip Morris stock. Suppose that the monthly percent returns on Philip Morris stock from June 1990 to July 2001 are contained in a *JMP* data table named **ta1\_012.jmp**. Open the *JMP* data table.

1. Select **Analyze** ⇒ **Modeling** ⇒ **Time Series**.
  - a. Select **Monthly Return** and press **Y, Time Series**, and **OK**.
  - b. Press the red triangle and select **Smoothing Model** ⇒ **Simple Exponential Smoothing**.



2. In the model specification area, select **Custom** from the **Constraints** pull-down menu.
  - a. Select **Fixed** from the **Level** menu.
  - b. Enter a **Weight** of **0.5**.
  - c. Press **Estimate**.



Forecasts for simple exponential smoothing models are constructed in *JMP* using the same command as for autoregressive models, **Save Columns**. Let's have *JMP* compute the forecasted return for next month (August 2001). We change the number of forecast periods to 1 first, though.

3. a. From the **Time Series** report menu, select **Number of Forecast Periods**, enter **1**, and press **OK**.
- b. From the **Simple Exponential Smoothing** red triangle menu, select **Save Columns**.

	Actual Monthly	Row	Predicted Monthly Return	Std Err Pred Monthly Return	Residual Monthly	Upper CL (0.95) Monthly Return	Lower CL (0.95) Monthly Return
132	-2.7	132	3.65191202	9.06481126	-6.351912	21.4186156	-14.114792
133	-8.1	133	0.47595601	9.06481126	-8.575956	18.2426596	-17.290748
134	4.2	134	-3.812022	9.06481126	8.01202199	13.9546816	-21.578726
135	•	135	0.193989	9.06481126	•	17.9606926	-17.572715

A new *JMP* data table containing predicted monthly returns and 95% confidence interval estimates is displayed. The last row of the new data table contains the forecasted value for August 2001 from our model, 0.193989.

## 18.2.4 Spline Fits

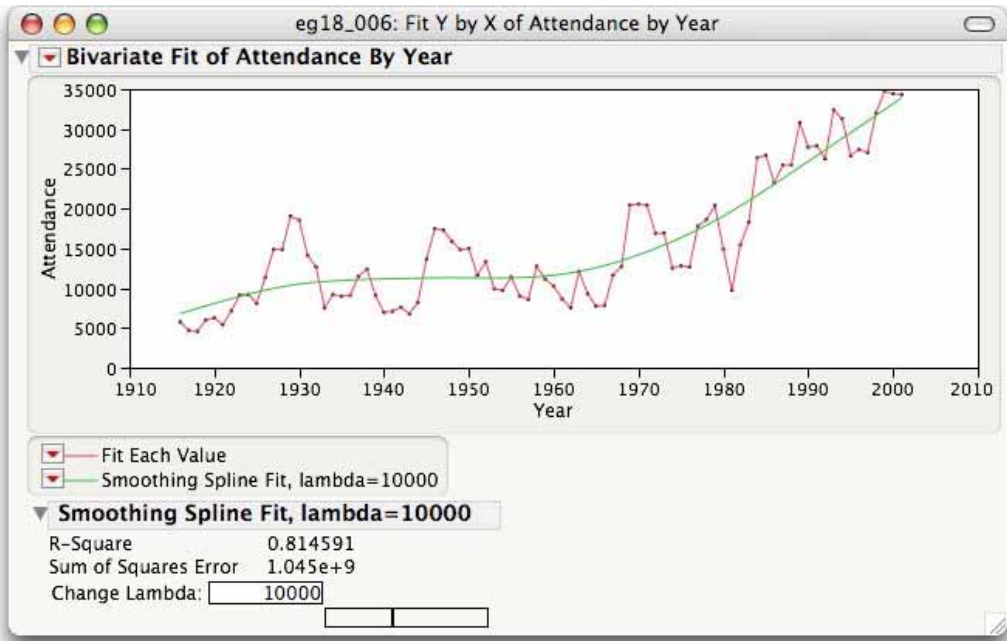
Spline curves can be fit in *JMP* using the **Fit Spline** command in **Fit Y by X** platform.

### Example 18.7 Chicago Cubs attendance per game

We illustrate with the home attendance data from the Chicago Cubs. Open the *JMP* data table with the attendance records, **eg18\_006.jmp**.

1. Select **Analyze** ⇒ **Fit Y by X**.
  - a. Select **Attendance** and press **Y, Response**.
  - b. Select **Year**, press **X, Factor** and **OK**.
  - c. Press the red triangle menu and select **Fit Each Value**.
2. Press the red triangle menu again and select **Fit Spline** ⇒ **10000**.

✓ Show Points	
Fit Mean	
Fit Line	1000000, stiff
Fit Polynomial	▶ 100000
Fit Special...	10000
Fit Spline	▶ 1000
Fit Each Value	100
Fit Orthogonal	▶ 10
Density Ellipse	▶ 1
Nonpar Density	.1
Group By...	.01, flexible
Script	▶ Other...



To see the effect of decreasing the smoothing parameter, lambda, select the **Change Lambda** input field and enter a smaller number like 100.

## 18.3 Summary

Activity	Command
Exploring trends	<u>Analyze</u> ⇒ <u>Fit Y by X</u> ⇒ ... ⇒ <u>Fit Each Value</u>
Linear trends	... ⇒ <u>Fit Line</u>
Exponential trends	... ⇒ <u>Fit Special</u> ⇒ <u>Natural logarithm: log (y)</u>
Exploring seasonality	
Using indicator variables	<u>Analyze</u> ⇒ <u>Fit Model</u>
Using seasonality factors	<u>Tables</u> ⇒ <u>Summary</u>
Seasonally adjusting a time series	<u>Tables</u> ⇒ <u>Join</u>
Autocorrelation	<u>Analyze</u> ⇒ <u>Modeling</u> ⇒ <u>Time Series</u>
Autoregressive models	<u>Analyze</u> ⇒ <u>Modeling</u> ⇒ <u>Time Series</u> ⇒ ... ⇒ <u>ARIMA</u>
Moving Average models	<u>Column</u> ⇒ <u>New Column</u> ⇒ ... ⇒ <u>Formula</u>
Exponential Smoothing	<u>Analyze</u> ⇒ <u>Modeling</u> ⇒ <u>Time Series</u> ⇒ ... ⇒ <u>Smoothing Models</u> ⇒ <u>Simple Exponential Smoothing</u>
Spline Fits	<u>Analyze</u> ⇒ <u>Fit Y by X</u> ⇒ ... ⇒ <u>Fit Each Value</u> ⇒ <u>Fit Spline</u>

## Chapter 1 Exercises

**1.5** Here are the scores on the first exam in an introductory statistics course for 30 students in one section of the course:

80	73	92	85	75	98	93	55	80	90	92	80	87	90	72
65	70	85	83	60	70	90	75	75	58	68	85	78	80	93

Use these data to make a stemplot. Then use the stemplot to describe the distribution of the first-exam scores for this course.

**1.7** Refer to the first exam scores from Exercise 1.5 (reproduced below) and this histogram you produced in Exercise 1.6. Now make a histogram for these data using classes 40 – 59, 60 – 79, and 80 – 100. Compare this histogram with the one that you produced in Exercise 1.6.

80	73	92	85	75	98	93	55	80	90	92	80	87	90	72
65	70	85	83	60	70	90	75	75	58	68	85	78	80	93

**1.19** Email spam is the curse of the Internet. Here is a compilation of the most common types of spam:

Type of spam	Percent
Adult	14.5
Financial	16.2
Health	7.3
Leisure	7.8
Products	21.0
Scams	14.2

Make two bar graphs of these percents, one with bars ordered as in the table (alphabetical), and the other with bars in order from tallest to shortest. Comparisons are easier if you order the bars by height. A bar graph ordered from tallest to shortest is sometimes called a **Pareto chart**, after the Italian economist who recommended this procedure.

**1.23** People with diabetes must monitor and control their blood glucose level. The goal is to maintain “fasting plasma glucose level. The goal is to maintain “fasting plasma glucose” between about 90 and 130 milligrams per deciliter (mg/dl). Here are the fasting plasma glucose levels for 18 diabetics enrolled in a diabetes control class, five months after the end of the class:

141	158	112	153	134	95	96	78	148
172	200	271	103	172	359	145	147	255

Make a stemplot of these data and describe the main features of the distribution. (You will want to trim and also split stems.) Are there outliers? How well is the group as a whole achieving the goal for controlling glucose levels?

**1.29** The *One-Variable Statistical Calculator* applet on the text CD and Web site will make stemplots and histograms. It is intended mainly as a learning tool rather than as a replacement for a statistical software. The histogram function is particularly useful because you can change the number of classes by dragging with the mouse. The tornado damage data from Table 1.5 are available in the applet. Choose this data set and go to the “Histogram” tab.

- Sketch the default histogram that the applet first presents. If the default graph does not have nine classes, drag it to make a histogram with nine classes and sketch the result. This should agree with your histogram in part (b) of the previous exercise.
- Make a histogram with the greatest number of classes that the applet allows. Sketch the results.
- Drag the graph until you find the histogram that you think best pictures the data. How many classes did you choose? Sketch your final histogram.

**1.31** Table 1.7 contains data on the mean annual temperatures (degrees fahrenheit) for the years 1951 to 2000 at two locations in California: Pasadena and Redding. Make time plots of both time series and compare their main features. You can see why discussions of climate change often bring disagreement.

<b>Table 1.7 Mean annual temperatures Fahrenheit in two California cities</b>					
Mean Temperature			Mean Temperature		
Year	Pasadena	Redding	Year	Pasadena	Redding
1951	62.27	62.02	1976	64.23	63.51
1952	61.59	62.27	1977	64.47	63.89
1953	62.64	62.06	1978	64.21	64.05
1954	62.88	61.65	1979	63.76	60.38
1955	61.75	62.48	1980	65.02	60.04
1956	62.93	63.17	1981	65.80	61.95
1957	63.72	62.42	1982	63.50	59.14
1958	65.02	65.04	1983	64.19	60.66
1959	65.69	63.07	1984	66.06	61.72
1960	64.48	63.50	1985	64.44	60.50
1961	64.12	63.97	1986	65.31	61.76
1962	62.82	62.42	1987	64.58	62.94

1963	63.71	63.29	1988	65.22	63.70
1964	62.76	63.29	1989	64.53	61.50
1965	63.03	63.32	1990	64.96	62.22
1966	64.25	64.51	1991	65.60	62.73
1967	64.36	64.21	1992	66.07	63.59
1968	64.15	63.40	1993	65.16	61.55
1969	63.51	63.77	1994	64.63	61.63
1970	64.08	64.30	1995	65.43	62.62
1971	63.59	62.23	1996	65.76	62.93
1972	64.53	63.06	1997	66.72	62.48
1973	63.46	63.75	1998	64.12	60.23
1974	63.93	63.80	1999	64.85	61.88
1975	62.36	62.66	2000	66.25	61.58

**1.57** C-reactive protein (CRP) is a substance that can be measured in the blood. Values increase substantially within 6 hours of an infection and reach a peak within 24 to 48 hours after. In adults, chronically high values have been linked to an increased risk of cardiovascular disease. In a study of apparently healthy children aged 6 to 60 months in Papua, New Guinea, CRP was measured in 90 children. The units are milligrams per liter (mg/l). Here are the data from a random sample of 40 of these children:

0.00	0.00	30.61	46.70	22.82	0.00	5.36	59.76	0.00	20.78
3.90	5.62	0.00	0.00	0.00	0.00	0.00	0.00	0.00	7.10
5.64	3.92	73.20	0.00	0.00	4.81	5.66	15.74	0.00	7.89
8.22	6.81	0.00	26.41	3.49	9.57	0.00	0.00	9.37	5.53

- Find the five-number summary for these data.
- Make a boxplot.
- Make a histogram.
- Write a short summary of the major features of this distribution. Do you prefer the boxplot or the histogram for these data?

**1.59** In the Papua New Guinea study that provided the data for the previous two exercises, the researchers also measured serum retinol. A low value of this variable can be an indicator of vitamin A deficiency. Below are the data on the same sample of 40 children from this study. The units are micromoles per liter ( $\mu\text{mol/l}$ ).

1.15	1.36	.38	.34	.35	.37	1.17	.97	.97	.67
.31	.99	.52	.70	.88	.36	.24	1.00	1.13	.31
1.44	.35	.34	1.90	1.19	.94	.34	.35	.33	.69
.69	1.04	.83	1.11	1.02	.56	.82	1.20	.87	.41

Analyze these data. Use the questions in the previous two exercises as a guide.

**1.61** Figure 1.16 (page 25) is a histogram of the tuition and fees charged by the 56 four-year colleges in the state of Massachusetts. Here are those charges (in dollars), arranged in increasing order:

4,123	4,186	4,324	4,342	4,557	4,884	5,397	6,129
6,963	6,972	8,232	13,584	13,612	15,500	15,934	16,230
16,696	16,700	17,044	17,500	18,550	18,750	19,145	19,300
19,410	19,700	19,700	19,910	20,234	20,400	20,640	20,875
21,165	21,302	22,663	23,550	24,324	25,840	26,965	27,522
27,544	27,904	28,011	28,090	28,420	28,420	28,900	28,906
28,950	29,060	29,338	29,392	29,600	29,624	29,630	29,875

Find the five-number summary and make a boxplot. What distinctive feature of the histogram do these summaries miss? Remember that numerical summaries are not a substitute for looking at the data.

**1.63** Table 1.5 (page 25) shows the average property damage caused by tornadoes over a 50-year period in each of the states. The distribution is strongly skewed to the right.

- Give the five-number summary. Explain why you can see from these five numbers that the distribution is right-skewed.
- A histogram or stemplot suggests that a few states are outliers. Show that there are *no* suspected outliers according to the  $1.5 \times \text{IQR}$  rule. You see once again that a rule is not a substitute for plotting your data.
- Find the mean property damage. Explain why the mean and median differ so greatly for this distribution.

**1.81** How does regular running affect heart rate? The RUNNERS data set, described in detail in the Data Appendix, contains heart rates for four groups of people:

- Sedentary females
- Sedentary males
- Female runners (at least 15 miles per week)
- Male runners (at least 15 miles per week)

The heart rates were measured after 6 minutes of exercise on a treadmill. There are 200 subjects in each group. Give a complete comparison of the four distributions, using both graphs and numerical summaries. How would you describe the effect of running on heart rate? Is the effect different for men and women?

**1.139** The length of human pregnancies from conception to birth varies according to a distribution that is approximately Normal with mean 266 days and standard deviation 16 days.

- (a) What percent of pregnancies last less than 240 days (that's about 8 months)?
- (b) What percent of pregnancies last between 240 and 270 days (roughly between 8 months and 9 months)?
- (c) How long do the longest 20% of pregnancies last?



## Chapter 2 Exercises

**2.7** Here are the data for the second test and the final exam for the same students as in Exercise 2.6:

<b>Second-test score</b>	158	163	144	162	136	158	175	153
<b>Final-exam score</b>	145	140	145	170	145	175	170	160

- Explain why you should use the second-test score as the explanatory variable.
- Make a scatterplot and describe the relationship.
- Why do you think the relationship between the second-test score and the final-exam score is stronger than the relationship between the first-test score and the final-exam score?

**2.15** Often the percent of an animal species in the wild that survive to breed again is lower following a successful breeding season. This is part of nature's self-regulation, tending to keep population size stable. A study of merlins (small falcons) in northern Sweden observed the number of breeding pairs in an isolated area and the percent of males (banded for identification) who returned the next breeding season. Here are data for nine years:

<b>Pairs</b>	28	29	29	29	30	32	33	38	38
<b>Percent</b>	82	83	70	61	69	58	43	50	47

- Why is the response variable the *percent* of males that return rather than the *number* of males that return?
- Make a scatterplot. To emphasize the pattern, also plot the mean response for years with 29 and 38 breeding pairs and draw lines connecting the mean responses for the six values of the explanatory variable.
- Describe the pattern. Do the data support the theory that a smaller percent of birds survive following a successful breeding season?

**2.17** We often describe our emotional reaction to social rejection as “pain.” A clever study asked whether social rejection causes 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. Subjects were first included and then deliberately excluded from social activity while increases in blood flow in their brains were measured. After each activity, the subjects filled out questionnaires that assessed how excluded they felt.

Below 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 activity in the anterior cingulate cortex, a region of the brain that is activated by physical pain.

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			

**2.19** Management theory says that the value of a business should depend on its operating income, the income produced by the business after taxes. (Operating income excludes income from sales of assets and investments, which don't reflect the actual business.) Total revenue, which ignores costs, should be less important. Table 2.1 shows the values, operating incomes, and revenues of an unusual group of businesses: the teams in the National Basketball (NBA). Professional sports teams are generally privately owned, often by very wealthy individuals who may treat their team as a source of prestige rather than a business.

<b>Table 2.1 NBA team businesses</b>			
<b>Team</b>	<b>Value (\$ millions)</b>	<b>Revenue (\$ millions)</b>	<b>Income (\$ millions)</b>
Los Angeles Lakers	447	149	22.8
New York Knicks	401	160	13.5
Chicago Bulls	356	119	49
Dallas Mavericks	338	117	-17.7
Philadelphia 76ers	328	109	2
Boston Celtics	290	97	25.6
Detroit Pistons	284	102	23.5
San Antonio Spurs	283	105	18.5
Phoenix Suns	282	109	21.5
Indiana Pacers	280	94	10.1
Houston Rockets	278	82	15.2
Sacramento Kings	275	102	-16.8
Washington Wizards	274	98	28.5
Portland Trail Blazers	272	97	-85.1
Cleveland Cavaliers	258	72	3.8
Toronto Raptors	249	96	10.6
New Jersey Nets	244	94	-1.6
Utah Jazz	239	85	13.8
Miami Heat	236	91	7.9
Minnesota Timberwolves	230	85	6.9
Memphis Grizzlies	227	63	-19.7

Denver Nuggets	218	75	7.9
New Orleans Hornets	216	80	21.9
Los Angeles Clippers	208	72	15.9
Atlanta Hawks	202	78	-8.4
Orlando Magic	199	80	13.1
Seattle Supersonics	196	70	2.4
Golden State Warriors	188	70	7.8
Milwaukee Bucks	174	70	-15.1

- (a) Plot team value against revenue. There are several outliers. Which teams are these, and in what way are they outliers? Is there a positive association between value and revenue? Is the pattern roughly linear?
- (b) Now plot value against operating income. Are the same teams outliers? Does revenue or operating income better predict the value of an NBA team?

**2.21** Metabolic rate, the rate at which the body consumes energy, is important in studies of weight gain, dieting, and exercise. The table below gives data on the lean body mass and resting metabolic rate for 12 women and 7 men who are subjects in a study of dieting. Lean body mass, given in kilograms, is a person's weight leaving out all fat. Metabolic rate is measured in calories burned per 24 hours, the same calories used to describe the energy content of foods. The researchers believe that lean body mass is an important influence on metabolic rate.

- (a) Make a scatterplot of the data, using different symbols or colors for men and women.
- (b) Is the association between these variables positive or negative? How strong is the relationship? Does the pattern of the relationship differ for women and men? How do the male subjects as a group differ from the female subjects as a group?

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

**2.23** Table 2.3 (reproduced below) shows the progress of world record times (in seconds) for the 10,000 meter run up to mid-2004. Concentrate on the women's world record

times. Make a scatterplot with year as the explanatory variable. Describe the pattern of improvement over time that your plot displays.

<b>Women's Record Times</b>			
1967	2286.4	1982	1895.3
1970	2130.5	1983	1895.0
1975	2100.4	1983	1887.6
1975	2041.4	1984	1873.8
1977	1995.1	1985	1859.4
1979	1972.5	1986	1813.7
1981	1950.8	1993	1771.8
1981	1937.2		

**2.25** Table 2.3 shows the progress of world record times (in seconds) for the 10,000-meter run for both men and women.

<b>Table 2.3</b>					
<b>Men</b>				<b>Women</b>	
<b>Record Year</b>	<b>Time (Seconds)</b>	<b>Record Year</b>	<b>Time (Seconds)</b>	<b>Record Year</b>	<b>Time (Seconds)</b>
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	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
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	2004	1580.3		

- (a) Make a scatterplot of world record time against year, using separate symbols for men and women. Describe the pattern for each sex. Then compare the progress of men and women.

- (b) Women began running this 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?

**2.27** Fidelity Investments, like other large mutual funds companies, offers many “sector funds” that concentrate their investments in narrow segments of the stock market. These funds often rise or fall by much more than the market as a whole. We can group them by broader market sector to compare returns. Here are percent total returns for 23 Fidelity “Select Portfolios” funds for the year 2003, a year in which stocks rose sharply:

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	

- (a) Make a plot of total return against market (space the four market sectors equally on the horizontal axis). Compute the mean return for each sector, add the means to your plot, and connect means with line segments.
- (b) Based on the data, which of these market sectors were the best places to invest in 2003? Hindsight is wonderful.
- (c) Does it make sense to speak of a positive or negative association between market sector and total return?

**2.31** Here are the data for the second test and the final exam for the same students as in Exercise 2.6 (and 2.30):

<b>Second-test score</b>	158	163	144	162	136	158	175	153
<b>Final-exam score</b>	145	140	145	170	145	175	170	160

Find the correlation between these two variables.

**2.33** Examine the data in Exercise 2.31 and add a ninth student who has low scores on the second test and the final exam, and fits the overall pattern of the other scores in the data set. Calculate the correlation and compare it with the correlation that you calculated in Exercise 2.31. Write a short summary of your findings.

**2.39** Table 2.1 (page 98) gives the values of the 29 teams in the National Basketball Association along with their total revenues and operating incomes. You made Scatterplots of value against both explanatory variables in Exercise 2.19.

- (a) Find the correlations of team value with revenue and with operating income. Do you think that the two values of  $r$  provide a good first comparison of what the plots show about predicting value?
- (b) Portland is an outlier in the plot of values against income. How does  $r$  change when you remove Portland? Explain from the position of this point why the change has the direction it does.

**2.45** Table 1.10 (reproduced below) gives the city and highway gas mileage for 21 two-seater cars, including the Honda Insight gas-electric hybrid car.

- (a) Make a scatterplot of highway mileage  $y$  against city mileage  $x$  for all 21 cars. There is a strong positive linear association. The Insight lies far from the other points. Does the Insight extend the linear pattern of the other cars, or is it far from the line they form?
- (b) Find the correlation between city and highway mileages both without and with the Insight. Based on your answer to (a), explain why  $r$  changes in this direction when you add the Insight.

City	Hwy	City	Hwy
17	24	9	13
20	28	15	22
20	28	12	17
17	25	22	28
18	25	16	23
12	20	13	19
11	16	20	26
10	16	20	29
17	23	15	23
60	66	26	32
9	15		

**2.51** Table 1.9 (page 29) reports data on 78 seventh-grade students. We expect a positive association between IQ and GPA. Moreover, some people think that self-concept is related to school performance. Examine in detail the relationships between GPA and the two explanatory variables IQ and self-concept. Are the relationships roughly linear? How strong are they? Are there unusual points? What is the effect of removing these points?

**Table 1.9**

Educational data for 78 seventh-grade students

OBS	GPA	IQ	Gender	Self-concept	OBS	GPA	IQ	Gender	Self-concept
1	7.94	111	M	67	43	10.76	123	M	64
2	8.292	107	M	43	44	9.763	124	M	58
3	4.643	100	M	52	45	9.41	126	M	70
4	7.47	107	M	66	46	9.167	116	M	72
5	8.882	114	F	58	47	9.348	127	M	70
6	7.585	115	M	51	48	8.167	119	M	47
7	7.65	111	M	71	50	3.647	97	M	52
8	2.412	97	M	51	51	3.408	86	F	46
9	6	100	F	49	52	3.936	102	M	66
10	8.833	112	M	51	53	7.167	110	M	67
11	7.47	104	F	35	54	7.647	120	M	63
12	5.528	89	F	54	55	0.53	103	M	53
13	7.167	104	M	54	56	6.173	115	M	67
14	7.571	102	F	64	57	7.295	93	M	61
15	4.7	91	F	56	58	7.295	72	F	54
16	8.167	114	F	69	59	8.938	111	F	60
17	7.822	114	F	55	60	7.882	103	F	60
18	7.598	103	F	65	61	8.353	123	M	63
19	4	106	M	40	62	5.062	79	M	30
20	6.231	105	F	66	63	8.175	119	M	54
21	7.643	113	M	55	64	8.235	110	M	66
22	1.76	109	M	20	65	7.588	110	M	44
24	6.419	108	F	56	68	7.647	107	M	49
26	9.648	113	M	68	69	5.237	74	F	44
27	10.7	130	F	69	71	7.825	105	M	67
28	10.58	128	M	70	72	7.333	112	F	64
29	9.429	128	M	80	74	9.167	105	M	73
30	8	118	M	53	76	7.996	110	M	59
31	9.585	113	M	65	77	8.714	107	F	37
32	9.571	120	F	67	78	7.833	103	F	63
33	8.998	132	F	62	79	4.885	77	M	36
34	8.333	111	F	39	80	7.998	98	F	64
35	8.175	124	M	71	83	3.82	90	M	42
36	8	127	M	59	84	5.936	96	F	28
37	9.333	128	F	60	85	9	112	F	60
38	9.5	136	M	64	86	9.5	112	F	70
39	9.167	106	M	71	87	6.057	114	M	51
40	10.14	118	F	72	88	6.057	93	F	21
41	9.999	119	F	54	89	6.938	106	M	56

**2.59** Here are the data for the second test and the final-exam scores (again).

<b>Second-test score</b>	158	163	144	162	136	158	175	153
<b>Final-exam score</b>	145	140	145	170	145	175	170	160

- Plot the data with the second-test scores on the  $x$  axis and the final-exam scores on the  $y$  axis.
- Find the least-squares regression line for predicting the final-exam score using the second-test score.
- Graph the least-squares regression line on your plot.

**2.61** Examine the data in Exercise 2.31 and add a ninth student who has low scores on the second test and the final exam, and fits the overall pattern of the other scores in the data set. Recalculate the least-squares regression line with this additional case and summarize the effect it has on the least-squares regression line.

**2.67** Exercise 2.17 gives data from a study that shows that social exclusion causes “real pain.” That is, activity in the area of the brain that responds to physical pain goes up as distress from social exclusion goes up. Your scatterplot in Exercise 2.17 shows a moderately strong linear relationship.

- 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.
- On your plot, show the “up and over” lines that predict brain activity for social distress score 2.0. Use the equation of the regression line to get the predicted brain activity level. Verify that it agrees with your plot
- What percent of the variation in brain activity among these subjects is explained by the straight-line relationship with social distress score?

**2.69** Table 2.4 (reproduced below) gives data on the growth of icicles at two rates of water flow. You examined these data in Exercise 2.24. Use least-squares regression to estimate the rate (centimeters per minute) at which icicles grow at these two flow rates. How does flow rate affect growth?

Run 8903				Run 8905			
Time (min)	Length (cm)	Time (min)	Length (cm)	Time (min)	Length (cm)	Time (min)	Length (cm)
10	0.6	130	18.1	10	0.3	130	10.4
20	1.8	140	19.9	20	0.6	140	11.0
30	2.9	150	21.0	30	1.0	150	11.9



40	4.0	160	23.4	40	1.3	160	12.7
50	5.0	170	24.7	50	3.2	170	13.9
60	6.1	180	27.8	60	4.0	180	14.6
70	7.9			70	5.3	190	15.8
80	10.1			80	6.0	200	16.2
90	10.9			90	6.9	210	17.9
100	12.7			100	7.8	220	18.8
110	14.4			110	8.3	230	19.9
120	16.6			120	9.6	240	21.1

**2.73** Compute the mean and the standard deviation of the metabolic rates and lean body masses in Exercise 2.21 (page 98) and the correlation between these two variables. Use these values to find the slope of the regression line of metabolic rate on lean body mass. Also find the slope of the regression line of lean body mass on metabolic rate. What are the limits for each of the two slopes?

**2.81** Unfortunately, the main product of the decay of the pesticide fenthion is fenthion sulfoxide, which is also toxic. Here are the data on the total concentration of fenthion and fenthion sulfoxide in the same specimens of olive oil described in the previous exercise:

Days stored		Concentration			
28	1.03	1.03	.99	.99	.99
84	1.05	1.04	1.00	.99	.99
183	1.03	1.02	1.01	.98	.98
273	1.07	1.06	1.03	1.03	1.02
365	1.06	1.02	1.01	1.01	.99

- Plot the natural logarithm of concentration against days stored. Notice that there are several pairs of identical data points. Does the pattern suggest that the model of simple exponential decay describes the data reasonably well, at least over this interval of time? Explain your answer.
- Regress the logarithm of concentration on time. Use your result to estimate the value of the constant  $k$ .

**2.85** Here are the average monthly temperatures for Chicago Illinois:

Month	1	2	3	4	5	6
Temperature (Fahrenheit)	21.0	25.4	37.2	48.6	58.9	68.6

Month	7	8	9	10	11	12
Temperature (Fahrenheit)	73.2	71.7	64.4	52.8	40.0	26.6

In this table, months are coded as integers, with January corresponding to 1 and December corresponding to 12.

- Plot the data with month on the  $x$  axis and temperature on the  $y$  axis. Describe the relationship
- Find the least-squares regression line and add it to the plot. Does the line give a good fit to the data? Explain your answer.
- Calculate the residuals and plot them versus month. Describe the pattern and explain what the residual plot tells you about the relationship between temperature and month in Chicago.
- Do you think you would find the similar pattern if you plotted the same kind of data for another city.
- Would your answer to part (d) change if the other city was Melbourne, Australia? Explain why or why not.

**2.87** A study of nutrition in developing countries collected data from the Egyptian village of Nahya. Here are the mean weights (in kilograms) for 170 infants in Nahya who were weighed each month during their first year of life:

Age (months)	1	2	3	4	5	6	7	8	9	10	11	12
Weight (kg)	4.3	5.1	5.7	6.3	6.8	7.1	7.2	7.2	7.2	7.2	7.5	7.8

- Plot weight against time.
- A hasty user of statistics enters the data into software and computes the least-squares line without plotting the data. The result is

**The regression equation is**  
**Weight = 4.88 + 0.267 age**

Plot this line on your graph. Is it an acceptable summary of the overall pattern of growth? Remember that you can calculate the least-squares line for *any* set of two-variable data. It's up to you to decide if it makes sense to fit a line.

- Fortunately, the software also prints out the residuals from the least-squares line. In order of age along the rows, they are

-0.85	-0.31	0.02	0.35	0.58	0.62
0.45	0.18	-0.08	-0.35	-0.32	-0.28

Verify that the residuals have sum zero (except for roundoff error). Plot the residuals against age and add a horizontal line at zero. Describe carefully the pattern that you see.

**2.97** Table 1.10 gives the city and highway gas mileages for 21 two-seater cars, including the Honda Insight gas-electric hybrid car. In Exercise 2.45 you investigated the influence of the Insight on the correlation between city and highway mileage.

<b>Fuel economy (miles per gallon) for 2004 model vehicles</b>					
Two-Seater Cars			Minicompact Cars		
Model	City	Highway	Model	City	Highway
Acura NSX	17	24	Aston Marin Vanquish	12	19
Audi TT Roadster	20	28	Audi TT Coupe	21	29
BMW Z4 Roadster	20	28	BMW 325CI	19	27
Cadillac XLR	17	25	BMW 330CI	19	28
Chevrolet Corvette	18	25	BMW M3	16	23
Dodge Viper	12	20	Jaguar XK8	18	26
Ferrari 360 Modena	11	16	Jaguar XKR	16	23
Ferrari Maranello	10	16	Lexus SC 430	18	23
Ford Thunderbird	17	23	Mini Cooper	25	32
Honda Insight	60	66	Mitsubishi Eclipse	23	31
Lamborghini Gallardo	9	15	Mitsubishi Spyder	20	29
Lamborghini Murcielago	9	13	Porsche Cabriolet	18	26
Lotus Esprit	15	22	Porsche Turbo 911	14	22
Maserati Spyder	12	17			
Mazda Miata	22	28			
Mercedes-Benz SL 500	16	23			
Mercedes-Benz SL 600	13	19			
Nissan 350Z	20	26			
Porsche Boxster	20	29			
Porsche Carrera 911	15	23			
Toyota MR2	26	32			

- Make a scatterplot of highway mileage (response) against city mileage (explanatory) for all 21 cars.
- Use software or a graphing calculator to find the regression line for predicting highway mileage from city mileage and also the 21 residuals for this regression. Make a residual plot with a horizontal line at zero. (The “stacks” in the plot are due to the fact that mileage is measured only to the nearest mile per gallon.)
- Which car has the largest positive residual? The largest negative residual?
- The Honda Insight, an extreme outlier, does not have the largest residual in either direction. Why is this not surprising?

**2.111** The Census Bureau provides estimates of numbers of people in the United States classified in various ways. Let's look at college students. The following table gives us data to examine the relation between age and full-time and part-time status. The numbers in the table are expressed as thousands of U.S. college students.

Age	Full-time	Part-time
15-19	3388	389
20-24	5238	1164
25-34	1703	1699
35 and over	762	2045

- What is the U.S. Census Bureau estimate of the number of full-time college students aged 15 to 19?
- Give the joint distribution of age and status for this table.
- What is the marginal distribution of age? Display the results graphically.
- What is the marginal distribution of status? Display the results graphically.

**2.119** A market research firm conducted a survey of companies in its state. They mailed a questionnaire to 300 small companies, 300 medium-sized companies, and 300 large companies. The rate of nonresponse is important in deciding how reliable survey results are. Here are the data on response to this survey.

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

- What is the overall percent of nonresponse?
- Describe how nonresponse is related to the size of business. (Use percents to make your statements precise.)
- Draw a bar graph to compare the nonresponse percents for the three size categories.
- Using the total number of responses as a base, compute the percent of responses that come from each of small, medium, and large businesses.
- The sampling plan was designed to obtain equal numbers of responses from small, medium, and large companies. In preparing an analysis of the survey results, do you think it would be reasonable to proceed as if the responses represented companies of each size equally?

**2.121** Cocaine addiction can be difficult to overcome. Since addicts derive pleasure for the drug, one proposed aid is to provide an antidepressant drug. A 3-year study with 72 chronic cocaine users compared an antidepressant drug called desipramine with lithium and a placebo. (Lithium is a standard drug to treat cocaine addiction. A placebo is a

tablet with no effects that tastes and looks like the antidepressant drug. It is used so the antidepressant drug can be seen.) One-third of the subjects, chosen at random, received each treatment. Here are the results:

	Cocaine Relapse	
	Yes	No
Desipramine	10	14
Lithium	18	6
Placebo	20	4

Compare the effectiveness of the three treatments in preventing relapse. Use percents and draw a bar graph. Write a brief summary of your conclusions.

**2.145** Some statistical methods require that the individuals from a regression line have a Normal distribution. The residuals for the nonexercise activity are given in Exercise 2.83. Is their distribution close to Normal? Make a Normal quantile plot to find out.

**2.159** Mountain View University has professional schools in business and law. Here is a three-way table of applicants to these professional schools, categorized by gender, school, and admission decision.

Business			Law		
	Admit			Admit	
Gender	Yes	No	Gender	Yes	No
Male	400	200	Male	90	110
Female	200	100	Female	200	200

- Make a two-way table of gender by admission decision for the combined professional schools by summing entries in the three-way table.
- From your two-way table, compute separately the percents of male and female applicants admitted. Male applicants are admitted to Mountain View's professional schools at a higher rate than female applicants.
- Now compute separately the percents of male and female applicants admitted by the business school and by the law school.
- Explain carefully, as if speaking to a skeptical reporter, how it can happen that Mountain View appears to favor males when this is not true within each of the professional schools.

### Chapter 3 Exercises

**3.27** Doctors identify “chronic tension-type headaches” as headaches that occur almost daily for at least six months. Can antidepressant medications or stress management training reduce the number and severity of these headaches? Are both together more effective than either alone? Investigators compared four treatments: antidepressant alone, placebo alone, antidepressant plus stress management, and placebo plus stress management. Outline the design of the experiment. The headache sufferers named below have agreed to participate in the study. Use software or Table B at line 151 to randomly assign the subjects to the treatments.

Anderson	Archberger	Bezawada	Cetin	Cheng
Chronopoulou	Codrington	Daggy	Daye	Engelbrecht
Guha	Hatfield	Hua	Kim	Kumar
Leaf	Li	Lipka	Lu	Martin
Mehta	Mi	Nolan	Olbricht	Park
Paul	Rau	Saygin	Shu	Tang
Towers	Tyner	Vassilev	Wang	Watkins
Xu				

**3.33** A maker of fabric for clothing is setting up a new line to “finish” the raw fabric. The line will use either metal rollers or natural-bristle rollers to raise the surface of the fabric; a dyeing cycle time of either 30 minutes or 40 minutes; and a temperature of either 150 or 175 degrees Celsius. An experiment will compare all combinations of these choices. Four specimens of fabric will be subjected to each treatment and scored for quality.

- (a) What are the factors and the treatments? How many individuals (fabric specimens) does the experiment require?
- (a) Outline a completely randomized design for this experiment. (You need to actually do the randomization.)

**3.51** The walk to your statistics class takes about 10 minutes, about the amount of time needed to listen to three songs on your iPod. You decide to take a simple random sample of songs from a Billboard list of Rock Songs. Here is the list:

1	Miss Murder	2	Animal I Have Become	3	Steady As She Goes	4	Dani California
5	The Kill (Bury Me)	6	Original Fire	7	When You Were Young	8	MakeD – Sure
9	Vicarious	10	The Diary of Jane				

Select the three songs for your iPod using a simple random sample.

**3.57** You are planning a report on apartment living in a college town. You decide to select 5 apartment complexes at random for in-depth interviews with residents. Select a simple random sample of 5 of the following apartment complexes. If you use Table B, start at line 137.

1	Ashley Oaks	2	Country View	3	Mayfair Village
4	Bay Pointe	5	Country Villa	6	Nobb Hill
7	Beau Jardin	8	Crestview	9	Pemberly Courts
10	Bluffs	11	Del-Lynn	12	Peppermill
13	Brandon Place	14	Fairington	15	Pheasant Run
16	Briarwood	17	Fairway Knolls	18	Richfield
19	Brownstone	20	Fowler	21	Sagamore Ridge
22	Burberry	23	Franklin Park	24	Salem Courthouse
25	Cambridge	26	Georgetown	27	Village Manor
28	Chauncey Village	29	Greenacres	30	Waterford Court
31	Country Squire	32	Lahr House	33	Williamsburg

**3.61** The Census Bureau divides the entire country into “census tracts” that contain about 4000 people. Each tract is in turn divided into small “blocks,” which in urban areas are bounded by local streets. An SRS of blocks from a census tract is often the next-to-last stage in a multistage sample. Figure 3.10 shows part of the census tract 8051.12, in Cook County, Illinois, west of Chicago. The 44 blocks in this tract are divided into three “block groups.” Group 1 contains 6 blocks numbered 1000 to 1005; Group 2 (outlined in Figure 3.10) contains 12 blocks numbered 2000 to 2011; Group 3 contains 26 blocks numbered 3000 to 3025. Use Table B, beginning at line 135, to choose an SRS of 5 of the 44 blocks in this census tract. Explain carefully how you labeled the blocks.

## Chapter 4 Exercises

**4.1** Use Table B. We can use the random digits in Table B in the back of the text to simulate tossing a fair coin. Start at line 109 and read the numbers from left to right. If the number is 0, 1, 2, 3, or 4, you will say that the coin toss resulted in a head; if the number is a 5, 6, 7, 8, or 9, the outcome is tails. Use the first 20 random digits on line 109 to simulate 20 tosses of a fair coin. What is the actual proportion of heads in your simulated sample? Explain why you did not get exactly 10 heads.

**4.7** The basketball player Shaquille O'Neal makes about half of his free throws over an entire season. Use Table B or the *Probability* applet to simulate 100 free throws shot independently by a player who has probability 0.5 of making each shot.

- (a) What percent of the 100 shots did he hit?
- (b) Examine the sequence of hits and misses. How long was the longest run of shots made? Of shots missed? (Sequences of random outcomes often show runs longer than our intuition thinks likely.)



## Chapter 6 Exercises

**6.71** Refer to Exercise 6.26. In addition to the computer computing mpg, the driver also recorded the mpg by dividing the miles driven by the number of gallons at each fill-up. The following data are the differences between the computer's and the driver's calculations for that random sample of 20 records. The driver wants to determine if these calculations are different. Assume the standard deviation of a difference to be  $\sigma = 3.0$ .

5.0	6.5	-0.6	1.7	3.7	4.5	8.0	2.2	4.9	3.0
4.4	0.1	3.0	1.1	1.1	5.0	2.1	3.7	-0.6	-4.2

- State the appropriate  $H_0$  and  $H_A$  to test this suspicion.
- Carry out the test. Give the  $p$ -value, and then interpret the result in plain language.

**6.111** Refer to the previous exercise. Note that in the east-west direction, the average location was 113.8. Use the *Power* applet to find the power for the alternative  $\mu = 110$ .

**6.113** Example 6.16 gives a test of a hypothesis about the SAT scores of California high school students based on an SRS of 500 students. The hypotheses are  $H_0: \mu = 450$  and  $H_A: \mu > 450$ . Assume that the population standard deviation is  $\sigma = 100$ . The test rejects  $H_0$  at the 1% level of significance when  $z \geq 2.326$ , where

$$z = \frac{\bar{x} - 450}{100 / \sqrt{500}}$$

Is this test sufficiently sensitive to usually detect an increase of 10 points in the population mean SAT score? Answer this question by calculating the power of the test against the alternative  $\mu = 460$ .

**6.119** Refer to the previous exercise. Do the simulations and report the results for 90% confidence.

**6.121** Patients with chronic kidney failure may be treated by dialysis, using a machine that removes toxic wastes from the blood, a function normally performed by the kidneys. Kidney failure and dialysis can cause other changes, such as retention of phosphorus that must be corrected by changes in diet. A study of the nutrition of dialysis patients measured the level of phosphorus in the blood of several patients on six occasions. Here are the data for one patient (in milligrams of phosphorus per deciliter of blood:

5.4	5.2	4.5	4.9	5.7	6.3
-----	-----	-----	-----	-----	-----

The measurements are separated in time and can be considered an SRS of the patient's blood phosphorus level. Assume that this level varies Normally with  $\sigma = 0.9$  mg/dl.

- (a) Give a 95% confidence interval for the mean blood phosphorus level.
- (b) The normal range of phosphorus in the blood is considered to be 2.6 to 4.8 mg/dl. Is there strong evidence that this patient has a mean phosphorus level that exceeds 4.8?

**6.123** Many food products contain small quantities of substances that would give an undesirable taste or smell if they are present in large amounts. An example is the “off-odors” caused by sulfur compounds in wine. Oenologists (wine experts) have determined the odor threshold, the lowest concentration of a compound that the human nose can detect. For example, the odor threshold for dimethyl sulfide (DMS) is given in the oenology literature as 25 micrograms per liter of wine ( $\mu\text{g/l}$ ). Untrained noses may be less sensitive, however. Here are the DMS odor thresholds for 10 beginning students of oenology:

31	31	43	36	23	34	32	30	20	24
----	----	----	----	----	----	----	----	----	----

Assume (this is not realistic) that the standard deviation of the odor threshold for untrained noses is known to be a  $\sigma = 7 \mu\text{g/l}$ .

- (a) Make a stemplot to verify that the distribution is roughly symmetric with no outliers. (A Normal quantile plot confirms that there are no systematic departures from Normality.)
- (b) Give a 95% confidence interval for the mean DMS odor threshold among all beginning oenology students.
- (c) Are you convinced that the mean odor threshold for beginning students is higher than the published threshold, 25  $\mu\text{g/l}$ ? Carry out a significance test to justify your answer.

## Chapter 7 Exercises

**7.29** Children in a psychology study were asked to solve some puzzles and were then given feedback on their performance. Then they were asked to rate how luck played a role in determining their scores. This variable was recorded on a 1 to 10 scale with 1 corresponding to very lucky and 10 corresponding to very unlucky. Here are the scores for 60 children:

1	10	1	10	1	1	10	5	1	1	8	1	10	2	1
9	5	2	1	8	10	5	9	10	10	9	6	10	1	5
1	9	2	1	7	10	9	5	10	10	10	1	8	1	6
10	1	6	10	10	8	10	3	10	8	1	8	10	4	2

- Use graphical methods to display the distribution. Describe any unusual characteristics. Do you think that these would lead you to hesitate before using the Normality-based methods of this section?
- Give a 95% confidence interval for the mean luck score.

**7.35** Refer to Exercise 7.24. In addition to the computer calculating mpg, the driver also recorded the mpg by dividing the miles driven by the amount of gallons at fill-up. The driver wants to determine if these calculations are different.

<b>Fill-up</b>	1	2	3	4	5	6	7	8	9	10
<b>Computer</b>	41.5	50.7	36.6	37.3	34.2	45.0	48.0	43.2	47.7	42.2
<b>Driver</b>	36.5	44.2	37.2	35.6	30.5	40.5	40.0	41.0	42.8	39.2
<b>Fill-up</b>	11	12	13	14	15	16	17	18	19	20
<b>Computer</b>	43.2	44.6	48.4	46.4	46.8	39.2	37.3	43.5	44.3	43.3
<b>Driver</b>	38.8	44.5	45.4	45.3	45.7	34.2	35.2	39.8	44.9	47.5

- State the appropriate  $H_0$  and  $H_A$ .
- Carry out the test. Give the  $p$ -value, and then interpret the result.

**7.39** Dual-energy X-ray absorptimetry (DXA) is a technique for measuring bone health. One of the most common measures is a total body bone mineral content (TBBMC). A highly skilled operator is required to take the measurements. Recently, a new DXA machine was purchased by a research lab and two operators were trained to take the measurements. TBBMC for eight subjects was measured by both operators. The unit is grams (g). A comparison of the means for the two operators provides a check on the training they received and allows us to determine if one of the operators is producing measurements that are consistently higher than the other. Here are the data:

				Subject				
Operator	1	2	3	4	5	6	7	8

1	1.328	1.342	1.075	1.228	.939	1.004	1.178	1.286
2	1.323	1.322	1.073	1.233	.934	1.019	1.184	1.304

- Take the difference between the TBBMC recorded for Operator 1 and the TBBMC for Operator 2. Describe the distribution of these differences.
- Use a significance test to examine the null hypothesis that the two operators have the same mean. Be sure to give the test statistic with its degrees of freedom, the  $P$ -value, and your conclusion.
- The sample here is rather small, so we may not have much power to detect differences of interest. Use a 95% confidence interval to provide a range of differences that are compatible with these data.
- The eight subjects used for this comparison were not a random sample. In fact, they were friends of the researchers whose ages and weights were similar to the types of people who would be measured this DXA. Comment on the appropriateness of this procedure for selecting a sample, and discuss any consequences regarding the interpretation of the significance test and confidence interval results.

**7.43** Refer to the IQ test scores for fifth-grade students in Table 1.3 (page 13). Give numerical and graphical summaries of the data and compute a 95% confidence interval. Comment on the validity of the interval.

IQ test scores for 60 randomly chosen fifth-grade students									
145	139	126	122	125	130	96	110	118	118
101	142	134	124	112	109	134	113	81	113
123	94	100	136	109	131	117	110	127	124
106	124	115	133	116	102	127	117	109	137
117	90	103	114	139	101	122	105	97	89
102	108	110	128	114	112	114	102	82	101

**7.51** Suppose that the bone researchers in Exercise 7.39 wanted to be able to detect an alternative mean difference of .002. Find the power for this alternative for a sample size of 15. Use the standard deviation that you found in Exercise 7.39 for these calculations.

**7.81** The study of 584 longleaf pine trees in the Wade Tract in Thomas County, Georgia, had several purposes. Are trees in one part of the tract more or less like trees in any other part of the tract or are there differences? In Example 6.1 we examined how the trees were distributed in the tract and found that the pattern was not random. In this exercise we will examine the sizes of the trees. In Exercise 7.25 we analyzed the sizes, measured as diameter at breast height (DBH), for a randomized sample of 40 trees. Here we divide the tract into northern and southern halves and take random samples of 30 trees from each half. Here are the diameters in centimeters (cm) of the sampled trees:

North	27.8	14.5	39.1	3.2	58.8	55.5	25.0	5.4	19.0	30.6
	15.1	3.6	28.4	15.0	2.2	14.2	44.2	25.7	11.2	46.8
	36.9	54.1	10.2	2.5	13.8	43.5	13.8	39.7	6.4	4.8
South	44.4	26.1	50.4	23.3	39.5	51.0	48.1	47.2	40.3	37.4
	36.8	21.7	35.7	32.0	40.4	12.8	5.6	44.3	52.9	38.0
	2.6	44.6	45.5	29.1	18.7	7.0	43.8	28.3	36.9	51.6

- Use a back-to-back stemplot and side-by-side boxplots to examine the data graphically. Describe the patterns in the data.
- Is it appropriate to use the methods of this section to compare the mean DBH of the trees in the north half of the tract with the mean DBH of trees in the south half? Give reasons for your answer.
- What are appropriate null and alternative hypotheses for comparing the two samples of tree DBHs? Give reasons for your choices.
- Perform the significance test. Report the test statistic, the degrees of freedom, and the  $P$ -value. Summarize your conclusion.
- Find a 95% confidence interval for the difference in mean DBHs. Explain how this interval provides additional information about this problem.

**7.83** A market research firm supplies manufacturers with estimates of the retail sales of their products from samples of retail stores. Marketing managers are prone to look at the estimate and ignore sampling error. Suppose that an SRS of 70 stores this month shows mean sales of 53 units of a small appliance, with standard deviation 15 units. During the same month last year, an SRS of 55 stores gave mean sales of 50 units, with standard deviation 18 units. An increase from 50 to 53 is a rise of 6%. The marketing manager is happy because sales are up 6%.

- Explain what is wrong with your friend's procedure and why.
- Suppose he reported  $t = 1.70$  with a  $P$ -value of 0.06. What is the correct  $P$ -value that he should report.

**7.89** Refer to the Wade Tract DBH data in Exercise 7.81 (page 471), where we compared a sample of trees from the northern half of the tract with sample from the southern half. Because the standard deviations for the two samples are quite close, it is reasonable to analyze these data using the pooled procedures. Perform the significance test and find the 95% confidence interval for the difference in means using these methods. Summarize your results and compare them with what you found in Exercise 7.81.

**7.91** Use the Wade Tract data in Exercise 7.81 to calculate the software approximation to the degrees of freedom using the formula on page 460. Verify your calculation with software.

**7.105** The diameters of trees in the Wade Tract for random samples selected from the north and south portions of the tract are compared in Exercise 7.81 (page 471). Are there statistically significant differences in the standard deviations for these two parts of the tract? Perform the significance test and summarize the results. Does the Normal assumption appear reasonable for these data?

**7.125** A retailer entered into an exclusive agreement with a supplier who guaranteed to provide all products at competitive prices. The retailer eventually began to purchase supplies from the other vendors who offered better prices. The original supplier filed with a legal action claiming violation of the agreement. In defense, the retailer had an audit performed on a random sample of invoices. For each audited invoice, all purchases made from other suppliers were examined and the prices were compared with those offered by the original supplier. For each invoice, the percent of purchases for which the alternate supplier offered a lower price than the original supplier was recorded. Here are the data:

0	100	0	100	33	34	100	48	78	100	77	100	38
68	100	79	100	100	100	100	100	100	89	100	100	

Report the average of the percents with a 95% margin of error. Do the sample invoices suggest that the original supplier's prices are not competitive on the average?

**7.133** Table 1.2 (page 10) gives literacy rates for men and women in 17 Islamic nations. Is it proper to apply the one-sample  $t$  method to these data to give a 95% confidence interval for the mean literacy rate of Islamic men? Explain your answer.

## Chapter 8 Exercises

**8.3** A 1993 nationwide survey by the National Center for Education Statistics reports that 72% of all undergraduates work while enrolled in school. You decide to test whether this percent is different at your university. In your random sample of 100 students, 77 said they were currently working.

- Give the null and alternative hypotheses.
- Carry out the significance test. Report the test statistic and  $p$ -value.
- Does it appear that the percent of students working at your university is different at the  $\alpha = 0.05$  level?

**8.11** Gambling is an issue of great concern to those involved in Intercollegiate athletics. Because of this, the National Collegiate Athletic Association (NCAA) surveyed student-athletes concerning their gambling-related behaviors. There were 5594 Division I male athletes in the survey. Of these, 3547 reported participation in some gambling behavior. This included playing cards, betting on games of skill, buying lottery tickets, and betting on sports.

Find the sample proportion and the large-sample margin of error for 95% confidence. Explain in simple terms the 95%.

**8.19** A survey of 1280 student loan borrowers found that 192 had loans totaling more than \$30,000 for their undergraduate education. Give a 95% confidence interval for the proportion of all student loan borrowers who have loans of more than \$30,000 for their undergraduate education.

**8.23** For a study of unhealthy eating behaviors, 267 college women aged 18 to 25 years were surveyed. Of these, 69% reported that they had been on a diet sometime during the past year. Give a 95% confidence interval for the true proportion of college women aged 18 to 25 years in this population who dieted last year.

**8.29** The South African mathematician John Kerrich, while a prisoner of war during World War II, tossed a coin 10,000 times and obtained 5067 heads.

- Is this significant evidence at the 5% level that the probability that Kerrich's coin comes up heads is not 0.5? Use a sketch of the standard Normal distribution to illustrate the  $p$ -value.
- Use a 95% confidence interval to find the range of probabilities of heads that would not be rejected at the 5% level.

**8.35** A study was designed to compare two energy drink commercials. Each participant was shown the commercials in random order and asked to select the better one. Commercial A was selected by 45 out of 100 women and 80 out of 140 men. Give an

estimate of the difference in gender proportions that favored Commercial A. Also construct a large-sample 95% confidence interval for this difference.

**8.37** Refer to Exercise 8.35. Test that the proportions of women and men that liked Commercial A are the same versus the two-sided alternative at the 5% level.

**8.51** Different kinds of companies compensate their key employees in different ways. Established companies may pay higher salaries, while new companies may offer stock options that will be valuable if the company succeeds. Do high-tech companies tend to offer stock options more often than other companies? One study looked at a random sample of 200 companies. Of these, 91 were listed in the *Directory of Public High High Technology Corporations* and 109 were not listed. Treat these two groups as SRSs of high-tech and non-high-tech companies. Seventy three of the high-tech companies and 75 of the non-high-tech companies offered incentive stock options to key employees.

- Give a 95% confidence interval for the difference in the proportions of the two types of companies that offer stock options.
- Compare the two groups of companies with a significance test.
- Summarize your analysis and conclusions.

**8.63** In Exercise 8.48 the effects of a reduction in air pollution on wheezing was examined by comparing the one-year change in symptoms in a group of residents who lived on congested streets with a group of residents who lived on congested streets with a group who lived in an area that had been congested but from which the congestion was removed when a bypass was built. The effect of the reduction in air pollution was assessed by comparing the proportions of residents in the two groups who reported that their wheezing symptoms improved. Here are some additional data from the same study:

	Bypass		Congest	
Symptom	N	Improved	N	Improved
Number of wheezing attacks	282	45	163	21
Wheezing disturbs sleep	282	45	164	12
Wheezing limits speech	282	12	164	4
Wheezing affects activities	281	26	165	13
Winter cough	261	15	156	14
Winter phlegm	253	12	144	10
Consulted	247	29	140	18



doctor				
--------	--	--	--	--

The table gives the number of subjects in each group and the number reporting improvement. So, for example, the proportion who reported improvement in the number of wheezing attacks was 21/163 in the congested group.

- The reported sample sizes vary from symptom to symptom. Give possible reasons for this and discuss the possible impact on the results.
- Calculate the difference in the proportion for each symptom. Make a table of symptoms ordered from highest to lowest based on these differences. Include the estimates of the differences and the 95% confidence intervals in the table. Summarize your conclusions.
- Can you justify a one-sided alternative in this situation? Give reasons for your answer.
- Perform a significance test to compare the two groups for each of the symptoms. Summarize the results.
- Reanalyze the data using only the data from the bypass group. Give confidence intervals for the proportions that reported improved symptoms. Compare results with those you presented in part (b). Use your analyses of the data in this exercise to discuss the importance of a control group in studies such as this.

**8.65** To devise effective marketing strategies it is helpful to know the characteristics of your customers. A study compared demographic characteristics of people who use the Internet for travel arrangements and of people who do not. Of 1132 Internet users, 643 had completed college. Among the 852 nonusers, 349 had completed college.

- Do users and nonusers differ significantly in the proportion of college graduates?
- Give a 95% confidence interval for the difference in the proportions.

**8.67** Refer to the previous two exercises. Give the total number of users and the total number of nonusers for the analysis of education. Do the same for the analysis of income. The difference is due to respondents who chose “Rather not say” for the income question. Give the proportions of “Rather not say” individuals for users and nonusers. Perform a significance test to compare these and give a 95% confidence interval for the difference. People are often reluctant to provide information about their income. Do you think that this amount of nonresponse for the income question is a serious limitation of the study?

## Chapter 9 Exercises

**9.7** The consensus Bureau provides estimates of numbers of people in the United States classified in various ways. Let's look at college students. The following table gives us data to examine the relation between age and full-time or part-time status. The numbers in the table are expressed as thousands of U.S. college students.

	Status	
Age	Full-time	Part-time
15-19	3553	329
20-24	5710	1215
25-34	1825	1864
35 and over	901	1983

- Give the joint distribution of age and status for this table.
- What is the marginal distribution of age? Display the results graphically.
- What is the marginal distribution of status? Display the results graphically.
- Compute the conditional distribution of age for each of the two status categories. Display the results graphically.
- Write a short paragraph describing the distributions and how they differ.

**9.11** Cocaine addiction is difficult to overcome. Addicts have been reported to have a significant depletion of stimulating neurotransmitters and thus continue to take cocaine to avoid feelings of depression and anxiety. A 3-year study with 72 chronic cocaine users compared an antidepressant drug called desipramine with lithium and a placebo. (Lithium is a standard drug to treat cocaine addiction. A placebo is a substance containing no medication, used so that the effect of being in the study but not taking any drug can be seen.) One-third of the subjects, chosen at random, received each treatment. Following are the results:

Treatment	Cocaine relapse?	
	Yes	No
Desipramine	10	14
Lithium	18	6
Placebo	20	4

- Compare the effectiveness of the three treatments in preventing relapse using percents and a bar graph. Write a brief summary.
- Can we comfortably use the chi-square test to test the null hypothesis that there is no difference between treatments? Explain.
- Perform the significance test and summarize the results.

**9.15** A study examined patterns and characteristics of volunteer-service for young people from high school through early adulthood. Here are some data that can be used to compare males and females on participation in unpaid volunteer service or community service and motivation participation:

	Participants			
	Motivation			
Gender	Strictly Voluntary	Court-ordered	Other	Non-participants
Men	31.1%	2.1%	6.3%	59.7%
Women	43.7%	1.1%	6.5%	48.7%

Note that the percents in each row sum to 100%

- Graphically compare the volunteer-service profiles for men and women. Describe any differences that are striking.
- Find the proportion of men who volunteer. Do the same for women. Refer to the section on relative risk in Chapter 8 (page 51) and the discussion on page 535 of this chapter. Compute the relative risk of being a volunteer for females versus males. Write a clear sentence contrasting females and males using relative risk as your numerical summary.

**9.17** As part of the 1999 College Alcohol Study, students who drank alcohol in the last year were asked if drinking ever resulted in missing a class. The data are given in the following table:

	Drinking Status		
	Nonbinger	Occasional Binger	Frequent Binger
Missed Class			
No	4617	2047	1176
Yes	446	915	1959

- Summarize the results of this table graphically and numerically.
- What is the marginal distribution of drinking status? Display the results graphically.
- Compute the relative risk of missing a class for occasional bingers versus nonbingers and for frequent bingers versus nonbingers. Summarize these results.
- Perform the chi-square test for this two-way table. Give the test statistic, degrees of freedom, the  $p$ -value, and your conclusion.

**9.19** The ads in the study described in the previous exercise were also classified according to the age group of the intended readership. Here is a summary of the data:

Magazine readership age group		
Model dress	Young adult	Mature adult
Not sexual	72.3%	76.1%
Sexual	27.2%	23.9%
Number of ads	1006	503

Using parts (a) and (b) in the previous exercise as a guide, analyze these data and write a report summarizing your work.

- Give the joint distribution of age and status for this table.
- What is the marginal distribution of age? Display the results graphically.
- What is the marginal distribution of status? Display the results graphically.
- Compute the conditional distribution of age for each of the two status categories. Display the results graphically.

**9.23** A survey of student-athletes that asked questions about gambling behavior classified students according to the National Collegiate Athletic Association (NCAA) division. For male student-athletes, the percents who reported wagering on collegiate sports are given here along with the numbers of respondents in each division:

Division	I	II	III
Percent	17.2%	21.0%	24.4%
Number	5619	2957	4089

- Use a significance test to compare the percents for the three NCAA divisions. Give details and a short summary of your conclusion.
- The percents in the table above are given in the NCAA report, but the numbers of male student-athletes in each division who responded to the survey question are estimated based on other information in the report. To what extent do you think this has an effect on the results? (*Hint:* Rerun your analysis a few times, with slightly different numbers of students but the same percents.)
- Some student-athletes may be reluctant to provide this kind of information, even in a survey where there is no possibility that they can be identified. Discuss how this fact may affect your conclusions.
- The chi-square test for this set of data assumes that the responses of the student-athletes are independent. However, some of the students are at the same school and even on the same team. Discuss how you think this might affect the results.

**9.29** A task force set up to examine retention of students in the majors that they chose when starting college examined data on transfers to other majors. Here are some data giving counts of students classified by initial major and the area that they transferred to:

		Area Transferred to		
Initial Major	Engineering	Management	Liberal arts	Total
Biology	13	25	158	398
Chemistry	16	15	19	114
Mathematics	3	11	20	72
Physics	9	5	14	61

Complete the table by computing the values for the “other” column. Write a short paragraph explaining what conclusions you can draw about the relationship between initial major and area transferred to. Be sure to include numerical and graphical summaries as well as the details of your significance test.

- Analyze the data for the dogs and the cats separately. Be sure to include graphical and numerical summaries. Is there evidence to conclude that the source of the animal is related to whether or not the pet is brought to an animal shelter?
- Write a discussion comparing the results for the cats with those for the dogs.
- These data were collected using a telephone interview with pet owners in Mishawaka, Indiana. The animal shelter was run by the Humane Society of Saint Joseph County. The control group data were obtained by a random digit dialing telephone survey. Discuss how these facts relate to your interpretation of the results.

**9.33** Euthanasia of healthy but unwanted pets by animal shelter is believed to be the leading cause of death for cats and dogs. A study designed to find factors associated with bringing a cat to an animal shelter compared data on cats that were brought to an animal shelter with data on cats from the same county that were not brought in. One of the factors examined was the source of the cat: the categories were private owner or breeder, pet store, and other (includes born in home, stray, and obtained from a shelter). This kind of study is called a **case-control study** by epidemiologists. Here are the data:

		Source	
Group	Private	Pet store	Other
Cases	124	16	76
Controls	219	24	203

The same researchers did a similar study for dogs. The data are given in the following table:

		Source	
Group	Private	Pet store	Other
Cases	188	7	90
Controls	518	68	142

**9.41** The 2005 National Survey of Student Engagement reported on the use of campus services during the first year of college. In terms of academic assistance (for example tutoring, writing lab), 43% never used the services, 35% sometimes used the services,, 15% often used the services, and 7% very often used the services. You decide to see if your large university has this same distribution. You survey first-year students and obtain the counts 79, 83, 36, and 12 respectively. Use a goodness of fit test to examine how well your university reflects the national average.

## Chapter 10 Exercises

**10.13** How well does the number of beers a student drinks predict his or her blood alcohol content? Sixteen student volunteers at Ohio State University drank a randomly assigned number of 12-ounce cans of beer. Thirty minutes later, a police officer measured their blood alcohol content (BAC). Here are the data:

Student	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Beers	5	2	9	8	3	7	3	5	3	5	4	6	5	7	1	4
BAC	.10	.03	.19	.12	.04	.095	.07	.06	.02	.05	.07	.10	.085	.09	.01	.05

The students were equally divided between men and women and differed in weight and usual drinking habits. Because of this variation, many students don't believe that number of drinks predicts blood alcohol well.

- Make a scatterplot of the data. Find the equation of the least-squares regression line for predicting blood alcohol from number of beers and add this line to your plot. What is  $r^2$  for these data? Briefly summarize what your data analysis shows.
- Is there significant evidence that drinking more beers increases blood alcohol on the average in the population of all students? State hypotheses, give a test statistic and  $P$ -value, and state your conclusion.
- Steve thinks he can drive legally 30 minutes after he drinks 5 beers. The legal limit is  $BAC = .08$ . Give a 90% confidence interval for Steve's BAC. Can he be confident he won't be arrested if he drives and is stopped?

**10.21** Exercise 2.144 (page 163) gives the modulus of elasticity (MOE) and modulus of rupture (MOR) for 32 plywood specimens. Because measuring MOR involves breaking the wood but measuring MOE does not, we would like to predict the destructive test result, MOR, using the nondestructive test result, MOE

- Describe the distribution of MOR using graphical and numerical summaries. Do the same for MOE.
- Make a plot of the two variables. Which should be plotted on the  $x$  axis? Give a reason for your answer.
- Give the statistical model for this analysis, run the analysis, summarize the results, and write a short summary of your conclusion.
- Examine the assumptions needed for the analysis. Are you satisfied that there are no serious violations that would cause you to question the validity of your conclusions?

**10.25** In exercise 7.26 (page 442) we examined the distribution of C-reactive protein (CRP) in a sample of 40 children from Papua New Guinea. Serum retinol values for the same children were studied in Exercise 7.28. One important question that can be

addressed with these data is whether or not infections, as indicated by CRP, cause a decrease in the measured values of retinol, low values of which indicate a vitamin A deficiency. The data are given in Table 10.5.

CRP	Retinol	CRP	Retinol	CRP	Retinol	CRP	Retinol	CRP	Retinol
0	1.15	30.61	0.97	22.82	0.24	5.36	1.19	0	0.83
3.9	1.36	0	0.67	0	1	0	0.94	0	1.11
5.64	0.38	73.2	0.31	0	1.13	5.66	0.34	0	1.02
8.22	0.34	0	0.99	3.49	0.31	0	0.35	9.37	0.56
0	0.35	46.7	0.52	0	1.44	59.76	0.33	20.78	0.82
5.62	0.37	0	0.7	0	0.35	12.38	0.69	7.1	1.2
3.92	1.17	0	0.88	4.81	0.34	15.74	0.69	7.89	0.87
6.81	0.97	26.41	0.36	9.57	1.9	0	1.04	5.53	0.41

- Examine the distributions of CRP and serum retinol. Use graphical and numerical methods.
- Forty percent of the CRP values are zero. Does this violate any assumption that we need to do a regression analysis of CRP to predict serum retinol? Explain your answer.
- Run the regression, summarize the results, and write a short paragraph explaining your conclusions.
- Explain the assumptions needed for your results to be valid. Examine the data with respect to these assumptions and report your results.

**10.27** In Exercise 7.119 (page 482) we looked at the distribution of tartrate resistant acid phosphate (TRAP), a biomarker for bone resorption. Table 10.7 gives values for this biomarker and a measure of bone resorption VO-. Analyze these data using the questions in the previous exercise as a guide.

**10.29** Refer to the TRAP and VO- data in Exercise 10.27. Reanalyze these data using the logs of both TRAP and VO-. Summarize your results and compare them with those you obtained in Exercise 10.27.

**10.33** How is the flow of investors' money into stock mutual funds related to the flow of money into bond mutual funds? Here are data on the net new money flowing into stock and bond mutual funds in the years 1985 to 2000, in billions of dollars. "Net" means that funds flowing out are subtracted from those flowing in. If more money leaves than arrives, the net flow will be negative. To eliminate the effect of inflation, all dollar amounts are in "real dollars" with constant buying power equal to that of a dollar in the year 2000.



Year	1985	1986	1987	1988	1989	1990	1991	1992
Stocks	12.8	34.6	28.8	-23.3	8.3	17.1	50.6	97.0
Bonds	100.8	161.8	10.6	-5.8	-1.4	9.2	74.6	87.1
Year	1993	1994	1995	1996	1997	1998	1999	2000
Stocks	151.3	133.6	140.1	238.2	243.5	165.9	194.3	309.0
Bonds	84.6	-72.0	-6.8	3.3	30.0	79.2	-.62	-48.0

- (a) Make a scatterplot with cash flow into stock funds as the explanatory variable. Find the least-squares line for predicting net bond investments from net stock investments. What do the data suggest?
- (b) Is there statistically significant evidence that there is some straight-line relationship between the flows of cash into bond funds and stock funds? (State hypotheses, give a test statistic and its  $P$ -value, and state your conclusion.)
- (c) What fact about the scatterplot explains why the relationship described by the least-squares line is not significant?

**10.37** We assume that our wages will increase as we gain experience and become more valuable to our employers. Wages also increase because of inflation. By examining a sample of employees at a given point in time, we can look at part of the picture. How does length of service (LOS) relate to wages? Table 10.8 gives data on the LOS in months and wages for 60 women who work in Indiana banks. Wages are yearly total income divided by the number of weeks worked. We have multiplied wages by a constant for reasons of confidentiality.

Table 10.8 Bank wages and length of service (LOS)					
Wages	LOS	Wages	LOS	Wages	LOS
48.3355	94	64.1026	24	41.2088	97
49.0279	48	54.9451	222	67.9096	228
40.8817	102	43.8095	58	43.0942	27
36.5854	20	43.3455	41	40.7000	48
46.7596	60	61.9893	153	40.5748	7
59.5238	78	40.0183	16	39.6825	74
39.1304	45	50.7143	43	50.1742	204
39.2465	39	48.8400	96	54.9451	24
40.2037	20	34.3407	98	32.3822	13
38.1563	65	80.5861	150	51.7130	30
50.0905	76	33.7163	124	55.8379	95
46.9043	48	60.3792	60	54.9451	104
43.1894	61	48.8400	7	70.2786	34
60.5637	30	38.5579	22	57.2344	184
97.6801	70	39.2760	57	54.1126	156
48.5795	108	47.6564	78	39.8687	25
67.1551	61	44.6864	36	27.4725	43

38.7847	10	45.7875	83	67.9584	36
51.8926	68	65.6288	66	44.9317	60
51.8326	54	33.5775	47	51.5612	102

- Plot wages versus LOS. Describe the relationship. There is one woman with relatively high wages for her length of service. Circle this point and do not use it in the rest of this exercise.
- Find the least-squares line. Summarize the significance test for the slope. What do you conclude?
- State carefully what the slope tells you about the relationship between wages and length of service.
- Give a 95% confidence interval for the slope.

**10.39** The Leaning Tower of Pisa is an architectural wonder. Engineers concerned about the tower's stability have done extensive studies of its increasing tilt. Measurements of the lean of the tower over time provide much useful information. The following table gives measurements for the years 1975 to 1987. The variable "lean" represents the differences between where a point on the tower would be if the tower were straight and where it actually is. The data are coded as tenths of a millimeter in excess of 2.9 meters, so that the 1975 lean, which was 2.9642 meters, appears in the table as 642. Only the last two digits of the year were entered into the computer.

Year	75	76	77	78	79	80	81	82	83	84	85	86	87
Lean	642	644	656	667	673	688	696	698	713	717	725	742	757

- Plot the data. Does the trend in lean over time appear to be linear?
- What is the equation of the least-squares line? What percent of the variation in lean is explained by this line?
- Give a 99% confidence interval for the average rate of change (tenths of a millimeter per year) of the lean.

**10.41** Refer to exercise 10.39.

- How would you code the explanatory variable for the year 2009?
- The engineers working on the Leaning Tower of Pisa were most interested in how much the tower would lean if no corrective action was taken. Use the least-squares equation to predict the tower's lean in the year 2009.
- To give a margin of error for the lean in 2009, would you use a confidence interval for a mean response or a prediction interval? Explain your choice.

**10.53** The SAT and the ACT are the two major standardized tests that colleges use to evaluate candidates. Most students take just one of these tests. However, some students

take both. Table 10.9 gives the scores of 60 students who did this. How can we relate the two tests?

SAT	ACT	SAT	ACT	SAT	ACT	SAT	ACT
1000	24	870	21	1090	25	800	21
1010	24	880	21	860	19	1040	24
920	17	850	22	740	16	840	17
840	19	780	22	500	10	1060	25
830	19	830	20	780	12	870	21
1440	32	1190	30	1120	27	1120	25
490	7	800	16	590	12	800	18
1050	23	830	16	990	24	960	27
870	18	890	23	700	16	880	21
970	21	880	24	930	22	1020	24
920	22	980	27	860	23	790	14
810	19	1030	23	420	21	620	18
1080	23	1220	30	800	20	1150	28
1000	19	1080	22	1140	24	970	20
1030	25	970	20	920	21	1060	24

- Plot the data with SAT on the  $x$  axis and ACT on the  $y$  axis. Describe the overall pattern and any unusual observations.
- Find the least-squares regression line and draw it on your plot. Give the results of the significance test for the slope.
- What is the correlation between the two tests?

## Chapter 11 Exercises

**11.27** Let's consider developing a model to predict total score based on the peer review score (PEER), faculty-to-student ratio (FtoS), and citations-to-faculty ratio (CtoF).

- (a) Using numerical and graphical summaries, describe the distribution of each explanatory variable.
- (b) Using numerical and graphical summaries, describe the relationship between each pair of explanatory variables.

**11.29** Now consider a regression model using all three explanatory variables.

- (a) Write out the statistical model for this analysis, making sure to specify all assumptions.
- (b) Run the multiple regression model and specify the fitted regression equation.
- (c) Generate a 95% confidence interval for each coefficient. Should any of these intervals contain 0? Explain.
- (d) What percent of the variation in total score is explained by this model? What is the estimate for  $\sigma$ ?

**11.31** Consider the following five variables for each nation: LSI, life-satisfaction score, an index of happiness; GINI, a measure of inequality in the distribution of income; CORRUPT, a measure of corruption in government; LIFE, the average life expectancy; and DEMOCRACY, a measure of civil and political liberties.

- (a) Using numerical and graphical summaries, describe the distribution of each variable.
- (b) Using numerical and graphical summaries, describe the relationship between each pair of variables.

**11.35** Let's use regression methods to predict VO+, the measure of bone formation.

- (a) Since OC is a biomarker of bone formation, we start with a simple linear regression using OC as the explanatory variable. Run the regression and summarize the results. Be sure to include an analysis of the residuals.
- (b) Because the processes of bone formation and bone resorption are highly related, it is possible that there is some information in the bone resorption variables that can tell us something about bone formation. Use a model with both OC and TRAP, the biomarker of bone resorption, to predict VO+. Summarize the results. In the context of this model, it appears that TRAP is a better predictor of bone formation, VO+, than the biomarker of bone formation, OC. Is this view consistent with the pattern of relationships that you described in the previous exercise? One possible explanation is that,

while all of these variables are highly related, TRAP is measured with more precision than OC.

**11.37** Because the distributions of VO+, VO-, OC, and TRAP tend to be skewed, it is common to work with logarithms rather than the measured values. Using the questions in the previous three exercises as a guide, analyze the log data.

**11.41** Use four congeners, PCB52, PCB118, PCB138, and PCB180, in a multiple regression to predict PCB.

- (a) Write the statistical model for this analysis. Include all assumptions.
- (b) Run the regression and summarize the results.
- (c) Examine the residuals. Do they appear to be approximately Normal? When you plot them versus each of the explanatory variables, are any patterns evident?

**11.43** Run a regression to predict PCB using the variables PCB52, PCB118, and PCB138. Note that this is similar to the analysis that you did in Exercise 11.41, with the change that PCB180 is not included as an explanatory variable.

- (a) Summarize the results.
- (b) In this analysis, the regression coefficient for PCB118 is not statistically significant. Give the estimate of the coefficient and the associated P-value.
- (c) Find the estimate of the coefficient for PCB118 and the associated P-value for the model analyzed in Exercise 11.41.
- (d) Using the results in parts (b) and (c), write a short paragraph explaining how the inclusion of other variables in a multiple regression can have an effect on the estimate of a particular coefficient and the results of the associated significance test.

**11.51** For each of the four variables in the CHEESE data set, find the mean, median, standard deviation, and interquartile range. Display each distribution by means of a stemplot and use a Normal quantile plot to assess Normality of the data. Summarize your findings. Note that when doing regressions with these data, we do not assume that these distributions are Normal. Only the residuals from our model need to be (approximately) Normal. The careful study of each variable to be analyzed is nonetheless an important first step in any statistical analysis.

**11.53** Perform a simple linear regression analysis using Taste as the response variable and Acetic as the explanatory variable. Be sure to examine the residuals carefully. Summarize your results. Include a plot of the data with the least-squares regression line.

Plot the residuals versus each of the other two chemicals. Are any patterns evident? (The concentrations of the other chemicals are lurking variables for the simple linear regression.)

**11.55** Repeat the analysis of Exercise 11.53 using Taste as the response variable and Lactic as the explanatory variable.

**11.57** Carry out a multiple regression using Acetic and H<sub>2</sub>S to predict Taste. Summarize the results of your analysis. Compare the statistical significance of Acetic in this model with its significance in the model with Acetic alone as a predictor (Exercise 11.53). Which model do you prefer? Give a simple explanation for the fact that Acetic alone appears to be a good predictor of Taste, but with H<sub>2</sub>S in the model, it is not.

**11.59** Use the three explanatory variables Acetic, H<sub>2</sub>S, and Lactic in a multiple regression to predict Taste. Write a short summary of your results, including an examination of the residuals. Based on all of the regression analyses you have carried out on these data, which model do you prefer and why?

## Chapter 12 Exercises

**12.29** Does bread lose its vitamins when stored? Small loaves of bread were prepared with flour that was fortified with a fixed amount of vitamins. After baking, the vitamin C content of two loaves was measured. Another two loaves were baked at the same time, stored for one day,, and then the vitamin C content was measured. In a similar manner, two loaves were stored for three, five, and seven days before measurements were taken. The units are milligrams of vitamin C per hundred grams of flour (mg/100 g). Here are the data:

Condition	Vitamin C (mg/100 g)	
Immediately after baking	47.62	49.79
One day after baking	40.45	43.46
Three days after baking	21.25	22.34
Five days after baking	13.18	11.65
Seven days after baking	8.51	8.13

- Give a table with sample size, mean, standard deviation, and standard error for each condition.
- Perform a one-way ANOVA for these data. Be sure to state your hypotheses, the test statistic with degrees of freedom, and the  $p$ -value.
- Summarize the data and the means with a plot. Use the plot and the ANOVA results to write a short summary of your conclusions.

**12.31** Refer to Exercise 12.29. Measurements of the amounts of vitamin A (beta-carotene) and vitamin E I each loaf are given below. Use the analysis of variance method to study the data for each of these vitamins.

Condition	Vitamin A (mg/100 g)		Vitamin E (mg/100 g)	
Immediately after baking	3.36	3.34	94.6	96.0
One day after baking	3.28	3.20	95.7	93.2
Three days after baking	3.26	3.16	97.4	94.3
Five days after baking	3.25	3.36	95.0	97.7
Seven days after baking	3.01	2.92	92.3	95.1

**12.33** Refer to Exercise 12.29. Write a report summarizing what happens to vitamins A, C, and E after bread is baked. Include appropriate statistical inference results and graphs.

**12.35** Different varieties of the tropical flower *Heliconia* are fertilized by different species of hummingbirds. Over time, the lengths of the flowers and the form of the hummingbirds' beaks have evolved to match each other. Here are data on the lengths in millimeters of three varieties of these flowers on the island of Dominica:

H.bihai							
47.12	46.75	46.81	47.12	46.67	47.43	46.44	46.64
48.07	48.34	48.15	50.26	50.12	46.34	46.94	48.36

H. caribaea red							
41.90	42.01	41.93	43.09	41.47	41.69	39.78	40.57
39.63	42.18	40.66	37.87	39.16	37.40	38.20	38.07
38.10	37.97	38.79	38.23	38.87	37.78	38.01	

H. caribaea yellow							
36.78	37.02	36.52	36.11	36.03	35.45	38.13	37.1
35.17	36.82	36.66	35.68	36.03	34.57	34.63	

Do a complete analysis that includes description of the data and a significance test to compare the mean lengths of the flowers for the three species.

**12.39** Kudzu is a plant that was imported to the United States from Japan and now covers over seven million acres in the South. The plant contains chemicals called isoflavones that have been shown to have beneficial effects on bones. One study used three groups of rats to compare a control group with rats that were fed wither a low dose or a high dose of isoflavones from kudzu. One of the outcomes examined was the bone mineral density in the femur (in grams per square centimeter). Here are the data:

Treatment	Bone mineral density (g/cm <sup>2</sup> )					
<b>Control</b>	0.228	0.221	0.234	0.220	0.217	0.228
	0.209	0.221	0.204	0.220	0.203	0.219
	0.218	0.245	0.210			
<b>Low dose</b>	0.211	0.220	0.211	0.233	0.219	0.233
	0.226	0.228	0.216	0.225	0.200	0.208
	0.198	0.208	0.203			
<b>High dose</b>	0.250	0.237	0.217	0.206	0.247	0.228
	0.245	0.232	0.267	0.261	0.221	0.219
	0.232	0.209	0.203			

- Use graphical and numerical methods to describe the data.
- Examine the assumptions necessary for ANOVA. Summarize your findings.
- Use a multiple-comparisons method to compare the three groups.

**12.41** Refer to the previous exercise. Use the Bonferroni or another multiple-comparisons procedure to compare the group means. Summarize the results and support your conclusions with a graph of the means.



**12.43** Refer to the previous exercise. Use the Bonferroni or another multiple-comparisons procedure to compare the group means. Summarize the results and support your conclusions with a graph of the means.

**12.47** Many studies have suggested that there is a link between exercise and healthy bones. Exercise stresses the bones and this causes them to get stronger. One study examined the effect of jumping on the bone density of growing rats. There were three treatments: a control with no jumping, a low-jump condition (the jump was 30 centimeters), and a high jump condition (the jump was 60 centimeters). After 8 weeks of 10 jumps per day, 5 days per week, the bone density of the rats (expressed in  $\text{mg}/\text{cm}^3$ ) was measured. Here are the data:

Group	Bone density ( $\text{mg}/\text{cm}^3$ )									
Control	611	621	614	593	593	653	600	554	603	569
Low jump	635	605	638	594	599	632	631	588	607	596
High jump	650	622	626	626	631	622	643	674	643	650

- Make a table giving the sample size, mean, and standard deviation for each group of rats. Is it reasonable to pool the variances?
- Run the analysis of variance. Report the F statistic with its degrees of freedom and  $p$ -value. What do you conclude?

**12.51** One way to repair serious wounds is to insert some material as a scaffold for the body's repair cells to use as a template for new tissue. Scaffolds made from extracellular material (ECM) are particularly promising for this purpose. Because they are made from biological material, they serve as an effective scaffold and are then resorbed. Unlike biological material that includes cells, however, they do not trigger tissue rejection reactions in the body. One study compared 6 types of scaffold material. Three of these were ECMs and the other three were made of inert materials. There were three mice used per scaffold type. The response measure was the percent of glucose phosphorylated isomerase (Gpi) cells in the region of the wound. A large value is good, indicating that there are many bone marrow cells sent by the body to repair the tissue.

Material	Gpi (%)		
ECM1	55	70	70
ECM2	60	65	65
ECM3	75	70	75
MAT1	20	25	20
MAT2	5	10	5
MAT3	10	15	10

- (a) Make a table giving the sample size, mean, and standard deviation for each of the six types of material. Is it reasonable to pool the variances? Note that the sample sizes are small and the data are rounded.
- (b) Run the analysis of variance. Report the  $F$  statistic with its degrees of freedom and  $P$ -value. What do you conclude?

**12.53** Refer to Exercise 12.25. There are two comparisons of interest to the experimenter: They are (1) Placebo versus the average of the 2 low-dose treatments; and (2) the difference between High A and Low A versus the difference between High B and Low B.

- (a) Express each contrast in terms of the means ( $\mu$ 's) of the treatments.
- (b) Give estimates with standard errors for each of the contrasts.
- (c) Perform the significance tests for the contrasts. Summarize the results of your tests and your conclusions.

## Chapter 13 Exercises

**13.25** One way to repair serious wounds is to insert some material as a scaffold for the body's repair cells to use as a template for new tissue. Scaffolds made from extracellular material (ECM) are particularly promising for this purpose. Because they are made from biological material, they serve as an effective scaffold and are then resorbed. Unlike biological material that includes cells, however, they do not trigger tissue rejection reactions in the body. One study compared 6 types of scaffold material. Three of these were ECMs and the other three were made of inert materials. There were three mice used per scaffold type. The response measure was the percent of glucose phosphorylated isomerase (Gpi) cells in the region of the wound. A large value is good, indicating that there are many bone marrow cells sent by the body to repair the tissue. In Exercise 12.51 we analyzed the data for rats whose tissues were measured 4 weeks after the repair. The experiment included additional groups of rats who received the same types of scaffold but were measured at different times. The data in the table below are for 4 weeks and 8 weeks after the repair:

- Make a table giving the sample size, mean, and standard deviation for each of the material-by-time combinations. Is it reasonable to pool the variances? Because the sample sizes in this experiment are very small, we expect a large amount of variability in the sample standard deviations. Although they vary more than we would prefer, we will proceed with the ANOVA.
- Make a plot of the means. Describe the main features of the plot.
- Run the analysis of variance. Report the  $F$  statistics with degrees of freedom and  $p$ -values for each of the main effects and the interaction. What do you conclude?

Material	4 weeks			6 weeks		
ECM1	55	70	70	60	65	65
ECM2	60	65	65	60	70	60
ECM3	75	70	75	70	80	70
MAT1	20	25	25	15	25	25
MAT2	5	10	5	10	5	5
MAT3	10	15	10	5	10	10

**13.31** One step in the manufacture of large engines requires that holes of very precise dimensions be drilled. The tools that do the drilling are regularly examined and are adjusted to ensure that the holes meet the required specifications. Part of the examination involves measurement of the diameter of the drilling tool. A team studying the variation in the sizes of the drilled holes selected this measurement procedure as a possible cause of variation in the drilled holes. They decided to use a designed experiment as one part of this examination. Some of the data are given in Table 13.2 reproduced below. The diameters in millimeters (mm) of five tools were measured by the same operator at three times (8:00 a.m., 11:00 a.m., and 3:00 p.m.). The person taking the measurements could

not tell which tool was being measured, and the measurements were taken in random order.

Tool	Time	Diameter (mm)		
1	1	25.030	25.030	25.032
1	2	25.028	25.028	25.028
1	3	25.026	25.026	25.026
2	1	25.016	25.018	25.016
2	2	25.022	25.020	25.018
2	3	25.016	25.016	25.016
3	1	25.005	25.008	25.006
3	2	25.012	25.012	25.014
3	3	25.010	25.010	25.008
4	1	25.012	25.012	25.012
4	2	25.018	25.020	25.020
4	3	25.010	25.014	25.018
5	1	24.996	24.998	24.998
5	2	25.006	25.006	25.006
5	3	25.000	25.002	24.999

- Make a table of means and standard deviations for each of the  $5 \times 3$  combinations of the two factors.
- Plot the means and describe how the means vary with tool and time. Note that we expect the tools to have slightly different diameters. These will be adjusted as needed. It is the process of measuring the diameters that is important.
- Use a two-way ANOVA to analyze these data. Report the test statistics, degrees of freedom, and  $p$ -values for the significance tests.

**13.37** The PLANTS1 data set in the Data Appendix gives the percent of nitrogen in four different species of plants grown in a laboratory. *Leucaena leucocephala*, *Acacia saligna*, *Prosopis juliflora*, and *Eucalyptus citriodora*. The researchers who collected these data were interested in commercially growing these plants in parts of the country of Jordan where there is very little rainfall. To examine the effect of water, they varied the amount per day from 50 millimeters (mm) to 650 mm in 100 mm increments. There were nine plants per species-by-water combination. Because the plants are to be used primarily for animal food, with some parts that can be consumed by people, a high nitrogen content is very desirable.

- Find the means for each species-by-water combination. Plot these means versus water for the four species, connecting the means for each species by lines. Describe the overall pattern.
- Find the standard deviations for each species-by-water combination. Is it reasonable to pool the standard deviations for this problem? Note that with

sample sizes of size 9, we expect these standard deviations to be quite variable.

- (c) Run the two-way analysis of variance. Give the results of the hypothesis tests for the main effects and the interaction.

OBS	Species	Water	pctnit
001	1	1	3.644
002	1	1	3.500
003	1	1	3.509
004	1	1	3.137
005	1	1	3.100

**13.39** Refer to Exercise 13.37. Run a separate one-way analysis of variance for each water level. If there is evidence that the species are not all the same, use a multiple-comparisons procedure to determine which pairs of species are significantly different. In what way, if any, do the differences appear to vary by water level? Write a short summary of your conclusions.

**13.41** Refer to Exercise 13.37. Additional data collected by the same researchers according to a similar design are given in the PLANTS2 data set in the Data Appendix. Here there are two response variables. They are fresh biomass and dry biomass. High values for both of these variables are desirable. The same four species and seven levels of water are used for this experiment. Here, however, there are four plants per species-by-water combination. Analyze each of the response variables in the PLANTS2 data set using the outline from Exercise 13.37.

**13.43** Perform the tasks described in Exercise 13.39 for the two response variables in the PLANTS2 data set.

**13.47** Refer to the data given for the change-of-majors study in the data set MAJORS described in the Data Appendix. Analyze the data for HSE, the high school English grades. Your analysis should include a table of sample sizes, means, and standard deviations; Normal quantile plots; a plot of the means; and a two-way ANOVA using sex and major as the factors. Write a short summary of your conclusions.

**13.49** Refer to the data given for the change-of-majors study in the data set MAJORS described in the Data Appendix. Analyze the data for SATV, the SAT Verbal score. Your analysis should include a table of sample sizes, means, and standard deviations; Normal quantile plots; a plot of the means; and a two-way ANOVA using sex and major as the factors. Write a short summary of your conclusions.

## Chapter 14 Exercises

**14.1** If you deal one card from a standard deck, the probability that the card is a heart is 0.25. Find the odds of drawing a heart.

**14.3** A study was designed to compare two energy drink commercials. Each participant was shown two commercials, A and B, in random order and asked to select the better one. There were 100 women and 140 men who participated in the study. Commercial A was selected by 45 women and by 80 men. Find the odds of selecting Commercial A for the men. Do the same for the women.

**14.5** Refer to Exercise 14.3. Find the log odds for the men and the log odds for the women.

**14.7** Refer to Exercises 14.3 and 14.5. Find the logistic regression equation and the odds ratio.

**14.13** Refer to Exercise 14.11. Use  $x = 1$  for women and  $x = 0$  for men.

- (a) Find the estimates  $b_0$  and  $b_1$ .
- (b) Give the fitted logistic regression model.
- (c) What is the odds ratio for men versus women?

**14.15** Refer to Example 14.8. Suppose that you wanted to report a 99% confidence interval for  $\beta_1$  and its standard error, find the 95% confidence interval for the odds ratio and verify that this agrees with the interval given by the software.

**14.21** Different kinds of companies compensate their key employees in different ways. Established companies may pay higher salaries, while new companies may offer stock options that will be valuable if the company succeeds. Do high-tech companies tend to offer stock options more often than other companies? One study looked at a random sample of 200 companies. Of these, 91 were listed in the *Directory of Public High Technology Corporations*, and 109 were not listed. Treat these two groups as SRSs of high-tech and non-high-tech companies. Seventy-three of the high-tech companies and 75 of the non-high-tech companies offered incentive stock options to key employees.

- (a) What proportion of the high-tech companies offer stock options to their key employees? What are the odds?
- (b) What proportion of the non-high-tech companies offer stock options to their key employees? What are the odds?
- (c) Find the odds ratio using the odds for the high-tech companies in the numerator. Describe the result in a few sentences.

**14.23** Refer to Exercises 14.21 and 14.23. Software gives 0.3347 for the standard error of  $b_1$ .

- (a) Find the 95% confidence interval for  $\beta_1$ .
- (b) Transform your interval in (a) to a 95% confidence interval for the odds ratio.
- (c) What do you conclude?

**14.25** There is much evidence that high blood pressure is associated with increased risk of death from cardiovascular disease. A major study of this association examined 3338 men with high blood pressure and 2676 men with low blood pressure. During the period of the study, 21 men from the low-blood-pressure group and 55 in the high-blood-pressure group died from cardiovascular disease.

- (a) Find the proportion of men who died from cardiovascular disease in the high-blood-pressure group. Then calculate the odds.
- (b) Do the same for the low-blood-pressure group.
- (c) Now calculate the odds ratio with the odds for the high-blood-pressure group in the denominator. Describe the result in words.

**14.33** To devise effective marketing strategies it is helpful to know the characteristics of your customers. A study compared demographic characteristics of people who use the Internet for travel arrangements and of people who do not. Of 1132 Internet users, 643 had completed college. Among the 852 nonusers, 349 had completed college. Model the log odds of using the Internet to make travel arrangements with an indicator variable for having completed college as the explanatory variable. Summarize your findings.

**14.37** Refer to the previous exercise. Run the same analysis using Lactic as the explanatory variable.

**14.39** Use a logistic regression to predict HIGPA using the three high school grade summaries as explanatory variables.

- (a) Summarize the results of the hypothesis test that the coefficients for all three explanatory variables are zero.
- (b) Give the coefficient for high school math grades with a 95% confidence interval. Do the same for the two other predictors in this model.
- (c) Summarize your conclusions based on parts (a) and (b).

**14.41** Run a logistic regression to predict HIGPA using the three high school grade summaries and the two SAT scores as explanatory variables. We want to produce an analysis that is similar to that done for the case study in Chapter 11.

- (a) Test the null hypothesis that the coefficients of the three high school grade summaries are zero; that is, test  $H_0: \beta_{\text{HSM}} = \beta_{\text{HSS}} = \beta_{\text{HSE}} = 0$
- (b) Test the null hypothesis that the coefficients of the two SAT scores are zero; that is, test  $H_0: \beta_{\text{SATM}} = \beta_{\text{SATV}} = 0$ .
- (c) What do you conclude from the tests in (a) and (b)?



## Chapter 15 Exercises

**15.3** Refer to Exercise 15.1. State appropriate null and alternative hypotheses for this setting and calculate the value of  $W$ , the test statistic.

<b>Group A</b>	552	448	68	243	30
<b>Group B</b>	329	780	560	540	240

**15.5** Refer to Exercises 15.1 and 15.3. Find  $\mu_W$ ,  $\sigma_W$ , and the standardized rank sum statistic. Then give the approximate p-value using the Normal approximation. What do you conclude?

**15.7** A study of early childhood education asked kindergarten students to retell two fairy tales that had been read to them earlier in the week. The 10 children in the study included 5 high-progress readers and 5 low-progress readers. Each child told two stories. Story 1 had been read to them; Story 2 had been read and illustrated with pictures. An expert listened to a recording of the children and assigned a score for certain uses of language. Here are the data:

Child	Progress	Story 1 score	Story 2 score
1	High	.55	.80
2	High	.57	.82
3	High	.72	.54
4	High	.70	.79
5	High	.84	.89
6	Low	.40	.77
7	Low	.72	.49
8	Low	.00	.66
9	Low	.36	.28
10	Low	.55	.38

Is there evidence that the scores of high-progress readers are higher than those of low-progress readers when they retell a story they have heard without pictures (Story 1)?

- Make Normal quantile plots for 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?

**15.11** How quickly do synthetic fabrics such as polyester decay in landfills? A researcher buried polyester strips in the soil for different lengths of time, then dug up the strips and measured the force required to break them. Breaking strength is easy to measure and is a good indicator of decay. Lower strength means the fabric has decayed. Part of the study involved burying 10 polyester strips in well-drained soil in the summer. Five of the strips, chosen at random, were dug up after 2 weeks; the other 5 were dug up after 16 weeks. Here are the breaking strengths in pounds:

<b>2 weeks</b>	118	126	126	120	129
<b>16 weeks</b>	124	98	110	140	110

- Make a back-to-back stemplot. Does it appear reasonable to assume that the two distributions have the same shape?
- Is there evidence that the breaking strengths are lower for the strips buried longer?

**15.25** Can the full moon influence behavior? A study observed at nursing home patients with dementia. The number of incidents of aggressive behavior was recorded each day for 12 weeks. Call a day a “moon day” if it is the day of a full moon or the day before or after a full moon. Here are the average numbers of aggressive incidents for moon days and other days for each subject:

<b>Patient</b>	<b>Moon days</b>	<b>Other days</b>
1	3.33	0.27
2	3.67	0.59
3	2.67	0.32
4	3.33	0.19
5	3.33	1.26
6	3.67	0.11
7	4.67	0.30
8	2.67	0.40
9	6.00	1.59
10	4.33	0.60
11	3.33	0.65
12	0.67	0.69
13	1.33	1.26
14	0.33	0.23
15	2.00	0.38

The matched pairs  $t$  test (Example 7.7) gives  $P < 0.000015$  and a permutation test (Example 16.14) gives  $P = 0.0001$ . Does the Wilcoxon signed rank test, based on ranks rather than means, agree that there is strong evidence that there are more aggressive behaviors on moon days?

**15.29** How accurate are radon detectors of a type sold to homeowners? To answer this question, university researchers placed 12 detectors in a chamber that exposed them to 105 picocuries per liter (pCi/l) of radon. The detector readings are as follows:

91.9	97.8	111.4	122.3	105.4	95.0
103.8	99.6	96.6	119.3	104.8	101.7

We wonder if the median reading differs significantly from the true value 105.

- (a) Graph the data, and comment on skewness and outliers. A rank test is appropriate.
- (b) We would like to test the hypotheses about the median reading from home radon detectors:

$$H_0: \text{median} = 105$$

$$H_a: \text{median} \neq 105$$

To do this, apply the Wilcoxon signed rank statistic to the differences between the observations and 105. (This is the one-sample version of the test.) What do you conclude?

**15.31** Exercise 7.32 presents the data below on the weight gains (in kilograms) of adults who were fed an extra 1000 calories per day for 8 weeks.

- (a) Use a rank test to test the null hypothesis that the median weight gain is 16 pounds, as theory suggests. What do you conclude?

Subject	Before	After
1	55.7	61.7
2	54.9	58.8
3	59.6	66.0
4	62.3	66.2
5	74.2	79.0
6	75.6	82.3
7	70.7	74.3
8	53.3	59.3
9	73.3	79.1
10	63.4	66.0
11	68.1	73.4
12	73.7	76.9
13	91.7	93.1
14	55.9	63.0
15	61.7	68.2

16	57.8	60.3
----	------	------

**15.33** Many studies suggest that exercise causes bones to get stronger. One study examined the effect of jumping on the bone density of growing rats. Ten rats were assigned to each of three treatments: a 60-centimeter “high jump,” a 30-centimeter “low jump,” and a control group with no jumping. Here are the bone densities (in milligrams per cubic centimeter) after 8 weeks of 10 jumps per day:

Group	Bone density (mg/cm <sup>3</sup> )									
Control	611	621	614	593	593	653	600	554	603	569
Low jump	635	605	638	594	599	632	631	588	607	596
High jump	650	622	626	626	631	622	643	674	643	650

- (c) Do the Kruskal-Wallis test. Explain the distinction between the hypotheses tested by Kruskal-Wallis and ANOVA.

**15.43** Different varieties of the tropical flower *Heliconia* are fertilized by different species of hummingbirds. Over time, the lengths of the flowers and the form of the hummingbirds’ beaks have evolved to match each other. Here are data on the lengths in millimeters of three varieties of these flowers on the island of Dominica:

<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	37.10	35.17	36.82	36.66	35.68
36.03	34.57	34.63			

Do a complete analysis that includes description of the data and a rank test for the significance of the differences in lengths among the three species.

**15.47** As in ANOVA, we often want to carry out a **multiple-comparisons** procedure following a Kruskal-Wallis test to tell us *which* groups differ significantly.<sup>29</sup> Here is a simple method: If we carry out  $k$  tests at fixed significance level  $0.05/k$ , the probability of *any* false rejection among the  $k$  tests is always no greater than 0.05. That is, to get overall significance level 0.05 for all of  $k$  comparisons, do each individual comparison at the  $0.05/k$  level. In Exercise 15.43 you found a significant difference among the lengths of three varieties of the flower *Heliconia*. Now we will explore multiple comparisons.

- (a) Write down all of the pairwise comparisons we can make, for example, *bihai* versus *caribaea* red. There are three possible pairwise comparisons.
- (b) Carry out three Wilcoxon rank sum tests, one for each of the three pairs of flower varieties. What are the three two-sided  $P$ -values?
- (c) For purposes of multiple comparisons, any of these three tests is significant if its  $P$ -value is no greater than  $0.05/3 = 0.0167$ . Which pairs differ significantly at the overall 0.05 level?

## Chapter 16 Exercises

*In order to solve problems in Chapter 16, a JMP script is needed. You can download this script on the IPS 6e Book Companion Site at [www.whfreeman.com/ips6e](http://www.whfreeman.com/ips6e).*

**16.11** Here is an SRS of 20 of the guinea pig survival times from Exercise 16.10:

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

We expect the sampling distribution of  $\bar{x}$  to be less close to Normal for samples of size 20 than for samples of size 72 from a skewed distribution. These data include some extreme high outliers.

- Create and inspect the bootstrap distribution of the sample mean for these data using 1000 resamples. Is it less close to Normal than your distribution from the previous exercise?
- Compare the bootstrap standard errors for your two runs. What accounts for the larger standard error for the smaller sample?

**16.13** Return to or create the bootstrap distribution resamples on the sample mean for the audio file lengths in Exercise 16.8. In Example 7.11, the  $t$  confidence interval for the average length was constructed.

- Inspect the bootstrap distribution. Is a bootstrap  $t$  confidence interval appropriate? Explain why or why not.
- Construct the 95% bootstrap  $t$  confidence interval.
- Compare the bootstrap results with the  $t$  confidence interval reported in Example 7.11.

**16.16** Return to or re-create the bootstrap distribution of the sample mean for the 8 listening times in Exercise 16.6.

- Although the sample is small, verify using graphs and numerical summaries of the bootstrap distribution that the distribution is reasonably Normal and that the bias is small relative to the observed  $\bar{x}$ .
- The bootstrap  $t$  confidence interval for the population mean  $\mu$  is therefore justified. Give the 95% bootstrap  $t$  confidence interval for  $\mu$ .
- Give the usual  $t$  95% interval and compare it with your interval from (b).

**16.19** For Example 16.5 we bootstrapped the 25% trimmed mean of the 50 selling prices in Table 16.1. Another statistic whose sampling distribution is unknown to us is the standard deviation  $s$ . Bootstrap  $s$  for these data. Discuss the shape and bias of the bootstrap distribution. Is the bootstrap  $t$  confidence interval for the population standard

deviation  $\sigma$  justified? If it is, give a 95% confidence interval.

Table 16.1 - Selling prices for Seattle real estate, 2002 (\$1000s)									
142	175	197.5	149.4	705	232	50	146.5	155	1850
132.5	215	116.7	244.9	290	200	260	449.9	66.407	164.95
362	307	266	166	375	244.95	210.95	265	296	335
335	1370	256	148.5	987.5	324.5	215.5	684.5	270	330
222	179.8	257	252.95	149.95	225	217	570	507	190

**16.39** The distribution of the 72 guinea pig survival times in Table 1.8 (page 29) is strongly skewed. In Exercise 16.17 (page 16-22) you found a bootstrap  $t$  confidence interval for the population mean  $\mu$ , even though some skewness remains in the bootstrap distribution. Bootstrap the mean lifetime and give all four bootstrap 95% confidence intervals:  $t$ , percentile, BCa, and tilting. Make a graphical comparison by drawing a vertical line at the original sample mean  $\bar{x}$  and displaying the four intervals horizontally, one above the other. Discuss what you see. Do bootstrap  $t$  and percentile agree? Do the more accurate intervals agree with the two simpler methods?

Table 1.8 - Survival times (days) of Guinea pigs in a medical experiment									
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	121	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								

**16.41** Exercise 16.11 (page 16-12) gives an SRS of 20 of the 72 guinea pig survival times in Table 1.8. The bootstrap distribution of  $\bar{x}$  from this sample is clearly right-skewed. Give a 95% confidence interval for the population mean  $\mu$  based on these data and a method of your choice. Describe carefully how your result differs from the intervals in Exercise 16.39, which use the full sample of 72 survival times.

## Chapter 17 Exercises

**17.13** A meat-packaging company produces 1-pound packages of ground beef by having a machine slice a long circular cylinder of ground beef as it passes through the machine. The timing between consecutive cuts will alter the weight of each section. Table 17.3, reproduced below, gives the weight of 3 consecutive sections of ground beef taken each hour over two 10-hour days. Past experience indicates that the process mean is 1.03 and the weight varies with  $\sigma = 0.02$  lb.

Sample	Weight (pounds)			$\bar{x}$	$s$
1	0.999	1.071	1.019	1.030	0.0373
2	1.030	1.057	1.040	1.043	0.0137
3	1.024	1.020	1.041	1.028	0.0108
4	1.005	1.026	1.039	1.023	0.0172
5	1.031	0.995	1.005	1.010	0.0185
6	1.020	1.009	1.059	1.029	0.0263
7	1.019	1.048	1.050	1.039	0.0176
8	1.005	1.003	1.047	1.018	0.0247
9	1.019	1.034	1.051	1.035	0.0159
10	1.045	1.060	1.041	1.049	0.0098
11	1.007	1.046	1.014	1.022	0.0207
12	1.058	1.038	1.057	1.051	0.0112
13	1.006	1.056	1.056	1.039	0.0289
14	1.036	1.026	1.028	1.030	0.0056
15	1.044	0.986	1.058	1.029	0.0382
16	1.019	1.003	1.057	1.026	0.0279
17	1.023	0.998	1.054	1.025	0.0281
18	0.992	1.000	1.067	1.020	0.0414
19	1.029	1.064	0.995	1.029	0.0344
20	1.008	1.040	1.021	1.023	0.0159

- Calculate the center line and control limits for an  $\bar{x}$  chart.
- What are the center line and control limits for an  $s$  chart for this process?
- Create the  $\bar{x}$  and  $s$  charts for these 20 consecutive samples.
- Does the process appear to be in control? Explain.

**17.16** Table 17.5 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  $\mu = 11.5$  and standard deviation  $\sigma = 0.2$ .

- Make both  $\bar{x}$  and  $s$  charts for these data based on the information given about the process.



(b) At some point, the within-sample process variation increased from  $\sigma = 0.2$  to  $\sigma = 0.4$ . About where in the 20 samples did this happen? What is the effect on the  $s$  chart? On the  $\bar{x}$  chart?

(c) At that same point, the process mean changed from  $\mu = 11.5$  to  $\mu = 11.7$ . What is the effect of this change on the  $s$  chart? On the  $\bar{x}$  chart?

**17.19** Figure 17.10 reproduces a data sheet from the floor of a factory that makes electrical meters. The sheet shows measurements of the distance between two mounting holes for 18 samples of size 5. The heading informs us that the measurements are in multiples of 0.0001 inch above 0.6000 inch. That is, the first measurement, 44, stands for 0.6044 inch. All the measurements end in 4. Although we don't know why this is true, it is clear that in effect the measurements were made to the nearest 0.001 inch, not to the nearest 0.0001 inch.

Calculate  $\bar{x}$  and  $s$  for the first two samples. The data file *ex17\_19* contains  $\bar{x}$  and  $s$  for all 18 samples. Based on long experience with this process, you are keeping control charts based on  $\mu = 43$  and  $\sigma = 12.74$ . Make  $s$  and  $\bar{x}$  charts for the data in Figure 17.10 and describe the state of the process.

**17.31** The  $\bar{x}$  and  $s$  control charts for the mesh-tensioning example (Figures 17.4 and 17.7) were based on  $\mu = 275$  mV and  $\sigma = 43$  mV. Table 17.1 gives the 20 most recent samples from this process.

(a) Estimate the process  $\mu$  and  $\sigma$  based on these 20 samples.

(b) Your calculations suggest that the process  $\sigma$  may now be less than 43 mV. Explain why the  $s$  chart in Figure 17.7 (page 17-15) suggests the same conclusion. (If this pattern continues, we would eventually update the value of  $\sigma$  used for control limits.)

**17.39** Do the losses on the 120 individual patients in Table 17.7 appear to come from a single Normal distribution? Make a Normal quantile plot and discuss what it shows. Are the natural tolerances you found in the Exercise 17.34 trustworthy?

**17.43** Make a Normal quantile plot of the 85 distances in data file *ex17\_19* that remain after removing sample 5. How does the plot reflect the limited precision of the measurements (all of which end in 4)? Is there any departure from Normality that would lead you to discard your conclusions from Exercise 17.39?

**17.53** Table 17.1 gives 20 process control samples of the mesh tension of computer monitors. In Example 17.13, we estimated from these samples that  $\hat{\mu} = \bar{\bar{x}} = 275.065$  mV and  $\hat{\sigma} = s = 38.38$  mV.

(a) The original specifications for mesh tension were LSL = 100 mV and USL = 400 mV. Estimate  $C_p$  and  $C_{pk}$  for this process.

- (b) A major customer tightened the specifications to  $LSL = 150$  mV and  $USL = 350$  mV. Now what are  $C_p$  and  $C_{pk}$ ?

**17.75** An egg farm wants to monitor the effects of some new handling procedures on the percent of eggs arriving at the packaging center with cracked or broken shells. In the past, roughly 2% of the eggs were damaged. A machine will allow the farm to inspect 500 eggs per hour. What are the initial center line and control limits for a chart of the hourly percent of damaged eggs?