Introduction to JMP INTRO

This manual accompanies Statistics and Data Analysis by Peck, Olsen, and Devore. It is intended to be used in conjunction with the text, so each chapter of this book corresponds to a chapter in the main text. You’ll find examples from each chapter worked out here, intended to show you how to use JMP INTRO for all the problems in the text.

This book is not intended to be a complete user’s guide to JMP INTRO. If you have questions about specific capabilities of JMP INTRO, refer to the online help or the User's Guide to JMP INTRO, a PDF installed by default with the JMP INTRO package.

About JMP INTRO

JMP INTRO is the version of SAS Institute’s award-winning JMP Statistical Discovery software tailor-made for the introductory statistics student. JMP INTRO is easy to learn and easy to use. All of the statistics are accessible in a familiar, point-and-click format, and the statistical concepts are supported with both graphs and appropriate numerical results. In addition, all the data tables, graphs, and charts are dynamically linked together, allowing for interactive exploration of patterns and outliers, whenever they present themselves. It is hoped that this visualization makes learning statistics more fun—and easier—than it has ever been before.

You may be using another product from the JMP family, JMP Professional or JMP IN. The things you read in this manual are equally applicable to those products. There may be a slight difference in the menus, but the commands you need are identical in all the products.

Statistical Analyses in JMP INTRO

JMP INTRO is different from most other statistical packages. It was designed for the statistical consumer, rather than the statistician. To use it efficiently, you need to know three big ideas. Learn them and the entire product makes sense.

Idea 1: Think about the number of variables in the analysis, not the name of the analysis.

JMP is designed for the statistical consumer rather than the statistician. Therefore, some of the technical names are softened a bit, so that you can think more about the context of the problem rather than the names of statistical methods.

To successfully analyze data in JMP, think only of how many variables you are analyzing.

- If you are analyzing one variable, use the Distribution command.
4 \ Introduction to JMP INTRO

- If you are analyzing two variables, use the \textit{Fit Y by X} command.
- If you are analyzing more than two variables, use the \textit{Fit Model} command.

This structure is deceptively simple, and its simplicity often takes people by surprise. After all, there are several chapters in every statistics book on one-variable analyses (histograms, testing means, etc.), and there are even more chapters on two-variable analysis (linear regression, two-by-two tables, ANOVA, logistic regression). Many people expect to find a distinct menu item for each of these methods. JMP reduces this complex list to a few commands.

\textbf{Idea 2: Tell JMP what kind of data is in each column, and appropriate graphs and statistics appear automatically.}

In \textit{Statistics and Data Analysis}, you will see methods for dealing with several types of data. The data type often determines what analysis is appropriate. As an example, consider the statistical concept of the \textit{average} (called the \textit{mean} by statistical types). Undoubtedly, you have encountered the concept of an average in your schooling so far.

The average is an appropriate measure for variables looked at individually—but only certain kinds. Grades, for example, can be averaged because they are measured on a continuous scale. Temperatures, too, are continuous, so finding an average temperature makes sense. However, you may have a variable that represents gender, which can have the values male or female. It doesn't make sense to find the “average” gender. Averages are not an appropriate statistic for these so-called categorical variables.

By specifying the type of a variables, JMP can discern when an average is appropriate. When you tell JMP to analyze a single variable, it produces an average when the variable is continuous, and it doesn't when the variable is not continuous. Although this is a simple example (who would even try to compute the average gender?), many more complicated analyses are not so obvious. JMP takes this all into account and only computes statistics that are appropriate.

Here is a sample JMP data set, named \textit{Denim.jmp}, showing two different variable types. (This data table is installed by default with JMP INTRO. Look for a folder called \textit{Sample Data}).

\begin{verbatim}
Continuous variables are numeric and measured on a continuous scale. For example, temperature measurements are often on a continuous scale, limited only by the exactness of the measuring instrument. In this example, the variables \textit{Size of Load} and \textit{Starch Content} are continuous variables.

Ordinal variables are measured on a discrete scale. There is an implicit order in the measuring scale, although the results are not necessarily numerical. For example, the age of people is often recorded on an ordinal scale—seldom do people report their age as 24.461 years, yet it is obvious that some ages are
\end{verbatim}
older than others. In the Denim.jmp data set, Thread Wear, with values “low”, “moderate”, and “severe”, is an ordinal variable.

Nominal variables simply name data. There is no order in the scale. People's names, for example, are represented as a nominal variable in JMP INTRO. Method is a nominal variable in the Denim.jmp data set. Ordinal and Nominal variables are often referred to collectively as categorical variables.

The modeling type of a variable is specified in the panel on the left of the data table. Continuous variables have a blue C, ordinal variables a green O, and nominal variables a red N.

To change the variable type, click on the C, N, or O to reveal a menu. Select the variable type from the menu.

To see the effect of the modeling type, we present a simple one-variable analysis. As stated in Idea 1 above, single-variable analyses are accomplished through the Distribution command.

Select Analyze > Distribution.

In the dialog that appears,

Select Size of Load and press the Y, Columns button.

This is called specifying the role of the variable.
A report with appropriate results appears. Since Size of Load is a continuous variable, you see an appropriate graphic (called a histogram), and appropriate statistics. The mean (average) is appropriate, and you see it in the Moments section of the report below the histogram.

Repeat the analysis with a nominal variable. Repeat exactly the same steps, but use the Method variable instead.

Select Analyze > Distribution.
In the dialog that appears, select Method and press the Y, Columns button.

Click OK.

Note the difference in the results. Since the Method variable is nominal, averages are not appropriate. The report no longer contains a section giving the mean. It does, however, show frequencies—the number of observations in each group, which is an appropriate statistic for this type of variable.

**Idea 3: To complete some analyses, launch the appropriate platform, then look for red triangles.**

Red triangles in JMP signify actions. When you click on a red triangle, a menu appears that lets you choose more options. In general, these options allow you to specify further analysis options, letting you dig deeper and explore other statistics as you wish.

As an example,

Click on the red triangle in the Method analysis you’ve just completed.
You now see options that are appropriate for this analysis. For example, you can turn the mosaic plot off. Or, you can compute confidence intervals, or do a test on the probabilities. (Don’t worry if these terms are unfamiliar—they’ll become familiar as you progress through the book.)

The important thing to remember is the general steps that are needed to complete an analysis.

- Select an analysis platform
- In the launch dialog, specify the roles of the variables
- Look for red triangles in the resulting report to further the analysis.

**Summary**

That’s all there is to it. Keeping these three big ideas in mind, you’ll be able to use JMP like a professional.

Idea 1: Think about the number of variables in the analysis, not the name of the analysis.

Idea 2: Tell JMP what kind of data is in each column, and appropriate graphs and statistics appear automatically.

Idea 3: To complete an analysis, launch the appropriate platform, then look for red triangles.
This chapter introduces you to some techniques for making sense of data. Summary tools include a dot plot and a bar chart.

In future chapters, you use JMP INTRO’s built-in analysis functions to conduct analyses of data. This chapter is an exception, because you use a built-in graphing function for one example and a script for the other.

**Example 1.11**

**Why Students Drop Out**

The article “So Close, yet So Far: Predictors of Attrition in College Seniors” (*J. College Student Development* (1998): 343-348) examined the reasons that college seniors leave their college programs before graduating. Forty-two college seniors at a large public university who dropped out prior to graduation were interviewed and asked the main reason for discontinuing enrolment at the university. Data consistent with that given in the article is summarized in the accompanying frequency distribution.

We want to produce a bar chart of this data.
- Select **Graph > Chart**.
- Select **Reason** in the column list and click the **X, Level** button.
- Select **Frequency** in the columns list and click the **Statistics** drop-down button.
- Select **Data** from the list of statistics that appears.
Click OK.

This produces the bar chart shown here.
Example 1.12

Graduation Rates for NCAA Division I Schools in California and Texas

The *Chronicle of Higher Education* (Almanac Issue, Aug. 31, 2001) reported graduation rates for NCAA Division I schools. The rates are the percent of full time freshmen in fall 1993 who had earned a bachelor's degree by August 1999. Data from the two largest states, the 20 Division I schools in California and the 19 in Texas, is given here.

We want to produce a dot plot of this data. Dot plots are not a built-in function of JMP (which uses histograms and stem-and-leaf plots, which you learn about in later chapters). Even though the function is not built in, JMP can still display the graph by running a script (provided with the book).

To open a script,

- Select Open Script from the JMP Starter.
- Navigate to the folder containing the `utilDotPlot.jsl` script and click Open.

The script opens in a text window.
To run the script,

远景 Select Edit > Run Script.

远景 The script brings up a dialog asking about the variables you want to plot.

远景 Click on Rate in the columns list and click the Y, Response button.

远景 Click OK.

远景 This produces the desired dot plot.

远景 To produce dot plots of separate schools,

远景 Open the script utilStackedDotPlot.jsl.

远景 The script opens in a text window.

远景 Select Edit > Run Script.

远景 This brings up the Select Columns dialog.
Assign Rate to Y, Response and State to Group.
Click OK.
The produces two stacked dot plots.
Most of your work in this chapter illustrates good statistical practice, not just the steps on using technology. Even though technology can be extremely important in analyzing a study, it cannot take the place of good planning and data collection. Software may make flawless computations, but if the data is not sound, the analyses are not either. This chapter shows you how to use JMP to gather data in a statistically sound way—a crucial step in any experiment or study.

Example 2.3

Selecting a Random Sample of Glass Soda Bottles

Breaking strength is an important characteristic of glass soda bottles. If the strength is too low, a bottle may burst—not a desirable outcome. Suppose that we want to measure the breaking strength of each bottle in a random sample of size $n = 3$ selected from four crates containing a total of 100 bottles (the population). Each crate contains five rows of five bottles each. We can identify each bottle with a number from 1 to 100 by numbering across the rows, starting with the top of crate 1. Use JMP to randomly sample five of the bottles.

The JMP Formula Editor

Any column in a JMP data table can hold a formula. Formulas are used when the contents of the column are calculated from other quantities. Frequently, these calculations are done using other columns of the data table—summing them, averaging them, or transforming them. In addition, there are several random number functions that can be used to designate samples from a larger population.

To select the random sample of five bottles, complete the following steps.

1. Create a new data table. On Windows or Linux, select File > New > Data Table. On Macintosh, select File > New Table.

Alternatively, you can select New Data Table from the File tab on the JMP Starter.
A new data table opens with a single column and no rows.

Rename the existing Column 1 name to a more meaningful name, like Sample Number.

- Click once in the header of the column (which highlights it). Then, type the name Sample Number and hit Enter.

- Add five rows to the table by selecting Rows > Add Rows and typing 5 into the dialog box that appears.

- Right-click (Windows and Linux) or Control-click (Macintosh) on the column header (the area around the name) and select Formula from the menu that appears.
This opens the Formula Editor, shown here.

Now, enter a formula to choose a random integer between 1 and 100.

- In the groups of functions, use the scroll bar to scroll down the list until you see Random.
- Click on Random, then select Random Integer from the menu that appears.
Type 100 as the argument to the Random Integer function.
Click OK.

Your data table should now fill with random integers, pulled from a uniform distribution of integers from 1 to 100. Since the column named Sample Number is determined by a formula, a formula icon appears in the columns list to the left of the data grid.

If you want to re-sample the numbers (that is, generate a new set of five random numbers),
Right-click on the column header and select Formula.
Click the Apply button.
Graphical Methods for Describing Data

JMP teems with graphics. In fact, one of the design goals of JMP was to have a graphic with every statistic. Seeing data in a graphical way (we think) is that important. Therefore, most of the time, graphs appear with your analyses automatically, without you having to ask for them. In all other cases, graphs are readily available from the report surface.

Example 3.1

Perceived Risk of Smoking

The article “Most Smokers Wish They Could Quit” (Gallup Poll Analyses, November 21, 2002) noted that smokers and nonsmokers perceive the risks of smoking differently. The accompanying relative frequency table summarizes responses regarding the perceived harm of smoking for each of three groups—a sample of 241 smokers, a sample of 261 former smokers, and a sample of 502 non-smokers. To analyze this data, we wish to construct a comparative bar chart.

Table 3.1 Perceived Risk of Smoking

<table>
<thead>
<tr>
<th>Perceived Risk of Smoking</th>
<th>Smokers</th>
<th>Former Smokers</th>
<th>Non Smokers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Very Harmful</td>
<td>0.6</td>
<td>0.78</td>
<td>0.86</td>
</tr>
<tr>
<td>Somewhat Harmful</td>
<td>0.3</td>
<td>0.16</td>
<td>0.1</td>
</tr>
<tr>
<td>Not Too Harmful</td>
<td>0.07</td>
<td>0.04</td>
<td>0.03</td>
</tr>
<tr>
<td>Not Harmful At All</td>
<td>0.03</td>
<td>0.02</td>
<td>0.01</td>
</tr>
</tbody>
</table>

In JMP, you have to enter data from tables like this—so-called two-way tables—in a row-wise format. There are twelve cells in the table, so you need twelve rows in the JMP data table. You need one column to hold the perceived risks, one to hold the smoking status, and one to hold the actual frequency data.

- Open a new JMP data table. (Remember, you can do this using the File menu or the JMP Starter).
- Add 12 rows using **Rows > Add Rows**.
Add 3 columns using Cols > Add Multiple Columns. Put 3 as the number of columns to add.

Name the three columns Perceived Risk of Smoking, Type, and Frequency.

Enter the data as shown in the following data table.

Note: When you created a new data table, all three columns were designated as continuous. However, two of these three columns contain categorical data. When you start entering the data, you’ll be presented with a dialog box asking about changing the format or length of the column. Select Change in each case, and all will be well.
We can now produce the bar chart.

Select **Graph > Chart**.

Click on **Frequency** in the columns list to select it.

Click the **Statistics** button and select Data from the menu that appears.

Select **Type** in the columns list and click **X, Level**.

Select **Perceived Risk of Smoking** in the columns list and select **X, Level**.

The dialog should look like this.

Click **OK**.

The chart appears.
Graphical Methods for Describing Data

Note that the order of the variables entered in the X,Level role is important. If you entered Perceived Risk of Smoking first and Type second, the graph would look like this:
This shows that the first X, Level variable delineates the individual bars. The second groups the bars together.

Example 3.4

Birds That “Fish”

Night herons and cattle egrets are species of birds that feed on aquatic prey in shallow water. These birds stalk submerged prey while wading in shallow water, and then strike rapidly downward through the water in an attempt to catch the prey. The article “Cattle Egrets Are Less Able to Cope with Light Refraction Than Are Other Herons” (Animal Behaviour (1999): 687-94) gave data on outcome when 240 cattle egrets attempted to capture submerged prey. The data is summarized in the accompanying frequency distribution.

Table 3.2 Capture Attempts

<table>
<thead>
<tr>
<th>Outcome</th>
<th>Frequency</th>
<th>Relative Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prey caught on first attempt</td>
<td>103</td>
<td>0.43</td>
</tr>
<tr>
<td>Prey caught on second attempt</td>
<td>41</td>
<td>0.17</td>
</tr>
<tr>
<td>Prey caught on third attempt</td>
<td>2</td>
<td>0.01</td>
</tr>
</tbody>
</table>
We analyze this data using a bar chart. First, enter the data into a JMP data table using the following steps. Note that we only enter the Outcome and Frequency—the Relative Frequency can be calculated using a formula if it's needed.

- Select **File > New > Table** (Windows and Linux) or **File > New** (Macintosh)
- Double-click to the right of Column 1 to add a second column.
- Click in the column header for Column 1 and type **Outcome**.
- Press the Tab key to move to Column 2's name.
- Type **Frequency** as the name for the second column.

To produce the pie chart,

- Click on the icon beside **Outcome** in the Columns panel (on the left of the data table, circled in the picture above) and select **Nominal** from the menu that appears.
- Select **Rows > Add Rows** to add four rows to the data table.
- Type in the data as shown in the above table.
- To produce the pie chart,
  - Select **Graph > Chart**.
  - Click on **Frequency** in the columns list, click the **Statistics** drop-down, and select **Data**.
  - Add **Outcome** to the **X, Level** role.
  - In the Options panel, change **Vertical** to **Pie**.

<table>
<thead>
<tr>
<th>Outcome</th>
<th>Frequency</th>
<th>Relative Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prey not caught</td>
<td>94</td>
<td>0.39</td>
</tr>
</tbody>
</table>

Table 3.2 Capture Attempts
Click **OK**.

This produces the pie chart shown here.

Click on the slices of the chart.
This causes the number of data points in each slice to appear.
You can quickly change this to a bar chart by selecting **Vertical** or **Horizontal** from the platform menu.
Example 3.8

Binge Drinking

The use of alcohol by college students is of great concern, not only to those in the academic community, but also because of potential health and safety consequences to society at large. The article “Health and Behavioral Consequences of Binge Drinking in College” (J. of the Amer. Med. Assoc. (1994): 1672-1677) reported on a comprehensive study of heavy drinking on campuses across the country. A binge episode was defined as five or more drinks in a row for males and four or more for females. A portion of this data is shown here, with the variable name Episodes.
The analysis for this data includes a Stem-and-Leaf plot. Stem-and-Leaf plots are found in the Distribution platform, which is the typical platform used for analyzing a single variable.

1. Select Analyze > Distribution.
2. Select Episodes in the columns list and click Y, Columns.

3. Click OK.

When the Distribution report shows,
4. Select Stem and Leaf from the drop-down menu beside Episodes.

This appends a stem-and-leaf plot to the report, as shown here.
Graphical Methods for Describing Data

Note: as with many displays in JMP, the stem-and-leaf report is “live”. You can click on a leaf in the report, and the corresponding data point highlights in the data table and all other open displays.

Example 3.17

Mercury Contamination

Mercury contamination is a serious environmental concern. Mercury levels are particularly high in certain types of fish. Citizens of the Republic of Seychelles, a group of islands in the Indian Ocean, are among those who consume the most fish in the world. The article “Mercury Content of Commercially Important Fish of the Seychelles, and Hair Mercury Levels of a Selected Part of the Population” (Environ Research (1983): 305-312) reported observations on mercury content (ppm) in the hair of 40 fishermen. A portion of the data is shown here.

To produce a histogram of this data, use the Distribution platform.

Select Analyze > Distribution.

Assign Mercury Content as the Y, Columns variable.

Click OK.

This produces the histogram and reports shown here.
Some people prefer to see the histogram and reports in a horizontal format. To change the orientation, select **Display Options > Horizontal** from the drop-down menu beside Mercury Content.
The layout changes to the appearance shown here.

If you want to change the width of the histogram bars,
- Select the Hand tool ( ) from the toolbar or Tools menu.
- Click and drag in a direction perpendicular to the axis to make the bars wider and narrower.
The growth and decline of forests is a matter of great public and scientific interest. The article "Relationships Among Crown Condition, Growth, and Stand Nutrition in Seven Northern Vermont Sugarbushes" (Canad. J. of Forest Res. (1995): 386-397) included a scatter plot of mean crown dieback (%), which is one indicator of growth retardation, and soil pH (higher pH corresponds to less acidic soil). The observations are shown in a JMP data table here.
We want to produce a scatter plot of these data. Since we are analyzing two variables, use the Fit Y by X platform.

- Select **Analyze > Fit Y by X**.

- Assign **soil pH** as the X, **Factor** variable and **mean crown dieback** as the Y, **Response**.

The dialog should look like this:

![Fit Y by X dialog](image)

- Click **OK**.

A scatterplot appears.

![Scatterplot](image)

For reports or presentations, some people prefer larger markers than the default dots shown here. To make the markers larger,

- Right click in the plot and select **Marker Size > Large** from the menu that appears.
This makes the dots larger.
Numerical Methods for Describing Data

A common task in a statistical analysis is to find “typical” values of a variable. The objective is often to summarize the data in a single number, or group of numbers, to describe the location and spread of the data. This chapter shows you how to compute some typical summary statistics: the mean, standard deviation, and quartiles. It also illustrates a common statistical graph, the box plot. All of these concepts are accessed through JMP’s Distribution command.

Example 4.3

Number of Visits to a Class Website

Forty students were enrolled in a section of STAT 130, a general education course in statistical reasoning, during Fall quarter 2002 at Cal Poly, San Luis Obispo. The instructor made course materials, grades, and lecture notes available to students on a class website, and course management software kept track of how often each student accessed any of the web pages on the class site. One month after the course began, the instructor requested a report that indicated how many times each student had accessed a web page on the class site. There were 40 observations, a subset of which are shown in a JMP data table here.

Note that the Number of Visits variable is designated as a continuous variable. By doing so, we assure that the Distribution platform gives appropriate statistics—means, medians, standard deviations, histograms, and box plots. This example focuses on the mean and the histogram.
Numerical Methods for Describing Data

Select Analyze > Distribution.

Assign Number of Visits as the Y, Columns variable.

Click OK.

To make the plot appear horizontal (not necessary, but it saves some page space here),

Select Display Options > Horizontal Layout from the red triangle menu next to Number of Visits.

The sample mean is 23.1. However, some statisticians would argue that the mean is not an appropriate measure of center for this data set, since there are two large outliers. The outliers are easy to see on the box plot above the histogram. Therefore, some would prefer to use the median (shown in Example 4.4 of your textbook). In this example, the median is 13.

Example 4.8

Acrylamide Levels in French Fries

Research by the FDA shows that acrylamide (a possible cancer-causing substance) forms in high carbohydrate foods cooked at high temperatures and that acrylamide level can vary widely even within the same brand of food. (Associated Press, Dec. 6, 2002). FDA scientists analyzed McDonald’s French Fries purchased at seven different locations and found the following acrylamide levels, shown here as a JMP data table.
We wish to compute the mean and standard deviation of this data set. In JMP, this is accomplished by using the Distribution platform.

1. Select Analyze > Distribution.
2. Assign Acrylamide Level as the Y, Columns variable.
3. Click OK.

The resulting report shows the mean and standard deviation in the Moments section of the report.

If needed, the median, lower quartile, and upper quartile are shown in the Quantiles section of the report.
Example 4.11

Golden Rectangles

The accompanying data come from an anthropological study of rectangular shapes (Lowie's Selected Papers in Anthropology, Cora Dubois, Ed., Berkeley, Calif: Univ. of Calif. Press, 1960: 137-142). Observations were made on the variable width/length for a sample of \( n = 20 \) beaded rectangles used in Shoshoni Indian leather handicrafts.

To examine this data, we want to construct a box plot. This is produced automatically (for continuous variables) in the Distribution platform.

- Select **Analyze > Distribution**.
- Assign **width/length** as the Y, **Columns** variable.
- Click **OK**.

The default distribution report is shown here, with the box plot (called an *Outlier Box Plot* in JMP) showing to the right of the histogram.
In some cases, you don’t want to see the histogram, but only the box plot. You can further customize the report by (for example), turning off the Quantiles and Moments outlines, and making the report appear horizontal.

- Select Histogram Options > Histogram from the drop-down menu next to the width/length variable name.
- Select Horizontal Layout, Quantiles, and Moments (in any order) from the Display Options menu found on the drop-down menu next to the width/length variable name.

This produces the simple report shown here. This is often useful when you have several variables to compare.
Summarizing Bivariate Data

Chapter 4 showed you how to summarize a single variable. This chapter goes one step further, showing you how to summarize two variables (a.k.a. bivariate data). We examine correlation, regression lines, and residual plots.
Since we are dealing with two variables, we generally use the Fit Y by X platform.

Example 5.4

Is Foal Weight related to Mare Weight?

Foal weight is an indicator of health, so it is of interest to breeders of thoroughbred horses. Is foal weight related to the weight of the mare (mother)? The accompanying data are from the article “Suckling Behavior Does Not Measure Milk Intake in Horses” (Animal Behaviour (1999): 673-68).

We want to use JMP to calculate the correlation coefficient for this data.
- Select Analyze > Fit Y by X.
- Assign Foal Weight as Y, Response and Mare Weight as X, Factor.

The dialog should appear like this.
Click OK.

When the scatterplot appears,

Select Density Ellipse > .90 from the drop-down menu at the top of the report.

(We picked 0.90 arbitrarily. Choosing any of the density ellipses result in the appearance of the Correlation report shown below.)

This reveals a Correlation report, although it is initially closed. To open the report,

Click the triangle next to the Correlation outline.

This reveals the complete report, shown here.
You can now read off the correlation, 0.001348.

Example 5.6

Time to Defibrillator Shock and Heart Attack Survival Rate

Studies have shown that people who suffer sudden cardiac arrest (SCA) have a better chance of survival if a defibrillator shock is administered very soon after cardiac arrest. How is survival rate related to the time between when cardiac arrest occurs and when the defibrillator shock is delivered? This question is addressed in the paper “Improving Survival from Sudden Cardiac Arrest—The Role of Home Defibrillators” (University of Michigan, Feb. 2002). The accompanying data (shown in a JMP data table) gives survival rate (percent) and mean call-to-shock time (minutes) for a cardiac rehabilitation center (where cardiac arrests occurred while victims were hospitalized and so the call-to-shock times tend to be short) and for four communities of different sizes.

We want to compute the least-squares regression line for this data.

- Select Analyze > Fit Y by X.
- Assign Call-to-shock time as X, Factor and Survival Rate as Y, Column.
- Click OK.
5 Summarizing Bivariate Data

This reveals the scatter plot shown here.

A line seems to be a reasonable model, so
Select **Fit Line** from the platform drop-down menu.
This adds the regression line to the graph, and gives estimates for the slope and intercept in the Linear Fit table below.

To predict the survival time at $x = 5$, there are two options.
Using the crosshair tool

- Select the crosshair tool ( ) from the toolbar or Tools menu.
- Click and hold the crosshair over the graph, moving it to correspond to the point on the line where $x = 5$.
- Read the value for $y$ off the graph.

Using a formula

You can save the prediction formula back to the data table, then add a data point corresponding to $x = 5$.

- Using the menu beside Linear Fit, select Save Predicteds.
This adds a column to the data table named *Predicted Survival Rate*. Note that there is a formula icon next to the column's name in the columns panel to the left of the data grid.

To add a row to the data table,

- Click in the space below row 5 of the data table.
- Type a 5 in the resulting cell.
Hit the Enter key.

The predicted survival rate for this point is appended to the **Predicted Survival Rate** column.

---

**Example 5.10**

**Tennis Elbow**

One factor in the development of tennis elbow is the impact-induced vibration of the racket and arm at ball contact. Tennis elbow is thought to be related to various properties of the tennis racket used. The accompanying data is a subset of that analyzed in the article “Transfer of Tennis Racket Vibrations into the Human Forearm” (*Med. And Sci. in Sports and Exercise* (1992): 1134-1140). Measurements on racket resonance frequency (Hz) and sum of peak-to-peak accelerations (a characteristic of arm vibration in m/sec/sec) are given for \( n = 14 \) different rackets.
Since there are two variables, we analyze the data using the Fit Y by X platform.

- Select **Analyze > Fit Y by X**.
- Assign **Acceleration** to the **Y, Response** role and **Resonance** as the **X, Factor** role.
- Click **OK**.

The resulting scatter plot is shown here.

To fit a regression line to this data,

- Select **Fit Line** from the platform drop-down list.

A good idea is to examine the residuals from this plot.

- Select **Plot Residuals** from the drop down list next to **Linear Fit** (below the graph).
This appends a residual plot to the bottom of the report.

Note that one observation seems to monopolize the fit, because it is so far separated from the rest of the points.
Summarizing Bivariate Data

Hover the mouse over the outlying point in the residual plot.

Click once on the outlying point to highlight it.

From the main menu bar, select Rows > Exclude/Unexclude.

From the Bivariate drop-down menu, select Fit Line as you did above.

This reveals that observation 14 is the one that is different from the others. It may be a good idea to remove this point and re-fit the line.
You can get a residual plot for this new line as you did above.

- Select **Plot Residuals** from the second (lower) Linear Fit drop-down menu below the scatter plot.
This chapter focuses on a single example from probability theory.

Example 6.32 (POD) or 6.10(DP)

Suppose that couples who wanted children were to continue having children until a boy is born. Assuming that each newborn child is equally likely to be a boy or a girl, would this behavior change the proportion of boys in the population? This question was posed to readers of *American Statistician*, and many answered incorrectly. We use JMP to create a simulation to estimate the long-run proportion of boys in the population if families were to continue to have children until they have a boy. This proportion is an estimate of the probability that a randomly selected child from this population is a boy. Note that every sibling group would have exactly one boy.

To create this simulation, we use a random number generator to generate numbers from 0 to 9. The odd numbers (1, 3, 5, 7, or 9) represent a male birth. If the birth is not a boy, we continue selecting numbers until a boy is born.

To accomplish this simulation, we need a data table to store our results.

- Create a new data table with a single column, called **Children**.
- Right-click on the column header of this column and select **Formula** from the menu that appears.

The formula editor that appears is normally used to enter formulas using point-and-click menus and buttons. However, you can actually enter programs into the formula editor as well. These programs are in JMP’s scripting language—too complicated a topic to get into here, but a small program for this simulation is easy.

- Double-click in the editing area where you see “no formula”.

This changes the formula editor into text editing mode. You can now type formulas directly rather than use the menus to generate them.

Type in the following short program, however, when you enter it, place it all on one line—the separate lines here are so that the program can be easily explained.

\[
t = \text{Random Integer}(10) - 1; \\
n = 0; \\
\text{While}(\text{Mod}(t, 2) == 0, \\
t = \text{Random Integer}(10) - 1; \\
n++;)
\]

Hit the Enter key on the keyboard.

The completed formula should look like this.

Click OK.

Here's what the program does:
6 Probability

• The first line picks a random number from 0 to 9. (The Random Integer (10) function actually picks integers from 1 to 10. We subtract one so that the integers are 0 to 9).

• The second line initializes a counter (named n) to hold the number of times we had to pick the random number until an odd one (male birth) appeared.

• The third line begins a while loop. It uses the Mod function to check that the random number is even. Mod(t, 2) is the remainder when the number is divided by 2, so when this is zero, the number is even. If the number is even, another random integer is picked and the counter (n) is increased by 1. This process continues until an odd number (male birth) appears.

• The last line, containing only an n, is the number that is returned by this formula and placed in the data table column.

When you've returned to the data table, add eight rows to the table.

☐ Select Rows > Add Rows from the main menu bar.

☐ Enter 8 in the dialog that appears.

Your data table fills with eight instances of the simulation. Note: Your data table will probably have different numbers than the one pictured here. That's randomness!

We now want to discover the proportion of these births that are male. Because of the way we set up the simulation, we know that there's only one male birth represented in each row. So, the proportion we are interested in is the number of rows in the data table (representing the number of males) divided by the total of the Children column (representing the number of total births).

To keep a running total of this proportion, we add another row to the data table, with a formula.

☐ Double-click in the column header area to the right of the Children column.

☐ Click once on the new column to select it.

☐ Type the word Proportion to rename the column.

☐ Right-click in the column and select Formula from the menu that appears.

We now want to enter the following formula.

\[
\frac{\text{Row(1)}}{\sum_{i=1}^{\text{Row(1)}} \text{Children}_i}
\]

Follow the steps here to enter this formula.

☐ In the functions list, select Row > Row
Hit the divide key (/)

In the functions list, select Statistical > Summation

In the formula, click on the Nrow() function, the upper index of the summation function

In the functions list, select Row > Row

Click on the body of the summation function

In the columns list, click Children

In the function list, select Row > Subscript
Probability

Type the letter “i” and hit return

Click OK to close the formula editor

The column Proportion contains the proportion of male births for all rows up to and including the entry's row. For example, the entry in the fifth row represents the proportion of male births considering rows 1, 2, 3, and 4.

It is interesting to observe the long-term proportion, after simulating many rows.

Select Rows > Add Rows and add 100 rows to the data table.

Scroll the data table so that the last row is visible.

The long-term proportion seems to hover around 50%. You can see this graphically as well.
6 Probability

Select Graph > Overlay Plot.

Assign Proportion to the Y role.

Click OK.

In the platform drop-down menu, select Y Options > Connect Points.

This connects the points together, as shown here.
6 Probability

Drag the right-side border to make the plot wider, or click and drag on the axes to modify their limits.
This chapter shows some introductory material on statistical inference. We show, in a single example, how to calculate values from a normal distribution.

**Example 7.25(POD) or 7.13(DP)**

In poor countries, the growth of children can be an important indicator of general levels of nutrition and health. Data from the article “The Osteological Paradox: Problems in Inferring Prehistoric Health from Skeletal Samples” (Current Anthropology (1992): 343-370) suggests that a reasonable model for the probability distribution of the continuous numerical variable \( x \) = height of a randomly selected five-year-old child is a normal distribution with a mean of \( \mu = 100 \text{cm} \) and standard deviation \( \sigma = 6 \text{cm} \). What proportion of the heights is between 94 cm and 112 cm?

To calculate this answer, we must find \( P(94 < x < 112) \).

The textbook shows how this is an equivalent problem to calculating \( P(-1.00 < z < 2.00) \) from a Normal distribution with mean zero and standard deviation 1. Normally, you look up the values corresponding to \(-1.00 \) and \( 2.00 \) from a table, but here we use JMP to calculate the values.

To request this information, we must open a script window.

- On Windows or Linux, select **File > New > Script**.
- On Macintosh, hold down the Shift key and select **File > New Script**.

This creates a new script window. Commands to calculate the values from the Normal Distribution are typed in here.

To get the value for the lower bound,

- Type `Normal Distribution(-1)`

- From the main menu bar, select **Edit > Run Script**. [Alternatively, use the shortcut key Ctrl+R (PC and Linux) or ⌘R (Macintosh).]
This runs the script. JMP displays the results in a log window.

```plaintext
// /*
// Normal Distribution(-1)
// */
// 0.15865529383248777
```

In the original script window (not the Log), change the –1 to 2 to find the upper limit.
Run the script.

```plaintext
// /*
// Normal Distribution(-1)
// */
// 0.15865529383248777
// */
// Normal Distribution(2)
// */
// 0.9724588653186585
```

You can even use JMP to compute the difference.
In the original script window (not the log), type Normal Distribution(2)-Normal Distribution(-1)
Run the script.

```plaintext
// /*
// Normal Distribution(-1)
// */
// 0.15865529383248777
// */
// Normal Distribution(2)
// */
// 0.9724588653186585
// */
// Normal Distribution(2)-Normal Distribution(-1)
// */
// 0.8138046141240977
```

So, about 82% of the students fall between 94 and 112 cm.
Chapter 8

There are no examples for chapter 8.
Estimation Using a Single Sample

This chapter further discusses inferential statistics, using sample data to determine something about the real world. We use sample statistics to estimate their corresponding population parameters. In addition, we use measures of variation from the sample to estimate variation in the population, and we use that estimate to construct a confidence interval.

Example 9.2

Internet Use by College Students

The article “Online Extracurricular Activity” (USA Today, March 13, 2000) reported the results of a study of college students conducted by a polling organization called The Student Monitor. One aspect of computer use examined in this study was the number of hours per week spent on the Internet. Suppose that the following observations represent the number of Internet hours per week reported by twenty college students (these data are compatible with summary values given in the article). They are shown here as a JMP data table.

We want to estimate the mean of the population of college students. To do so, we examine the sample mean, sample median, and a trimmed mean. Since we are analyzing one variable, we use the Distribution platform.
Select Analyze > Distribution.
Assign Hours as the Y, Columns role.
Click OK.

The following report appears.

Here, you can easily read off the mean (7.075) or the median (7.125).

Calculating a trimmed mean involves another step. We must select the rows we want to include in the mean and create another data table from them.

Use the Window menu to bring the data table to the front.
Click and drag in the area containing the row numbers to select rows 3 through 18.
From the main menu bar, select Tables > Subset.

Click OK to accept the default options.

When the new data table appears,

Select Analyze > Distribution from the main menu bar.

Select Hours from the column selection list and click Y, Columns.

Click OK.

A new report appears, showing the mean of the subsetted values (the trimmed mean). Note that the N tells us that this is a mean of only 16 values.
Example 9.5

Violence in the Workplace

An Associated Press article on potential violent behavior reported the results of a survey of 750 workers who were employed full time (San Luis Obispo Tribune, September 7, 1999). Of those surveyed, 125 indicated that they were so angered by a coworker during the past year that he or she felt like hitting the person (but didn’t). Assuming that it is reasonable to regard this sample of 750 as a random sample from the population of full-time workers, we can use this information to construct an estimate of \( \pi \), the true proportion of full-time workers so angered in the last year that they wanted to hit a colleague.

To compute the confidence interval, we use the formula

\[
p \pm (z \text{ critical value}) \sqrt{\frac{p(1-p)}{n}}
\]

so we need the \( z \) critical value from the Normal distribution. We use the script window to calculate the value.

- Select File > New > Script (Windows and Linux) or, holding down the Shift key, select File > New Script (Macintosh).
- Enter the text Normal Quantile (0.10)
- Run the script by pressing Control+R (Windows and Linux) or ⌘R (Mac).
Estimation Using a Single Sample

The result appears in a log window.

Repeat the process for Normal Quantile (0.90)
The results appear in the log window.

So, the critical value is

\[ 0.167 \pm (1.645)(0.014) = 0.167 \pm 0.022 = (-.145, 0.189). \]

Example 9.9

Walking a Straight Line

A study of the ability of individuals to walk in a straight line ("Can We Really Walk Straight?" *Amer. J. of Physical Anthropology* (1992): 19-27) reported the accompanying data on cadence (strides per second) for a sample of \( n = 20 \) randomly selected healthy men. The data are presented in a JMP data table.

First, we want to determine if the data are normal. Since we are analyzing a single variable, we use the Distribution platform.
Select Analyze > Distribution.

Assign Cadence to the Y, Columns role.

Click OK.

When the report appears,

Select **Normal Quantile Plot** from the menu beside the Cadence variable.

This appends a Normal Quantile Plot (referred to in the textbook as a Normal probability plot).
Since all (although the criteria is just most) of the points fall between the dotted confidence interval boundaries, we can conclude that the data is reasonably Normally distributed.

To find the 99% $t$ critical value, submit a script.

Select File > New > Script (Windows and Linux) or, while holding the Shift key down, select File > New Script (Macintosh).

Type in the script as shown here.

Select Edit > Run Script.

JMP reveals the answer, 2.86, in the log. Select Window > Log if the log is not showing.

Complete the calculation using this value.
Estimation Using a Single Sample

\[
0.926 \pm (2.86) \frac{0.0809}{\sqrt{20}}
\]

\[
=(0.874, 0.978)
\]

With 99% confidence, we estimate the population mean cadence to be between 0.874 and 0.978.
Hypothesis Testing Using a Single Sample

Previous chapters have looked at ways of finding the value of an unknown population parameter. In this chapter, we look at the formal methods for forming and testing a hypothesis.

Example 10.11

Credit Card Debt

The article “Credit Cards and College Students: Who Pays, Who Benefits?” (J. College Student Development (1998): 50-56) described a study of credit card payment practices of college students. According to the authors of the article, the credit card industry asserts that at most 50% or college students carry a credit card balance from month to month. However, the authors of the article report that, in a random sample of 310 college students, 217 carried a balance each month. Does this sample provide sufficient evidence to reject the industry claim? We answer this question by carrying out a hypothesis test using a 0.05 significance level.

See the textbook for the proper steps in conducting an hypothesis test. JMP is useful in finding the $z$-value corresponding to the calculated $z$-statistic. In this case, $z = 7.14$ (which is obviously so far out in the tail of the distribution that the test is significant, but we go through the motions here anyway).

Select File > New > Script (Windows and Linux) or, while pressing the Shift key, select File > New Script (Macintosh).

Type the following line into the script window that appears.

```
Select Edit > Run Script.
```

The answer that appears in the JMP Log window (0.9999999999995334) is the proportion of the area to the left of the statistic, which is obviously significant.
Example 10.14

Personal Use of Company Technology

One concern employers have about the use of technology is the amount of time that employees spend each day making personal use of company technology, such as personal phone, e-mail, Internet, and computer games. The Associated Press (September 7, 1999) reported that a management consultant believes that, on average, workers spend 75 minutes a day making personal use of technology. Suppose that the CEO of a large corporation wanted to determine whether the average amount of time spent in personal use of technology for her employees was greater than the reported 75 minutes. Each person in a random sample of 10 employees was contacted and asked about daily personal use of company technology. (Participants would probably have to be guaranteed anonymity to obtain useful responses.) The resulting data is shown in a JMP data table here.

We want summary statistics for the data. Since we are analyzing a single variable, we use the Distribution platform.

Select Analyze > Distribution.

Select Time from the list of columns and push the Y, Columns button.

Click OK.

The report shown here appears.
This shows that \( n = 10 \), the mean is 74.8, and the standard deviation is 9.45.

Does this data provide evidence that the mean for this company is greater than 75 minutes? Carry out a hypothesis test of \( H_0: \mu = 75 \) against \( H_a: \mu > 75 \) with \( \alpha = 0.05 \).

First, we must check that the data are approximately normal. To do so, use the Normal quantile plot.

Select **Normal Quantile Plot** from the drop-down menu found in the outline bar beside **Time**.
Hypothesis Testing Using a Single Sample

Since most (or, in this case, all) of the points fall within the dotted confidence lines on the Normal probability plot, we can conclude that it is plausible that the data are Normally distributed. We do notice some skewness in the box and whiskers plot, but it is not enough to worry about.

To test that this is different from the hypothesized mean of 75, use the Test Mean command.

Select Test Mean from the menu beside Time (the same one containing Normal Quantile Plot).

In the dialog that appears,

- Type 75

We do not know the true standard deviation, so leave the second edit field blank.

- Click OK.
A $t$-test report is shown at the bottom of the existing Distribution report.

The appropriate $t$-test statistic shows beside $\text{Prob > } t$ (since we are testing a one-sided alternative). The report shows this value to be 0.526 (slightly different from the one in the text, due to rounding). Therefore, we cannot reject the null hypothesis.

JMP provides an interesting illustration of the $p$-value concept in this context.

- Beside $\text{Test Mean = Value}$, select $\text{Pvalue Animation}$ from the red triangle drop-down menu.
This brings up a window that graphically shows the $t$-test. By changing the values in this display, you can pose questions about the $t$-test.

Click the **High Side** button so that the display corresponds to the one-sided alternative we’re considering here. The shaded area represents the $p$-value, which is shown on the plot as well.

What if the sample size were different, that is, smaller or larger?

Click the 10 showing beside **Sample Size**, and enter 5. Observe the changes to the graph and the $p$-value.

Change the 5 to a 30, and again observe what happens.

What if the hypothesized mean were further away from the calculated mean?

Click and drag on the handle in the plot to move the means apart, and observe the changes in the $p$-value.
Chapter 10 showed you how to use JMP for hypothesis testing of a single statistic. This chapter expands on that topic by extending the technique to two populations or treatments.

Example 11.2

To assess the impact of oral contraceptive use on bone mineral density (BMD), researchers in Canada carried out a study comparing BMD for women who had used oral contraceptives for at least three months to BMD for women who had never used oral contraceptives (“Oral Contraceptive Use and Bone Mineral Density in Premenopausal Women,” Canadian Medical Association Journal (2001): 1023-1029) Data consistent with summary quantiles given in the paper appear in the accompanying JMP table (the actual sample sizes for the study were much larger).

There is one important thing to note about the JMP data table. Although there are two columns, they do not represent “Never Used” and “Used”. Instead, the first column holds a label that says which group each row belongs to, and the second column holds the actual BMD measurement. All JMP tables are built in the same way—rows represent individual cases.

The authors of the paper believed that these two groups formed a representative sample of the population under study. They wanted to compare the means for the two groups to see if they were significantly different from each other (at the 0.05 level). In this case, our analysis involves two columns (the Used? column and the BMD column), so we use the Fit Y by X platform for this analysis.
Select Analyze > Fit Y By X.
Select the BMD column and click the Y, Response button.
Select the Used? column and click the X, Factor button.
Click OK.

The following side-by-side dot plot appears.

To do the t-test,
Select Means/Anova/t test from the drop-down menu on the Oneway title bar.1

The t-test report appears under the dot plot.

1. If you are using JMP IN or the professional version of JMP, this command appears as t test. The Means/Anova/Pooled t command is for a different test.
Note: If a question mark appears for any of these numbers, double-click the question mark and change the format from Fixed Dec to Best.

There are two t-tests in this report—one with a heading “Assuming Equal Variances” and the other with a heading “UnEqual Variances.” Which is appropriate here?

In the calculations for the t-test in the textbook, the equation used is

$$ t = \frac{(1.08 - 1.00) - 0}{\sqrt{\frac{0.16^2}{10} + \frac{0.14^2}{10}}} $$

Note that this uses the variances of the two groups individually, rather than considering all the data as a single group and computing a single variance. So, we are using unequal variances, and therefore read off the lower section of the report.
In general, the unequal variance \( t \)-test is the correct test to examine. Unless you have some prior knowledge that the variances are indeed equal, or you have an extremely small sample size (like, fewer than four observations in each group), use the unequal variance test.

The unequal variance report gives us the \( t \)-statistic value \( 1.135 \) and the \( p \)-value \( 0.2716 \). We therefore cannot reject the null hypothesis that the mean is different between the two groups.

**Example 11.4**

**Effect of Talking on Blood Pressure**

Does talking elevate blood pressure, contributing to the tendency for blood pressure to be higher when measured in a doctor's office than when measured in a less stressful environment (called the "white coat" effect)? The article "The Talking Effect and 'White Coat' Effect in Hypertensive Patients: Physical Effort or Emotional Content" (Behavioral Medicine (2001):149-157) describes a study in which patients with high blood pressure were randomly assigned to one of two groups. Those in the first group (the talking group) were asked questions about their medical history and about the sources of stress in their lives in the minutes prior to measuring blood pressure. Those in the second group (the counting group) were asked to count aloud from 1 to 100 four times prior to having blood pressure measured. The accompanying JMP data table shows values for diastolic blood pressure (mm Hg) are consistent with summary quantiles appearing in the paper.

Since we are analyzing two groups, we use the Fit Y By X platform.

- Select **Analyze > Fit Y By X**.
- Assign **Talking Status** to **X, Factor** and **Diastolic BP** as **Y, Response**.
- Click **OK**.

Side-by-side dotplots appear.
11 Comparing Two Populations or Treatments

To overlay box plots on them (as the text does),

- Select Display Options > Box Plots from the platform drop-down menu.

There are no outliers in either data set, and the box plots seem reasonably symmetric, suggesting that the assumption of approximate normality is reasonable.

To conduct the \( t \)-test that the means are different,

- Select Means/Anova/test from the platform drop-down menu.

Since we are using the standard deviations from the individual groups (as opposed to pooling them), use the UnEqual Variances section of the report.
You can see that the confidence interval for the difference is (1.04, 11.95). This is slightly different from the results in the text, because JMP uses the fractional degrees of freedom (13.77 in this case) rather than truncating the value to the integer 13.

Example 11.8

Lactic Acid in the Blood After Exercise

The effect of exercise on the amount of lactic acid in the blood was examined in the article “A Descriptive Analysis of Elite-Level Racquetball” (Research Quarterly for Exercise and Sport (1991): 109-114). Eight males were selected at random from those attending a weeklong training camp. Blood lactate levels were measured before and after playing three games of racquetball, as shown in the accompanying table. We will use this data to estimate the mean change in blood lactate level using a 95% confidence interval.
Note that the Difference column is computed using a formula. To see the formula stored in the column,

- Right-click on the column header and select Formula from the menu that appears.

This opens the formula editor, showing the formula

\[
\text{Before} - \text{After}
\]

This is a paired situation, since the same experimental unit was used for two separate measurements. In JMP, there are two ways to compute a confidence interval (or a paired t-test): Using the Distribution platform and using the Matched Pairs platform.

To analyze this using the Distribution platform,

- Select Analyze > Distribution and assign Difference as the Y, Column variable.

- Click OK.

You can read the confidence interval from the resulting report.
To estimate the confidence interval using the Matched Pairs platform,

- Select Analyze > Matched Pairs.
- Assign Before and After as the Y, Paired Response.
- Click OK.
The confidence interval can be read off the resulting report.

Example 11.10

AIDS and Housing Availability

This example does not contain any actual data, so would be more easily analyzed using a calculator.
11 Comparing Two Populations or Treatments
Previous chapters have been devoted to methods appropriate to numerical data. This chapter extends your statistical toolbox to include methods for categorical data.

Example 12.7

Risky Behavior

The article “Factors Associated with Sexual Risk-Taking Behaviors Among Adolescents” (J. Marriage and Family (1994): 622-632) examined the relationship between gender and contraceptive use by sexually active teens. Each person in a random sample of sexually active teens was classified according to gender and contraceptive use (with three categories: rarely or never use, use sometimes or most of the time, and always use) resulting in a 3 by 2 table. Data consistent with percentages given in the article are summarized in the following table.

Table 12.1 Risk-taking by gender

<table>
<thead>
<tr>
<th>Contraceptive Use</th>
<th>Gender</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Female</td>
<td>Male</td>
</tr>
<tr>
<td>Rarely/Never</td>
<td>210</td>
<td>350</td>
</tr>
<tr>
<td>Sometimes/Most Times</td>
<td>190</td>
<td>320</td>
</tr>
<tr>
<td>Always</td>
<td>400</td>
<td>530</td>
</tr>
</tbody>
</table>

Entering this data into JMP requires that you think in terms of variables and observations. Variables are columns in JMP, observations are rows. There are two variables in this study (Gender and Contraceptive Use), which will each occupy their own column. In addition, we need a third column to hold the frequencies—the cell counts—in the table. The correct completed JMP table is shown here.
Since this analysis involves two variables, we use the Fit Y By X Platform.

1. Select Analyze > Fit Y By X.
2. Assign Gender to Y, Response, Contraceptive Use to X, Factor, and Count to Freq.
3. The launch dialog should look like this.

   ![Example of Fit Y By X launch dialog]

4. Click OK.

   A contingency table report appears.
The upper part of the report is a Mosaic Plot. It shows the distribution of contraceptive use by males and females. If contraceptive use is the same between the genders, we would expect to see the dividing lines line up.

Below the Mosaic Plot is the contingency table. It shows Count, Total%, Col% and Row% (in that order) for each cell. Along the margins of the table are the marginal totals. For example, the marginal total for rarely/never used contraceptives is 560; the column total for females is 800.

Below the Contingency table is the textual output where you find the values for the chi-square statistical test. The test shown in your main text is the Pearson chi-square, which in this case has the value 6.572, whose associated $p$-value is 0.035. This is a statistically significant finding, so we conclude that there is a difference between genders.
Simple linear regression and correlation were introduced in chapter 5 as descriptive techniques. This chapter extends these bivariate techniques by illustrating methods of constructing hypothesis tests on the correlations and regression slopes.

Example 13.2

Mother's Age and Babies Birth Weight

Medical researchers have noted that adolescent females are much more likely to deliver low birth weight babies than are adult females. Because low birth weight babies have higher mortality rates, there have been a number of studies examining the relationship between birth weight and mother's age for babies born to young mothers. One such study is described in the article “The Risk of Teen Mothers Having Low Birth Weight Babies: Implications of Recent Medical Research for School Health Personnel” (*J. of School Health* (1998): 271-274). The accompanying JMP data table shows maternal age (in years) and birth weight of babies (in grams).

This is an analysis of two variables, so we use the Fit Y By X platform

- Select Analyze > Fit Y By X.
- Assign Birth Weight as Y, Response and Maternal Age as X, Factor.
- Click OK.

A scatterplot like the one shown here appears.
The data look reasonably linear, so fit a regression line to the data.

From the platform drop-down menu, select **Fit Line**.
The regression model appears right below the Linear Fit outline bar. It is

Birth Weight = -1163.45 + 245.15\times Maternal Age

To find an estimate of the birth weight for a maternal age of 18, you can substitute directly into the equation

Birth Weight = -1163.45 + 245.15\times 18 = 3249.25 \text{ grams}

Alternatively, you can store the prediction formula into the data table and add a row corresponding to Maternal Age = 18.

Select Save Predicteds from the Linear Fit drop-down menu.
This creates a new column in the data table, and the new column contains a formula that represents the prediction equation.

- Click below the last row in the data table to add another row.
- Enter 18 as the Maternal Age for this new row.

Observe that a new value appears in the Predicted Birth weight column, representing the prediction for the new value.

Note that this prediction represents two values: the estimate of a baby born to an individual mother at maternal age 18, and an estimate of the mean value of all babies born to 18-year-old mothers. The distinction between the two will become clear in a later section.
Example 13.4

Athletic Performance and Cardiovascular Fitness

Is cardiovascular fitness (as measured by time to exhaustion running on a treadmill) related to an athlete's performance in a 20-km ski race? The accompanying JMP data table shows treadmill time to exhaustion (in minutes) and 20-km ski time (also in minutes). The data was taken from the article “Physiological Characteristics and Performance of Top U.S. Biathletes” (Medicine and Science in Sports and Exercise (1995): 1302-1310).

The analysis involves two variables, so we use the Fit Y By X platform.
- Select Analyze > Fit Y By X.
- Assign Treadmill Time as X, Factor and Ski Time as Y, Response.
- Click OK.

To fit a regression line,
- Select Fit Line from the platform drop-down menu.
Note the regression equation, reported just beneath the Linear Fit outline node. In this case, the slope can be interpreted as the average change in ski time associated with a one-minute increase in treadmill time. The scatterplot shows a negative correlation, which agrees with the regression equation.

The value of $r^2$ is shown in the Summary of Fit section of the report, labeled Rsquare.

To see the 95% confidence interval of the slope estimate, you have to reveal some hidden columns in the parameter estimates report.

Right-click in the Parameter Estimates report
Select Lower 95% and Upper 95% from the menu that appears.
This reveals two columns in the parameter estimates report that show 95% confidence intervals for both the intercept and the slope.

The slope has confidence interval (-3.67, -0.996).

Example 13.6

Landslides and Timber Growth

Landslides are common events in tree-growing regions of the Pacific Northwest, so their effect on timber growth is of special concern to foresters. The article “Effects of Landslide Erosion on Subsequent Douglas Fir Growth and Stocking Levels in the Western Cascades, Oregon” (Soil Science Soc. of Amer. J
(1984) 667-671) reported on the results of a study in which growth in a landslide area was compared with growth in a previously-cut area. The JMP data table here shows tree age (in years) and 5-year growth (in cm).

For simple regression fits, we generally use the Fit Y By X platform. However, this example requires us to compute standardized residuals (called Studentized Residuals by JMP), which are not available in the Fit Y By X platform. Therefore, we use the more general Fit Model platform for this data.

- Select **Analyze > Fit Model**.
- Assign **Growth** as the Y variable.
- Assign **Tree Age** as a factor in the model by selecting its name in the columns list and clicking the **Add** button.

The completed Fit Model dialog should look like this.
Click Run Model.

This fits a linear regression line to the data. There’s a lot of output, with the important numbers highlighted here.

A residual plot shows below the text reports. To save these residuals to the data table,

- Select Save Columns > Residuals from the platform drop-down menu.
This creates a new column in the data table named Residual Growth.

You can also save the standardized (Studentized) residuals to the data table.

Select **Save Columns > Studentized Residuals** from the platform drop down menu.

This creates a new column in the data table called Studentized Resid Growth.
To check the normality of these residuals, use the Distribution platform.

1. Select **Analyze > Distribution**.
2. Select **Residual Growth** and **Studentized Resid Growth** as **Y, Columns** variables.
3. Click **OK**.
We want to apply the following commands to both plots in the report, so hold down the Control (PC and Linux) or ⌘ (Macintosh) key while accessing the menus. This broadcasts the command to all the individual reports.

Select **Normal Quantile Plot** from the platform drop-down menu.

This produces two Normal Quantile plots. Since the points fall within the dotted confidence interval lines, we conclude that both sets of residuals are approximately normal.

**Example 13.11**

Physical characteristics of sharks are of interest to surfers and scuba divers, as well as marine researchers. The accompanying data table shows length (feet) and jaw width (in) for 44 sharks reflects data shown in various issues of *Skin Diver* and *Scuba News.*
To perform a simple regression, we normally use the Fit Y By X platform.

- Select **Analyze > Fit Y By X**.
- Assign shark length as **X, Factor** and jaw length as **Y, Response**.
- Click **OK**.

A scatterplot appears.
To fit the regression line to the data,

Select Fit Line from the platform drop-down menu

This adds a line to the scatterplot and appends textual information.

The linear fit equation is seen as

\[
\text{jaw width} = 0.688 - 0.963 \times \text{shark length}.
\]

In addition, you can read off the values for \(SS(\text{Resid}) = 79.49\), \(SS(\text{Total}) = 339.02\), and \(r^2 = 0.766\).
The purpose of this exercise is to illustrate the difference between a confidence interval for the mean prediction of an \(x\)-level, and the confidence interval for an individual prediction for a given \(x\)-level. In the Fit Y By X platform, we can overlay curves to show these two intervals.

By default, the confidence intervals are at a level of 95\%. In the main text, however, the first example uses a 90\% confidence interval, so first we change the default.

From the Linear Fit drop down menu, select **Set Alpha Level > 0.10**.

Now, overlay the confidence intervals on the graph.

From the same Linear Fit drop-down menu, select **Confid Curves Fit** and **Confid Curves Indiv**. This produces the output shown here.
It’s now easy to see that the confidence limits for predicting the mean at a value of $x$ are smaller than predicting an interval. To think of it another way, suppose you were asked to predict the average weight of a group of attendees at a large party. Then suppose you were asked to predict the average weight of a specific individual. Which prediction would you be more sure of? This is essentially the difference between confidence intervals on a mean and confidence intervals on a prediction.

You cannot actually produce the confidence limits in the Fit Y By X report. If you need to actually see the confidence limits for each individual observation, you must use the Fit Model platform.

1. Select Analyze > Fit Model.
2. Assign jaw width as Y.
3. Add shark length as an effect in the model by selecting shark length in the columns list and clicking the Add button.

The completed Fit Model dialog should look like this.
Click **Run Model**.

As in the Fit Y By X report, the Fit Model report uses a default 95% level for confidence intervals. To enter a new confidence interval, hold the Shift key down while selecting a command.

手持 Shift 键并选择 **Save Columns > Mean Confidence Interval** 从平台下拉菜单。

在弹出的对话框中，输入 0.10。

手持 Shift 键并选择 **Save Columns > Indiv Confidence Interval**。

在弹出的对话框中，输入 0.10。

点击 **OK**。

这将为数据表添加两列，分别保存90%置信区间。要保存单个预测的置信区间，

手持 Shift 键并选择 **Save Columns > Indiv Confidence Interval**。

在弹出的对话框中，输入 0.10。

点击 **OK**。
Past chapters have explored the simple linear regression model. This chapter extends those techniques, allowing for more than one x-variable.

In general, simple linear regression involves two variables (a factor and a response, or an x and a y), so is analyzed with the Fit Y By X platform. Multiple Regression involves more than one x, so in total it involves at least three variables. For these cases, we use the Fit Model platform.

Example 14.6

Soil and Sediment Characteristics

Soil and sediment adsorption, the extent to which chemicals collect in a condensed form on the surface, is an important characteristic because it influences the effectiveness of pesticides and various agricultural chemicals. The article “Adsorption of Phosphates, Arsenate, Methanarsenate, and Calcodylate by Lake and Stream Sediments: Comparisons with Soils” (J. of Environ. Qual. (1984): 499-504) presented the accompanying data consisting of \( n = 13 \) \((x_1, x_2, y)\) triples and proposed the model

\[
y = a + b_1x_1 + b_2x_2 + e
\]

for relating

\[
y = \text{phosphate adsorption index}
\]

\[
x_1 = \text{amount of extractable iron}
\]

\[
x_2 = \text{amount of extractable aluminum}
\]

To fit a multiple regression model in JMP,
Select Analyze > Fit Model.
Assign Phosphate as the Y.
While holding down the Shift key, select both Iron and Aluminum in the columns list. Clicking with the Shift key held down allows you to select several variables.
Click on the Add button to add the variables as effects in the analysis.

The completed Fit Model dialog should look like this.

Click Run Model.
This gives you the output shown here.
Focus on the Parameter Estimates outline node. This gives you the estimated coefficients for the parameters in the model.

Intercept = -0.7351
Iron coefficient = 0.113
Aluminum Coefficient = 0.349

Thus, we estimate that the average change in Phosphate associated with a one-unit increase in Iron while Aluminum remains fixed is 0.11273. A similar interpretation applies to the aluminum coefficient.

The regression function is

\[(\text{estimated mean } y) = -7.351 + 0.113(\text{Iron}) + 0.349(\text{Aluminum})\]
14 Multiple Regression Analysis

Example 14.17

Modeling the Price of Industrial Properties

The paper “Using Multiple Regression Analysis in Real Estate Appraisal” (The Appraisal Journal (2001): 424-430) reported the accompanying data for a random sample of nine large industrial properties. A primary objective was to relate the price of the property to various other characteristics of the property. The variables shown are

- $y$ = price per square foot
- $x_1$ = size of the building
- $x_2$ = age of building (years)
- $x_3$ = quality of location (measured on a scale of 1, very poor, to 4, very good)
- $x_4$ = land-to-building ratio

The main textbook begins with a discussion of variable selection. Automated variable selection procedures are not available in JMP INTRO. To explore different models, you must manually make them with the Fit Model platform and record the results, like you did in “Example 14.6” on page 109.

However, if you have access to JMP IN (a higher-level student product) or the professional version of JMP, you can see all possible models.

- Select Analyze > Fit Model.
- Assign Price to $Y$ and add all other variables as effects in the model.
- Change the Personality to Stepwise.

The completed Fit Model dialog should look like this.
Click Run Model.
This produces the Stepwise Regression control panel.
Select All Possible Models from the drop-down menu at the top of the report.

This produces a report of all possible models.
You can fit a linear model using any JMP product, including JMP INTRO. The book suggests fitting a model with both Size and Age as variables.

1. Select **Analyze > Fit Model**.
2. Assign Price as Y and add Size and Age as effects in the model.
3. Click **Run Model**.

This produces output similar to that shown below.
You can read the model from the Parameter Estimates section of the report. In this case, it does not appear that either the Size or Age effects are significant at the 0.05 or 0.10 levels.

**Example 14.18**

**Durable Press Rating of Cotton Fabric**

The accompanying data was taken from the article “Applying Stepwise Multiple Regression Analysis to the Reaction of Formaldehyde with Cotton Cellulose (Textile Research Journal (1984) 157-165). The dependent variable

\[ y = \text{durable press rating} \]

is a quantitative measure of wrinkle resistance. The four independent variables used in the model building process are

\[ x_1 = \text{HCH (formaldehyde) concentration} \]
Multiple Regression Analysis

\[ x_2 = \text{catalyst ratio} \]
\[ x_3 = \text{curing temperature} \]
\[ x_4 = \text{curing time} \]

In addition to these variables, the investigators considered the as potential predictors \( x_1^2, x_2^2, x_3^2, x_4^2, \) and all six interactions for a total of \( p = 14 \) candidates.

JMP INTRO does not allow for automatic variable selection. However, if you have access to JMP IN (a higher-level student product) or the professional version of JMP, complete the following steps.

- Select **Analyze > Fit Model**.
- Assign **Press Rating** to the **Y** role.
- Select **HCH, catalyst, Temp, and Time** in the columns list.
- Click the **Macros** button and select **Factorial to Degree**.
This adds all four effects, plus their two-factor interactions.

- Again, select HCH, catalyst, Temp, and Time in the columns list.
- Click the Macros button and select Polynomial to Degree.
  
This adds the squares of each effect. You should now have 14 effects in the model.

- Change the Personality to Stepwise.
- Click Run Model.

When the stepwise control panel appears,

- Select All Possible Models from the platform drop-down list.

This adds a (lengthy) report showing all possible models. The textbook uses Mallow’s \( C_p \) statistic, which is initially hidden in this report.

- Right-click in the All Possible Models report and select Columns > Cp.

This adds the \( C_p \) values to the report.

The text recommends fitting a six-predictor model, involving catalyst, \((HCH)^2\),\((catalyst)^2\), \(HCH*Temp\), \(HCH*Time\), and \((Catalyst*Temp)\). However, this model cannot be fit in JMP, since interaction and power terms require their lower-order terms to also be in the model.
The Analysis Of Variance

Earlier in the text, methods you encountered methods to compare two means using the t test. In this chapter, we cover methods to compare more than two means, using the statistical procedure called \textit{analysis of variance} or ANOVA.

Example 15.4

The article “The Soundtrack of Recklessness: Musical Preferences and Reckless Behavior Among Adolescents” (\textit{J. Adolescent Research} (1992): 313-331) described a study whose purpose was to determine whether adolescents who preferred certain types of music reported higher rates of reckless behaviors, such as speeding, drug use, shoplifting, and unprotected sex. Independently chosen random samples were selected from each of the four groups of students with different musical preferences at a large high school: (1) acoustic/pop, (2) mainstream rock, (3) hard rock, and (4) heavy metal. Each student in these samples was asked how many times they had engaged in various reckless activities during the last year. A portion of the data on driving over 80mph that is consistent with summary quantiles given in the article is shown here. (The sample sizes in the article were much larger, but for purposes of this example, we use \(n_1 = n_2 = n_3 = n_4 = 20\).)
15 The Analysis Of Variance

Note that there are 80 rows, since there are 80 data points. Each row corresponds to an observation, with a column labeling the type of music and a column labeling the number of reported incidents. Since this analysis involves two variables, the Fit Y By X platform is used.

- Select Analyze > Fit Y By X.
- Assign Incidents as the Y, Response and Type as the X, Factor.
- Click OK.

The report shown here appears.

In the textbook, means are compared using box plots. To add these to the plot,
- Select Display Options > Box Plots from the platform drop-down menu as shown here.
Box plots appear on the report.

The box plots look reasonably symmetric, and there are no outliers, so we proceed with the analysis of variance.

*Select** **Means/Anova/t test** from the platform drop-down menu.
The $F$-Ratio is shown as 5.19 with a computed $p$-value of 0.002g. We can therefore conclude that the mean number of times driving over 80mph does not appear to be the same for all four musical preference groups. Note that this doesn’t tell us which groups are significantly different.

**Example 15.6**

**Sleep Time**

A biologist wished to study the effects of ethanol on sleep time. A sample of 20 rats, matched for age and other characteristics, was selected, and each rat was given an oral injection having a particular concentration of ethanol per body weight. The rapid eye movement (REM) sleep time for each rat was then recorded for a 24-hour period, with the results shown in the accompanying table.

Since this analysis involves two variables, we use the Fit Y By X platform.

- Select **Analyze > Fit Y By X**.
- Assign **REM Time** as Y, Response and **Group** as X, Factor.
- Click **OK**.

To compute the ANOVA,

- Select **Means/Anova/t test** from the platform drop-down menu.
You can see the Tukey-Kramer comparisons both graphically and in a table.

- Select Compare Means > All Pairs, Tukey HSD.
There are several methods to see which means are significantly different.

Click on the lowest circle in the comparison circles plot.

This circle corresponds to group 3. Note that its circle turns red, as well as the circle for group 2. This tells you that the means for these two groups are not significantly different. The circles for groups 0 and 1 are gray. This indicates that the mean for group 3 is significantly different from these two groups.
You can continue clicking on circles to make all the comparisons.

Alternatively, look at the Means Comparisons text report. In the section of the report that lists Abs(Dif)-Mean, look for positive numbers. Positive numbers say that the means are significantly different. So, you can see that group 3 has positive numbers corresponding to groups 0 and 1, so group 3 is significantly different from these groups.

Finally, you can look at the connected letters portion of the report. Levels that are not significantly different are connected by the same letter. In this example, group 3 shares a letter (C) with group 2, so the means of these two groups are not significantly different. Group 3 does not share a letter with groups 0 or 1, so you can conclude that their means are significantly different from group 3.

Example 15.10

Comparing Four Stool Designs

In the article “The Effects of a Pneumatic Stool and a One-Legged Stool on Lower Limb Joint Load and Muscular Activity During Sitting and Rising” (Ergonomics (1993): 519-535), the following data is given on the effort (measured on the Borg scale) required by a subject to rise from a sitting position for each of four different stools. Because it was suspected that different people could exhibit large differences in effort, even for the same type of stool, a sample of nine people was selected and each tested on all four stools, yielding the following data.
Since this analysis involves two variables, we use the Fit Y by X platform.

1. Select **Analyze > Fit Y By X**.
2. Assign **Stool Type** as **X, Factor** and **Effort** as **Y, Response**.
3. Click **OK**.

When the report appears,

1. Select **Means/Anova/t test** from the platform drop-down menu.

The report shown here appears.
The $p$-value is a small 0.002, so we can conclude that the mean amount of effort required to arise from the four stools is not the same.

Example 15.14

Effect of Soil Type and Pipe Coating on Corrosion

When metal pipe is buried in soil, it is desirable to apply a coating to retard corrosion. Four different coatings are under consideration for use with pipe that will ultimately be buried in three types of soil. An experiment to investigate the effects of these coatings and soils was carried out by first selecting the 12 pipe segments and applying each coating to three segments. The segments were then buried in soil for a specified period in such a way that each soil type received one piece with each coating. The resulting data (depth of corrosion) is shown here.
This analysis involves three variables—two factors and one response. Therefore, we use the Fit Model platform.

1. Select Analyze > Fit Model.
2. Assign Corrosion to the Y role.
3. Select both Coating and Soil (Shift click to select more than one variable) and click the Add button to add them as factors in the model.

The completed Fit Model dialog box should look like this.

1. Click the Run Model button to display the analysis report.
The $p$-value for the whole-model test is 0.2664, which says that we have no evidence for a difference in corrosion based on coating or soil type.
15 The Analysis Of Variance