Dear Professor or Other Supplement Recipient:

Cengage Learning has provided you with this product (the “Supplement”) for your review and, to the extent that you adopt the associated textbook for use in connection with your course (the “Course”), you and your students who purchase the textbook may use the Supplement as described below. Cengage Learning has established these use limitations in response to concerns raised by authors, professors, and other users regarding the pedagogical problems stemming from unlimited distribution of Supplements.

Cengage Learning hereby grants you a nontransferable license to use the Supplement in connection with the Course, subject to the following conditions. The Supplement is for your personal, noncommercial use only and may not be reproduced, or distributed, except that portions of the Supplement may be provided to your students in connection with your instruction of the Course, so long as such students are advised that they may not copy or distribute any portion of the Supplement to any third party. Test banks, and other testing materials may be made available in the classroom and collected at the end of each class session, or posted electronically as described herein. Any material posted electronically must be through a password-protected site, with all copy and download functionality disabled, and accessible solely by your students who have purchased the associated textbook for the Course. You may not sell, license, auction, or otherwise redistribute the Supplement in any form. We ask that you take reasonable steps to protect the Supplement from unauthorized use, reproduction, or distribution. Your use of the Supplement indicates your acceptance of the conditions set forth in this Agreement. If you do not accept these conditions, you must return the Supplement unused within 30 days of receipt.

All rights (including without limitation, copyrights, patents, and trade secrets) in the Supplement are and will remain the sole and exclusive property of Cengage Learning and/or its licensors. The Supplement is furnished by Cengage Learning on an “as is” basis without any warranties, express or implied. This Agreement will be governed by and construed pursuant to the laws of the State of New York, without regard to such State’s conflict of law rules. Thank you for your assistance in helping to safeguard the integrity of the content contained in this Supplement. We trust you find the Supplement a useful teaching tool.

SPSS® is a registered trademark of IBM Corporation, in the United States and other countries.
## Contents*

Chapter 1: Getting Started with SPSS ................................................................. 1

Chapter 2: Turning Data into Information .......................................................... 11

Chapter 3: Relationships Between Quantitative Variables .............................. 37

Chapter 4: Relationships Between Categorical Variables ............................... 47

Chapter 5: Sampling: Surveys and How to Ask Questions .............................. 48

Chapter 6: Gathering Useful Data for Examining Relationships .................. 53

Chapter 8: Random Variables ........................................................................... 57

Chapter 9: Understanding Sampling Distributions: Statistics as Random Variables ... 67

Chapter 10: Estimating Proportions with Confidence ..................................... 77

Chapter 11: Estimating Means with Confidence ............................................. 84

Chapter 12: Testing Hypotheses about Proportions ....................................... 102

Chapter 13: Testing Hypotheses about Means .............................................. 112

Chapter 14: Inference about Simple Regression .......................................... 126

Chapter 15: More about Inference for Categorical Variables ......................... 146

Chapter 16: Analysis of Variance ................................................................. 162

* Chapter 7 has been omitted from this guide because it contains no material relevant to SPSS
Chapter 1

Getting Started with SPSS

1.1 Overview

This manual is a basic introduction to SPSS. It is meant to be used in addition to the textbook *Mind on Statistics* by Utts and Heckard. The manual is by no means meant to be a complete guide to SPSS. Each of the statistical applications in this manual is explained by means of an example in the book. Short introductions to the problems are given, but the reader is referred back to the book for additional information on these examples. This chapter covers the basic ideas, steps, and commands of SPSS Statistics 21.0 by IBM. In later chapters, knowledge of these steps will be assumed.

1.2 Starting SPSS

Open SPSS by either using the Start menu on your computer or by double-clicking on the SPSS icon. A window will open up in addition to the SPSS application. The window is shown below:

![SPSS Window](image)

The default option is to immediately open a data file. By clicking **OK**, you will be directed to the SPSS Data Directory and asked which data set to open. You may also opt to select the Type in data option. When clicking **OK**, you will be left with an empty spreadsheet that is ready for you to enter data. An alternative to this option is to click **Cancel**. If, in the future, you wish to always start with a blank spreadsheet, you may choose to check the **Don’t show this dialog in the future** option at the bottom left-hand side.
SPSS is a spreadsheet-based program. The main window, called the IBM SPSS Statistics Data Editor, contains the spreadsheet. It is shown below:

All commands are given using the menu bar at the top of the screen. The File and Edit menus contain the usual submenus available in most Windows applications. The remainder of the menus is used for statistical analysis. When a command is given to SPSS, the output will appear in a separate window, called the PASW Statistics Viewer. You can switch between the Editor and the Viewer by using the Window menu or by using the buttons on the bottom of the computer screen.

1.3 Opening an Existing Data File

In this manual, many existing data sets will be used. The command “Open the data set filename.sav” will be used. Most data sets are available on the website. The steps necessary to open the data set handheight.sav are shown below:

1. **Steps** to open a data set:
   a. To open a data set, we use the File menu.
   b. Scroll down to the Open submenu and select the Data option. An alternative is to use the “Open data document” button at the top left-hand corner of the window:
c. The usual **Open File** window opens up. The window is shown below:

![Open File Window](image)

- Find the location of the SPSS data sets.
- Select the file **handheight** and click **Open**. A section of the handheight.sav data set is shown below:

![Handheight Data Set](image)
Each dataset has two different views that we can use to learn about the data. The first one is called the **Data View**. The **Data View** of the handheight dataset is repeated below:

In the **Data View** we can see which variables make up the dataset and what values are observed in the dataset. Each column represents a variable and each row represents a unit in the dataset.

The second view is called the **Variable View**. The **Data View** of the handheight dataset is shown below:

In the **Variable View** we can see the properties of the variables in the dataset. Note that here each row represents a variable! Properties that we can see here are the type of variable, how many decimals are being used, any possible value labels that are being used, and the measurement level of the variables.
1.4 Creating a New Data File

In this manual some data sets will need to be created. The command “Open a new data file” will be used. Most data sets that need to be created are small and the data is listed in the corresponding section of the textbook. The steps necessary to open a new data set are shown below:

1. **Steps** to open a new data set:
   a. To open a new data set, we use the **File** menu.
   b. Scroll down to the **New** submenu and select the **Data** option. A blank spreadsheet will appear. This spreadsheet is ready to receive data. Each column will represent a different variable. How to treat different types of variables is described in Section 1.5.

1.5 Preparing a New Data File

A hypothetical data set is shown below. The steps needed to enter this data into SPSS follow.

<table>
<thead>
<tr>
<th>haircolor</th>
<th>gender</th>
<th>time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blond</td>
<td>1</td>
<td>15.8</td>
</tr>
<tr>
<td>Brown</td>
<td>2</td>
<td>13.2</td>
</tr>
<tr>
<td>Black</td>
<td>1</td>
<td>12.8</td>
</tr>
<tr>
<td>Blond</td>
<td></td>
<td>13.4</td>
</tr>
<tr>
<td>Red</td>
<td>2</td>
<td>12.1</td>
</tr>
<tr>
<td>Black</td>
<td>1</td>
<td>14.8</td>
</tr>
</tbody>
</table>

1. **Steps** to naming variables:
   a. Click on the **Variable View** tab at the bottom of the spreadsheet. The window is shown below:
b. You can now enter the name of the variables you wish to include in your dataset in the first column (called Name). Note that SPSS variable names cannot include spaces or special characters. An example for the hypothetical dataset is shown below:

![SPSS Data Editor](image1.png)

SPSS automatically fills in the information in some of the remaining columns. These will need to be adjusted accordingly.

2. **Steps** to changing variable type:
   a. The type of variable is indicated in column 2: Type. The default is Numeric. If you wish to change variable type, click on the word Numeric. A blue square with three dots will appear:

![Variable Type](image2.png)

b. Click on the blue square. A window opens up. The window is shown below:

![Variable Type Window](image3.png)
c. Select the desired option. In the hypothetical data set, the variable haircolor is not made up of numbers, but names. We call this a string variable. You may change the Type of this variable to String.

d. Click OK.

3. Steps to changing variable size:
   a. The width of variables is indicated in column 3: Width. The number of decimals used is indicated in column 4: Decimals. The default is set to a maximum of 8 with 2 decimals. If you wish to change this, you can click on the respective numbers and use the arrows that appear:

   b. In the hypothetical data set, the variable gender is coded as 1 and 2. We do not need any decimals, so you may change this to 0. We also know we will never perform any mathematical operations on the 1’s and 2’s, so we only need, say, 2 spaces for it. Change the Decimals to 0 and change the Width to 2.

   c. The size for a string variable will be adjusted automatically to accommodate longer names if necessary.

4. Steps to adding labels to the variables:
   a. The fifth column is called Label. In this column you can enter a full name for the variables. This will be the name that will show up in the output. In the variable label, we may also use spaces and symbols. In the hypothetical data set, the time represents the time it took the person to solve a particular complicated calculus problem. As a Name we might enter time and as a Label we might use Solving Time. Name the other variables in a similar way:
5. **Steps** to adding labels to the values:
   a. SPSS often requires a categorical variable to be treated as numerical. This leads us to the need of coding. In the hypothetical data set, the variable gender is coded as 1 and 2. We could have entered it as Male and Female, but for later use in an analysis, the 1’s and 2’s are a better option. We can now assign labels to the numerical values: 1 = Male and 2 = Female. Labels are assigned in column 6 (named **Values**).
   b. In the row for the variable Gender, click on the **Values** column, where it currently says **None**.
   c. A small blue square with three dots will appear. Click on the square.
   d. A window opens up. The window is shown below:

   ![Image of Value Labels window]

   e. In the **Value** box, enter the number 1. In the **Label** box, enter **Male**. Click **Add**.
   f. In the **Value** box, enter the number 2. In the **Label** box, enter **Female**. Click **Add**.
   g. When you are all finished labeling the categories, click **OK**. The SPSS output will now show Male versus Female when the variable Gender is used, instead of 1 versus 2.

6. **Steps** to treat missing values:
   a. Missing values are treated in column 7 (named **Missing**). Datasets often have values missing from it, when people forget to answer a question or wish not to answer a certain question. When we enter the data (see Section 1.6) we may just leave the space blank. SPSS will consider this a missing value. However, often we choose to use a certain numerical value to indicate the data is missing. We usually choose a value that is not possible for that particular variable, such as 999 for the variable age. It is also possible to use two different values for two different reasons: ‘Question was left blank’ (missing) and ‘Person refused to answer question’ (omitted). We can now assign numerical values to these missing values:
   b. For each of the three variables we will use the code 999 to indicate a value is missing. In each row, click on the **Missing** column, where it currently says **None**.
   c. A small blue square with three dots will appear. Click on the square.
   d. A window opens up. The window is shown to the right:
   e. Click on the bullet in front of **Discrete missing values**.
   f. Type the value **999** in the box and click **OK**.
   g. Repeat these steps for all variables.

7. **Steps** to change measurement scale:
   a. The measurement scale of each variable can be indicated in column 10 (named **Measure**). When we analyze data, the analysis technique we use depends on the scale of measurement. It is important to have the correct scale assigned to each variable both for graphing and analyzing purposes.
b. When the variable haircolor was changed to be a string variable, SPSS automatically changed the measurement scale from Unknown to Nominal. But we need to adjust the other two variables manually.

c. The variable gender consists of the numbers 1 and 2. However, we cannot perform any mathematical operations on these (what is the average gender?). Gender is considered to be measured on the nominal scale. Click in the column Measure on the word Unknown. A small drop-down menu will appear. The menu is shown to the right:

d. Give the variable the appropriate measurement level (Nominal) by clicking on it.

e. Categorical variables that have categories that have a specific order need to be given the measurement level Ordinal. Numerical variables (variables on which we can perform mathematical operations) need to be given the measurement level Scale in this column.

f. Your spreadsheet should now look like this:

Note: The columns Columns, Align, and Role may remain unchanged. Some dialogs support predefined roles that can be used to pre-select variables for analysis. When you open one of these dialogs, variables that meet the role requirements will be automatically displayed in the destination list(s). Available roles are: Input (predictor or independent variable), Target (output or dependent variable), Both (input and output), or None.

1.6 Entering Data into the Data Editor

1. Steps to enter the data:
   a. Click on the tab Data View. Your spreadsheet should now look like this:
b. Begin typing the data in the first column. At this point you need to remember how you set up the variables (see Section 1.5). Enter the hair colors as words. Use the second column for the 1’s and 2’s representing gender and the third column for the times. Don’t forget to enter 999 for the missing value!
c. The spreadsheet should look like this:

![Spreadsheet Image]

1.7 Saving your Work

1. **Steps** to saving a data file:
   a. To save a data file, we use the File menu. Make sure you are in the SPSS Data Editor.
   b. Scroll down to the Save As... option. The usual Save As window opens up. An alternative is to use the “Save this document” button at the top left-hand corner of the window.
   c. Name your data set and save it in a location where you will be able to find it next time.

2. **Steps** to saving graphs and output:
   a. Graphs and other output can be saved in two ways: 1) as an output viewer file (.spv file) and 2) copy the work into a word processing file such as a Word file.
   b. To save the output file itself, we use the File menu. Make sure you are in the IBM SPSS Statistics Viewer.
   c. Scroll down to the Save As... option. The usual Save As window opens up.
   d. Name your output file and save it in a location where you will be able to find it next time.
   e. To copy (some of) the output into a Word file, click on the sections you wish to copy.
   f. Use the Edit menu and scroll down to the Copy option or right-click on the output and select Copy.
   g. You may now paste the output into a Word document.
   h. In case of numerical output, use the Edit menu and scroll down to the Paste Special option. Select the Picture (Windows Metafile) option. Click OK.
   i. In case of graphs, use the regular Paste command.
   j. Save the Word file in the usual manner.
Chapter 2

Turning Data into Information

2.1 Raw Data

Opening and creating datasets are covered in Chapter 1.

2.2 Types of Variables

The Problem – Randomly Pick S or Q
A group of college students is given a list of 8 questions:
1. What is your sex (male or female)?
2. How many hours did you sleep last night?
3. Randomly pick a letter -- S or Q.
4. What is your height in inches?
5. Randomly pick a number between 1 and 10.
6. What’s the fastest you’ve ever driven a car (mph)?
7. What is your right handspan in centimeters?
8. What is your left handspan in centimeters?
For questions 7 and 8 a centimeter ruler was provided. For question 3, some students were asked to randomly pick a letter -- S or Q, others were asked to pick Q or S.

a. Open the data set pennstate1.sav.
b. The Data View of this dataset should look like this:
The variables in the file are:
- sex
- hrrsleep
- sqpick
- height
- randnumb
- fastest
- rtspan
- lftspan
- form

Each of these variables represents the responses to these questions for the selected students. From a statistical standpoint, these variables can be labeled as categorical or quantitative. In SPSS we treat variables as 1 of several types. The variable type can be seen by using the VariableView tab at the bottom of the IBM SPSS Statistics Data Editor screen. The variable sex is recorded as Male or Female. It is treated as a string variable. Other string variables are sqpick (S or Q) and form (SorQ or QorS). All other variables in the data set are treated as numeric. There is no distinction between discrete and continuous variables.

To view all types of data used in SPSS, click on the variable type of one of the variables. A gray box with 3 dots in it will appear. Click on the box. A window opens up. The window is shown on the next page:

Here the type of data can be specified and/or changed, when typing in your own data, or when recoding an existing data set.
2.3 Summarizing One or Two Categorical Variables

Example 2.1 – Seatbelt Use by Twelfth-Graders
How often do you wear a seatbelt when driving a car? This is one of many questions asked in a biennial nationwide survey of American high school students. Survey questions concern potentially risky behaviors such as cigarette smoking, alcohol use, and so on. For the question about seatbelt use when driving, possible answers were Always, Most times, Sometimes, Rarely, and Never. The data is recorded in the file youthrisk03.sav.

1. **Steps** to prepare the data:
   a. Open the data set youthrisk03.sav.

2. **Steps** to create a frequency table with 1 categorical variable:
   a. To obtain a frequency table listing the number of students who answered the seatbelt question with each of the 5 possible answers, we use the **Analyze** menu.
   b. Scroll down to the **Descriptive Statistics** submenu and select the **Frequencies** option. A window opens up. The window is shown below:

   ![Frequencies window](image)

   c. Select the variable **Seatbelt** and move it into the **Variable(s)** box, as shown above.
   d. Click **OK**.

   **The SPSS Output**

   **Frequencies**
   **Statistics**

<table>
<thead>
<tr>
<th>seatbelt</th>
<th>N</th>
<th>Valid</th>
<th>Missing</th>
</tr>
</thead>
<tbody>
<tr>
<td>seatbelt</td>
<td></td>
<td>3042</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>seatbelt</th>
<th>Frequency</th>
<th>Percent</th>
<th>Valid Percent</th>
<th>Cumulative Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Valid</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1_Never</td>
<td>115</td>
<td>3.8</td>
<td>3.8</td>
<td>3.8</td>
</tr>
<tr>
<td>2_Rarely</td>
<td>249</td>
<td>8.2</td>
<td>8.2</td>
<td>12.0</td>
</tr>
<tr>
<td>3_Sometimes</td>
<td>414</td>
<td>13.6</td>
<td>13.6</td>
<td>25.6</td>
</tr>
<tr>
<td>4_Mosttimes</td>
<td>578</td>
<td>19.0</td>
<td>19.0</td>
<td>44.6</td>
</tr>
<tr>
<td>5_Always</td>
<td>1686</td>
<td>55.4</td>
<td>55.4</td>
<td>100.0</td>
</tr>
<tr>
<td>Total</td>
<td>3042</td>
<td>100.0</td>
<td>100.0</td>
<td></td>
</tr>
</tbody>
</table>
3. **Steps** to create a frequency table with 2 categorical variables:
   a. To obtain a frequency table of the number of students who answered the seatbelt question with each of the 5 possible answers, separated by gender, we use the **Analyze** menu.
   b. Scroll down to the **Descriptive Statistics** submenu and select the **Crosstabs** option. A window opens up. The window is shown below:
   
   ![Crosstabs Window](image)
   
   c. Select the variable **Gender** and move it into the **Row(s)** box.
   d. Select the variable **Seatbelt** and move it into the **Column(s)** box.
   e. Click on **Cells**. A second window opens up. The window is shown to the right:
   f. Select the **Row** option under the **Percentages** heading, as shown. Click **Continue**.
   g. Click **OK**.
The SPSS Output

Crosstabs

Case Processing Summary

<table>
<thead>
<tr>
<th></th>
<th>Valid</th>
<th></th>
<th>Missing</th>
<th></th>
<th>Total</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N</td>
<td>Percent</td>
<td>N</td>
<td>Percent</td>
<td>N</td>
<td>Percent</td>
</tr>
<tr>
<td>gender * seatbelt</td>
<td>3042</td>
<td>100.0%</td>
<td>0</td>
<td>.0%</td>
<td>3042</td>
<td>100.0%</td>
</tr>
</tbody>
</table>

gender * seatbelt Crosstabulation

<table>
<thead>
<tr>
<th>gender</th>
<th>seatbelt</th>
<th>Count</th>
<th>1_Never</th>
<th>2_Rarely</th>
<th>3_Sometimes</th>
<th>4_Mosttimes</th>
<th>5_Always</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Count</td>
<td>25</td>
<td>84</td>
<td>167</td>
<td>276</td>
<td>915</td>
<td>1467</td>
<td></td>
</tr>
<tr>
<td></td>
<td>% within gender</td>
<td>1.7%</td>
<td>5.7%</td>
<td>11.4%</td>
<td>18.8%</td>
<td>62.4%</td>
<td>100.0%</td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>Count</td>
<td>90</td>
<td>165</td>
<td>247</td>
<td>302</td>
<td>771</td>
<td>1575</td>
<td></td>
</tr>
<tr>
<td></td>
<td>% within gender</td>
<td>5.7%</td>
<td>10.5%</td>
<td>15.7%</td>
<td>19.2%</td>
<td>49.0%</td>
<td>100.0%</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>Count</td>
<td>115</td>
<td>249</td>
<td>414</td>
<td>578</td>
<td>1686</td>
<td>3042</td>
<td></td>
</tr>
<tr>
<td></td>
<td>% within gender</td>
<td>3.8%</td>
<td>8.2%</td>
<td>13.6%</td>
<td>19.0%</td>
<td>55.4%</td>
<td>100.0%</td>
<td></td>
</tr>
</tbody>
</table>

Example 2.3 – Humans Are Not Good Randomizers

As part of the survey described in Section 2.2, the students were asked to “Randomly pick a number between 1 and 10.” The data is recorded in the file pennstate1.sav.

1. **Steps** to prepare the data:
   a. Open the data set pennstate1.sav.

2. **Steps** to create a Pie Chart of one categorical variable:
   a. To create a pie chart, we use the **Graphs** menu.
   b. Click on the **Chart Builder** option. You may receive the following warning box reminding you to be sure you have set each measurement level properly and value labels for each categorical variable.
You can click on OK and a window opens up. The window is shown below:

c. In the **Gallery** at the bottom of the window, select the **Pie/Polar** option:
Drag the **Pie Chart** to the **Chart Preview** at the top of the window:

![Pie Chart](image)

d. Select the variable **randnumb** and move it into the **Slices by?** box.

e. Click on **Element Properties** on the right side of the window:

![Element Properties](image)

Note: Often times the **Element Properties** window will automatically open when creating a graph. If this is the case, this step (e) may be skipped.

f. In the **Statistics** box in the Element Properties box, select **Percentage()** from the pull down menu:

![Statistics](image)

g. Click **Apply** at the bottom of the **Element Properties** window.
h. Click **OK** in the **Chart Builder** window.

![Pie Chart]

The SPSS Output

Note: To obtain the percentages for each slice, use the following steps:
- Double click on the graph. The **Chart Editor** opens up.
- Click on the **Elements** menu and scroll down to the **Show Data Labels** option.
- Close the **Chart Editor**. The graph will look like:

![Pie Chart with Percentages]
3. **Steps** to create a Bar Graph of one categorical variable:
   a. To create a bar graph, we use the **Graphs** menu.
   b. Click on the **Chart Builder** option.
   c. In the **Gallery** at the bottom of the window, select the **Bar** option.
   d. Drag the first type of bar graph (**Simple Bar**) to the **Chart Preview** at the top of the window.
   e. Select the variable **randnumb** and move it into the **X-Axis?** box.
   f. If you wish to display percentages, rather than counts, on the y-axis, select **Percentage()** in the **Statistics** box in the **Element Properties** window. Don’t forget to click **Apply**.
   g. Click **OK**.

![The SPSS Output](image)
**Example 2.4 – Lighting the Way to Nearsightedness**
A survey of 479 children found that those who had slept with a nightlight or in a fully lit room before the age of 2 had a higher incidence of nearsightedness later in childhood. The raw data consisted of two categorical variables, each with three categories. The data can be found in Example 2.2 of the book. The data is not recorded in an existing file, so the data set needs to be created.

1. **Steps** to prepare the data:
   a. Open a new data file.
   b. Use the **Variable View** to create three variables: **count**, **sleep**, and **eyesight**.
   c. We do not need decimals, so you may change the number of decimals to **0**.
   d. The variable **count** will need to be given the appropriate measurement level (**Scale**). The other two variables will need to be given the appropriate measurement level (**Ordinal**).
   e. The variable **sleep** will be a categorical variable indicating the sleeping circumstances (Darkness, Nightlight, Full Light). We will enter those as 1, 2, and 3. Give the values 1, 2, and 3 appropriate labels in the **Values** column.
   f. Repeat the above step for the variable **eyesight** with the categories No Myopia, Myopia, and High Myopia.
   g. Use the **Data View** to enter the 9 frequencies (155, 153, 34, 15, 72, 36, 2, 7, 5) into the first variable **count**.
   h. In the second column, indicate how the children slept: 1 = Darkness, 2 = Nightlight, 3 = Full Light (if the order indicated above is used, type 1, 2, 3, 1, 2, 3, 1, 2, 3).
   i. In the third column, indicate the incidence of nearsightedness: 1 = No Myopia, 2 = Myopia, 3 = High Myopia (if the order indicated above is used, type 1, 1, 1, 2, 2, 2, 3, 3, 3).
   j. Save the file.

2. **Steps** to prepare the data for a clustered bar graph:
   a. The count variable needs to be designated as a weighting variable. For this we use the **Data** menu.
   b. Scroll down to the **Weight Cases**… option. A window opens up. The window is shown to the right:
   c. Select the **Weight cases by** option.
   d. Select the variable **count** and move it into the **Frequency Variable** box.
   e. Click **OK**.

![Weight Cases](image)

3. **Steps** to create a Bar Graph of two categorical variables:
   a. To create a bar graph, we use the **Graphs** menu.
   b. Click on the **Chart Builder** option.
   c. In the **Gallery** at the bottom of the window, select the **Bar** option.
   d. Drag the second type of bar graph (**Clustered Bar**) to the **Chart Preview** at the top of the window.
   e. Select the variable **sleep** and move it into the **X-Axis?** box.
   f. Select the variable **eyesight** and move it into the **Cluster on X: set color** box.
   g. To display percentages on the y-axis, select **Percentage()** in the **Statistics** box in the **Element Properties** window.
   h. Click on the **Set Parameters** button. A small window will open.
   i. From the drop-down menu, select **Total for Each X-Axis Category** as shown above.
   j. Click **Continue**.
   k. Click **Apply** in the **Element Properties** window.
   l. Click **OK** in the **Chart Builder** window.
2.4 Exploring Features of Quantitative Data with Pictures

Example 2.5 – Right Handspans
As part of the survey described in Section 2.2, the students were asked to measure the span of their right hands (in cm). The data is recorded in the file pennstate1.sav.

1. Steps to prepare the data:
   a. Open the data set pennstate1.sav.

2. Steps to create a dotplot:
   a. To create a dotplot, we use the **Graphs** menu.
   b. Click on the **Chart Builder** option.
   c. In the **Gallery** at the bottom of the window, select the **Scatter/Dot** option.
   d. Drag the sixth graph (**Simple Dot Plot**) to the **Chart Preview** at the top of the window.
   e. Select the variable **rtspan** and move it into the **X-Axis?** box.
   f. To separate the graph between males and females, click on the **Groups/Point ID** tab (the **Gallery** will disappear):
g. Check the box by **Rows panel variable**.

h. Select the variable **sex** and move it into the **Panel?** box:

i. Click **OK**.

**The SPSS Output**
3. **Steps** to obtain the five-number summary:
   a. To obtain the five-number summary, we use the **Analyze** menu.
   b. Scroll down to the **Descriptive Statistics** submenu and select the **Frequencies** option. A window opens up. The window is shown to the right:
   c. Select the variable *rtspan* and move it into the **Variable(s)** box.
   d. Remove the checkmark in front of **Display frequency tables at the bottom** of the window.
   e. Click on the **Statistics** button. A second window opens up. The window is shown below:
   
   ![Frequencies window]
   
   f. Select the **Quartiles** at the top of the window and the **Minimum** and **Maximum** at the bottom of the window, as shown above. Click **Continue**.
   g. Click **OK**.

**The SPSS Output**

```
rtspan

<table>
<thead>
<tr>
<th>Statistics</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>190</td>
</tr>
<tr>
<td>Valid</td>
<td></td>
</tr>
<tr>
<td>Missing</td>
<td>0</td>
</tr>
<tr>
<td>Minimum</td>
<td>12.50</td>
</tr>
<tr>
<td>Maximum</td>
<td>26.00</td>
</tr>
<tr>
<td>Percentiles</td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>20.0000</td>
</tr>
<tr>
<td>50</td>
<td>21.5000</td>
</tr>
<tr>
<td>75</td>
<td>22.5000</td>
</tr>
</tbody>
</table>
```
4. **Steps** to obtain the five-number summary separated by gender:
   a. To obtain the five-number summary separate for males and females, we use the **Data** menu.
   b. Scroll down to the **Split File** option. A window opens up. The window is shown below:
   c. Select the **Compare Groups** option.
   d. Select the variable **sex** and move it into the **Groups Based on** box.
   e. Click on **OK**.
   f. Now repeat the steps as described in part (3) above.

   **The SPSS Output**

   **Statistics**

<table>
<thead>
<tr>
<th></th>
<th>Female</th>
<th></th>
<th>Male</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N</td>
<td>Valid</td>
<td>N</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>103</td>
<td>87</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Minimum</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>12.50</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Maximum</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>23.25</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Percentiles</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>25</td>
<td>19.0000</td>
<td></td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>20.0000</td>
<td></td>
</tr>
<tr>
<td></td>
<td>75</td>
<td>21.0000</td>
<td></td>
</tr>
</tbody>
</table>
Pictures for Quantitative Data

The Problem – Right Handspans
As part of the survey described in Section 2.2, the students were asked to measure the span of their right hands (in cm). The data for the female students is recorded in the file pennstate1F.sav.

1. **Steps** to prepare the data:
   a. Open the data set pennstate1F.sav.

2. **Steps** to create a Histogram:
   a. To create a histogram, we use the **Graphs** menu.
   b. Click on the **Chart Builder** option.
   c. In the **Gallery** at the bottom of the window, select the **Histogram** option.
   d. Drag the first graph (**Simple Histogram**) to the **Chart Preview** at the top of the window.
   e. Select the variable **rtspan** and move it into the **X-Axis?** box.
   f. Click **OK**.

The SPSS Output
Note: to change the class widths and make the graph match the one in the book, follow the following steps:

a. Double click on the graph. The Chart Editor opens up.
b. Click on the bars of the graph. In the Properties window (if this does not automatically open, double click on the bars of the graph) click on the Binning tab.
c. To define the width of the bars, select the Custom option under X-axis.
d. Enter a width of 1.
e. Also check the option Custom value for anchor and enter 11.5.
f. Click Apply.
g. To remove the legend from the graph, click on it and hit the delete key.
h. Close the Chart Editor window.

The SPSS Output

### Steps to create a Stem-and-leaf plot:

a. To create a stem-and-leaf plot, we will use the Analyze menu and not the Graphs menu.
b. Under the Analyze menu, scroll down to the Descriptive Statistics submenu and select the Explore option.
c. A window opens up. The window is shown to the right:
d. Select the variable \texttt{rtspan} and move it into the \textit{Dependent List} box.
e. Select the \textit{Plots} option under the \textit{Display} heading on the bottom left of the window.
f. Click on the \textit{Plots} button. A second window opens up. The window is shown below:

![Image of the window with the Plots option selected]

g. Select the \texttt{None} option under the \textit{Boxplots} heading and the \texttt{Stem-and-leaf} option under the \textit{Descriptive} heading, as shown above. Click \textit{Continue}.
h. Click \textit{OK}.

The SPSS Output

\texttt{rtspan}

\texttt{rtspan Stem-and-Leaf Plot}

<table>
<thead>
<tr>
<th>Frequency</th>
<th>Stem &amp; Leaf</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.00</td>
<td>Extremes (=&lt;13.0)</td>
</tr>
<tr>
<td>1.00</td>
<td>16 . 0</td>
</tr>
<tr>
<td>.00</td>
<td>16 .</td>
</tr>
<tr>
<td>2.00</td>
<td>17 . 00</td>
</tr>
<tr>
<td>1.00</td>
<td>17 . 5</td>
</tr>
<tr>
<td>7.00</td>
<td>18 . 000000</td>
</tr>
<tr>
<td>4.00</td>
<td>18 . 5558</td>
</tr>
<tr>
<td>15.00</td>
<td>19 . 000000000000000000000000000000000022</td>
</tr>
<tr>
<td>4.00</td>
<td>19 . 5555</td>
</tr>
<tr>
<td>23.00</td>
<td>20 . 0000000000000000000000000000000022</td>
</tr>
<tr>
<td>6.00</td>
<td>20 . 555555</td>
</tr>
<tr>
<td>17.00</td>
<td>21 . 000000000000000000000000000000000022</td>
</tr>
<tr>
<td>5.00</td>
<td>21 . 55557</td>
</tr>
<tr>
<td>8.00</td>
<td>22 . 000000000000000000000000000000000022</td>
</tr>
<tr>
<td>6.00</td>
<td>22 . 555555</td>
</tr>
<tr>
<td>2.00</td>
<td>23 . 02</td>
</tr>
</tbody>
</table>

Stem width: 1.00
Each leaf: 1 case(s)
4. **Steps** to create a Dotplot:
   These steps have been described on p. 21-22.

5. **Steps** to construct a Boxplot:
   a. Open the data set pennstate1.sav.
   b. To construct boxplots, we will use the **Graphs** menu.
   c. Click on the **Chart Builder** option.
   d. In the **Gallery** at the bottom of the window, select the **Boxplot** option.
   e. Drag the first graph (**Simple Boxplot**) to the **Chart Preview** at the top of the window.
   f. Select the variable **rtspan** and move it into the **Y-Axis?** box.
   g. Select the variable **sex** and move it into the **X-Axis?** box.
   h. Click **OK**.

![The SPSS Output](image)

### 2.5 Numerical Summaries of Quantitative Variables

**Example 2.14 – Fastest Driving Speed for Men**
As part of the survey described in Section 2.2, the students were asked “What is the fastest you have ever driven a car?” The answers were given in miles per hour (mph). The data for the male students are recorded in the file pennstate1M.sav.

1. **Steps** to prepare the data:
   a. Open the data set pennstate1M.sav.
2. **Steps** to obtain summary measures:
   a. To calculate summary measures, we use the **Analyze** menu.
   b. Scroll down to the **Descriptive Statistics** submenu and select the **Descriptives** option. A window opens up. The window is shown below:
   c. Select the variable **fastest** and move it into the **Variable(s)** box.
   d. Click on the **Options** button.
   e. Another window opens up. The window is shown to the right:
   f. Select the summary measures you wish to calculate. Click **Continue**.
   g. Click **OK**.

   ![Descriptives Window](image)

   ![Descriptives Options Window](image)

   **The SPSS Output**

   **Descriptives**

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>Range</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Mean</th>
<th>Std. Deviation</th>
<th>Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>fastest</td>
<td>87</td>
<td>95</td>
<td>55</td>
<td>150</td>
<td>107.4</td>
<td>17.434</td>
<td>303.941</td>
</tr>
<tr>
<td>Valid N (listwise)</td>
<td>87</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

3. **Steps** to obtain more summary measures:
   a. To calculate summary measures, we will use the **Analyze** menu.
   b. Scroll down to the **Descriptive Statistics** submenu and select the **Frequencies** option.
c. A window opens up. The window is shown below:

![Window showing variable selection for fastest]


d. Select the variable **fastest** and move it into the **Variable(s)** box.

e. Click on the **Statistics** button.

f. Another window opens up. The window is shown to the right:

g. Select the summary measures you wish to calculate. You can enter the percentiles you wish to calculate. Click on **Add** after each value, as shown here.

h. Click **Continue**.

i. In the **Frequencies** window, uncheck the **Display frequency tables** option at the bottom left of the window:

![Frequency window with uncheck option]

j. Click **OK**.
### The SPSS Output

#### Frequencies

<table>
<thead>
<tr>
<th>Statistics</th>
<th>fastest</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>Valid: 87</td>
</tr>
<tr>
<td></td>
<td>Missing: 0</td>
</tr>
<tr>
<td>Mean</td>
<td>107.40</td>
</tr>
<tr>
<td>Median</td>
<td>110.00</td>
</tr>
<tr>
<td>Mode</td>
<td>110</td>
</tr>
<tr>
<td>Std. Deviation</td>
<td>17.434</td>
</tr>
<tr>
<td>Variance</td>
<td>303.941</td>
</tr>
<tr>
<td>Range</td>
<td>95</td>
</tr>
<tr>
<td>Minimum</td>
<td>55</td>
</tr>
<tr>
<td>Maximum</td>
<td>150</td>
</tr>
<tr>
<td>Percentiles</td>
<td>5: 80.00</td>
</tr>
<tr>
<td></td>
<td>10: 85.00</td>
</tr>
<tr>
<td></td>
<td>25: 95.00</td>
</tr>
<tr>
<td></td>
<td>50: 110.00</td>
</tr>
<tr>
<td></td>
<td>75: 120.00</td>
</tr>
<tr>
<td></td>
<td>90: 126.00</td>
</tr>
<tr>
<td></td>
<td>95: 140.00</td>
</tr>
</tbody>
</table>
2.7 Bell-Shaped Distributions and Standard Deviations

The Problem – Students’ Heights and Parents’ Heights
As part of a survey, students in a statistics class were asked for their own height and for the height of their father and mother. The data is recorded in the file UCDavis2.sav.

1. **Steps** to prepare the data:
   a. Open the data set UCDavis2.sav.

2. **Steps** to create a Histogram with a normal curve superimposed:
   a. To create a histogram, we use the **Graphs** menu.
   b. Click on the **Chart Builder** option:

   ![Chart Builder Interface]

   c. In the **Gallery** at the bottom of the window, select the **Histogram** option.
   d. Drag the first graph (Simple Histogram) to the **Chart Preview** at the top of the window.
   e. Select the variable **momheight** and move it into the X-Axis? box.
f. To display the normal distribution curve in the graphs, check the box **Display normal curve** in the **Element Properties** window.

![Element Properties Window]

- **Display normal curve**

- **Variable:** monthheight
  - **Statistic:** Histogram

- **Error Bars Represent**
  - Confidence intervals
  - Standard error
  - Standard deviation

- **Bar Style:** Bar

- **Apply**

- **Cancel**

- **Help**

- **Bar1**
  - X-Axis1 (Bar1)
  - Y-Axis1 (Bar1)

- Click **Apply** in the **Element Properties** window.

- Click **OK** in the **Chart Builder** window.

- **Chart Builder** window

- **Apply**

- **Cancel**

- **Help**

- **Chart Builder**

- **OK**
3. **Steps** to calculate z-scores:
   a. To transform data, we use the **Transform** menu.
   b. Scroll down to the **Compute Variable** option.

(Note the outlier on the left!)
c. Type a name for the transformed variable in the **Target Variable** box, e.g. *z_momheight*.
d. In the **Numeric Expression** box type the formula for the z-score. Use the variable *momheigh* for the *x*-variable, and use the mean and standard deviation obtained with the histogram in step 2:

```
[meanheigh 63.744, 3.185]
```
e. Click **OK**.
f. The new variable *z_momheight* will appear as a new variable in the **Data Editor**.

---

**The Data Editor**

![Data Editor Image]
Note: When we create a histogram of the z-scores, we see that the mean and standard deviation are now 0 and 1 respectively.
Chapter 3

Relationships Between Quantitative Variables

3.1 Looking for Patterns with Scatterplots

Example 3.1 – Height and Handspan
The heights (in inches) and fully stretched handspans (in cm) of 167 college students are recorded in the file handheight.sav. A connection between the 2 variables is to be studied.

1. Steps to prepare the data:
   a. Open the data set handheight.sav.

2. Steps to create a scatterplot:
   a. To create a scatterplot, we use the Graphs menu.
   b. Click on the Chart Builder option.
   c. In the Gallery at the bottom of the window, select the Scatter/Dot option.
   d. Drag the first graph (Simple Scatter) to the Chart Preview at the top of the window.
   e. Select the variable handspan and move it into the Y-axis? box.
   f. Select the variable height and move it into the X-axis? box.
   g. Click OK.

The SPSS Output
Note: To adjust the scale on the Y-axis, so the smallest observation can be more easily seen, follow the following steps:

a. Double click on the graph. The Chart Editor opens up.
b. Click on the Y-axis of the graph. In the Properties window (if this does not automatically open, double click on the Y-axis) click on the Scale tab.
c. Because we only wish to change the minimum value on the Y-axis, remove the checkmark in front of Minimum and change the value to 14 under the Custom heading.
d. Click Apply.
e. Close the Chart Editor window.

3. **Steps** to create a scatterplot indicating two groups:
   a. To create a scatterplot, we use the Graphs menu.
   b. Click on the Chart Builder option.
   c. In the Gallery at the bottom of the window, select the Scatter/Dot option.
d. Drag the second graph (Grouped Scatter) to the Chart Preview at the top of the window.
e. Select the variable handspan and move it into the Y-axis? box.
f. Select the variable height and move it into the X-axis? box.
g. Select the variable sex and move it into the Set Color box.
h. Click OK.

Note: The default method to indicate the different groups is to use two different colors of the same symbol. To change the plotting symbol for one of the groups, use the following steps:
   a. Double click on the graph. The Chart Editor opens up.
   b. Double click on the plotting symbol in the legend that you wish to change.
   c. In the Properties window, select a different plotting symbol under the Type option. You may also change the color and size at this moment.
   d. You may repeat this for the other symbol(s).
   e. Click Apply:
   f. Close the SPSS Chart Editor.
3.2 Describing Linear Patterns with a Regression Line

Example 3.7 – Driver Age and the Maximum Legibility Distance of Highway Signs
In a study of the legibility and visibility of highway signs, a Pennsylvania research firm determined the maximum distance at which each of 30 drivers could read the newly designed sign. The 30 participants in the study ranged in age from 18 to 82 years old. The government agency that funded the research hoped to improve highway safety for older drivers and wanted to examine the relationship between age and the sign legibility distance. The data is recorded in the file signdist.sav.

1. Steps to prepare the data:
   a. Open the data set signdist.sav.

2. Steps to create a scatterplot with the regression line superimposed:
   a. To create a scatterplot, we use the Graphs menu.
   b. Click on the Chart Builder option.
   c. In the Gallery at the bottom of the window, select the Scatter/Dot option.
   d. Drag the first graph (Simple Scatter) to the Chart Preview at the top of the window.
   e. Select the variable distance and move it into the Y-axis? box.
   f. Select the variable age and move it into the X-axis? box.
   g. Click OK.
   h. Double click on the graph. The Chart Editor opens up.
   i. Under the Elements menu scroll down to Fit Line at Total:

   j. The regression line will automatically be drawn on the graph.
   An alternative to this option is to click on the Fit Line at Total button at the top of the Chart Editor window and then Close the SPSS Chart Editor.
3. **Steps** to find the equation of the least squares regression line and other regression summaries:
   a. To find the equation and produce other regression summaries, we can use the **Analyze** menu.
   b. Scroll down to the **Regression** submenu and select the **Linear** option.
   c. A window opens up. The window is shown below:
   d. Select the variable **distance** and move it into the **Dependent** box.
   e. Select the variable **age** and move it into the **Independent(s)** box.
   f. Click **OK**.
The SPSS Output

Regression

Variables Entered/Removeda

<table>
<thead>
<tr>
<th>Model</th>
<th>Variables Entered</th>
<th>Variables Removed</th>
<th>Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>age²</td>
<td>.</td>
<td>Enter</td>
</tr>
</tbody>
</table>

a. All requested variables entered.
b. Dependent Variable: distance

Model Summary

<table>
<thead>
<tr>
<th>Model</th>
<th>R</th>
<th>R Square</th>
<th>Adjusted R Square</th>
<th>Std. Error of the Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.801a</td>
<td>.642</td>
<td>.629</td>
<td>49.762</td>
</tr>
</tbody>
</table>

a. Predictors: (Constant), age

ANOVar

<table>
<thead>
<tr>
<th>Model</th>
<th>Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>F</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Regression</td>
<td>1</td>
<td>124332.843</td>
<td>60.211</td>
<td>.000b</td>
</tr>
<tr>
<td></td>
<td>Residual</td>
<td>26</td>
<td>2476.215</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>27</td>
<td>193566.667</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

a. Dependent Variable: distance
b. Predictors: (Constant), age

Coefficientsa

<table>
<thead>
<tr>
<th>Model</th>
<th>Unstandardized Coefficients</th>
<th>Standardized Coefficients</th>
<th>t</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>B</td>
<td>Std. Error</td>
<td>Beta</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>(Constant)</td>
<td>576.682</td>
<td>23.471</td>
<td></td>
</tr>
<tr>
<td></td>
<td>age</td>
<td>-3.007</td>
<td>.424</td>
<td>-7.086</td>
</tr>
</tbody>
</table>

a. Dependent Variable: distance

The output consists of 4 boxes. The least squares regression equation may be obtained from the last (fourth) box in the output labeled Coefficients. The estimates are both given in the first column labeled B. The y-intercept is given in the first row labeled (Constant) and the slope is given in the second row labeled age. For this example the least squares regression equation is \( \hat{y} = 576.682 - 3.007x \).
4. **Steps** to find the residuals:
   a. To find the residuals, we use the **Analyze** menu.
   b. Scroll down to the **Regression** submenu and select the **Linear** option.
   c. Select the variable **distance** and move it into the **Dependent** box.
   d. Select the variable **age** and move it into the **Independent(s)** box. If continuing from part (2), these will already be filled in.
   e. Click on **Save**. A window opens up. The window is shown to the right:
   f. Under the **Residuals** heading select the **Unstandardized** option. Click **Continue**.
   g. Click **OK**.
   h. You will once again get the regression output in the **Output** window. The residuals will appear in the **Data Editor** window as a variable named **RES_1**.
   i. The **Data Editor** is shown below:
5. **Steps** to find the sum of squared errors:
   a. Obtain the regression equation as described in part (3).
   b. Part of the output is called the ANOVA table.
   c. The **Sum of Squared Errors** is given, by default, in this table. It is located in the column named **Sum of Squares** and the row named **Residual**:

   ![ANOVA Table]

   a. Dependent Variable: distance
   b. Predictors: (Constant), age

3.3 Measuring Strength and Direction with Correlation

**Example 3.11 – Driver Age and the Sign Legibility Distance**
Continued from Section 3.2.

1. **Steps** to find the correlation coefficient:
   a. The **correlation coefficient** may be obtained from the second box in the regression output labeled Model Summary. This part of the output is shown below:

   ![Model Summary Table]

   a. Predictors: (Constant), age

   b. The magnitude of the correlation coefficient is given in the first column labeled R. Note that this is only the magnitude of the correlation coefficient. Since the sign of the correlation coefficient equals the sign of the slope, check with the slope if the correlation coefficient should be positive or negative.

   c. For this example, the correlation coefficient should be negative, since we have a negative slope. The correlation coefficient is \( r = -0.801 \).

2. **Steps** to find the correlation coefficient:
   a. The **correlation coefficient** may also be obtained independent from the regression output. To find the correlation, we use the Analyze menu.
   b. Scroll down to the Correlate submenu and select the Bivariate option.
c. A window opens up. The window is shown below:

![SPSS Bivariate Correlations Window]

d. Select the variables age and distance and move them into the Variables box.
e. Click OK.

The SPSS Output

**Correlations**

<table>
<thead>
<tr>
<th></th>
<th>age</th>
<th>distance</th>
</tr>
</thead>
<tbody>
<tr>
<td>age</td>
<td>1</td>
<td>-.801**</td>
</tr>
<tr>
<td>Sig. (2-tailed)</td>
<td></td>
<td>.000</td>
</tr>
<tr>
<td>N</td>
<td>30</td>
<td>30</td>
</tr>
<tr>
<td>distance</td>
<td>-.801**</td>
<td>1</td>
</tr>
<tr>
<td>Sig. (2-tailed)</td>
<td>.000</td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>30</td>
<td>30</td>
</tr>
</tbody>
</table>

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

f. The correlation can be found in either the top-right cell or the bottom-left cell of the output. In this case, the sign of the correlation is given as well. The correlation coefficient is $r = -0.801$. 
Example 3.16 – Driver Age and the Sign Legibility Distance
Continued from previous example.

The regression output for this example (= Example 3.7) is repeated below:

**The SPSS Output**

**Regression**

<table>
<thead>
<tr>
<th>Model</th>
<th>Variables Entered</th>
<th>Variables Removed</th>
<th>Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>age</td>
<td>.</td>
<td>Enter</td>
</tr>
</tbody>
</table>

Notes:
- The least squares regression equation may be obtained from the fourth box in the output labeled **Coefficients**. The estimates are both given in the first column labeled B. The y-intercept is given in the first row labeled (Constant) and the slope is given in the second row labeled age. For this example the least squares regression equation is \( \hat{y} = 576.682 - 3.007x \).
- The value of \( r^2 \) may be obtained from the second box in the output labeled **Model Summary**. The value is listed in the second column labeled **R Square**.
- Various Sums of Squares may be obtained from the third box in the output labeled **ANOVA** in the first column labeled **Sum of Squares**.
  - The SSR may be obtained from the first row labeled Regression. For this example SSR = 124,332.643.
  - The SSE may be obtained from the second row labeled Residual. For this example SSE = 69,334.024.
  - The SSTO may be obtained from the third row labeled Total. For this example SSTO = 193,666.667.
- We can check that SSTO = SSE + SSR.
Chapter 4

Relationships Between Categorical Variables

4.1 Displaying Relationships Between Categorical Variables

- The reader is referred to Section 2.3 part 3 in this manual for the steps to display relationships between categorical variables in a table.
- The reader is referred to Section 2.3 Example 2.4 in this manual for the steps to display relationships between categorical variables with a nested bar chart.

4.4 Assessing the Statistical Significance of a 2x2 Table

- The reader is referred to Section 15.2 in this manual for the steps to assess statistical significance of 2x2 tables.
- The reader is referred to Section 15.1 in this manual for the steps to obtain $p$-values for chi-squared statistics.
Chapter 5

Sampling: Surveys and How to Ask Questions

5.3 Simple Random Sampling

On the second day of classes, students in two statistics classes at the University of California at Davis were given a survey. One of the questions was, “In a typical week, about how much time do you spend watching television?” The data is recorded in the file UCDavis1.sav.

The Problem – Taking a Simple Random Sample

How do we take a simple random sample of 10 television watching amounts?

1. Steps to prepare the data:
   a. Open the data set UCDavis1.sav.

2. Steps to assign ID labels to all students:
   a. To assign the labels we use the Transform menu.
   b. Scroll down to the Compute Variable option.
   c. A window opens up. The window is shown below:

   ![Image of Compute Variable window]

   d. Type a variable name in the Target Variable box, such as ID.
   e. In the Function group box, select All.
   f. In the Functions and Special Variables box select $\text{SCasenum}$ and move it in the Numeric Expression box by clicking on the up arrow.
   g. Click OK.
h. A new variable has now been added to the **Data Editor** called **ID**. Each case has received a number between 1 and 173. Use the **Variable View** to change the number of decimals to **0**. Part of the window is shown below:

![Variable View](image)

3. **Steps** to take the simple random sample:
   a. To select a random sample, we use the **Data** menu.
   b. Scroll down to the **Select Cases** option.
   c. A window opens up. The window is shown below:

![Select Cases](image)

d. Select the **Random sample of cases** option.
e. Click on the **Sample** button. A second window opens up. The window is shown to the right:
f. Select the second option: **Exactly … cases from the first … cases**. Fill in the blanks as 10 cases from all cases. Note that the overall sample size equals 173, so use 173 in the second box, as shown above. Click **Continue**.
g. Click **OK**.
h. A **filter** variable has now been added to the **Data Editor**. Part of the window is shown below:

![Select Cases: Random Sample window]

i. Those students selected to be in the sample receive a 1, the remaining students receive a 0. The case numbers for these remaining students are crossed out. Use the values of the variable `tv` that are not crossed out as the first simple random sample of 10 television watching amounts. Any calculations you do will only include these 10 cases.

j. To view the 10 selected values all in a row, we may sort the data. The previously selected cases will end up as the first 10 cases. To do this, we use the **Data** menu.

k. Scroll down to **Sort Cases**.
l. A window opens up. The window is shown to the right:
m. Select the filter variable **10 from the first 173 cases** and move it into the **Sort by** box.
n. Select the **Descending** option under the **Sort Order** heading.
o. Click **OK**.
p. The previously selected cases are now the first 10 cases. The ID numbers for these 10 selected students are now listed at the top of the ID column. Part of the Data Editor window is shown below.

![Data Editor Window](image)

Notes:
Every time you repeat this procedure, the results will be different. The random number seed in SPSS changes every time you take a random sample.
- To make it so that you can repeat the process, you may set your own random number seed. Make sure you change the seed each time you use the random sample function. Under the Transform menu, use the Random Number Generators option.
- A window opens up. The window is shown below:

![Random Number Generators Window](image)

- To set the value, select the Active Generator Initialization option.
- Select the Fixed Value option and enter the seed value of your choice. Make sure to record what random number seed you used, so that the next time you wish to repeat the procedure, you can use the same seed.
- Make sure to return the settings to Random if you wish to create different samples next time!
• With random number seed of 65432 the ID numbers of the sampled students are: 14, 62, 65, 71, 79, 83, 118, 119, 136, and 160.
• Part of the **Data Editor** window is shown below:
Chapter 6

Gathering Useful Data for Examining Relationships

6.2 Designing a Good Experiment

Example 6.3 – Randomly Assigning Children to Weight-lifting Groups

Case Study 6.2 involves a randomized experiment in which 43 children were assigned to one of three treatment groups. Children in Group 1 were asked to perform weight-lifting repetitions with a heavy load, Group 2 performed more repetitions but with a moderate load, and Group 3 served as a control group and did not lift weights. According to the study report there were 15 children assigned to Group 1, 16 to Group 2 and the remaining 12 to Group 3. Assuming no blocks or pairs were used, how would we use SPSS to assign the children to the treatment groups?

1. **Steps to prepare the data:**
   a. Open a new data file.
   b. Type a 1 in the 30th location in the first column. Dots (periods) should appear in the first 29 spaces. Now type a 1 in the 43rd location of the first column. The idea is to create a temporary variable which contains 43 spots.

2. **Steps to assign ID labels to all children:**
   a. To assign the labels we use the **Transform** menu.
   b. Scroll down to the **Compute Variable** option.
   c. Type a variable name in the **Target Variable** box, such as **ID**.
   d. In the **Function group** box, select **All**.
   e. In the **Functions and Special Variables** box select **$Casenum** and move it in the **Numeric Expression** box by clicking on the up arrow.
   f. Click **OK**.
   g. A new variable has now been added to the **Data Editor** called **ID**. Use the **Variable View** to change the number of decimals to 0. Each case has received a number between 1 and 43.
   h. You may delete the column called **VAR00001**.
   i. Save the file. We are essentially creating a dummy variable of 43 cases to represent the 43 children.
3. **Steps** to take the first random sample:
   a. To select a random sample, we use the **Data** menu.
   b. Scroll down to the **Select Cases** option.
   c. Select the **Random sample of cases** option.
   d. Click on the **Sample** button. A second window opens up.
   e. Select the second option: **Exactly ... cases from the first ... cases**. Fill in the blanks as **15** cases from all cases. Note that the overall sample size equals 43, so use **43** in the second box. Click **Continue**.
   f. Click **OK**.
   g. A **filter** variable has now been added to the **Data Editor**. Those children selected to be in group 1 receive a **1**, the remaining children receive a **0**. The case numbers for these remaining children are crossed out.
   h. Part of the **Data Editor** window is shown below:

   ![Data Editor Window](image)

   i. Use the values of the variable **ID** that are not crossed out as the first random sample of 15 children to go into group 1.

4. **Steps** to take the second random sample:
   a. Before taking the second random sample, we need to sort the data. In order to be able to keep track of the first sample, change the name of the first filter variable in the **Variable View** (e.g. **first**).
   b. When sorting the data, we need the previously selected cases to end up as the last 15 cases. To do this, we use the **Data** menu.
   c. Scroll down to **Sort Cases**. A window opens up. The window is shown to the right:
   d. Select the variable **first** and move it into the **Sort by** box.
   e. Select the **Ascending** option under the **Sort Order** heading.
   f. Click **OK**.
g. The previously selected cases are now the last 15 cases. Part of the **Data Editor** window is shown below:

![Data Editor Window](image)

h. We can now select a random sample of 16 cases from the remaining 28 cases. To select the random sample, we use the **Data** menu.

i. Scroll down to the **Select Cases** option.

j. A window opens up. Select the **Random sample of cases** option.

k. Click on the **Sample** button.

l. Select the second option: **Exactly … cases from the first … cases**. Fill in the blanks as **16** cases from the first **28** cases.

m. Click **Continue**.

n. Click **OK**.

o. A new **filter** variable has now been added to the **Data Editor**. To keep track of the second sample, change the name of the filter variable in the **Variable View** (e.g. **second**). Those children selected to be in group 2 receive a **1**, the remaining children receive a **0**.
p. The case numbers for these remaining children are crossed out. Part of the Data Editor window is show below:

![Data Editor Window](image)

q. Use the values of the variable ID that are not crossed out as the second random sample of 16 children to go into group 2 and the remaining children to the control group. Sorting the data again (now in descending order by variable second) may come in useful.

r. Save the file.

Notes:
- Every time you repeat this procedure, the results will be different. The random number seed in SPSS changes every time you take a random sample. To make it to that you can repeat the process, you may set your own random number seed. See chapter 5 for instructions.
- With a random number seed of 2346 the samples are:
  Group 1: 5, 8, 10, 11, 12, 13, 15, 16, 25, 32, 35, 36, 37, 38, 39
  Group 2: 2, 3, 4, 6, 9, 14, 17, 19, 21, 23, 27, 29, 30, 31, 33, 41
  Group 3: 1, 7, 18, 20, 22, 24, 26, 28, 34, 40, 42, 43
Chapter 8

Random Variables

8.4 Binomial Random Variables

The Problem – Listing Binomial probabilities
The random variable $X$ follows the Binomial distribution with $n = 10$ and $p = 0.25$: $B(10, 0.25)$.

1. **Steps** to prepare the data:
   a. Open a new data file.
   b. Give the variable a name in the **Variable View** tab, e.g. $x$.
   c. Set the number of decimals to 0.
   d. In the **Data View** tab, enter the values 0 through 10 (the possible values of this Binomial random variable) in the column named $x$.
   e. Save the file.

2. **Steps** to calculate Binomial probabilities:
   a. To calculate the Binomial probabilities we use the **Transform** menu.
   b. Scroll down to the **Compute Variable** option. A window will open up. The window is shown below:

   c. Type a variable name to represent the probabilities in the **Target Variable** box, e.g. binprob.
   d. In the **Function Group** box, select **PDF & Noncentral PDF**.
   e. In the **Functions and Special Variables** box, select **Pdf.Binom** and move it in the **Numeric Expression** box by clicking on the up arrow.
f. The three question marks need to be replaced by the value of x, the sample size and the success probability. Place the cursor in the location of the first question mark. Click on the variable x and move it into the **Numeric Expression** box by clicking on the right arrow. Place the cursor on the location of the second question mark and type the sample size (10). Place the cursor on the location of the third question mark and type the success probability (0.25).

g. Click **OK**.

h. A new variable has now been added to the **Data Editor** called **binprob** containing the Binomial probabilities. The **Data Editor** is shown below:

![Data Editor](image)

i. To increase the number of significant digits, use the **Variable View** and change the **Decimals** to 3 or 4.

![Variable View](image)
Example 8.16 – Number of Girls in Ten Births
Let the random variable $X$ represent the number of girls born in 10 births. The random variable $X$ follows the Binomial distribution with $n = 10$ and $p = 0.488$: $B(10, 0.488)$. The value of $p$ is based on birth records in the United States.

1. **Steps** to calculate Binomial probabilities:
   a. Repeat the steps from part 1 above.
   b. Repeat the steps from part 2 above with $p = 0.488$. Call the variable `birthprob`.
   c. Repeat the steps from part 2 above with $p = 0.488$ but in the **Function Group** box select CDF & Noncentral CDF and in the **Functions and Special Variables** box select Cdf.Binom. Call the variable `cumprob`.
   d. Once again, you may want to increase the number of decimals using the **Variable View**.
   e. Part of the **Data Editor** is shown below:

   ![Data Editor](image)

   **Notes:**
   - To find $P(X = 7)$, you may use the value in the column `birthprob` next to $x = 7$. The answer is 0.1062.
   - To find $P(X \leq 7)$, you may use the value in the column `cumprob` next to $x = 7$. The answer is 0.9533.
   - To find $P(X \geq 7)$, we need to realize that $P(X \geq 7) = 1 - P(X < 7) = 1 - P(X \leq 6)$. Use the value in the column `cumprob` next to $x = 6$ and subtract from 1. The answer is $1 - 0.8471 = 0.1529$. 

59
8.5 Continuous Random Variables

Example 8.19 – Time Spent Waiting for the Bus
A bus arrives at the bus stop every 10 minutes. If a person arrives at the bus stop at a random time, how long will the person have to wait for the next bus? Let the random variable $X =$ waiting time until the next bus arrives. The random variable $X$ is said to follow a uniform distribution between 0 and 10 minutes. A graph of the density curve can be found in Chapter 8 of the book. Can we find the probability that the waiting time $X$ is somewhere between 5 and 7 minutes?

1. **Steps** to prepare the data:
   a. Open a new data file.
   b. Give the variable a name in the **Variable View** tab, e.g. $x$.
   c. Set the number of decimals to 0.
   d. In the **Data View** tab, enter the values 5 and 7 in the column named $x$.
   e. Save the file.

2. **Steps** to calculate uniform cumulative probabilities:
   a. To calculate uniform probabilities we use the **Transform** menu.
   b. Scroll down to the **Compute Variable** option. A window will open up. The window is shown below:
   c. Type a variable name to represent the probabilities in the **Target Variable** box, e.g. **prob**.
   d. In the **Function Group** box select **CDF & Noncentral CDF**.
   e. In the **Functions and Special Variables** box scroll down until you find the function **Cdf.Uniform**. Select this function and move it into the **Numeric Expression** box by clicking on the up arrow.
   f. The three question marks need to be replaced by the value of $x$, the min and the max. Place the cursor in the location of the first question mark. Click on the variable $x$ and move it into the **Numeric Expression** box by clicking on the right arrow. Place the cursor on the location of the second question mark and type the min (0). Place the cursor on the location of the third question mark and type the max (10).
   g. Click **OK**.
h. A new variable has now been added to the Data Editor called prob, containing the Uniform Cumulative probabilities.

![Data Editor screenshot]

Notes:

- To find $P(5 \leq X \leq 7)$, we need to realize that $P(5 \leq X \leq 7) = P(X \leq 7) - P(X \leq 5)$. Use the values in the column cumprob for these cumulative probabilities. The answer is $0.70 - 0.50 = 0.20$. 
8.6 Normal Random Variables

Example 8.23 – Standard Normal Probabilities
The random variable $Z$ is used to denote the random variable that follows the standard normal distribution: $N(0, 1)$. Can we find the probability that $Z$ is less than 1.82? And less than −2.59? And less than 1.31? And less than −2.00? And can we find the probability that $Z$ is greater than 1.31?

1. Steps to prepare the data:
   a. Open a new data file.
   b. Give the variable a name in the Variable View tab, e.g. $z$.
   c. In the Data View tab, enter the values 1.82, 1.31, −2.00, and −2.59 in the column named $z$.
   d. Save the file.

2. Steps to calculate standard normal probabilities:
   a. To calculate standard normal probabilities we use the Transform menu.
   b. Scroll down to the Compute Variable option. A window will open up. The window is shown below:

   c. Type a variable name to represent the cumulative probabilities in the Target Variable box, e.g. prob.
   d. In the Function Group box select CDF & Noncentral CDF.
   e. In the Functions and Special Variables box scroll down until you find the function Cdfnorm. NB: This is not the same as Cdf.normal! Be careful to pick the correct one.
   f. Select this function and move it into the Numeric Expression box by clicking on the up arrow.
   g. The question mark needs to be replaced by the value of $z$. Place the cursor in the location of the first question mark. Click on the variable $z$ and move it into the Numeric Expression box by clicking on the right arrow.
   h. Click OK.
i. A new variable has now been added to the **Data Editor** called **prob**. It contains the Standard Normal Cumulative probabilities.

j. You may want to increase the number of decimals using the **Variable View**.

k. The Data Editor is shown below:

![Data Editor](image)

**Notes:**
- To find \( P(Z < 1.82) \), note that \( P(Z < 1.82) = P(Z \leq 1.82) \). To find it, use the value in the column **prob** next to \( z = 1.82 \). The answer is \( P(Z \leq 1.82) = 0.9656 \).
- We find \( P(Z < 1.31) \) in a similar way and see that \( P(Z \leq 1.31) = 0.9049 \).
- To find \( P(Z > 1.31) \), we need to realize that \( P(Z > 1.31) = 1 - P(Z \leq 1.31) \). The answer is \( P(Z > 1.31) = 1 - 0.9049 = 0.0951 \).
- If many areas above need to be found, a new variable can be created which equals \( 1 - \text{prob} \).

**Example 8.24 – Normal Probabilities**

The random variable \( X \) is used to denote the height of college women. \( X \) is said to follow the normal distribution with mean 65 inches and standard deviation 2.7 inches: \( N(65, 2.7) \). Can we find the probability that a randomly selected college woman is shorter than 62 inches? Can we find the probability that a randomly selected college woman is taller than 70 inches?

1. **Steps** to prepare the data:
   - a. Open a new data file.
   - b. Give the variable a name in the **Variable View** tab, e.g. **height**.
   - c. In the **Data View** tab, enter the values 62 and 70 in the column named **height**.
   - d. Save the file.
2. **Steps** to calculate normal probabilities:
   a. To calculate normal probabilities we use the **Transform** menu.
   b. Scroll down to the **Compute Variable** option. A window will open up. The window is shown below:

   ![Image of the Compute Variable window](image)

   c. Type a variable name to represent the cumulative probabilities in the **Target Variable** box, e.g. `prob`.
   d. In the **Function Group** box select **CDF & Noncentral CDF**.
   e. In the **Functions and Special Variables** box scroll down until you find the function `Cdf.Normal`.
   f. Select this function and move it into the **Numeric Expression** box by clicking on the up arrow.
   g. The three question mark needs to be replaced by the value of \(x\), the mean, and the standard deviation. Place the cursor in the location of the first question mark. Click on the variable `height` and move it into the **Numeric Expression** box by clicking on the right arrow. Place the cursor on the location of the second question mark and type the mean (65). Place the cursor on the location of the third question mark and type the standard deviation (2.7).
   h. Click **OK**.
   i. A new variable has now been added to the **Data Editor** called `prob`. It contains the Normal Cumulative probabilities.
   j. You may want to increase the number of decimals using the **Variable View**.

**Notes:**
- To find \(P(X < 62)\), we need to realize that \(P(X < 62) = P(X \leq 62)\). To find it, use the value in the column `prob` next to `height = 62`. The answer is \(P(X < 62) = 0.1333\).
- To find \(P(X > 70)\), we need to realize that \(P(X > 70) = 1 - P(X \leq 70)\). For the probability, use the value in the column `prob` next to `height = 70`. The answer is \(P(X > 70) = 1 - 0.9680 = 0.0310\).
- If many areas **above** need to be found, a new variable can be created which equals \(1 - prob\).
Example 8.27 – Normal Percentiles
Suppose that the blood pressures of men aged 18 to 29 years old can be described with a normal curve having mean 120 and standard deviation 10: N(120, 10). Can we find the 75th percentile?

1. **Steps** to prepare the data:
   a. Open a new data file.
   b. Give the variable a name in the Variable View tab, e.g. `area`.
   c. In the Data View tab, enter the value 0.75 in the column named `area`.
   d. Save the file.

2. **Steps** to calculate normal percentiles:
   a. To calculate normal percentiles we use the Transform menu.
   b. Scroll down to the Compute Variable option. A window will open up. The window is shown below:
   c. Type a variable name to represent the blood pressure in the Target Variable box, e.g. `bldpress`.
   d. In the Function Group box select Inverse DF.
   e. In the Functions and Special Variables box scroll down until you find the function `Idf.Normal`. Select this function and move it into the Numeric Expression box by clicking on the up arrow.
   f. The three question mark needs to be replaced by the area, the mean, and the standard deviation. Place the cursor in the location of the first question mark. Click on the variable `area` and move it into the Numeric Expression box by clicking on the right arrow. Place the cursor on the location of the second question mark and type the mean (120). Place the cursor on the location of the third question mark and type the standard deviation (10).
   g. Click OK.
h. A new variable has now been added to the Data Editor called `bldpress`. It contains the 75th percentile of the blood pressure distribution. The Data Editor is shown below:

![Data Editor screenshot](image)

i. For this example, the 75th percentile equals 126.74.
9.4  Sampling Distributions for Proportions

Example 9.4 – Possible Sample Proportions Favoring a Candidate

Suppose that of all voters in the United States, 40% are in favor of Candidate C for president. Pollsters take a random sample of 2400 voters. What proportion of the sample would be expected to favor Candidate C? The rule tells us that the proportion of voters in the sample who favor Candidate C is a random variable that has a normal distribution with a mean of 0.40 (40%) and a standard deviation of 0.01. Can we simulate this distribution? For the purpose of this simulation, we will do 400 repetitions.

1. **Steps** to prepare the data file:
   a. Open a new data file.
   b. Go to the **Data View** tab.
   c. Enter the value 1 (or any other value) as the first entry in the first column.
   d. Scroll down as far as possible. If it is possible to scroll all the way down to 400 at once, enter the value 1 in row 400. In many cases, you may only be able to scroll 40 lines at the time. Enter the value 1 every 40 lines, until you reach line 400. The goal here is to create a dummy variable containing 400 values.
   e. Use the **Variable View** tab to give the variable a name, e.g. dummy.
   f. Go back to the **Data View** tab.
   g. Save the file.
   h. Before starting the simulation, check that the random number seed is set to random. To do this, use the **Transform** menu and select the **Random Number Generators** option. In the **Active Generator Initialization** box, make sure that the **Random** option is selected. Click **OK**.

2. **Steps** to create the data:
   a. To simulate the voting data, we use the **Transform** menu.
   b. Scroll down to the **Compute Variable** option. A window opens up.
   c. Type a variable name to represent the proportions of the voting data in the **Target Variable**: box, e.g. prop.
   d. In the **Function Group** box select **Random Numbers**.
   e. In the **Functions and Special Variables** box select the function **Rv.Binom** and move it into the **Numeric Expression** box by clicking on the up arrow.
   f. The two question marks will need to be replaced by the sample size and the population proportion. Place the cursor in the location of the first question mark and type the value 2400. Place the cursor in the location of the second question mark and type the value 0.40.
   g. The Binomial function returns the number of voters in the sample. We are looking for the proportion of voters in the sample. Type /2400 after the **Rv.Binom** function to turn the count into a proportion.
h. The **Compute Variable** window is shown below:

![](image1.png)

i. Click **OK**. A new variable has now been added to the **Data Editor** called **prop**. It contains a series of proportions, all fairly close to 0.40.

j. You may wish to increase the number of decimals using the **Variable View**. You can delete the dummy variable at this time if you wish.

![](image2.png)
3. **Steps** to create a histogram of the simulation results:
   a. To create a histogram, we use the **Graphs** menu.
   b. Click on the **Chart Builder** option.
   c. In the **Gallery** at the bottom of the window, select the **Histogram** option.
   d. Drag the first graph (**Simple Histogram**) to the **Chart Preview** at the top of the window.
   e. Select the variable **prop** and move it into the **X-Axis?** box.
   f. To superimpose the normal distribution on the histogram, we use the **Element Properties** window, shown to the right:
   g. In the middle of the window, check **Display normal curve**.
   h. Click **Apply** in the **Element Properties** window.
   i. Click **OK** in the **Chart Builder** window.

![Chart Builder Image]

**The SPSS Output**

![SPSS Histogram Output]
9.5 Sampling Distributions for Means

Example 9.12 – Hypothetical Mean Weight Loss
Suppose we are interested in estimating the average weight loss for everyone who attends a national weight-loss clinic for ten weeks. Suppose the distribution of weight losses is roughly normal with a mean of 8 pounds and a standard deviation of 5 pounds. A random sample of $n = 25$ is to be used. The rule tells us that potential sample means could be anything from a normal curve with a mean of 8 pounds and a standard deviation of 1.0 pound. Can we simulate this distribution?

For the purpose of this simulation, we will only do 50 repetitions. More repetitions could be used.

1. **Steps** to prepare the data file:
   a. Open a new data file.
   b. Go to the **Data View** tab.
   c. Enter the value 1 (or any other value) on line 25 of the first column. The goal here is to create a dummy variable containing $n = 25$ values.
   d. Use the **Variable View** tab to give the variable a name, e.g. **dummy**.
   e. Save the file.
   f. Before starting the simulation, check that the random number seed is set to random. To do this, use the **Transform** menu and select the **Random Number Generators**... option. In the **Active Generator Initialization** box, make sure that the **Random** option is selected. Click **OK**.

2. **Steps** to create the data:
   a. To simulate the weight-loss data, we use the **Transform** menu.
   b. Scroll down to the **Compute Variable** option. A window opens up. The window is shown below:
   c. Type a variable name to represent the weight-loss data in the **Target Variable:** box, e.g. **x1**.
   d. In the **Function Group** box, select **Random Numbers**.
   e. In the **Functions and Special Variables** box, select the function **Rv.Normal** and move it into the **Numeric Expression** box by clicking on the up arrow.
   f. The two question marks will need to be replaced by the mean and the standard deviation. Place the cursor in the location of the first question mark and type the value 8. Place the cursor in the location of the second question mark and type the value 5.
g. We will now use the **Syntax** window, to copy the command 50 times. Click **Paste** (and *not* OK). The **Syntax** window will open up. It is shown below:

![Syntax window](image)

h. Copy the line ‘`COMPUTE x1=RV.NORMAL(8,5).`’ 49 times. Now change the variable names in each line to `x2` through `x50`. The Syntax window will look like this:

![Syntax window](image)

i. Highlight all lines in the **Syntax window** and click on the green triangle (Selection) at the top of the screen. SPSS will now execute all the commands.
Fifty new variables have now been added to the Data Editor called \( x_1 \) through \( x_{50} \) containing the random normal numbers. You may delete the dummy variable if you wish. Part of the Data Editor is shown below:

![Data Editor screenshot]

Note: Your answers will be different from the ones shown above.

3. **Steps** to create a histogram and summary measures of the simulation:
   a. To summarize the simulated weight-loss data, we use the Analyze menu.
   b. Scroll down to the Descriptive Statistics submenu and select the Descriptives option. A window opens up. The window is shown below:
   ![Descriptives window screenshot]
   c. Highlight the variables \( x_1 \) through \( x_{50} \) and move them (simultaneously) into the Variable(s) box.
   d. Click on the Options button. A second window opens up. It is shown to the right:
   e. Deselect the minimum, maximum, and standard deviation option, as we will only need the mean.
   f. Click Continue in the Options window.
   g. Click OK in the Descriptives window.
h. The SPSS Output will now show 50 sample means. Part of the output is shown below:

The SPSS Output

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>x1</td>
<td>25</td>
<td>8.6620</td>
</tr>
<tr>
<td>x2</td>
<td>25</td>
<td>7.4747</td>
</tr>
<tr>
<td>x3</td>
<td>25</td>
<td>7.4994</td>
</tr>
<tr>
<td>x4</td>
<td>25</td>
<td>6.8191</td>
</tr>
<tr>
<td>x5</td>
<td>25</td>
<td>6.6246</td>
</tr>
<tr>
<td>x47</td>
<td>25</td>
<td>9.1235</td>
</tr>
<tr>
<td>x48</td>
<td>25</td>
<td>8.6026</td>
</tr>
<tr>
<td>x49</td>
<td>25</td>
<td>7.1401</td>
</tr>
<tr>
<td>x50</td>
<td>25</td>
<td>7.8357</td>
</tr>
<tr>
<td>Valid N (listwise)</td>
<td>25</td>
<td></td>
</tr>
</tbody>
</table>

i. Double click on the output. Highlight the 50 means. Use the Edit menu and select Copy.

j. Open a new data file and paste the 50 sample means in the first column. Use the Edit menu and select Paste.
k. Use the Variable View tab to give the variable a name, e.g. samplemean. You can increase the decimals to 3 or 4 and select the appropriate measurement level (Scale).
l. Save the file.
m. Create a histogram (use the Graphs menu) and obtain descriptives (use the Analyze menu) of the variable samplemean.
n. Note that your answers will be different than the ones displayed here, due to the random-number seed.
The Problem – Areas and Probabilities for Student’s t-Distribution

The random variable \( t \) will be used to denote the random variable that follows the \( t \)-distribution with 24 degrees of freedom: \( t(24) \). Can we find the probability that \( t \) is less than 0.34? Can we find the probability that \( t \) is greater than 0.34?

1. **Steps** to prepare the data:
   a. Open a new data file.
   b. Give the variable a name in the **Variable View** tab, e.g. \( t \).
   c. In the **Data View** tab, enter the value 0.34 in the column named \( t \).
   d. Save the file.

### Descriptives

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Mean</th>
<th>Std. Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>samplemean</td>
<td>50</td>
<td>5.7439</td>
<td>10.3221</td>
<td>7.970722</td>
<td>0.9851279</td>
</tr>
<tr>
<td>Valid N (listwise)</td>
<td>50</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### 9.9 Preparing for Statistical Inference: Standardized Statistics
2. **Steps** to calculate $t$-distribution probabilities:
   a. To calculate the probabilities we use the **Transform** menu.
   b. Scroll down to the **Compute Variable** option. A window will open up. The window is shown below:

   ![Image of Compute Variable window]

   c. Type a variable name to represent the cumulative probabilities in the **Target Variable** box, e.g. `prob`.
   d. In the **Function Group** box select **CDF & Noncentral CDF**.
   e. In the **Functions and Special Variables** box select the function `Cdf.T` and move it into the **Numeric Expression** box by clicking on the up arrow.
   f. The two question marks will need to be replaced by the value of $t$ and the degrees of freedom. Place the cursor in the location of the first question mark. Click on the variable `t` and move it into the **Numeric Expression** box by clicking on the right arrow. Place the cursor in the location of the second question mark and type the value 24.
   g. Click **OK**.
   h. The cumulative probability will appear in the spreadsheet. You may want to increase the number of decimals using the **Variable View**.

   ![Image of IBM SPSS data view]

   Notes:
   - To find $P(t < 0.34)$, use the value in the column `prob` next to $t = 0.34$.
     The answer is $P(t < 0.34) = 0.6316$.
   - To find $P(t > 0.34)$, we need to realize that $P(t > 0.34) = 1 - P(t < 0.34)$. The answer is $P(t > 0.34) = 1 - 0.6316 = 0.3684$.
   - If many areas above need to be found, a new variable can be created which equals $1 - \text{prob}$.
3. **Steps** to calculate percentiles for the Student’s *t*-distribution:
   a. Use the **Variable View** tab to create a new variable by typing in a new name, e.g. *area*.
   b. Use the **Data View** to enter the percentile(s) to be found in the column named *area*, e.g. 0.05, 0.10 and 0.99.

c. To calculate percentiles we use the **Transform** menu.
d. Scroll down to the **Compute Variable** option. A window will open up. The window is shown below:

e. Type a variable name to represent the percentile in the **Target Variable** box, e.g. *tval*.
f. In the **Function Group** box select **Inverse DF**.
g. In the **Functions and Special Variables** box select the function *Idf.T* and move it into the **Numeric Expression** box by clicking on the up arrow.
h. The two question marks will need to be replaced by the area and the degrees of freedom. Place the cursor in the location of the first question mark. Click on the variable *area* and move it into the **Numeric Expression** box by clicking on the right arrow. Place the cursor in the location of the second question mark and type the value 24.
i. Click **OK**.

i. The percentile(s) for the *t*-distribution will appear in the spreadsheet. You may want to increase the number of decimals using the **Variable View**.
Chapter 10

Estimating Proportions with Confidence

The Problem – Finding the appropriate $z$-multiplier
For a specified level of confidence, how do we find the multiplier $z^*$?

1. **Steps** to prepare the data:
   a. Open a new data file.
   b. Give the variable a name in the Variable View tab, e.g. **conf**.
   c. Give the variable the appropriate measurement level (**Scale**).
   d. In the Data View tab, enter the desired confidence levels in the first column. For this example we will use the confidence levels in **Table 10.1**: 90, 95, 98, 99.
   e. Save the file.

2. **Steps** to calculate normal multipliers:
   a. To calculate the multipliers, we first need to calculate the significance level and the percentiles we need. For this, we use the **Transform** menu.
   b. Scroll down to the **Compute Variable** option. A window will open up. The window is shown below:
   
   ![Compute Variable Window]

   c. Type a variable name to represent the significance level in the **Target Variable** box, e.g. **alpha**.
   d. In the **Numeric Expression** box, type the expression $1 - (\text{conf}/100)$ as in the previous window.
   e. Click **OK**.
   f. Repeat steps a – e to calculate the percentiles. Name the variable **percentile** and use the expression $1 - (\text{alpha}/2)$. (Note: these 2 steps can also be performed by hand).
g. To calculate the multipliers we use the **Transform** menu one more time.
h. Scroll down to the **Compute Variable** option. A window will open up. The window is shown below:

![Compute Variable Window]

i. Type a variable name to represent the multiplier in the **Target Variable** box, e.g. *zstar*.
j. In the **Function Group** box select **Inverse DF**.
k. In the **Functions and Special Variables** box select the function **Idf.Normal** and move it into the **Numeric Expression** box by clicking on the up arrow.
l. The three question marks will need to be replaced by the percentile, the mean and the standard deviation. Place the cursor in the location of the first question mark. Click on the variable *percentile* and move it into the **Numeric Expression** box by clicking on the right arrow. Place the cursor in the location of the second question mark and type the value 0. Place the cursor in the location of the third question mark and type the value 1.
m. Click **OK**.

n. The multipliers will appear in the spreadsheet. You may want to increase the number of decimals to 3 or 4 using the **Variable View**.
Example 10.3 – Is There Intelligent Life on Other Planets?

In a 2008 Scripps Howard News/ Ohio University survey of 1003 randomly selected Americans, 56% of the sample said that it is either very or somewhat likely that there is intelligent life on other. Let’s use this sample information to calculate a 90% confidence interval for the proportion of all Americans who believe that it is either very or somewhat likely that there is intelligent life on other planets.

1. Steps to prepare the data:
   a. Open a new data file.
   b. In the Variable View tab:
      • give two variables a name e.g. opinion and frequency;
      • set the number of decimals to zero for both variables.
   c. In the Data View tab:
      • enter 562 (56% of 1003) and 441 (44% of 1003) under frequency;
      • enter 1 and 2 under opinion.
d. In the **Variable View** tab:
   - give the variables the appropriate measurement level: **Nominal** for **opinion** and **Scale** for **frequency**;
   - give the values of **opinion** appropriate names (1 = yes and 2 = no) by clicking on the cell in the **Values** column. A small window will open. It is shown below:

   ![Value Labels Window](image)

   - Enter the value 1 and the label *yes* and click on **Add**.
   - Repeat for the value 2 and *no*.
   - Click **OK**.

  e. To weight the values 1 (yes) and 2 (no) with the appropriate frequencies, we use the **Analyze** menu.
  f. Scroll down to the **Weight Cases** option. A small window will open. It is shown below:

   ![Weight Cases Window](image)

   - Select the option **Weight cases by**.
   - Select the variable **frequency** and move it into the **Frequency Variable** box.
   - Click **OK**.
   - Save the file.
2. **Steps** to construct a confidence interval for the proportion:
   a. To construct the confidence interval, we use the **Analyze** menu.
   b. Scroll down to the **Nonparametric Tests** submenu and select the **One sample** option. A window opens up. The window is shown below:

   ![One-Sample Nonparametric Tests window](image)

   c. Select the **Customize analysis** option.
   d. Click on the **Fields** tab at the top of the window. The window is shown below:

   ![One-Sample Nonparametric Tests window with Fields tab](image)

   e. Select the **Use custom field assignments** option, as shown.
   f. Select the variable **opinion** and move it into the **Test Fields** box.
g. Click on the **Settings** tab at the top of the window. The window is shown below:

![Settings window](image)

h. Select the **Custom tests** option, as shown.

i. Select the **Binomial test** and click on **Options**. A small window will open. It is shown below:

![Binomial Options window](image)

j. Under **Confidence Interval**, select **Clopper-Pearson (exact)**.

k. Specify a success value under **Define Success for Categorical Fields**, by typing a 1 in the box.

l. Click **OK**.
m. Back in the **One-Sample Nonparametric Tests** window, click on **Test Options**. The window is shown below:

![One-Sample Nonparametric Tests Window](image1)

n. Change the confidence level to **90%**.

o. Click **Run**.

**The SPSS Output**

a. This is output for a hypothesis test. We wish to see the confidence interval. Now double click on the output. A new window will open. It is shown below:

![Model Viewer Window](image2)

b. At the bottom of the window, select the **Confidence Interval Summary View**. The 90% Confidence Interval will now be shown.
Chapter 11

Estimating Means with Confidence

11.1 Introduction to Confidence Intervals for Means

Example 11.2 – Mean Hours per Day That Penn State Students Watch TV
In a class survey at Penn State, a question was “In a typical day, about how much time do you spend watching television?” How do we obtain a numerical summary of the responses and determine the standard error? The data are recorded in the file pennstate5.sav.

1. **Steps** to prepare the data:
   a. Open the data set pennstate5.sav.

2. **Steps** to obtain summary measures:
   a. To calculate summary measures, we use the Analyze menu.
   b. Scroll down to the Descriptive Statistics submenu and select the Descriptives option. A window opens up. The window is shown below:
   
   ![Descriptives Window](image)

   c. Select the variable tv and move it into the Variable(s) box.
   d. Click on the Options button.
   e. Another window opens up. The window is shown to the right:
   
   ![Descriptives: Options Window](image)

   f. Check the box in front of S.E. mean. Click Continue.
   g. Click OK.
Example 11.4 – Finding the $t^*$ Values for 24 Degrees of Freedom and 95% or 99% Confidence Intervals
Suppose you wish to find a 95% and a 99% confidence interval for the mean of a population based on a sample of $n = 25$ values. How do we find the multiplier $t^*$?

1. **Steps** to prepare the data:
   a. Open a new data file.
   b. Give the variable a name in the **Variable View** tab, e.g. `conf`.
   c. Give the variable the appropriate measurement level (**Scale**).
   d. In the **Data View** tab, enter the desired confidence levels in the first column. For this example we will use the confidence levels in **Table 10.1**: 95, 99.
   e. Save the file.

2. **Steps** to calculate normal multipliers:
   a. To calculate the multipliers, we first need to calculate the significance level and the percentiles we need. For this, we use the **Transform** menu.
   b. Scroll down to the **Compute Variable** option. A window will open up. The window is shown below:
c. Type a variable name to represent the significance level in the Target Variable box, e.g. alpha.
d. In the Numeric Expression box, type the expression \(1 - (\text{conf}/100)\) as in the previous window.
e. Click OK.
f. Repeat steps a – e to calculate the percentiles. Name the variable percentile and use the expression \(1 - (\text{alpha}/2)\). (Note: these 2 steps can also be performed by hand).
g. To calculate the multipliers we use the Transform menu one more time.
h. Scroll down to the Compute Variable option. A window will open up. The window is shown below:

![Compute Variable Window](image)

i. Type a variable name to represent the multiplier in the Target Variable box, e.g. tstar.
j. In the Function Group box select Inverse DF.
k. In the Functions and Special Variables box select the function \(\text{Idf.T}\) and move it into the Numeric Expression box by clicking on the up arrow.
l. The two question marks will need to be replaced by the percentile and the degrees of freedom. Place the cursor in the location of the first question mark. Click on the variable percentile and move it into the Numeric Expression: box by clicking on the right arrow. Place the cursor in the location of the second question mark and type the value 24.
m. Click OK.

The multipliers will appear in the spreadsheet. You may want to increase the number of decimals to 3 or 4 using the Variable View.
11.2 Confidence Intervals for One Population Mean

Example 11.5 – Are Your Sleeves Too Short? The Mean Forearm Length of Men
People are always interested in comparing themselves to others. If you are male, how does the length of your forearm, from elbow to wrist, compare to the average for the population of men? The forearm lengths (in cm) for a randomly selected sample of $n = 9$ men are 25.5, 24.0, 26.5, 25.5, 28.0, 27.0, 23.0, 25.0, 25.0.

1. **Steps** to prepare the data:
   a. Open a new data file.
   b. Give the variable a name in the Variable View tab, e.g. *length*.
   c. Give the variable the appropriate measurement level (Scale).
   d. In the Data View tab, enter the forearm lengths in the first column.
   e. Save the file.

2. **Steps** to construct a confidence interval for the mean:
   a. To construct the confidence interval, we use the Analyze menu.
   b. Scroll down to the Descriptive Statistics submenu and select the Explore option. A window opens up. The window is shown below:

   ![Explore window](image1)

   c. Select the variable *length* and move it into the Dependent List box.
   d. Under the Display heading at the bottom left of the window, select the Statistics option.
   e. Click on the Statistics button at the top of the window.
   f. Another window opens up. The window is shown to the right:

   ![Statistics window](image2)

   g. Select the Descriptives option, as shown. Select a confidence level. The default is 95%.
   h. In the Statistics window, click Continue.
   i. In the Explore window, click OK.
The SPSS Output

Explore

<table>
<thead>
<tr>
<th>Descriptives</th>
<th>Statistic</th>
<th>Std. Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>length</td>
<td>Mean</td>
<td>25.500</td>
</tr>
<tr>
<td></td>
<td>Lower Bound</td>
<td>24.331</td>
</tr>
<tr>
<td></td>
<td>Upper Bound</td>
<td>26.669</td>
</tr>
<tr>
<td></td>
<td>5% Trimmed Mean</td>
<td>25.500</td>
</tr>
<tr>
<td></td>
<td>Median</td>
<td>25.500</td>
</tr>
<tr>
<td></td>
<td>Variance</td>
<td>2.313</td>
</tr>
<tr>
<td></td>
<td>Std. Deviation</td>
<td>1.5207</td>
</tr>
<tr>
<td></td>
<td>Minimum</td>
<td>23.0</td>
</tr>
<tr>
<td></td>
<td>Maximum</td>
<td>28.0</td>
</tr>
<tr>
<td></td>
<td>Range</td>
<td>5.0</td>
</tr>
<tr>
<td></td>
<td>Interquartile Range</td>
<td>2.3</td>
</tr>
<tr>
<td></td>
<td>Skewness</td>
<td>.034</td>
</tr>
<tr>
<td></td>
<td>Kurtosis</td>
<td>-.095</td>
</tr>
</tbody>
</table>

Note: The confidence interval can be found in the upper section of the output box. For this example the 95% confidence interval is (24.331, 26.669).

Example 11.6 – How Much TV Do Penn State Students Watch?
In a class survey at Penn State, a question was “In a typical day, about how much time do you spend watching television?” How do we obtain the histogram of the data and a 99% confidence interval for the mean time Penn State students watch TV?
The data are recorded in the file pennstate5.sav.

1. **Steps** to prepare the data:
   a. Open the data set pennstate5.sav.

2. **Steps** to construct a confidence interval for the mean:
   a. To construct the confidence interval, we use the **Analyze** menu.
b. Scroll down to the **Descriptive Statistics** submenu and select the **Explore** option. A window opens up. The window is shown below:

![Explore window](image)

- Select the variable **tv** and move it into the **Dependent List** box.
- Under the **Display** heading at the bottom left of the window, select the **Both** option.
- Click on the **Statistics** button at the top of the window. Another window opens up. The window is shown below:

![Explore: Statistics window](image)

- Select the **Descriptives** option, as shown. Change the confidence level to **99%**.
- In the **Statistics** window, click **Continue**.
- In the **Explore** window, click on the **Plots** button at the top of the window. Another window opens up. The window is shown to the right:

![Explore: Plots window](image)

- Under the **Boxplots** heading, check the box in front of **None**.
- Under the **Descriptive** heading, check the box in front of **Histogram**. Uncheck the box in front of **Stem-and-leaf**.
- In the **Plots** window, click **Continue**.
- In the **Explore** window, click **OK**.
The SPSS Output

**Explore**

<table>
<thead>
<tr>
<th>Statistics</th>
<th>Mean</th>
<th>Std. Error</th>
<th>99% Confidence Interval for Mean</th>
<th>Lower Bound</th>
<th>Upper Bound</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.7683</td>
<td>2.4155</td>
</tr>
<tr>
<td>5% Trimmed Mean</td>
<td>1.5613</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Median</td>
<td>2.0000</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Variance</td>
<td>2.702</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Std. Deviation</td>
<td>1.6437</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Minimum</td>
<td></td>
<td>.00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Maximum</td>
<td>11.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Range</td>
<td>11.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Interquartile Range</td>
<td>2.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Skewness</td>
<td>1.921</td>
<td>.184</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kurtosis</td>
<td>7.003</td>
<td>.365</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: The confidence interval can be found in the upper section of the output box. For this example the 99% confidence interval is (1.7683, 2.4155).

**Example 11.7 – What Type of Students Sleep More?**

On the second day of classes in Spring 2000, a Monday, students in two statistics classes at the University of California at Davis were given a survey. One of the questions was, “How many hours of sleep did you get last night, to the nearest half hour?” The two classes, Statistics 10 and Statistics 13, are to be compared. The data is recorded in the file UCDavis1.sav.

1. **Steps** to prepare the data:
   a. Open the data set UCDavis1.sav.

2. **Steps** to check conditions graphically, before constructing a confidence interval for the mean:
   a. To construct boxplots, we will use the **Graphs** menu.
   b. Click on the **Chart Builder** option.
   c. In the **Gallery** at the bottom of the window, select the **Boxplot** option.
   d. Drag the first graph (**Simple Boxplot**) to the **Chart Preview** at the top of the window.
   e. Select the variable **sleep** and move it into the **Y-Axis?** box.
   f. Select the variable **class** and move it into the **X-Axis?** box.
   g. Click **OK**.
3. **Steps** to check conditions graphically with histograms:
   a. To create a histogram, we use the **Graphs** menu.
   b. Click on the **Chart Builder** option.
   c. In the **Gallery** at the bottom of the window, select the **Histogram** option.
   d. Drag the first graph (**Simple Histogram**) to the **Chart Preview** at the top of the window.
   e. Select the variable **sleep** and move it into the **X-Axis?** box.
   f. To separate the graph between the two classes, click on the **Groups/Point ID** tab (the **Gallery** will disappear)
   g. Check the box by **Rows panel variable**.
   h. Select the variable **class** and move it into the **Panel?** box.
   i. Click **OK**.
Notes:

- To display the normal distribution curve in the graphs, check the box Display normal curve in the Element Properties window when creating the graph.
- A third graph, called a Q-Q plot may be used for checking conditions. This option can be found under the Analyze menu and the Descriptive Statistics submenu.

4. **Steps** to construct confidence intervals for the mean:
   a. To construct the confidence intervals, we use the Analyze menu.
   b. Scroll down to the Descriptive Statistics submenu and select the Explore option. A window opens up:
   
   ![Explore window screenshot]

   c. Select the variable sleep and move it into the Dependent List box.
   d. Select the variable class and move it into the Factor List box.
   e. Under the Display heading at the bottom left of the window, select the Statistics option.
   f. Click on the Statistics button.
   g. A second window opens up. Select the Descriptives option.
   h. Select a confidence level. The default is 95%.
   i. In the Statistics window, click Continue.
   j. In the Explore window, click OK.

   **The SPSS Output (Edited for length)**

<table>
<thead>
<tr>
<th>class</th>
<th>Statistic</th>
<th>Std. Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>sleep</td>
<td>Mean</td>
<td>7.660</td>
</tr>
<tr>
<td></td>
<td>95% Confidence Interval for Mean Lower Bound</td>
<td>7.1052</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Upper Bound</td>
</tr>
<tr>
<td></td>
<td>5% Trimmed Mean</td>
<td>7.8222</td>
</tr>
<tr>
<td></td>
<td>Median</td>
<td>8.0000</td>
</tr>
<tr>
<td></td>
<td>Variance</td>
<td>1.807</td>
</tr>
<tr>
<td></td>
<td>Std. Deviation</td>
<td>1.34412</td>
</tr>
<tr>
<td>NonLib</td>
<td>Mean</td>
<td>6.6125</td>
</tr>
<tr>
<td></td>
<td>95% Confidence Interval for Mean Lower Bound</td>
<td>6.5310</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Upper Bound</td>
</tr>
<tr>
<td></td>
<td>5% Trimmed Mean</td>
<td>6.8044</td>
</tr>
<tr>
<td></td>
<td>Median</td>
<td>7.0000</td>
</tr>
<tr>
<td></td>
<td>Variance</td>
<td>3.002</td>
</tr>
<tr>
<td></td>
<td>Std. Deviation</td>
<td>1.73276</td>
</tr>
</tbody>
</table>
11.3 Confidence Interval for the Population Mean of Paired Differences

Example 11.9 – Screen Time -- Computer versus TV
The 25 students in a liberal arts course were given a survey that included questions on how many hours per week they watched television and how many hours a week they used the computer. The times spent on these two activities are to be compared. The data is recorded in the file UCDavis1.sav.

1. Steps to prepare the data:
   a. Open the data set UCDavis1.sav.

2. Steps to find the differences from the raw data:
   a. To calculate the differences, we use the Transform menu.
   b. Scroll down to the Compute Variable option. A window opens up. The window is shown below:

   ![Compute Variable window](image)

   c. Type a variable name in the box under the Target Variable heading, e.g. diff. This variable will contain the differences.
   d. Select the variable computer and move it into the Numeric Expression box.
   e. Type a minus sign (–) in the Numeric Expression box.
   f. Select the variable tv and move it into the Numeric Expression box.
   g. Click OK.
   h. The new variable will appear in the SPPS Data Editor. The variables computer and tv are both measured with 1 decimal, so we will only need 1 decimal for the variable diff as well. Use the Variable View to change the Decimals to 1. At this time you can also give the variable an appropriate label.
3. **Steps** to prepare the data:
   a. To analyze the students in the Liberal Arts class only, we need to select these cases only. For this we use the Data menu.
   b. Scroll down to the **Select Cases** option. A window opens up:
   c. Select the **If condition is satisfied** option. Click on the **If** button.
   d. A second window opens up.
e. The window is shown below:

```
96
```

f. Select the variable class and move it into the box on the right. Type the condition in this same box, as shown above: `class='LibArts'`. The accents are necessary because class is a string variable.

g. In the If window, click Continue.

h. In the Select Cases window, click OK.

i. A Filter variable has been added to the data Editor. The NonLib cases will be crossed out in the Data Editor window:
4. **Steps** to check conditions graphically, before calculating a confidence interval for paired data:
   a. Repeat the steps for creating a boxplot and a histogram described in parts 2 and 3 of example 11.7. Use the variable `diff`. In this case we are not comparing two groups, so you need to make a simple boxplot and a single histogram.

5. **Steps** to calculate a confidence interval for paired data using the differences:
   a. The confidence interval can be found in the same way described in part 4 of example 11.7. For this we use the Analyze menu.
   b. Scroll down to the Descriptive Statistics submenu and select the Explore option.
   c. Select the variable `diff` and move it in to the Dependent List box.
   d. Under the Display heading at the bottom left of the window, select the Statistics option.
   e. Click on the Statistics button.
   f. Another window opens up. Select the Descriptives option. Select a confidence level. The default is 95%.
   g. In the Statistics window, click Continue.
   h. In the Explore window, click OK.

**The SPSS Output**

**Explore**

<table>
<thead>
<tr>
<th>Descriptives</th>
<th>Statistic</th>
<th>Std. Error</th>
</tr>
</thead>
<tbody>
<tr>
<td><code>diff</code> Mean</td>
<td>5.360</td>
<td>3.0486</td>
</tr>
<tr>
<td>90% Confidence Interval for Mean</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lower Bound</td>
<td>.144</td>
<td></td>
</tr>
<tr>
<td>Upper Bound</td>
<td>10.576</td>
<td></td>
</tr>
<tr>
<td>5% Trimmed Mean</td>
<td>4.833</td>
<td></td>
</tr>
<tr>
<td>Median</td>
<td>4.000</td>
<td></td>
</tr>
<tr>
<td>Variance</td>
<td>232.344</td>
<td></td>
</tr>
<tr>
<td>Std. Deviation</td>
<td>15.2428</td>
<td></td>
</tr>
<tr>
<td>Minimum</td>
<td>-20.0</td>
<td></td>
</tr>
<tr>
<td>Maximum</td>
<td>40.0</td>
<td></td>
</tr>
<tr>
<td>Range</td>
<td>60.0</td>
<td></td>
</tr>
<tr>
<td>Interquartile Range</td>
<td>21.3</td>
<td></td>
</tr>
<tr>
<td>Skewness</td>
<td>.601</td>
<td>.464</td>
</tr>
<tr>
<td>Kurtosis</td>
<td>-.007</td>
<td>.902</td>
</tr>
</tbody>
</table>

Note:
- Using the Explore option, you may choose to select Both under the Display heading at the bottom left of the window. After clicking on the Plots button, you can select a Histogram and a Q-Q Plot. This way you can check assumptions and calculate the confidence interval all at once.

6. **Steps** to calculate a confidence interval for paired data without having to calculate the differences first:
   a. To analyze the students in the Liberal Arts class only, prepare the data as described in part 3 above.
   b. To construct the confidence interval, we use the Analyze menu.
   c. Scroll down to the Compare Means submenu and select the Paired-Samples T Test option. A window opens up.
d. The window is shown below:

![Paired-Samples T Test](image)

- The window is shown below:

- Select the variable **computer** and move it in to the **Paired Variables** box.
- Select the variable **tv** and move it in to the **Paired Variables** box. It will automatically appear next to the variable **computer**.
- Click on the **Options** button. Another window opens up. The window is shown below:

![Options window](image)

- Select a confidence level. The default is 95%. To calculate a 90% confidence interval, change it to **90%**.
- In the Options window, click **Continue**.
- In the **Paired-Samples T Test** window, click **OK**.

**The SPSS Output (Edited for length)**

**T-Test**

<table>
<thead>
<tr>
<th>Paired Samples Test</th>
<th>Paired Differences</th>
<th>90% Confidence Interval of the Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>Std. Deviation</td>
</tr>
<tr>
<td>Pair 1 computer - tv</td>
<td>5.3600</td>
<td>15.2428</td>
</tr>
</tbody>
</table>
11.4 Confidence Intervals for the Difference in Two Population Means (Independent Samples)

Example 11.14 – Pooled t-Interval for Difference Between Mean Female and Male Sleep Times

Students in an introductory statistics class filled out a survey on a variety of issues, including how much sleep they had the previous night. Let’s use the data to estimate how much more or less male students sleep than female students, on average. The class included 83 females and 65 males. Let’s assume these students are equivalent to a random sample of all students who take introductory statistics. How much difference is there between how long female and male students, represented by this sample, slept the previous night? To answer that question, a 95% confidence interval is to be calculated. The data is recorded in the file UCDavis1.sav.

1. **Steps** to prepare the data:
   a. Open the data set UCDavis1.sav.
   b. To analyze the students in the Statistical Methods class only, we need to select these cases only. For this we use the **Data** menu.
   c. Scroll down to the **Select Cases** option. A window opens up.
   d. Select the **If condition is satisfied** option.
   e. Click on the **If** button. A second window opens up. The window is shown below:

   ![Select Cases Window](image)

   f. Select the variable **class** and move it into the box on the right. Type the condition in this same box, as shown above: `class='NonLib'`. The accents are necessary because **class** is a string variable.
   g. In the **If** window, click **Continue**.
   h. In the **Select Cases** window, click **OK**.
   i. A **Filter** variable has been added to the **data Editor**. The **LibArts** cases will be crossed out in the **Data Editor** window.


2. **Steps** to check conditions before computing a confidence interval for the difference between two independent means:
   a. Repeat the steps for creating side-by-side boxplots and histograms described in parts 2 and 3 of example 11.7. Use the variables `sleep` and `sex`.

The SPSS Output
3. **Steps** to construct a confidence interval for the difference in means:
   a. To construct the confidence interval, we use the Analyze menu.
   b. Scroll down to the Compare Means submenu and select the Independent-Samples T Test option. A window opens up. The window is shown below:

![Image](image_url)

   c. Select the variable sleep and move it into the Test Variable(s) box.
   d. Select the variable sex and move it into the Grouping Variable box.
   e. Click on the Define Groups button. A second window opens up. The window is shown below:

![Image](image_url)

   f. Define the two groups as Female and Male. Make sure to use the same notation as in the Data Editor.
   g. In the Define Groups window, click Continue.
   h. To change the confidence level from the default of 95%, click on the Options button and change the level, if necessary.
   i. In the Independent-Samples T Test window, click OK.

   **The SPSS Output (Edited for length)**

   **T-Test**

<table>
<thead>
<tr>
<th>Independent Samples Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Levene's Test for Equality of Variances</td>
</tr>
<tr>
<td>F</td>
</tr>
<tr>
<td>sleep</td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

   **Notes:**
   - The confidence intervals for both the pooled and the unpooled versions are given in the output (by default).
   - Oftentimes, grouping variables such as gender are coded with 1’s and 2’s in the data set. When defining the groups (see step f) you need to know what values are used for the two groups you wish to compare.
Chapter 12

Testing Hypotheses About Proportions

12.2 Testing Hypotheses About a Population Proportion

Example 12.11 – The Importance of Order in Voting
In a student survey, a statistics teacher asked his students to “randomly pick a letter from these two choices—S or Q.” For about half of the students, the order of the letters S and Q was reversed so that the end of the instruction read “Q or S.” The purpose of the activity was to determine whether there might be a preference for choosing the first letter. A tendency to pick the first choice offered has been noted in several settings, including elections. Can we determine if the proportion of students who pick the first letter is greater than 0.50? The data is recorded in the file pennstate1.sav.

1. Steps to prepare the data:
   a. Open the data set pennstate1.sav.
   b. To create a new variable for testing, we use the Transform menu.
   c. Scroll down to the Compute Variable option. A window opens up. The window is shown below:

   ![Compute Variable Window](image)

   d. Create a new variable, e.g. first, which will contain zeros and ones. A one will represent those students who picked the first letter. A zero will indicate the student did not pick the first letter. Initially the variable first will be equal to zero everywhere. Type first in the Target Variable box and enter the value 0 in the Numerical Expression box.
   e. Click OK.
   f. The new variable first will appear in the SPSS Data Editor.
   g. Go back to the Transform menu and use the Compute option once again. The variable first should equal 1 whenever the student picked the letter Q on the form “Q or S”, or S on the form “S or Q”: Change the value 0 in the Numerical Expression box to the value 1.
h. Click on the **If** button. A second window opens up. The window is shown below:

![Compute Variable: If Cases](image)

i. Select the **Include if case satisfies condition** option. The variable list on the left and the box below the option will turn white.

j. Use the variable list and the operators displayed in the middle to carefully type the following expression:

```
((sqpick='S')&(form='SorQ'))|((sqpick='Q')&(form='QorS'))
```

Note that the symbol `|` represents “or” and the symbol `&` represents “and”.

k. In the **If** window, click **Continue**.

l. In the **Compute Variable** window, click **OK**.

m. SPSS will ask you if you would like to “Change the existing variable?”

n. Click **OK**.

o. The variable **first** now consists of zeros and ones only.

p. In the **Variable View** change the **Measure** to **Nominal** and change the **Decimals** to **0**.

![Variable View](image)
2. **Steps** to test hypotheses about a proportion:
   a. To perform the test, we use the **Analyze** menu.
   b. Scroll down to the **Nonparametric Tests** submenu and select the **One Sample** option. A window opens up. The window is shown below:

   ![Image of One-Sample Nonparametric Tests window](image1)

   c. Select the **Customize analysis** option and click on the **Fields** tab:

   ![Image of One-Sample Nonparametric Tests window with Fields tab](image2)

   d. Select the **Use custom field assignments** option.
   e. Select the variable **first** and move it into the **Test Fields** box.
f. Click on the **Settings** tab:

Select the **Customize tests** option and put a checkmark in front of the **Binomial test**.

g. Click on the **Options** button:

h. Enter the hypothesized proportion in the **Hypothesized proportion** box. The default is 0.50. For this example, this is the value we need.

i. Enter the value for a ‘success’ under the heading **Define Success for Categorical Fields**. In this example this is the value 1.

j. In the **Binomial Options** window, click **OK**.

k. Click **Run**.
3. Alternative steps to test hypotheses about a proportion:
   a. To perform the test, we use the Analyze menu.
   b. Scroll down to the Nonparametric Tests submenu, the Legacy Dialogs submenu, and select the Binomial option. A window opens up. The window is shown below:
   c. Select the variable first and move it into the Test Variable List box.
   d. Enter the hypothesized proportion in the Test proportion box. The default is 0.50. For this example, this is the value we need.
   e. If the testing variable is dichotomous, SPSS will automatically detect that. If the variable is not dichotomous, you need to define what value or what range of values are used for a success. In the Define Dichotomy box, define a Cut point. For this example this step is unnecessary.
   f. Click OK.

The SPSS Output

<table>
<thead>
<tr>
<th>Category</th>
<th>N</th>
<th>Observed Prop.</th>
<th>Test Prop</th>
<th>Exact Sig. (2-tailed)</th>
</tr>
</thead>
<tbody>
<tr>
<td>first</td>
<td>1</td>
<td>.60</td>
<td>.50</td>
<td>.007</td>
</tr>
<tr>
<td>Group 1</td>
<td>114</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Group 2</td>
<td>76</td>
<td>.40</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>150</td>
<td>1.00</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: This is a 2-sided $p$-value! For the one-sided $p$-value, divide 0.007 by 2!
4. **Steps** to prepare the data to test hypotheses about a proportion from summarized data:
   a. Open a new data file.
   b. Use the **Variable View** tab to create two variables: one to indicate the two groups (first letter chosen, second letter chosen) and one to contain the counts. Give both variables a name, e.g. **group** and **count**.
   c. For both variables, set the **Decimals** to **0**.
   d. For the variable **group**, set the **Measure** to **Nominal**.
   e. For the variable **count**, set the **Measure** to **Scale**.
   f. For the variable **group**, you can define value labels in the column **Values**, e.g. 1 = First letter and 2 = Second letter.
   g. Use the **Data View** to enter the groups (1 and 2) and the observed counts (both for those who picked the first letter and for those who didn’t). For this example the counts are 114 and 76.
   h. The variable **count** needs to be designated as a weighting variable. For this, we use the **Data menu**.
   i. Scroll down to the **Weight Cases** option. A window opens up. The window is shown below:

   ![Weight Cases Window](image)

   j. Select the **Weight cases by** option.
   k. Select the variable **count** and move it into the **Frequency Variable** box.
   l. Click **OK**.

5. **Steps** to test hypotheses about a proportion from summarized data:
   a. To test the hypotheses, we can now use either testing procedure described above. Both methods work in identical ways as before. Use the variable group as the testing variable.

   **The SPSS Output**

   ![SPSS Output](image)

   **Null Hypothesis**

<table>
<thead>
<tr>
<th>Hypothesis Test Summary</th>
<th>Test</th>
<th>Sig.</th>
<th>Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>The categories defined by group = First letter and Second letter occur with probabilities 0.5 and 0.5.</td>
<td>One-Sample Binomial Test</td>
<td>.007</td>
<td>Reject the null hypothesis.</td>
</tr>
</tbody>
</table>

   Asymptotic significances are displayed. The significance level is .05.

   **Binomial Test**

<table>
<thead>
<tr>
<th>Category</th>
<th>N</th>
<th>Observed Prop</th>
<th>Test Prop</th>
<th>Exact Sig. (2-tailed)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group 1</td>
<td>114</td>
<td>.60</td>
<td>.50</td>
<td>.007</td>
</tr>
<tr>
<td>Group 2</td>
<td>76</td>
<td>.40</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>190</td>
<td>1.00</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Notes about using the Binomial test in SPSS:
- The proportion in the null hypothesis \( p \) is referred to as the success probability.
- If you wish to use the Legacy Dialogues option, make sure to enter the counts with the number of successes in the sample first and the number of failures in the sample second.
- If the hypothesized value of \( p \) equals 0.50 (the default), SPSS will assume the alternative hypothesis is 2-sided and it will report a 2-sided \( p \)-value for both binomial tests.
- If the hypothesized value of \( p \) is different from 0.50, SPSS will assume the alternative hypothesis is 1-sided. The direction of extreme is determined by the sample data. If the sample proportion is larger than the hypothesized value, the direction of extreme used will be to the right. If the sample proportion is smaller than the hypothesized value, the direction of extreme used will be to the left. If the direction of extreme used is to the left, SPSS will indicate it in the output with a footnote. See example output below.

### Binomial Test

<table>
<thead>
<tr>
<th>Category</th>
<th>N</th>
<th>Observed Prop.</th>
<th>Test Prop.</th>
<th>Exact Sig. (1-tailed)</th>
</tr>
</thead>
<tbody>
<tr>
<td>group</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Group 1</td>
<td>15</td>
<td>.7</td>
<td>.8</td>
<td>.231*</td>
</tr>
<tr>
<td>Group 2</td>
<td>6</td>
<td>.3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>21</td>
<td>1.0</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* a. Alternative hypothesis states that the proportion of cases in the first group < .8.

---

### 12.4 Testing the Difference Between Two Population Proportions

**Example 12.17 – The Prevention of Ear Infections**

Based on its biochemical properties, Finnish researchers hypothesized that regular use of the sweetener xylitol might be useful for preventing ear infections in preschool children. In a randomized experiment, 165 children took five daily doses of placebo syrup, and 68 of these children got an ear infection. Another 159 children took five daily doses of xylitol, and 46 of these children got an ear infection during the study. Is this observed difference large enough to conclude that using xylitol reduces the risk of ear infection? The data is not recorded in a data file, so the data needs to be created.

1. **Steps** to prepare the data:  
   a. Open a new data file.  
   b. Use the **Variable View** tab to create three variables, e.g. freq, Xylitol, and Earinfection.  
   c. For all three variables, set the **Decimals** to 0.  
   d. For the variables Xylitol and Earinfection, set the **Measure** to Nominal.  
   e. For the variable freq, set the **Measure** to Scale.  
   f. For the variables Xylitol and Earinfection, you can define value labels in the column **Values**, e.g. 0 = No and 1 = Yes. These two variables will indicate whether someone received the drug (yes or no) and whether someone had an ear infection (yes or no).  
   g. Use the **Data View** to enter the observed counts (both for the children who got an ear infection and for those who didn’t) in the first column, including the counts for both groups in the same column: 68, 97, 46, 113.
h. Enter the values 0 and 1 for the two xylitol groups in the second column. Make sure to match these with the counts in column 1. If the counts were entered in the same order as above, enter: 0, 0, 1, 1.
i. Enter the values 0 and 1 in the third column, corresponding to whether the child had an ear infection or not. Make sure to match these with the counts in column 1. If the counts were entered in the same order as above, enter: 0, 1, 0, 1.
j. Save the file.

2. **Steps** to prepare count data for a hypothesis test:
   a. The count variable needs to be designated as a weighting variable. For this we use the Data menu.
   b. Scroll down to the Weight Cases option. A window opens up.
   c. Select the Weight cases by option.
   d. Select the variable freq and move it into the Frequency Variable box.
   e. Click OK.

3. **Steps** to conduct a hypothesis test:
   Note: SPSS cannot perform a z-test for two proportions. Performing a Chi-squared test, as shown below, is equivalent for the 2-sided alternative.
   a. To conduct a chi-square test, we use the Analyze menu.
   b. Scroll down to the Descriptive Statistics submenu and select the Crosstabs option. A window opens up. The window is shown to the right:
   c. Select the variable Xylitol and move it into the Row(s) box.
   d. Select the variable Earinfection and move it into the Column(s) box. Note that there is no specific order to these variables, so they may be reversed.
   e. Click on the Statistics button.
f. A window opens up. The window is shown below:

![Crosstabs Statistics Window]

- At the top left-hand corner, check the **Chi-square** option.
- In the **Statistics** window, click **Continue**.
- In the **Crosstabs** window, click **OK**.

### The SPSS Output

#### Crosstabs

**Xylitol * Ear Infection Crosstabulation**

<table>
<thead>
<tr>
<th></th>
<th>Ear Infection</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Xylitol</td>
<td>68</td>
<td>97</td>
</tr>
<tr>
<td></td>
<td>46</td>
<td>113</td>
</tr>
<tr>
<td>Total</td>
<td>114</td>
<td>210</td>
</tr>
</tbody>
</table>

**Chi-Square Tests**

<table>
<thead>
<tr>
<th>Test</th>
<th>Value</th>
<th>df</th>
<th>Asymp. Sig. (2-sided)</th>
<th>Exact Sig. (2-sided)</th>
<th>Exact Sig. (1-sided)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pearson Chi-Square</td>
<td>5.355</td>
<td>1</td>
<td>0.021</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Continuity Correction</td>
<td>4.830</td>
<td>1</td>
<td>0.028</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Likelihood Ratio</td>
<td>5.381</td>
<td>1</td>
<td>0.020</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fisher's Exact Test</td>
<td></td>
<td></td>
<td></td>
<td>0.027</td>
<td>0.014</td>
</tr>
<tr>
<td>Linear-by-Linear Association</td>
<td>5.339</td>
<td>1</td>
<td>0.021</td>
<td></td>
<td></td>
</tr>
<tr>
<td>N of Valid Cases</td>
<td>324</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- a. 0 cells (0%) have expected count less than 5. The minimum expected count is 55.94.
- b. Computed only for a 2x2 table
Notes:

- The **Pearson Chi-square** test statistic, found in the top row in the bottom box of the output labeled **Chi-Square Tests**, is equivalent to the square of the $z$-statistic. For this example the chi-square test statistic equals 5.355, so that the absolute value of the $z$-statistic equals $\sqrt{5.355} = 2.32$.
- The 2-sided $p$-value corresponding to the test statistic can be found in the same row, but in the third column. For this example the 2-sided $p$-value equals 0.021.
- In the case of a one-sided test, check the sample data. Depending on whether the first sample proportion is smaller or larger than the second sample proportion, the sign of the $z$-statistic will be negative or positive.
- In the case of a one-sided test, check if the data is consistent with the direction of extreme. If it is, the $p$-value reported in the output needs to be divided by 2. If the data is opposite from the direction of extreme, the correct $p$-value can be found by dividing the value from the output by 2 and subtracting that from 1.
- In this case the alternative hypothesis was 1-sided. The correct $p$-value is $0.021/2 = 0.0105$.
- If an existing data file is to be used, containing the raw data, the same procedure may be used omitting the weighting step.
Chapter 13

Testing Hypotheses About Means

13.2 Testing Hypotheses About One Population Mean

Example 13.1 – Normal Human Body Temperature for Young Adults
What is normal body temperature? A paper presented evidence that normal body temperature may be less than 98.6 degrees Fahrenheit, the long-held standard. A random sample of 16 human body temperatures is available. The data can be found in Section 13.2 of the book. The dataset bodytemp on the companion website also includes the body temperatures for the 16 healthy adults. If you do not have access to the website, the dataset will need to be created.

1. Steps to prepare the data:
   a. Open the data set bodytemp.sav or . . .
   b. Open a new data file. Use the Variable View tab to create a variable (e.g. temp) and use the Data View to enter the data. Make sure to give the variable the appropriate measurement level (Scale) and save the file.

2. Steps to check conditions graphically, before conducting a one-sample t-test for the mean:
   Boxplot:
   a. To construct a boxplot, we will use the Graphs menu.
   b. Click on the Chart Builder option. A window will open up. The window is shown below:
c. In the **Gallery** at the bottom of the window, select the **Boxplot** option.
d. Drag the third graph (**1-D Boxplot**) to the **Chart Preview** at the top of the window.
e. Select the variable **temp** and move it into the **X-Axis?** box.
f. Click **OK**.

### The SPSS Output

![Boxplot Diagram]

**Histogram:**

a. To create a histogram, we use the **Graphs** menu.
b. Click on the **Chart Builder** option.
c. In the **Gallery** at the bottom of the window, select the **Histogram** option.
d. Drag the first graph (**Simple Histogram**) to the **Chart Preview** at the top of the window.
e. Select the variable **temp** and move it into the **X-Axis?** box.
f. To display the normal distribution curve in the graphs, check the box **Display normal curve** in the **Element Properties** window.
g. In the **Element Properties** window, click **Apply**.
h. In the **Chart Builder** window, click **OK**.
Note:

- A third graph, called a Q-Q plot may be used for checking conditions. This option can be found under the Analyze menu and the Descriptive Statistics submenu.

3. Steps to conduct a one-sample $t$-test:
   a. To conduct a one-sample $t$-test, we use the Analyze menu.
   b. Scroll down to the Compare Means submenu and select the One-Sample T Test option. A window opens up. The window is shown below:

   ![One-Sample T Test window](image)

   c. Select the variable temp and move it into the Test Variable(s) box.
   d. Enter the hypothesized value (98.6) in the Test Value box, as shown above.
   e. Click OK.

Notes:

- In the output, a confidence interval for the mean is calculated as well. This is confidence interval for the difference between the mean and 98.6. To obtain a regular confidence interval for the mean, leave the test value to 0.
- To change the confidence level from the default of 95%, click on the Options button.
13.3 Testing Hypotheses About the Population Mean of Paired Differences

**Example 13.2 – Why Can’t the Pilot Have a Drink?**
Ten pilots performed tasks at a simulated altitude of 25,000 feet. Each pilot performed the tasks in a completely sober condition and, three days later, after drinking alcohol. The response variable is the time (in seconds) of useful performance of the tasks for each condition. The longer a pilot spends on useful performance, the better. The research hypothesis is that useful performance decreases with alcohol use. The data can be found in Section 13.3 of the book. The data is not recorded in an existing file, so the data set needs to be created.

1. **Steps** to prepare the data:
   a. Open a new data file.
   b. Use the **Variable View** tab to create 2 variables, e.g. *sober* and *alcohol*.
   c. Use the **Data View** tab to enter the data.
   d. Compute the differences using the **Transform** menu and the **Compute Variable**… option. Define a new variable *diff* to be *sober – alcohol*.
   e. Save the file.
2. **Steps** to check conditions before conducting a paired \( t \)-test:
   a. Repeat the steps for creating a boxplot and a histogram described in part 2 of Example 13.1. Use the variable \( \text{diff} \).

   **The SPSS Output**

   ![Boxplot and Histogram Output](image-url)
3. **Steps** to conduct a paired $t$-test:
   a. To conduct a paired $t$-test, we use the **Analyze** menu.
   b. Scroll down to the **Compare Means** submenu and select the **Paired-Samples T Test** option. A window opens up. The window is shown below:

   ![Paired-Samples T Test Window]

   c. Select the variable **sober** and move it into the **Paired Variables** box.
   d. Select the variable **alcohol** and move it into the **Paired Variables** box. It will automatically appear next to the variable **sober**.
   e. Click **OK**.

   **Note:**
   - In the output, a confidence interval for the mean difference is calculated as well. To change the confidence level from the default of 95%, click on the **Options** button.

   **The SPSS Output**

   **T-Test**

   **Paired Samples Statistics**

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>N</th>
<th>Std. Deviation</th>
<th>Std. Error Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pair 1 sober</td>
<td>546.60</td>
<td>10</td>
<td>238.812</td>
<td>75.519</td>
</tr>
<tr>
<td>alcohol</td>
<td>351.00</td>
<td>10</td>
<td>210.882</td>
<td>66.687</td>
</tr>
</tbody>
</table>

   **Paired Samples Correlations**

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>Correlation</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pair 1 sober &amp; alcohol</td>
<td>10</td>
<td>.480</td>
<td>.160</td>
</tr>
</tbody>
</table>

   **Paired Samples Test**

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Std. Deviation</th>
<th>Std. Error Mean</th>
<th>95% Confidence Interval of the Difference</th>
<th>t</th>
<th>df</th>
<th>Sig. (2-tailed)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pair 1 sober - alcohol</td>
<td>195.60</td>
<td>230.527</td>
<td>72.899</td>
<td>30.691 - 360.509</td>
<td>2.683</td>
<td>9</td>
<td>.025</td>
</tr>
</tbody>
</table>
Notes:
- The reported p-value is a 2-sided p-value.
- In the case of a one-sided test, check if the sign of the test statistic is consistent with the direction of extreme. If it is, the p-value reported in the output needs to be divided by 2. If the sign of the test statistic is opposite from the direction of extreme, the correct p-value can be found by dividing the value from the output by 2 and subtracting that from 1.
- In this example, the correct p-value is 0.025/2 = 0.0125.

13.4 Testing Hypotheses About the Difference Between Two Population Means (Independent Samples)

Example 13.4 – The Effect of a Stare on Driving Behavior
Social psychologists at the University of California at Berkley wanted to study the effect that staring at drivers would have on driver behavior. In a randomized experiment, the researchers either stared or did not stare at the drivers of cars stopped at a campus stop sign. The researchers timed how long it took each driver to proceed from the stop sign to a mark on the other side of the intersection. The times can be found in Example 11.11 of the book. The data is not recorded in an existing file, so the data set needs to be created.

1. Steps to prepare the data:
   a. Open a new data file.
   b. Use the Variable View tab to create 2 variables, e.g. time and stare.
   c. Set the Decimals of the variable time to 1.
   d. Set the Decimals of the variable stare to 0.
   e. Set the Measure of the variable time to Scale.
   f. Set the Measure of the variable stare to Nominal.
   g. The variable stare will be the grouping variable indicating whether the person was stared at or not. Assign labels to the values of the variable stare in the column Values: 0 = No, 1 = Yes.
   h. Use the Data View to enter the data. Note that the second column (the variable stare) will contain only zeros and ones.
   i. Save the file.
2. **Steps** to check conditions graphically, before conducting a two-sample $t$-test:

**Boxplots:**
- To construct boxplots, we will use the *Graphs* menu.
- Click on the *Chart Builder* option.
- In the *Gallery* at the bottom of the window, select the *Boxplot* option.
- Drag the first graph (*Simple Boxplot*) to the *Chart Preview* at the top of the window.
- Select the variable *time* and move it into the *Y-Axis?* box.
- Select the variable *stare* and move it into the *X-Axis?* box.
- Click *OK.*

The SPSS Output

![Boxplot](image)

**Histograms:**
- To create a histogram, we use the *Graphs* menu.
- Click on the *Chart Builder* option.
- In the *Gallery* at the bottom of the window, select the *Histogram* option.
- Drag the first graph (*Simple Histogram*) to the *Chart Preview* at the top of the window.
- Select the variable *time* and move it into the *X-Axis?* box.
- To separate the graph between the two groups, click on the *Groups/Point ID* tab (the *Gallery* will disappear).
- Check the box by *Rows panel variable.*
- Select the variable *stare* and move it into the *Panel?* box.
- To display the normal distribution curve in the graphs, check the box *Display normal curve* in the *Element Properties* window.
- In the *Element Properties* window, click *Apply.*
- In the *Chart Builder* window, click *OK.*
The SPSS Output

Note:
- A third graph, called a Q-Q plot may be used for checking conditions. This option can be found under the Analyze menu and the Descriptive Statistics submenu.

3. **Steps** to conduct a two-sample *t*-test (unpooled):
   a. To conduct a two-sample *t*-test, we use the Analyze menu.
   b. Scroll down to the Compare Means submenu and select the Independent-Samples T Test option. A window opens up. The window is shown below
   c. Select the variable time and move it into the Test Variable(s) box.
   d. Select the variable stare and move it into the Grouping Variable box.
   e. Click on the Define Groups button. A second window opens up.
f. The window is shown below:

![Define Groups window]

Define the two groups as 0 and 1.

h. In the Define Groups window, click Continue.

i. In the Independent-Samples T Test window, click OK.

Note that in the output, a confidence interval for the difference in means is calculated as well. To change the confidence level from the default of 95%, click on the Options button.

The SPSS Output

T-Test

<table>
<thead>
<tr>
<th>Group Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>stare</td>
</tr>
<tr>
<td>time</td>
</tr>
<tr>
<td>Yes</td>
</tr>
</tbody>
</table>

Independent Samples Test

<table>
<thead>
<tr>
<th>Levene's Test for Equality of Variances</th>
<th>t-test for Equality of Means</th>
<th>95% Confidence Interval of the Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>F</td>
<td>Sig.</td>
<td>t</td>
</tr>
<tr>
<td>Equal variances assumed</td>
<td>5.363</td>
<td>.029</td>
</tr>
<tr>
<td>Equal variances not assumed</td>
<td>2.412</td>
<td>21.596</td>
</tr>
</tbody>
</table>

Notes:
- The output reports both the test statistic for the unpoled and the pooled version of the independent samples test. In this case we need the unpoled version, so we will use the bottom row.
- The reported p-value is a 2-sided p-value.
- In the case of a one-sided test, check if the sign of the test statistic is consistent with the direction of extreme. If it is, the p-value reported in the output needs to be divided by 2. If the sign of the test statistic is opposite from the direction of extreme, the correct p-value can be found by dividing the value from the output by 2 and subtracting that from 1.
- In this example, the correct p-value is 0.025/2 = 0.0125.
Example 13.7 – Comparing Male and Female Sleep Time
On the second day of classes in Spring 2000, students in an introductory statistical methods class at the University of California at Davis were given a survey. One of the questions was, “How many hours of sleep did you get last night, to the nearest half hour?” The class consists of 83 females and 65 males. Who sleeps more on average, males or females? The data is recorded in the file UCDavis1.sav.

1. Steps to prepare the data:
   a. Open the file UCDavis1.sav.
   b. To analyze the students in the introductory statistical methods class only, we need to select these cases: For this, we use the Data menu.
   c. Scroll down to the Select Cases option. A window opens up. The window is shown below:
   
   ![Select Cases window](image)
   
   d. Select the If condition is satisfied option.
   e. Click on the If button. A second window opens up.
f. The window is shown below:

- Select the variable `class` and move it into the box on the right. Type the condition in this same box, as shown above: `class='NonLib'`. The accents are necessary because `class` is a string variable.
- In the If window, click Continue.
- In the Select Cases window, click OK. A filter variable will be added to the Data Editor. The LibArts cases will be crossed out in the Data Editor window.
2. **Steps** to check conditions before conducting a pooled two-sample *t*-test:
   a. Follow the steps in part 2 of Example 13.4. Use the variables `sleep` and `sex`.

   The SPSS Output

   ![Box plots and histograms for sleep by sex](image-url)
3. **Steps** to conduct a pooled two-sample $t$-test:
   a. Follow the steps in part 3 of Example 13.4.
   b. Select the variable **sleep** and move it into the **Test Variable(s)** box.
   c. Select the variable **sex** and move it into the **Grouping Variable** box.
   d. Define the two groups as **Female** and **Male**.

Note that in the output, a confidence interval for the difference in means is calculated as well. To change the confidence level from the default of 95%, click on the **Options** button.

**The SPSS Output**

**T-Test**

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>Mean</th>
<th>Std. Deviation</th>
<th>Std. Error Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>sleep</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>83</td>
<td>7.0151</td>
<td>1.75494</td>
<td>.19263</td>
</tr>
<tr>
<td>Male</td>
<td>65</td>
<td>6.5538</td>
<td>1.68199</td>
<td>.20863</td>
</tr>
</tbody>
</table>

**Independent Samples Test**

<table>
<thead>
<tr>
<th></th>
<th>Levene's Test for Equality of Variances</th>
<th>t-test for Equality of Means</th>
<th>95% Confidence Interval of the Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>F</td>
<td>Sig.</td>
<td>t</td>
</tr>
<tr>
<td>sleep</td>
<td>Equal variances assumed</td>
<td>F</td>
<td>.654</td>
</tr>
<tr>
<td></td>
<td>Equal variances not assumed</td>
<td>F</td>
<td>1.624</td>
</tr>
</tbody>
</table>

Notes:
- The output reports both the test statistic for the unpooled and the pooled version of the independent samples test. In this case we need the pooled version, so we will use the top row.
- The reported $p$-value is a 2-sided $p$-value.
- In the case of a one-sided test, check if the sign of the test statistic is consistent with the direction of extreme. If it is, the $p$-value reported in the output needs to be divided by 2. If the sign of the test statistic is opposite from the direction of extreme, the correct $p$-value can be found by dividing the value from the output by 2 and subtracting that from 1.
Chapter 14

Inference About Simple Regression

14.1 Sample and Population Regression Models

The Problem – Height and Handspan
The heights (in inches) and fully stretched handspans (in cm) of 167 college students are recorded in the file handheight.sav. The relationship between the handspan and height of these college students is to be studied. Can we predict the handspan of students based on their height?

1. Steps to prepare the data:
   a. Open the data set handheight.sav.

2. Steps to find the equation of the least squares regression line:
   a. To find the equation, we use the Analyze menu.
   b. Scroll down to the Regression submenu and select the Linear option. A window opens up. The window is shown below:
   c. Select the variable handspan and move it into the Dependent box.
   d. Select the variable height and move it into the Independent(s) box.
   e. Click OK.
The SPSS Output

Regression

Variables Entered/Removeda

<table>
<thead>
<tr>
<th>Model</th>
<th>Variables Entered</th>
<th>Variables Removed</th>
<th>Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>height*</td>
<td>.</td>
<td>Enter</td>
</tr>
</tbody>
</table>

a. All requested variables entered.
b. Dependent Variable: handspan

Model Summary

<table>
<thead>
<tr>
<th>Model</th>
<th>R</th>
<th>R Square</th>
<th>Adjusted R Square</th>
<th>Std. Error of the Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.740*</td>
<td>.547</td>
<td>.544</td>
<td>1.3009</td>
</tr>
</tbody>
</table>

a. Predictors: (Constant), height

ANOVAb

<table>
<thead>
<tr>
<th>Model</th>
<th>Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>F</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Regression</td>
<td>337.082</td>
<td>1</td>
<td>337.082</td>
<td>199.171</td>
</tr>
<tr>
<td></td>
<td>Residual</td>
<td>279.250</td>
<td>165</td>
<td>1.692</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>616.332</td>
<td>166</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

a. Predictors: (Constant), height
b. Dependent Variable: handspan

Coefficients*

<table>
<thead>
<tr>
<th>Model</th>
<th>Unstandardized Coefficients</th>
<th>Standardized Coefficients</th>
<th>t</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>B</td>
<td>Std. Error</td>
<td>Beta</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>(Constant)</td>
<td>-3.002</td>
<td>1.694</td>
<td>-1.772</td>
</tr>
<tr>
<td></td>
<td>height</td>
<td>.351</td>
<td>.025</td>
<td>.740</td>
</tr>
</tbody>
</table>

a. Dependent Variable: handspan

The least squares regression equation

The least squares regression equation may be obtained from the last (fourth) box in the output labeled Coefficients. The estimates are both given in the first column labeled B. The y-intercept is given in the first row labeled (Constant) and the slope is given in the second row labeled height. For this example the least squares regression equation is \( \hat{y} = -3.002 - .351x \).
The Problem – Girls’ heights and parents’ heights
When a girl is born can we use her parents’ heights to predict how tall she will be as an adult? The dataset ucdwomht.sav contains heights for 90 female students, along with the heights of the student’s mother and father.

1. **Steps** to prepare the data:
   a. Open the data set UCDwomht.sav.

2. **Steps** to find the multiple regression equation:
   a. To find the least squares regression equation, we use the Analyze menu.
   b. Scroll down to the Regression submenu and select the Linear option. A window opens up. The window is shown below:

   ![Linear Regression Window]

   c. Select the variable **height** and move it into the Dependent box.
   d. Select the variables **momheigh** and **dadheigh** and move them both into the Independent(s) box.
   e. Click **OK**.
The SPSS Output

Regression

Variables Entered/Removed a

<table>
<thead>
<tr>
<th>Model</th>
<th>Variables Entered</th>
<th>Variables Removed</th>
<th>Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>dadheigh, momheigh*</td>
<td>.</td>
<td>Enter</td>
</tr>
</tbody>
</table>

a. All requested variables entered.
b. Dependent Variable: height

Model Summary

<table>
<thead>
<tr>
<th>Model</th>
<th>R</th>
<th>R Square</th>
<th>Adjusted R Square</th>
<th>Std. Error of the Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.590a</td>
<td>.348</td>
<td>.333</td>
<td>2.0444</td>
</tr>
</tbody>
</table>

a. Predictors: (Constant), dadheigh, momheigh

ANOVA b

<table>
<thead>
<tr>
<th>Model</th>
<th>Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>F</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Regression</td>
<td>2</td>
<td>96.858</td>
<td>23.175</td>
<td>.000a</td>
</tr>
<tr>
<td></td>
<td>Residual</td>
<td>87</td>
<td>4.179</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>89</td>
<td>557.322</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

a. Predictors: (Constant), dadheigh, momheigh
b. Dependent Variable: height

Coefficients a

<table>
<thead>
<tr>
<th>Model</th>
<th>Unstandardized Coefficients</th>
<th>Standardized Coefficients</th>
<th>t</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>B</td>
<td>Std. Error</td>
<td>Beta</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>(Constant)</td>
<td>24.532</td>
<td>5.968</td>
<td>4.111</td>
</tr>
<tr>
<td></td>
<td>momheigh</td>
<td>.313</td>
<td>.087</td>
<td>.331</td>
</tr>
<tr>
<td></td>
<td>dadheigh</td>
<td>.291</td>
<td>.067</td>
<td>.394</td>
</tr>
</tbody>
</table>

a. Dependent Variable: height

The least squares regression equation

The least squares regression equation may be obtained from the last (fourth) box in the output labeled Coefficients. The estimates are all given in the first column labeled B. The y-intercept is given in the first row labeled (Constant) and the slope coefficients are given in the second and third rows labeled momheigh and dadheigh. For this example the least squares regression equation is:

Average height = –24.532 + 0.313 mom’s height + 0.291 dad’s height.
14.2 Estimating the Standard Deviation for Regression

Example 14.5 – Driver Age and Highway Sign-Reading Distance
In a study of the legibility and visibility of highway signs, a Pennsylvania research firm determined the maximum distance at which each of 30 drivers could read the newly designed sign. The 30 participants in the study ranged in age from 18 to 82 years old. The government agency that funded the research hoped to improve highway safety for older drivers and wanted to examine the relationship between \( x = \) the age of the driver and \( y = \) maximum distance the driver can read the highway sign (in feet). The data is recorded in the file signdist.sav.

1. **Steps** to prepare the data:
   a. Open the data set signdist.sav.

2. **Steps** to find the equation of the least squares regression line:
   a. To find the equation, we use the **Analyze** menu.
   b. Scroll down to the **Regression** submenu and select the **Linear** option. A window opens up.
   c. Select the variable **distance** and move it into the **Dependent** box.
   d. Select the variable **age** and move it into the **Independent(s)** box.
   e. Click **OK**.

The SPSS Output

**Regression**

<table>
<thead>
<tr>
<th>Variables Entered/Removed&lt;sup&gt;a&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
</tr>
<tr>
<td>1</td>
</tr>
</tbody>
</table>

a. All requested variables entered.
b. Dependent Variable: distance

**Model Summary**

<table>
<thead>
<tr>
<th>Model</th>
<th>R</th>
<th>R Square</th>
<th>Adjusted R Square</th>
<th>Std. Error of the Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.801&lt;sup&gt;a&lt;/sup&gt;</td>
<td>.642</td>
<td>.629</td>
<td>49.762</td>
</tr>
</tbody>
</table>

a. Predictors: (Constant), age

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

<table>
<thead>
<tr>
<th>Model</th>
<th>Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>F</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Regression</td>
<td>124332.643</td>
<td>1</td>
<td>124332.643</td>
<td>50.211</td>
</tr>
<tr>
<td></td>
<td>Residual</td>
<td>69334.024</td>
<td>28</td>
<td>2476.215</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>193666.667</td>
<td>29</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

a. Predictors: (Constant), age
b. Dependent Variable: distance
The least squares regression equation
The least squares regression equation may be obtained from the last (fourth) box in the output labeled Coefficients. The estimates are both given in the first column labeled B. The y-intercept is given in the first row labeled (Constant) and the slope is given in the second row labeled age. For this example the least squares regression equation is $\hat{y} = 576.682 - 3.007x$.

The correlation coefficient
The correlation coefficient may be obtained from the second box in the output labeled Model Summary. The magnitude of the correlation coefficient is given in the first column labeled R. Note that this is only the magnitude of the correlation coefficient. Since the sign of the correlation coefficient equals the sign of the slope, check with the slope if the correlation coefficient should be positive or negative. For this example the correlation coefficient should be negative, since we have a negative slope. The correlation coefficient is $-0.801$.

$r^2$
The value of $r^2$ may be obtained from the second box in the output labeled Model Summary. The value is listed in the second column labeled R Square. For this example $r^2 = 0.642$.

Sums of Squares
Various Sums of Squares may be obtained from the third box in the output labeled ANOVA in the first column labeled Sum of Squares. The SSE may be obtained from the second row labeled Residual and the SSTO may be obtained from the third row labeled Total. For this example SSE = 69334.024 and SSTO = 193,666.667.

Standard Deviation for Regression
The standard deviation may be obtained from the second box in the output labeled Model Summary. The standard deviation is given in the last column labeled Std. Error of the Estimate. For this example $s = 49.762$.

14.3 Inference About the Linear Regression Relationship

Example 14.3/7 – Driver Age and Highway Sign-Reading Distance
In a study of the legibility and visibility of highway signs, a Pennsylvania research firm determined the maximum distance at which each of 30 drivers could read the newly designed sign. The 30 participants in the study ranged in age from 18 to 82 years old. The government agency that funded the research hoped to improve highway safety for older drivers and wanted to examine the relationship between $x =$ the age of the driver and $y =$ maximum distance the driver can read the highway sign (in feet). The data is recorded in the file signdist.sav.

1. Steps to prepare the data:
   a. Open the data set signdist.sav.
2. **Steps** to test hypotheses about the population slope:
   a. Run the regression as described in Section 14.2.
   b. To perform a hypothesis test about the population slope we may use the fourth box of the regression output named **Coefficients**. This part of the output is show below:

   ![Coefficients Table]

   a. Dependent Variable: distance

   c. The test statistic and corresponding *p*-value for testing the hypotheses $H_0: \beta_1 = 0$ versus $H_a: \beta_1 \neq 0$ can be found in the two last columns of the output in the row labeled **age**. The $t$-statistic can be found in the column labeled **t** and the corresponding *p*-value in the column labeled **Sig**. For this example $t = -7.086$ and $p$-value $= 0.000$ (i.e. $p$-value $< 0.001$).

   Notes:
   - To test hypotheses with a different null-value, the test statistic may be obtained from the estimate and its standard error given in the same row of the output.
   - The reported *p*-value is a 2-sided *p*-value.
   - In the case of a one-sided test, check if the sign of the test statistic is consistent with the direction of extreme. If it is, the *p*-value reported in the output needs to be divided by 2. If the sign of the test statistic is opposite from the direction of extreme, the correct *p*-value can be found by dividing the value from the output by 2 and subtracting that from 1.

3. **Steps** to find a 95% confidence interval for the population slope:
   a. Run the regression as described in Section 14.2.
   b. Before clicking **OK**, click on the **Statistics** button:
c. A second window opens up. The window is shown below:

![Linear Regression: Statistics window]

- Under the **Regression Coefficients** heading, check the **Confidence intervals** option. Adjust the confidence level if needed.
- In the **Statistics** window, click **Continue**.
- In the **Linear Regression** window, click **OK**.

**Part of the SPSS Output**

**Regression**

<table>
<thead>
<tr>
<th>Model</th>
<th>Unstandardized Coefficients</th>
<th>Standardized Coefficients</th>
<th>95.0% Confidence Interval for B</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>B</td>
<td>Std. Error</td>
<td>Beta</td>
</tr>
<tr>
<td>1</td>
<td>(Constant)</td>
<td>576.682</td>
<td>23.471</td>
</tr>
<tr>
<td>age</td>
<td>-3.007</td>
<td>.424</td>
<td>-.801</td>
</tr>
</tbody>
</table>

a. Dependent Variable: distance

The 95% confidence interval for the slope can be found in the last 2 columns of the fourth box of the output (displayed above). For this example, the 95% confidence interval is (−3.876, −2.138).
4. **Steps** to test hypotheses about the population correlation coefficient:
   a. To test hypotheses about the correlation coefficient, we use the **Analyze** menu.
   b. Scroll down to the **Correlate** submenu and select the **Bivariate** option. A window opens up. The window is shown below:

   ![Bivariate Correlations Window](image)

   c. Select the variables **age** and **distance** and move them both into the **Variables** box.
   d. Under the heading **Correlation Coefficients** check the **Pearson** option (the default).
   e. Under the heading **Test of Significance** select the **Two-tailed** option (the default). If a one-sided test is to be performed, the **One-tailed** option may be selected.
   f. Click **OK**.

   **The SPSS Output**

   **Correlations**

<table>
<thead>
<tr>
<th></th>
<th>age</th>
<th>distance</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Pearson</strong></td>
<td>1.00</td>
<td>-0.801**</td>
</tr>
<tr>
<td>Sig. (2-tailed)</td>
<td>.000</td>
<td>.000</td>
</tr>
<tr>
<td>N</td>
<td>30</td>
<td>30</td>
</tr>
<tr>
<td><strong>distance</strong></td>
<td>-0.801**</td>
<td>1.00</td>
</tr>
<tr>
<td>Sig. (2-tailed)</td>
<td>.000</td>
<td>.000</td>
</tr>
<tr>
<td>N</td>
<td>30</td>
<td>30</td>
</tr>
</tbody>
</table>

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

   Note that the correlation coefficient and the \( p \)-value can be found in 2 locations, as the correlation between age and distance equals the correlation between distance and age. For this example, \( r = -0.801 \) and \( p \)-value = 0.000 (i.e. \( p \)-value < 0.001).
14.4 Predicting $y$ and Estimating Mean $y$ at a Specific $x$

Example 14.9 – Driver Age and Highway Sign-Reading Distance
Continued from section 14.3.

1. **Steps** to prepare the data:
   a. Open the data set signdist.sav.

2. **Steps** to calculate a prediction interval for specified values of $x$:
   Note: SPSS can only calculate prediction intervals for the observed values of $x$. If a different value of $x$ is to be used, hand calculations with the formula in the book will need to be utilized.
   a. Run the regression as described in Section 14.2.
   b. Before clicking **OK**, click on the **Save** button:
c. A second window opens up. The window is shown below:

d. Under the Prediction Intervals heading, check the Individual option.
e. Select a confidence level (95% is the default).
f. In the Save window, click Continue.
g. In the Linear Regression window, click OK.
h. The prediction intervals at the observed values of $x$ will appear in the Data Editor as 2 new variables labeled LICI_1 and UICI_1 (Lower and Upper Individual Confidence Interval).
Example 14.10 – Estimating Mean Weight of College Men at Various Heights
A sample of \( n = 43 \) college men in a Penn State statistics class participated in an experiment to model the relationship between \( x = \) height (in inches) and \( y = \) weight (in pounds). The data is recorded in the file wtheightM.sav.

1. **Steps** to prepare the data:
   a. Open the data set wtheightM.sav.

2. **Steps** to calculate a confidence interval for values of \( x \):
   Note: SPSS can only calculate confidence intervals for the mean of \( y \) at the observed values of \( x \). If a different value of \( x \) is to be used, hand calculations with the formula in the book will need to be utilized.
   a. Run the regression as described in Section 14.2. Use the variable weight as the **Dependent** and the variable height as the **Independent** variable.
   b. Before clicking **OK**, click on the **Save** button. A second window opens up.
c. The window is shown below:

d. Under the Prediction Intervals heading, check the Mean option.
e. Select a confidence level (95% is the default).
i. In the Save window, click Continue.
j. In the Linear Regression window, click OK.
f. The confidence intervals at the observed values of x will appear in the Data Editor as 2 new variables labeled LMCI_1 and UMCI_1 (Lower and Upper Mean Confidence Interval).
The Data Editor
14.5 Checking Conditions for Using Regression Models for Inference

Example 14.11 – Estimating Mean Weight of College Men at Various Heights
A sample of \( n = 43 \) college men in a Penn State statistics class participated in an experiment to model the relationship between \( x = \) height (in inches) and \( y = \) weight (in pounds). The data is recorded in the file wtheightM.sav.

1. **Steps** to prepare the data:
   a. Open the data set wtheightM.sav.

2. **Steps** to create residual plots:
   g. Run the regression as described in Section 14.2. Use the variable **weight** as the **Dependent** and the variable **height** as the **Independent** variable.
   a. Before clicking **OK**, click on the **Plots** button.
   b. A second window opens up. The window is shown below:
   
   ![Linear Regression: Plots window](image)

   c. To create a residual plot, we will create a scatterplot. Select the variable **ZRESID** (\( z \)-scores of the residuals) and move it into the **Y** box.
   d. Select the variable **ZPRED** (\( z \)-scores of the predicted values) and move it into the **X** box.
   e. If additional scatterplots are desired, click on the **Next** button and add them at this point.
   f. Under the **Standardized Residual Plots** heading, place a checkmark in front of **Histogram**. If desired, also check the option **Normal probability plot** to create a Q-Q plot of the residuals.
   g. In the **Plots** window, click **Continue**.
   h. In the **Linear Regression** window, click **OK**.
The SPSS Output

Regression

**Variables Entered/Removed**

<table>
<thead>
<tr>
<th>Model</th>
<th>Variables Entered</th>
<th>Variables Removed</th>
<th>Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>height*</td>
<td>.</td>
<td>Enter</td>
</tr>
</tbody>
</table>

a. All requested variables entered.
b. Dependent Variable: weight

**Model Summary**

<table>
<thead>
<tr>
<th>Model</th>
<th>R</th>
<th>R Square</th>
<th>Adjusted R Square</th>
<th>Std. Error of the Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.568*</td>
<td>.323</td>
<td>.307</td>
<td>24.001</td>
</tr>
</tbody>
</table>

a. Predictors: (Constant), height
b. Dependent Variable: weight

**ANOVA**

<table>
<thead>
<tr>
<th>Model</th>
<th>Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>F</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Regression</td>
<td>1</td>
<td>11277.167</td>
<td>19.577</td>
<td>.000*</td>
</tr>
<tr>
<td></td>
<td>Residual</td>
<td>41</td>
<td>576.027</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>42</td>
<td>34894.279</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

a. Predictors: (Constant), height
b. Dependent Variable: weight

**Coefficients**

<table>
<thead>
<tr>
<th>Model</th>
<th>Unstandardized Coefficients</th>
<th>Standardized Coefficients</th>
<th>t</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>B</td>
<td>Std. Error</td>
<td>Beta</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>(Constant)</td>
<td>-317.919</td>
<td>110.922</td>
<td>-2.866</td>
</tr>
<tr>
<td></td>
<td>height</td>
<td>6.996</td>
<td>1.581</td>
<td>.568</td>
</tr>
</tbody>
</table>

a. Dependent Variable: weight

**Residuals Statistics**

<table>
<thead>
<tr>
<th></th>
<th>Minimum</th>
<th>Maximum</th>
<th>Mean</th>
<th>Std. Deviation</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predicted Value</td>
<td>143.81</td>
<td>206.77</td>
<td>172.60</td>
<td>16.386</td>
<td>43</td>
</tr>
<tr>
<td>Residual</td>
<td>-45.779</td>
<td>58.213</td>
<td>0.000</td>
<td>23.713</td>
<td>43</td>
</tr>
<tr>
<td>Std. Predicted Value</td>
<td>-1.757</td>
<td>2.085</td>
<td>0.000</td>
<td>1.000</td>
<td>43</td>
</tr>
<tr>
<td>Std. Residual</td>
<td>-1.907</td>
<td>2.425</td>
<td>0.000</td>
<td>0.988</td>
<td>43</td>
</tr>
</tbody>
</table>

a. Dependent Variable: weight
Charts

Histogram
Dependent Variable: weight

Normal P-P Plot of Regression Standardized Residual
Dependent Variable: weight
Note:

- To add a line at $y = 0$, to represent the $x$-axis, double click the graph. The **Chart Editor** will open up:

![Chart Editor](image)

- To add the line, click on the **Add a reference line to the Y axis** button at the top of the screen. A horizontal line will be added to your graph.
- The **Properties** window will open up. It is shown below:

![Properties Window](image)

- Set the **Position** of the reference line to 0 (zero).
- In the **Properties** window, click **Apply**.
- Close the **Chart Editor**.
4. **Steps** to perform a transformation on the $y$-variable:
   a. To transform data, we use the **Transform** menu.
   b. Scroll down to the **Compute Variable** option. A window opens up. The window is shown below:

   ![Image of Compute Variable window]

   c. Type a name for the transformed variable in the **Target Variable** box. For this example, a natural log transformation on the $y$-variable will be used. The new variable is **lnweight**.
   d. Under the **Functions group** heading, select **Arithmetic**.
   e. Under the **Functions and Special Variables** heading, select **Ln** and move it into the **Numeric Expression** box by clicking on the up arrow.
   f. Replace the question mark in the **Ln** function by the variable **weight**, by selecting it and moving it into the **Numeric Expression** box by clicking on the right arrow.
   g. Click **OK**.
   h. The new variable **lnweight** will appear as a new variable in the **Data Editor**. You can now run a regression as described in Section 14.2, using the new variable **lnweight** as the **Dependent** variable.

   **The Data Editor**

   ![Image of Data Editor]
Chapter 15

More About Inference for Categorical Variables

15.1 The Chi-Square Test for Two-Way Tables

Example 15.1/3/4 – Ear Infections and Xylitol Sweetener

Xylitol is a food sweetener that may also have antibacterial properties. In an experiment conducted in Finland, researchers investigated whether the regular use of Xylitol could reduce the risk of a middle ear infection in children in daycare centers. A total of 533 children participated in the study. The data is not recorded in a data file, but the data is recorded in Table 15.1. The data needs to be created from the table.

1. Steps to prepare the data:
   a. Open a new data file.
   b. Use the Variable View tab to create three variables, e.g. freq, Xylitol, and Earinfection.
   c. For all three variables, set the Decimals to 0.
   d. For the variable freq set the Measure to Scale.
   e. For the variables Xylitol and Earinfection, set the Measure to Nominal.
   f. For the variable Xylitol define value labels in the column Values, e.g. 1 = Placebo, 2 = Xylitol gum, 3 = Xylitol lozenge.
   g. For the variable Earinfection define value labels in the column Values, e.g. 0 = No and 1 = Yes.
   h. Use the Data View to enter the observed counts in the first column: 129, 150, 137, 49, 29, 39.
   i. Enter the values 1, 2, and 3 for the three Xylitol groups in the second column. Make sure to match these with the counts in column 1 (if you used the order given above, use 1, 2, 3, 1, 2, 3).
   j. Enter the values 0 and 1 in the third column, corresponding to whether the child had an ear infection or not. Make sure to match these with the counts in column 1 (if you entered the values in the order given above, use 0, 0, 0, 1, 1, 1).
   k. Save the file.
2. **Steps** to prepare count data for a chi-square test:
   a. The frequency variable needs to be designated as a weighting variable. For this we use the **Data** menu.
   b. Scroll down to the **Weight Cases** option. A window opens up. The window is shown below:

   ![Weight Cases Window]

   c. Select the **Weight cases by** option.
   d. Select the variable `freq` and move it into the **Frequency Variable** box.
   e. Click **OK**.

3. **Steps** to conduct a chi-square test:
   a. To conduct a chi-square test, we use the **Analyze** menu.
   b. Scroll down to the **Descriptive Statistics** submenu and select the **Crosstabs** option. A window opens up. The window is shown below:

   ![Crosstabs Window]

   c. Select the variable **Xylitol** and move it into the **Row(s)** box.
   d. Select the variable **Earinfection** and move it into the **Column(s)** box. Note that there is no specific order to these variables, so they may be reversed.
   e. Click on the **Statistics** button. A window opens up.
f. The window is shown below:

![Crosstabs: Statistics](image)

At the top left-hand corner, check the **Chi-square** option.

h. In the **Statistics** window, click **Continue**.

i. Click on the **Cells** button. A second window opens up.

j. The window is shown below:

![Crosstabs: Cell Display](image)

k. Under the **Counts** heading, check both the **Observed** and the **Expected** options. If row or column percents are desired, those can be obtained at this point as well. To match the percentages in the book, check **Row**.

l. In the **Cell Display** window, click **Continue**.

m. In the **Crosstabs** window, click **OK**.

Note: If an existing data set is to be used, listing all the individual responses, please refer back to Chapter 4, Section 6.4.
The SPSS Output

Crosstabs

<table>
<thead>
<tr>
<th>Xylitol * Earinfection Crosstabulation</th>
<th>Earinfection</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No</td>
<td>Yes</td>
<td>Total</td>
</tr>
<tr>
<td>Xylitol Placebo</td>
<td>129</td>
<td>49</td>
<td>178</td>
</tr>
<tr>
<td>Expected Count</td>
<td>138.9</td>
<td>39.1</td>
<td>178.0</td>
</tr>
<tr>
<td>% within Xylitol</td>
<td>72.5%</td>
<td>27.5%</td>
<td>100.0%</td>
</tr>
<tr>
<td>Xylitol gum</td>
<td>150</td>
<td>29</td>
<td>179</td>
</tr>
<tr>
<td>Expected Count</td>
<td>139.7</td>
<td>39.3</td>
<td>179.0</td>
</tr>
<tr>
<td>% within Xylitol</td>
<td>83.8%</td>
<td>16.2%</td>
<td>100.0%</td>
</tr>
<tr>
<td>Xylitol lozenge</td>
<td>137</td>
<td>39</td>
<td>176</td>
</tr>
<tr>
<td>Expected Count</td>
<td>137.4</td>
<td>38.6</td>
<td>176.0</td>
</tr>
<tr>
<td>% within Xylitol</td>
<td>77.8%</td>
<td>22.2%</td>
<td>100.0%</td>
</tr>
<tr>
<td>Total</td>
<td>416</td>
<td>117</td>
<td>533</td>
</tr>
<tr>
<td>Expected Count</td>
<td>416.0</td>
<td>117.0</td>
<td>533.0</td>
</tr>
<tr>
<td>% within Xylitol</td>
<td>78.0%</td>
<td>22.0%</td>
<td>100.0%</td>
</tr>
</tbody>
</table>

Chi-Square Tests

<table>
<thead>
<tr>
<th></th>
<th>Value</th>
<th>df</th>
<th>Asymp. Sig. (2-sided)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pearson Chi-Square</td>
<td>6.690*</td>
<td>2</td>
<td>.035</td>
</tr>
<tr>
<td>Likelihood Ratio</td>
<td>6.772</td>
<td>2</td>
<td>.034</td>
</tr>
<tr>
<td>Linear-by-Linear Association</td>
<td>1.504</td>
<td>1</td>
<td>.220</td>
</tr>
<tr>
<td>N of Valid Cases</td>
<td>533</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 38.63.

The chi-square test statistic can be found in the bottom box of the output labeled Chi-Square Tests. The test statistic is listed in the first column labeled Value and the first row labeled Pearson Chi-Square. For this example the chi-square test statistic equals 6.690.

The p-value corresponding to the test statistic can be found in the same row, but in the third column labeled Asymp. Sig. For this example the p-value equals 0.035.
The Problem – Finding a \( p \)-value

Sometimes you may calculate a chi-square test statistic by hand. How do we find the corresponding \( p \)-value?

1. **Steps** to prepare the data:
   a. Open a new data set.
   b. Use the **Variable View** tab to create a new variable, e.g. \( \text{chi} \).
   c. Give the variable the appropriate measurement level (Scale).
   d. Use the **Data View** to enter the value of the chi-square test statistic in the column named \( \text{chi} \). For this exercise we will use the test statistic value of 6.690 with 2 degrees of freedom from Example 15.1.

2. **Steps** to find the \( p \)-value:
   a. To find the \( p \)-value we use the **Transform** menu.
   b. Scroll down to the **Compute Variable** option. A window opens up. The window is shown below:

   ![Compute Variable Window](image)

   c. Type a variable name in the **Target Variable** box, such as \( p\text{value} \).
   d. In the **Numeric Expression** box enter the number \( 1 \) and the minus sign (\(-\)).
   e. In the **Function Group** box select **CDF & Noncentral CDF**.
   f. In the **Functions and Special Variables** box select **Cdf.Chisq** and move it in the **Numeric Expression** box by clicking on the up arrow.
   g. The two question marks need to be replaced by the value of your chi-square test statistic and the appropriate degrees of freedom. Place the cursor in the location of the first question mark. Click on the variable \( \text{chi} \) and move it into the **Numeric Expression** box by clicking on the right arrow. Place the cursor on the location of the second question mark and type the degrees of freedom. For this exercise that is 2.
   h. Click **OK**.
   i. A new variable has now been added to the **Data Editor** called \( p\text{value} \). It contains the \( p \)-value.
   j. To increase the number of significant digits, use the **Variable View** and change the **Decimals** to 3 or 4. For this exercise the \( p \)-value equals 0.0353.
Note:
• If you wish to find multiple *p*-values at once, all with different numbers of degrees of freedom, you may create a variable called *chi* for the values of the chi-square test statistics and a separate variable called *df* for the degrees of freedom. When calculating the *p*-values, replace the first question mark with the variable *chi* and the second question mark with the variable *df*. All *p*-values will then appear in the third column called *pvalue*.

### 15.2 Analyzing 2x2 Tables

**Example 15.12 – Sheep, Goats, and ESP**

In parapsychology, the sheep-goat effect is that people who believe in ESP (sheep) tend to be more successful in ESP experiments than people who do not believe in it (goats). The effect is slight, but it has been observed in many forced-choice ESP experiments. In the experiment, the participant “guesses” which of several known possibilities will be selected by the experimenter. Students in a large statistics class used a coin to perform such an experiment. Students were categorized as sheep and goats and the numbers of correct and incorrect guesses in 10 coin tosses were recorded. More details can be found in Chapter 15 of the book. The data is not recorded in a data file, so the data needs to be created.

1. **Steps** to prepare the data:
   a. Open a new data file.
   b. Use the *Variable View* tab to create three variables, e.g. *count*, *group*, and *guess*.
   c. For all three variables, set the *Decimals* to 0.
   d. For the variable *count*, set the *Measure* to *Scale*.
   e. For the variables *group* and *guess*, set the *Measure* to *Nominal*.
   f. For the variable *group* define value labels in the column *Values*, e.g. 1 = Sheep, 2 = Goats.
   g. For the variable *guess* define value labels in the column *Values*, e.g. 1 = Yes and 2 = No.
   h. Use the *Data View* to enter the observed counts in the first column: 582, 389, 538, 411.
   i. Enter the values 1 and 2 for the groups in the second column. Make sure to match these with the counts in column 1 (if you used the order given above, use 1, 2, 1, 2).
j. Enter the values 1 and 2 in the third column, corresponding to whether the guess was successful or not. Make sure to match these with the counts in column 1 (if you entered the values in the order given above, use 1, 1, 2, 2).
k. Save the file.

2. Steps to conduct Fisher’s Exact test:
a. To conduct Fisher’s Exact test, we follow the exact same steps as for the Chi-square test in Example 15.1 in Section 15.1. For a 2x2 table, Fisher’s Exact test will automatically be calculated.
b. The count variable needs to be designated as a weighting variable. For this we use the Data menu.
c. Scroll down to the Weight Cases option.
d. Select the Weight cases by option. A small window will open. It is shown below:
e. Select the variable count and move it into the Frequency Variable box.
f. Click OK.
g. To conduct Fisher’s exact test, we use the Analyze menu.
h. Scroll down to the Descriptive Statistics submenu and select the Crosstabs option.
i. Select the 2 grouping variables (group and guess) and move them into the Row(s) box and the Column(s) box (in no specific order). To get the table as in the book, place group in the rows and guess in the columns.
j. At the right of the window, click on the Statistics button.
k. At the top left-hand corner, select the Chi-square option.
l. Click Continue.
m. To get the row percentages, click on the Cells button. Select the Rows option under Percentages.
n. Click Continue.
o. Click OK.
The SPSS Output

Crosstabs

Case Processing Summary

<table>
<thead>
<tr>
<th></th>
<th>Valid</th>
<th>Missing</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>Percent</td>
<td>N</td>
<td>Percent</td>
</tr>
<tr>
<td>group * guess</td>
<td>1920</td>
<td>100.0%</td>
<td>0</td>
</tr>
</tbody>
</table>

group * guess Crosstabulation

<table>
<thead>
<tr>
<th>group</th>
<th>Sheep</th>
<th>Count</th>
<th>Yes</th>
<th>No</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>% within group</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sheep</td>
<td>502</td>
<td>536</td>
<td>1120</td>
<td></td>
<td></td>
</tr>
<tr>
<td>% within group</td>
<td>52.0%</td>
<td>48.0%</td>
<td>100.0%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Goals</td>
<td>Count</td>
<td>389</td>
<td>411</td>
<td>800</td>
<td></td>
</tr>
<tr>
<td>% within group</td>
<td>48.6%</td>
<td>51.4%</td>
<td>100.0%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>Count</td>
<td>971</td>
<td>949</td>
<td>1920</td>
<td></td>
</tr>
<tr>
<td>% within group</td>
<td>50.6%</td>
<td>49.4%</td>
<td>100.0%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Chi Square Tests

<table>
<thead>
<tr>
<th></th>
<th>Value</th>
<th>df</th>
<th>Asymp. Sig. (2-sided)</th>
<th>Exact Sig. (2-sided)</th>
<th>Exact Sig. (1-sided)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pearson Chi-Square</td>
<td>2.083</td>
<td>1</td>
<td>.143</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Continuity Correction</td>
<td>1.950</td>
<td>1</td>
<td>.183</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Likelihood Ratio</td>
<td>2.083</td>
<td>1</td>
<td>.143</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Fisher's Exact Test</strong></td>
<td></td>
<td></td>
<td><strong>.152</strong></td>
<td><strong>.081</strong></td>
<td></td>
</tr>
<tr>
<td>Linear by Linear Association</td>
<td>2.081</td>
<td>1</td>
<td>.149</td>
<td></td>
<td></td>
</tr>
<tr>
<td>N of Valid Cases</td>
<td>1920</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 395.42.
b. Computed only for a 2x2 table.
Example 15.14 – Asthma Prevalence over Time
Suppose that a researcher records whether \( n = 500 \) children have asthma or not at age 13, and then again records whether these same 500 individuals have asthma 7 years later. Table 15.9 in e book contains the data for such a study. The researcher is interested in the difference in asthma rates at the two ages. The data is not recorded in a data file, so the data needs to be created.

1. **Steps** to prepare the data:
   a. Open a new data file.
   b. Use the **Variable View** tab to create three variables, e.g. **count**, **at13**, and **at20**.
   c. For all three variables, set the **Decimals** to **0**.
   d. For the variable **count**, set the **Measure** to **Scale**.
   e. For the variables **at13** and **at20**, set the **Measure** to **Nominal**.
   f. For the variables **at13** and **at20**, define value labels in the column **Values**, e.g. 0 = No, 1 = Yes.
   g. Use the **Data View** to enter the observed counts in the first column: 50, 8, 22, 420.
   h. Enter the values 0 and 1 in the second column. Make sure to match these with the counts in column 1 (if you used the order given above, use 1, 0, 1, 0).
   i. Enter the values 0 and 1 in the third column. Make sure to match these with the counts in column 1 (if you entered the values in the order given above, use 1, 1, 0, 0).
   j. The count variable needs to be designated as a weighting variable. For this we use the **Data** menu.
   k. Scroll down to the **Weight Cases** option.
   l. Select the **Weight cases by** option.
   m. Select the variable **count** and move it into the **Frequency Variable** box.
   n. Click **OK**.
   o. Save the file.

2. **Steps** to conduct McNemar’s test:
   a. To conduct McNemar’s Test, we use the **Analyze** menu.
   b. Scroll down to the **Nonparametric Tests** submenu and select the **Related Samples** option. A window opens up. The window is shown below:
c. Select the **Customize analysis** option and click on the **Fields** tab:

![Customize analysis](image1)

**Customize analysis** and **Fields** tabs are highlighted.

d. Select the **Use custom field assignments** option.

e. Select the variables **at13** and **at20** and move them into the **Test Fields** box.

f. Click on the **Settings** tab:

![Settings](image2)

**Settings** tab is highlighted.

g. Select the **Customize tests** option and put a checkmark in front of the **McNemar’s test**. Using the 0-1 notation, we can use the default settings of this test. If you have coded the data differently, you can define which value to use as a success by clicking on the **Define Success** button.

h. Click **Run**.
The SPSS Output

Hypothesis Test Summary

<table>
<thead>
<tr>
<th>Null Hypothesis</th>
<th>Test</th>
<th>Sig.</th>
<th>Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>The distributions of different</td>
<td>Related-Samples</td>
<td>.018</td>
<td>Reject the null</td>
</tr>
<tr>
<td>values across at13 and at20 are</td>
<td>McNemar Test</td>
<td></td>
<td>hypothesis.</td>
</tr>
<tr>
<td>equally likely.</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Asymptotic significances are displayed. The significance level is .05.

Note:
- If you double click on the output, you will find some more information about the test and a graphical display of the observed data:
15.3 Testing Hypotheses About One Categorical Variable: Goodness of Fit

Example 15.15 – The Pennsylvania Daily Number
The Pennsylvania Daily Number is a state lottery game in which the state constructs a three-digit number by drawing a digit between 0 and 9 from three different containers. If numbers are randomly selected from the first container, each value between 0 and 9 would be equally likely to occur. Data for 500 days is recorded. The data is not available in a data file, so the data needs to be created.

1. **Steps** to prepare the data:
   a. Open a new data file.
   b. Use the **Variable View** tab to create two variables, e.g. **count** and **digit**.
   c. For both variables, set the **Decimals** to 0.
   d. For the variable **count** set the **Measure** to **Scale**.
   e. For the variable **digit** set the **Measure** to **Nominal**.
   f. Use the **Data View** to enter the observed counts in the first column: 47, 50, 55, 46, 53, 39, 55, 55, 44, 56.
   g. Enter the values 0 through 9 in the second column.
   h. The count variable needs to be designated as a weighting variable. For this we use the **Data** menu.
   i. Scroll down to the **Weight Cases** option.
   j. Select the **Weight cases by** option.
   k. Select the variable **count** and move it into the **Frequency Variable** box.
   l. Click **OK**.
   m. Save the file.

2. **Steps** to conduct a chi-square goodness of fit test:
   a. To conduct a chi-square goodness of fit test, we use the **Analyze** menu.
   b. Scroll down to the **Nonparametric Tests** submenu and select the **One Sample** option. A window opens up. The window is shown below:
c. Select the **Customize analysis** option and click on the **Fields** tab:

![Image of the Fields tab with highlighted options]

- **Use custom field assignments** option.
- Select the variable **digit** and move them into the **Test Fields** box.
- Click on the **Settings** tab:

![Image of the Settings tab with highlighted options]

- **Customize tests** option and put a checkmark in front of the **Chi-Square test**.
i. Click on the **Options** button. A window will open up. The window is shown below:

![Chi Square Test Options](image)

j. Under the **Choose Test Option** heading, select the distribution specified in the null hypothesis. If all proportions are stated to be equal, select the **All categories have equal probability** option (the default). If specific proportions are stated in the null hypothesis, select the **Customize expected probability** option and enter the values in order. Make sure the proportions entered add up to 1. For this example, the numbers should be equally likely, so keep the **All categories have equal probability** option.

k. Click **OK** in the **Options** window.

l. Click **Run**.

### The SPSS Output

#### Hypothesis Test Summary

<table>
<thead>
<tr>
<th>Null Hypothesis</th>
<th>Test</th>
<th>Sig.</th>
<th>Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>The categories of digit occur with equal probabilities.</td>
<td>One-Sample Chi-Square Test</td>
<td>.736</td>
<td>Retain the null hypothesis.</td>
</tr>
</tbody>
</table>

Asymptotic significances are displayed. The significance level is .05.

### Notes:
- If a data file is available containing the raw data (e.g. a list of length 500 with the numbers 0 through 9), the same procedure may be used, omitting the weighting step.
- If you double click on the output, you will find some more information about the test and a graphical display of the observed data:
3. **Alternative steps** to conduct a chi-square goodness of fit test:
   a. To perform the test, we use the Analyze menu.
   b. Scroll down to the Nonparametric Tests submenu, the Legacy Dialogs submenu, and select the Chi-square option. A window opens up. The window is shown below:
   
   ![Chi-square Test Window](image)
   
   c. Select the variable digit and move it into the Test Variable List box.
   d. Under the heading Expected Values, select the All categories equal option. If specific proportions are stated in the null hypothesis, select the Values option and enter the values and corresponding probabilities in order. Click Add after each one. Make sure the proportions entered add up to 1.
   e. Click OK.
The SPSS Output

Chi-Square Test

Frequencies

<table>
<thead>
<tr>
<th>digit</th>
<th>Observed N</th>
<th>Expected N</th>
<th>Residual</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>47</td>
<td>50.0</td>
<td>-3.0</td>
</tr>
<tr>
<td>1</td>
<td>50</td>
<td>50.0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>55</td>
<td>50.0</td>
<td>5.0</td>
</tr>
<tr>
<td>3</td>
<td>46</td>
<td>50.0</td>
<td>-4.0</td>
</tr>
<tr>
<td>4</td>
<td>53</td>
<td>50.0</td>
<td>3.0</td>
</tr>
<tr>
<td>5</td>
<td>39</td>
<td>50.0</td>
<td>-11.0</td>
</tr>
<tr>
<td>6</td>
<td>55</td>
<td>50.0</td>
<td>5.0</td>
</tr>
<tr>
<td>7</td>
<td>55</td>
<td>50.0</td>
<td>5.0</td>
</tr>
<tr>
<td>8</td>
<td>44</td>
<td>50.0</td>
<td>-6.0</td>
</tr>
<tr>
<td>9</td>
<td>56</td>
<td>50.0</td>
<td>6.0</td>
</tr>
<tr>
<td>Total</td>
<td>500</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Test Statistics

<table>
<thead>
<tr>
<th>digit</th>
<th>Chi-square</th>
<th>df</th>
<th>Asymp. Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>6.040a</td>
<td>9</td>
<td>.736</td>
</tr>
</tbody>
</table>

a. 0 cells (.0%) have expected frequencies less than 5. The minimum expected cell frequency is 50.0.
Chapter 16

Analysis of Variance

16.1 Comparing Means with an ANOVA F-Test

Example 16.1 – Classroom Seat Location and Grade Point Average
Is it true that the best students sit in the front of the classroom, or is that a false stereotype? In surveys done in 2 statistics classes at the University of California at Davis, students reported their grade point average and indicated where they typically sit in the classroom: front, middle, or back. The data is recorded in the file UCDavis seats.sav.

1. Steps to prepare the data:
   a. Open the data set UCDavis seats.sav.

2. Steps to check conditions numerically before conducting a one-way analysis of variance:
   a. To obtain summary measures for each of the three groups, we use the Analyze menu.
   b. Scroll down to the Compare Means submenu and select the Means option. A window opens up. The window is shown below:

   ![Means Window]

   c. Select the variable GPA and move it into the Dependent List box.
   d. Select the variable Seat and move it into the Independent List box.
   e. Click on the Options button. A window opens up.
f. The window is shown below:

![Means: Options window](image)

Select the summary measures you wish to calculate. In case of the ANOVA, we choose for the mean, the standard deviation, and the variance.

h. In the Options window, click Continue.

i. In the Means window, click OK.

The SPSS Output

**Descriptives**

<table>
<thead>
<tr>
<th>GPA</th>
<th>N</th>
<th>Mean</th>
<th>Std. Deviation</th>
<th>Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Front</td>
<td>88</td>
<td>3.2029</td>
<td>.54906</td>
<td>.301</td>
</tr>
<tr>
<td>Middle</td>
<td>218</td>
<td>2.9853</td>
<td>.55766</td>
<td>.311</td>
</tr>
<tr>
<td>Back</td>
<td>78</td>
<td>2.9194</td>
<td>.51046</td>
<td>.281</td>
</tr>
<tr>
<td>Total</td>
<td>384</td>
<td>3.0217</td>
<td>.55452</td>
<td>.307</td>
</tr>
</tbody>
</table>
3. **Steps** to check conditions graphically before conducting a one-way analysis of variance:

**Boxplots:**
- To construct boxplots, we will use the **Graphs** menu.
- Click on the **Chart Builder** option.
- In the **Gallery** at the bottom of the window, select the **Boxplot** option.
- Drag the first graph (**Simple Boxplot**) to the **Chart Preview** at the top of the window.
- Select the variable **GPA** and move it into the **Y-Axis?** box.
- Select the variable **Seat** and move it into the **X-Axis?** box.
- Click **OK**.

![The SPSS Output](image)

**Histograms:**
- To create a histogram, we use the **Graphs** menu.
- Click on the **Chart Builder** option.
- In the **Gallery** at the bottom of the window, select the **Histogram** option.
- Drag the first graph (**Simple Histogram**) to the **Chart Preview** at the top of the window.
- Select the variable **GPA** and move it into the **X-Axis?** box.
- To separate the graph between the three groups, click on the **Groups/Point ID** tab (the **Gallery** will disappear).
- Check the box by **Rows panel variable**.
- Select the variable **Seat** and move it into the **Panel?** box.
- To display the normal distribution curve in the graphs, check the box **Display normal curve** in the **Element Properties** window.
- In the **Element Properties** window, click **Apply**.
- In the **Chart Builder** window, click **OK**.
4. **Steps** to conduct a one-way analysis of variance:
   a. To conduct the one-way analysis of variance, we use the **Analyze** menu.
   b. Scroll down to the **Compare Means** submenu and select the **One-Way ANOVA** option. A window opens up. The window is shown below:
   
   ![One-Way ANOVA Window](image)
   
   c. Select the variable **GPA** and move it into the **Dependent List** box.
   d. Select the variable **Seat** and move it into the **Factor** box.
   e. Click **OK**.
The SPSS Output

Oneway

ANOVA

<table>
<thead>
<tr>
<th></th>
<th>Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>F</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between Groups</td>
<td>3.994</td>
<td>2</td>
<td>1.997</td>
<td>6.688</td>
<td>.001</td>
</tr>
<tr>
<td>Within Groups</td>
<td>113.775</td>
<td>381</td>
<td>.299</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>117.769</td>
<td>383</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

5. **Steps** to perform multiple comparisons, using Tukey’s or Bonferroni’s method:
   a. Rerun the one-way ANOVA as described in step 4 above.
   b. Before clicking **OK**, click on the **Post Hoc** button:

   ![Image of SPSS Output](image1)

   c. A window opens up. The window is show below:

   ![Image of SPSS Output](image2)

   d. Check the **Tukey** option and the **Bonferroni** option as shown above.
   e. Select a significance level at the bottom of the window. The default is 0.05.
   f. In the **Post Hoc** window, click **Continue**.
   g. In the **One-Way ANOVA** window, click **OK**.
   h. Note that the same ANOVA output will appear, in addition to the pairwise comparisons.
The SPSS Output

Oneway

ANOVA

<table>
<thead>
<tr>
<th>GPA</th>
<th>Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>F</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between Groups</td>
<td>3.994</td>
<td>2</td>
<td>1.997</td>
<td>6.688</td>
<td>.001</td>
</tr>
<tr>
<td>Within Groups</td>
<td>113.775</td>
<td>381</td>
<td>.299</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>117.769</td>
<td>383</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Post Hoc Tests

Multiple Comparisons

Dependent Variable: GPA

<table>
<thead>
<tr>
<th>(I) Seat</th>
<th>(J) Seat</th>
<th>Mean Difference (I-J)</th>
<th>Std. Error</th>
<th>Sig.</th>
<th>95% Confidence Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Lower Bound</td>
</tr>
<tr>
<td>Tukey HSD</td>
<td>Front</td>
<td>.21759</td>
<td>.06902</td>
<td>.005</td>
<td>-.3800</td>
</tr>
<tr>
<td></td>
<td>Middle</td>
<td>-2.1759</td>
<td>.06902</td>
<td>.005</td>
<td>-.3800</td>
</tr>
<tr>
<td></td>
<td>Back</td>
<td>.28349</td>
<td>.08498</td>
<td>.003</td>
<td>.0835</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-.1037</td>
</tr>
<tr>
<td>Middle</td>
<td>Front</td>
<td>-2.8349</td>
<td>.08498</td>
<td>.003</td>
<td>-4.834</td>
</tr>
<tr>
<td></td>
<td>Back</td>
<td>-.06591</td>
<td>.07210</td>
<td>.632</td>
<td>-.2355</td>
</tr>
<tr>
<td>Back</td>
<td>Front</td>
<td>-.28349</td>
<td>.08498</td>
<td>.003</td>
<td>-4.878</td>
</tr>
<tr>
<td></td>
<td>Middle</td>
<td>-.06591</td>
<td>.07210</td>
<td>.632</td>
<td>-.2393</td>
</tr>
</tbody>
</table>

Bonferroni

<table>
<thead>
<tr>
<th>(I) Seat</th>
<th>(J) Seat</th>
<th>Mean Difference (I-J)</th>
<th>Std. Error</th>
<th>Sig.</th>
<th>95% Confidence Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Lower Bound</td>
</tr>
<tr>
<td>Tukey HSD</td>
<td>Front</td>
<td>.21759</td>
<td>.06902</td>
<td>.005</td>
<td>-.3800</td>
</tr>
<tr>
<td></td>
<td>Middle</td>
<td>-2.1759</td>
<td>.06902</td>
<td>.005</td>
<td>-.3800</td>
</tr>
<tr>
<td></td>
<td>Back</td>
<td>.28349</td>
<td>.08498</td>
<td>.003</td>
<td>.0835</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-.1037</td>
</tr>
<tr>
<td>Middle</td>
<td>Front</td>
<td>-2.8349</td>
<td>.08498</td>
<td>.003</td>
<td>-4.834</td>
</tr>
<tr>
<td></td>
<td>Back</td>
<td>-.06591</td>
<td>.07210</td>
<td>.632</td>
<td>-.2355</td>
</tr>
<tr>
<td>Back</td>
<td>Front</td>
<td>-.28349</td>
<td>.08498</td>
<td>.003</td>
<td>-4.878</td>
</tr>
<tr>
<td></td>
<td>Middle</td>
<td>-.06591</td>
<td>.07210</td>
<td>.632</td>
<td>-.2393</td>
</tr>
</tbody>
</table>

* The mean difference is significant at the 0.05 level.

Homogeneous Subsets

<table>
<thead>
<tr>
<th>GPA</th>
<th>Subset for alpha = 0.05</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N</td>
</tr>
<tr>
<td>Tukey HSD*</td>
<td>Back</td>
</tr>
<tr>
<td></td>
<td>Middle</td>
</tr>
<tr>
<td></td>
<td>Front</td>
</tr>
<tr>
<td>Sig.</td>
<td></td>
</tr>
</tbody>
</table>

Means for groups in homogeneous subsets are displayed.

a. Uses Harmonic Mean Sample Size = 104.271.

b. The group sizes are unequal. The harmonic mean of the group sizes is used. Type I error levels are not guaranteed.
The Problem – Finding a *p*-value
Sometimes you may calculate an *F* test statistic by hand. How do we find the corresponding *p*-value?

1. **Steps** to prepare the data:
   a. Open a new data set.
   b. Use the **Variable View** tab to create a new variable, e.g. **Fstat**.
   c. Give the variable the appropriate measurement level (**Scale**).
   d. Use the **Data View** to enter the value of the *F* test statistic in the column named **Fstat**. For this exercise we will use the test statistic value from Example 16.1: 6.688 with 2 and 381 degrees of freedom.
   e. Save the file.

2. **Steps** to find the *p*-value:
   a. To find the *p*-value we use the **Transform** menu.
   b. Scroll down to the **Compute Variable** option. A window opens up. The window is shown below:
   
   ![Compute Variable Window]

   c. Type a variable name in the **Target Variable** box, such as **pvalue**.
   d. In the **Numeric Expression** box enter the number **1** and the minus sign (**−**).
   e. In the **Function Group** box select **CDF & Noncentral CDF**.
   f. In the **Functions and Special Variables** box select **Cdf.F** and move it in the **Numeric Expression** box by clicking on the up arrow.
   g. The three question marks need to be replaced by the value of your *F* test statistic and the appropriate degrees of freedom. Place the cursor in the location of the first question mark. Click on the variable **Fstat** and move it into the **Numeric Expression** box by clicking on the right arrow. Place the cursor on the location of the second and third question marks and type the degrees of freedom. For this exercise that is 2 and 381.
   h. Click **OK**.
   i. A new variable has now been added to the **Data Editor** called **pvalue**. It contains the *p*-value.
   j. To increase the number of significant digits, use the **Variable View** and change the **Decimals** to 3 or 4. For this exercise the *p*-value equals 0.00140:
Note:

- If you wish to find multiple $p$-values at once, all with different numbers of degrees of freedom, you may create a variable called `Fstat` for the values of the $F$ test statistics and two separate variables called `df1` and `df2` for the degrees of freedom. When calculating the $p$-values, replace the first question mark with the variable `Fstat`, the second question mark with the variable `df1` and the third question mark with the variable `df2`. All $p$-values will appear in the fourth column called `pvalue`. 
16.3 Other Methods for Comparing Populations

Example 16.11/12/13 – Drinks per Week and Seat Location
In the survey described in the previous example, students were also asked “How many alcoholic beverages do you consume each week?” Responses are to be compared in the same three groups: students who typically sit in the front, middle, and back of the classroom. The data is recorded in the file UCDavis seats.sav.

1. **Steps** to prepare the data:
   a. Open the data set UCDavis seats.sav.

2. **Steps** to conduct a Kruskal-Wallis test:
   a. To perform the test, we use the **Analyze** menu.
   b. Scroll down to the **Nonparametric Tests** submenu and select the **Independent Samples** option. A window opens up. The window is shown below:

   ![Nonparametric Tests window](image)

   c. Select the **Customize analysis** option.
d. Click on the **Fields** tab:

![Fields Tab](image)

- e. Select the **Use custom field assignments** option.
- f. Select the variable **alcohol** and move it into the **Test Fields** box.
- g. Select the variable **seat** and move it into the **Groups** box.
- h. Click on the **Settings** tab:

![Settings Tab](image)

- i. Select the **Customize tests** option and put a checkmark in front of the **Kruskal-Wallis test**.
- j. Click **Run**.
The SPSS Output

### Hypothesis Test Summary

<table>
<thead>
<tr>
<th>Null Hypothesis</th>
<th>Test</th>
<th>Sig.</th>
<th>Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>The distribution of Alcoholic Drinks per Week is the same across categories of Seat.</td>
<td>Independent-Samples Kruskal-Wallis Test</td>
<td>.000</td>
<td>Reject the null hypothesis.</td>
</tr>
</tbody>
</table>

Asymptotic significances are displayed. The significance level is .05.

**Note:**
- If you double click on the output, you will find some more information about the test and a graphical display of the observed data:

- If you double click on the pull down menu at the bottom of the window, you can select the **Pairwise Comparisons** option, to obtain the following information on the pairwise comparisons:
4. **Steps** to conduct Mood’s median test:
   a. To conduct Mood’s median test, we can repeat the steps previously described in Step 2.
   b. Deselect the **Kruskal-Wallis** option and select **Median** as the test type:
   c. Click **Run**.
The SPSS Output

Hypothesis Test Summary

<table>
<thead>
<tr>
<th>Null Hypothesis</th>
<th>Test</th>
<th>Sig.</th>
<th>Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>The medians of Alcoholic Drinks per Week are the same across categories of Seat.</td>
<td>Independent-Samples Median Test</td>
<td>.000</td>
<td>Reject the null hypothesis.</td>
</tr>
</tbody>
</table>

Asymptotic significances are displayed. The significance level is .05.

Note:
- If you double click on the output, you will find some more information about the test and a graphical display of the observed data:
- Via the pull down menu at the bottom of the window, you can once again obtain information on the pairwise comparisons.
16.4 Two-Way Analysis of Variance

Example 16.15 –Happy Faces and Restaurant Tips
When a restaurant server writes a friendly note or draws a “happy face” on your restaurant check, is this just a
friendly act, or is there a financial incentive? Temple University psychologists conducted a randomized experiment to
investigate whether drawing a happy face on the back of a restaurant bill increased the average tip given to the
server. One female server and one male server in a Philadelphia restaurant either did or did not draw a happy face on
checks during the experiment. Tips are to be compared between male and female server and between drawing or not
drawing the happy face on the bill. The data is recorded in the file happyface.sav.

1. Steps to prepare the data:
   a. Open the data set happyface.sav.

4. Steps to conduct a two-way analysis of variance:
   a. To conduct the two-way analysis of variance, we use the Analyze menu.
   b. Scroll down to the General Linear Model submenu and select the Univariate option. A window opens up.
   The window is shown below:

   ![Univariate Window](image)

   c. Select the variable tippct and move it into the Dependent Variable box.
   d. Select the variables message and sex and move them into the Fixed Factor(s) box.
e. To also obtain the interaction plot, click on the Plots button. A window opens up. It is shown below:

![Univariate: Profile Plots window]

f. To match the graph in the book, select the variable `message` and move it into the **Horizontal Axis** box.

g. Select the variable `sex` and move it into the **Separate Lines** box. Note that the two variables could also be reversed. Usually we put the variable with the most categories on the horizontal axis.

h. Click **Add**. The window will now look like this:

![Univariate: Profile Plots window with 'message' and 'sex' selected]

i. In the Profile Plots window, click **Continue**.

j. In the **Univariate** window, click **OK**.
The SPSS Output

**Univariate Analysis of Variance**

### Between-Subjects Factors

<table>
<thead>
<tr>
<th></th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>message HappyFace</td>
<td>45</td>
</tr>
<tr>
<td>None</td>
<td>44</td>
</tr>
<tr>
<td>sex Female</td>
<td>45</td>
</tr>
<tr>
<td>Male</td>
<td>44</td>
</tr>
</tbody>
</table>

### Tests of Between-Subjects Effects

Dependent Variable: tippct

<table>
<thead>
<tr>
<th>Source</th>
<th>Type III Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>F</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Corrected Model</td>
<td>3072.396a</td>
<td>3</td>
<td>1024.132</td>
<td>9.325</td>
<td>.000</td>
</tr>
<tr>
<td>Intercept</td>
<td>55554.126</td>
<td>1</td>
<td>55554.126</td>
<td>505.821</td>
<td>.000</td>
</tr>
<tr>
<td>message</td>
<td>14.719</td>
<td>1</td>
<td>14.719</td>
<td>.134</td>
<td>.715</td>
</tr>
<tr>
<td>sex</td>
<td>2602.009</td>
<td>1</td>
<td>2602.009</td>
<td>23.691</td>
<td>.000</td>
</tr>
<tr>
<td>message * sex</td>
<td>438.705</td>
<td>1</td>
<td>438.705</td>
<td>3.994</td>
<td>.049</td>
</tr>
<tr>
<td>Error</td>
<td>9335.526</td>
<td>85</td>
<td>109.830</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>67998.428</td>
<td>89</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>12407.922</td>
<td>88</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

a. R Squared = .248 (Adjusted R Squared = .221)

---

**Estimated Marginal Means of tippct**

- Sex: Female, Male

---

message: HappyFace, None