TABLE OF CONTENTS

LABORATORY SESSION 1 GETTING STARTED WITH SPSS 4
LABORATORY SESSION 2 DESCRIBING DATA I 13
LABORATORY SESSION 3 DESCRIBING DATA II 18
LABORATORY SESSION 4 A FIRST LOOK AT SOME SOCIOLOGICAL DATA 26
LABORATORY SESSION 5 EXPLORING RELATIONSHIPS 32
LABORATORY SESSION 6 REGRESSIONS 40
LABORATORY SESSION 7 MORE ON RELATIONSHIPS 48
LABORATORY SESSION 8 A FIRST LOOK AT EXPERIMENTAL DESIGN 54
LABORATORY SESSION 9 CENTRAL LIMIT THEOREM 57
LABORATORY SESSION 10 BUILDING CONFIDENCE IN CONFIDENCE INTERVALS 63
LABORATORY SESSION 11 CONCEPTS IN HYPOTHESIS TESTING 66
LABORATORY SESSION 12 TWO GROUP INFERENCES 68
LABORATORY SESSION 13 POPULATION MEANS 71
LABORATORY SESSION 14 COMPARING TWO POPULATION MEANS 77
APPENDIX 82
DESCRIPTION OF THE DATA SETS 82
SOCIOLOGY DATA SET 83
ECONOMICS DATA SETS 87
PSYCHOLOGY DATA SETS 89
INDEX 90
LABORATORY SESSION 1

GETTING STARTED WITH SPSS

Instructions:
1. Read the following introduction to SPSS for Windows before coming to lab.

Introduction to SPSS

SPSS is a widely used statistical package. Versions of it run on a wide variety of computer systems. The version you will be using operates in the Microsoft Windows environment. SPSS provides a wide variety of procedures for data analysis. It allows the user to enter, edit and manipulate data. The data analysis features include exploratory data analysis, basic statistics, regression analysis, analysis of variance, multivariate analysis, and nonparametric statistics. You will only use a small portion of these capabilities in this course. In addition, SPSS has a macro capability that permits users to program commands to carry out procedures that are not part of SPSS. In SPSS macros are called scripts. A number of scripts have been written for you in order to make some laboratory tasks easier.

![Figure 1—A Typical SPSS Data View Window](image-url)

The user interface for the SPSS system varies depending on the platform on which it is run. However, the essential features--the data view and commands--are the same for all environments. This means that a user who is familiar
with SPSS on, say, a Macintosh should encounter little difficulty using the program in Microsoft Windows. This introduction will deal exclusively with the Microsoft Windows version. You start SPSS by clicking on the SPSS for Windows icon in the SPSS for Windows folder on the Programs menu in the Start menu. When you start SPSS for the first time you will see a screen much like the one that is shown in Figure 1.

You will note that there are two windows within the main SPSS windows. The small window shows SPSS files that you have most recently worked with. Since you are new to SPSS, click on Cancel. Briefly, the windows have the following functions.

- Data View—This is where data are stored.
- Variable View—This is where variable names and other properties can be defined. The laboratory activities will walk you through what to do in Variable View.
- Output—Most results of SPSS operations are displayed in an output window that opens after completion of the computations.

The currently active window is indicated in the usual manner by having the colored title bar. You move from one window to another by clicking anywhere in the window you wish to make active.

You stop SPSS in the manner typical of most Windows programs. That is click the close window button in Windows or click the EXIT item in the FILE menu.

**DATA VIEW**

All data that SPSS will process are stored in the Data View. It consists of a number of columns that are the variables as discussed in Workshop Statistics. The rows represent observational units (called cases in SPSS). There is no limit to the number of variables and/or cases that can be used. An empty Data View looks like that shown in Figure 2.
Figure 2—Data View

VARIABLE VIEW

The Variable View lets you create variable names and define the attributes of each variable. A portion of an empty Variable View is shown in Figure 3. The entries are:

- **Name**—You can type in a name for each variable. Variable names can be no longer than 8 characters and can consist only of letters, numbers and the underbar character (_).
- **Type**—Clicking on this box will bring up an ellipsis (...). Clicking on the ellipsis will open a dialogue box that allows you to specify what type the variable has. For this course all variables will be of type *string* or *type numeric*.
- **Width**—The numerical entry in this box gives how many spaces the entries in the Data View will be for this variable.
- **Decimals**—For numeric data this entry gives how many decimal places will be shown for this variable in the Data View.
- **Label**—Here is where you can give a descriptive title for the variable. Always give a variable a label. This makes all output much more readable.
- **Values**—This is used for variables which are categorical. You can specify an English label for each of the numerical values of a categorical variable.
- **Missing**—This allows you to specify which values for a variable indicate missing data. By default, SPSS assigns a period for missing data. However, sometimes data sets you might receive use special numerical
values to indicate missing data.

- **Columns**—The numerical value in this item gives how many spaces will be allocated for the variable in the Data View. This is different from width in that width limits the number of spaces for the actual number. Columns limits how many spaces will be visible in the Data View.
- **Align**—This entry either left aligns, centers, or right aligns the entries for the variable.
- **Measure**—This indicates what type of variable a numerical variable is. The available types are *scale*, *ordinal*, and *nominal*. A *scale* variable is a quantitative variable. An *ordinal* variable is a categorical variable where the categories have a natural order to them, such as poor, fair, good, better, best. A *nominal* variable is a categorical variable where there is no natural order to the categories, such as male, female.

![Variable View](image)

**Figure 3**—Variable View

**SPSS COMMANDS**

SPSS commands to process data can be run in three different ways. You will use two of them in this course. They are:

- **The menu system.** The menus that appear near the top of the Data View (*File, Edit, View, Data, etc.*) contain entries that allow the user to process data, manipulate data, and create new variables. This is by far the method that you will use most often.
- **Writing SPSS syntax.** Every command in the menu system has beneath it SPSS code, called syntax, that it generates to do the requested operation. In the menu system you never see this syntax. SPSS has many capabilities and options that are not included in the menu system. To access them you must write SPSS syntax and then run it. This is more for more advanced users of SPSS, and we will discuss it no further.
• Using SPSS scripts. SPSS has a scripting language that is based on the Visual Basic programming language. It allows you to write scripts that do computations and operations that are not part of SPSS. A number of scripts have been written for use with both this laboratory manual and your text.

Using Help

SPSS has an extensive help system that explains how to use all of the built-in commands and explains other aspects of the system. You access it by clicking Help > Topics item in the menu bar. Doing this causes the window shown in Figure 4 to appear.

![Help Dialogue Box](image)

Figure 4—Help Dialogue Box

The most useful items are INDEX and SEARCH. INDEX contains a listing of SPSS commands. SEARCH allows you to search SPSS help for keywords. If you are trying to find if SPSS has a command for performing a certain task, you can enter its name in either INDEX or SEARCH and looking at the entries that it finds. You can either double click on the entry or click on the entry and then click the DISPLAY button at the bottom of the window.

EXAMPLE

Suppose we wish to determine what SPSS command will sum up the values in a variable. Possible command names might be ADD, SUM or TOTAL. Typing SUM in the INDEX results in a dialogue box looks like that shown in Figure 5.

![Dialogue Box](image)

SUM appears in a number of different commands. You can then check each one out by either double clicking on the item or clicking the item once and then click Display.
Typing SUM in FIND brings a much longer list of hits as shown below. Again, however, you can check each one out to see which one is the one you want.

In either case you will find that the sum of a variable is an option in Analyze > Descriptive Statistics > Descriptives. There are other commands that do it also.

**Entering and Saving Data**

There are two ways to enter data into the Data View.

**ENTERING DATA**

You enter data directly into the Data View by first making the Data View window active. Then click on the cell in the first row of the variable where you want to place the data. Then simply enter the data by typing a value and pressing the ENTER key. Notice that the first variable created gets the name `Var0001`. Move to the first row of the next variable to receive data and enter values there in the same manner. It gets the name `Var0002`. If, for some reason, you want add new values in a column somewhere other than at the bottom, you can cut and paste entries as you would in a Word document or an Excel spreadsheet.

**RETRIEVING PREVIOUSLY SAVED DATA**

The OPEN DATA command enters the contents of a previously saved Data View into the current Data View. If you do not give an extension to the filename SPSS will retrieve the file with the specified FILENAME and a default extension of .SAV. The OPEN DATA command is invoked by selecting File > Open > Data, or by making the Data View active and clicking the open file button on the button bar. This brings up the dialogue box shown in
Figure 6. You select the file to be retrieved by clicking on the name of the file in the file name box. You can then retrieve the file in one of two ways. The first is to double click on the file name. The second is to click on the file name and then click on the OPEN button.

SAVING DATA
Data may be saved using the SAVE or SAVE AS commands. SAVE saves data in a file with the same name as the active Data View. SAVE AS saves the data in a file with the name and destination that you choose. The resulting file will contain, in coded form, the contents of all non-empty variables. A data file created by the SAVE or SAVE AS command will write a file with a default extension .SAV. A data file is stored using the SAVE AS command as follows. We assume that this is the first time the data file is to be saved. With the Data View active select File > Save As. This will bring up a dialogue box similar to the one shown in Figure 7. Make sure that you have selected a drive that you are permitted to write to. Then, in the FILE NAME box enter the name you wish to give this file. Clicking Save will write your data to a file.
Figure 7—Dialogue Box for SAVE DATA FILE AS

KEEPING A RECORD OF AN SPSS SESSION
The results of SPSS computations are shown in an Output window. The Output window can be saved in a manner similar to that for saving data. Make the Output window active and select either File > Save or File > Save As. Then proceed in the same way as you do to save a Data View. The default extension for saving output is .SPO. You can also print a hard copy of the Output window by either making the Output window active and clicking the print button on the button bar or by selecting File > Print.

If you wish to print or save just a portion of the contents of an Output window, do the following. One at a time highlight each item you do not wish to keep and delete each one. Then print or save in the usual manner

Editing and Manipulating Data

The Windows environment makes editing data particularly easy. The contents of any rectangular block of cells may be declared missing by clicking and dragging to highlight the desired cells and then pressing the DELETE key. An entire variable may be deleted by simply clicking on the variable name and pressing the DELETE key. An entire case may be deleted by clicking on the case number on the left edge of the Data View and pressing the DELETE key.

Data may be copied from one block of cells to another in the following manner. Click and drag to highlight the block of cells to be copied. In the EDIT menu click on COPY item. This placed the contents on the Clipboard. Now click on the cell that is the upper left-hand corner of the destination block of cells. Then click on the upper left cell where the data are to be pasted and select the PASTE item on the EDIT menu to put the cells in the new location. Be warned that if data are present in the new location, SPSS will simply overwrite whatever is in the target location. This is like using the overwrite feature in a word processor.

The contents of any single cell may be changed at any time by simply clicking on the desired cell and typing in the new value.
Laboratory Activities

Submit a copy of your SPSS Output windows along with answers to all the questions asked.

1. Start SPSS in the Data View and enter the following names in a variable. It will be given the name Var0001.
   Choi, DiCaprio, Hsu, Maravi, M. Miller, W. Miller, Rinaman, Voorhees
2. Enter the following in a second variable. For this variable a 1 indicates that the case is female and a 2 indicates that the case is male. It will be given the name Var0002.
   1, 2, 1, 2, 2, 2, 2, 2
3. Enter the following in a third variable. It will be given the name Var0003.
   17, 20, 34, 11, 15, 33, 83, 65
4. Go to Variable View by clicking the tab in the lower left corner of the Data View. For Var0001 give it the name name. Make sure that it is of type string, and the measure is nominal. Enter a label name in the label box.
5. Var0002 is a binary variable with no order to the categories. Make sure that it is of type numeric and its measure is nominal. Enter gender in the label box. Give it zero decimal places by clicking on Decimal and either typing 0 in the box or clicking the down arrow that appears until 0 is in the box. Now to give the variable value labels. Click on the value box and then click the ellipsis that appears to bring up the dialogue box shown below. Enter 1 in the Value box, Female in the Value Label box and click Add. Enter 2 in the Value box, Male in the Value Label box and click Add. Finally click OK to complete the assignment of value labels.
6. Variable Var0003 contains quantitative data. Give it a name of result. Since the data are whole numbers, give them 0 decimal places. Enter Number picked in the label box. Finally, make sure the data are of type scale.
7. Save the data in a file named LAB1.SAV.
8. We wish to compute the sum of the values in the variable result for the men and the women separately. Look up Split File in SPSS Help to see what it does. Then search through the menus to find the Split File command and run it to split the file according to gender. Splitting causes separate analysis to be done for each distinct group as defined by the grouping variable. You want to organize output by groups with the groups defined by values in the variable gender. Now compute the sums as described in the SPSS introduction for this laboratory session.
9. Open the data file LABOR.SAV.
10. Use Help with the keyword histogram to find out how to create a histogram and use SPSS to draw a histogram of the variable lhur.
11. Print a copy of your complete Output window.
LABORATORY SESSION 2

DESCRIBING DATA I

Instructions:
1. Read the following description of SPSS commands before coming to lab.
2. Bring your statistics text to lab.

SPSS Commands

You will be working with two SPSS graphs this session. They are stemplots and histograms. We shall illustrate their use on data from the American Film Institutes listing of the top 100 movies of all time. These data are stored in the file 100FILMS.SAV. The file consists of four variables—rank is the American Film Institutes ranking of the film, title is the name of the film, year is the year when the film was made, and oscar is a binary variable indicating whether or not the film won the Oscar for best picture. An entry of 1 indicates that the film won the best picture Oscar, and an entry of 0 indicates that it did not.

STEMPLOTS

Stemplots are created as a part of the Explore command. You invoke the command by selecting Analyze > Descriptive Statistics > Explore to bring up the following dialogue box.

![Explore Dialogue Box](image)

Figure 1—Dialogue Box for Explore

The box labeled Dependent List contains the variable(s) to be analyzed. The box labeled Factor List contains the variable(s) that identify groups in the data. The box labeled Label Cases By allows you to use a variable to identify observational units. We will not use them in this course.

Variables are selected for the boxes mentioned above by clicking on them and then clicking the right pointing arrow adjacent to the box where it is to go. Select year for the Dependent List box, and select oscar for the Factor List box.

The Explore command can compute a number of summary statistics and draw a number of graphs. You will be learning about the summary statistics later in the course. To have SPSS draw graphs make sure that either the Both
or the *Plots* radio button is selected in the *Display* area.

Click on *Plots* to bring up the following dialogue box.

![Figure 2—Dialogue Box for Plots](image)

Make sure that *Stem-and-leaf* and *Histogram* are selected and click *Continue*.

Now click *OK*, and the summary statistics and the graphs will appear in an Output window. There will be two stemplots and two histograms—one showing the distribution of years for films that won the Oscar and one showing the distribution of year for films that did not win the Oscar.

One feature of these plots that is not covered by your text is the leftmost column. It shows the number of observations that are in each stem.

**HISTOGRAM**

You have already seen how histograms can be drawn in SPSS using the *Explore* command. In addition to that, there are two other ways that histograms can be produced in SPSS. Again, for the sake of illustration, we will assume that we want a histogram for the pictures that won the Oscar and a separate one for the films that did not.

**METHOD 1:**
One method uses the *Histogram* command that can be found in the *Graphs* menu. However, to use it to plot separate histograms you will need to split the file using the values in *oscar*. Recall that you learned how to split a file in Laboratory Session 1. Once you have split the file select *Graphs > Histogram* to bring up the dialogue box shown in Figure 3.

Select *year* in the *Variable* box and click *OK*, and the histograms will be drawn in an Output window.

**METHOD 2:**
One problem with the histograms you have drawn thus far is that it is a little difficult to compare them. The intervals plotted are not the same. You can, however use interactive histograms to correct this problem. There are two ways to do this.
The first method begins with the file being split. Since the file is already split select **Graphs > Interactive > Histogram** to bring up the dialogue box in Figure 4.

You will notice that **oscar** does not appear in the variables box on the left. This is because it is being used to split...
the file. To draw the histograms click and drag **year** to the box with the left pointing arrow through it as is shown in Figure 4 and click **OK** to draw the histograms. They will appear side by side with the same scales in an Output window.

For the second method unsplit the file by selecting **Data > Split Cases** and clicking **Analyze all cases, do not split groups**. Now select **Graphs > Interactive > Histogram**. Notice that **oscar** appears in the list of variables. Click and drag **year** to the box with the left pointing arrow if it is not already there and click and drag **oscar** to the **Panel Variables** box. Panel variables are used to define groups. Now click **OK** to draw the side by side histograms with the same axes in an Output window as before.
Laboratory Activities

Submit a copy of your SPSS Output windows along with answers to all the questions asked and explanations that are asked for.

1. Load the SPSS save file LAB2.SAV. This data set contains samples from five different populations. Variables pop1 through pop5 contain samples of 100 observations from the five populations. Variables pop6 through pop10 contain samples of 1000 observations from the same populations, not necessarily in the same order as variables pop1 through pop5.
2. Draw a histogram for each of the ten variables.
3. Describe as best you can the shape of each of the distributions of the data in variables pop6 through pop10 in terms of the six properties listed on pages 43 and 44 of your statistics text.
4. Use your histograms to match each sample of 100 from a population with its corresponding sample of 1000. That is, one of the samples in pop6 through pop10 is from the same population as the sample in pop1, one of the samples in pop6 through pop10 is from the same population as the sample in pop2, and so forth. Give reasons for making your decisions.
5. Construct stemplots for the data in pop1 through pop10. Compare the information provided with your histogram for the same column. What, if anything is different about the way SPSS produces stemplots for the samples of size 100 compared to those for samples of size 1000?
6. Save a copy of your results for use in Laboratory Session 3.
LABORATORY SESSION 3

DESCRIBING DATA II

Instructions:
1. Review what you did in Laboratory Session 2.
2. Read the command description before coming to lab.
3. Bring your results from Laboratory Session 2 to lab.

SPSS Commands

In this laboratory you will use numerical summaries to gain information about data. There are a number of ways to generate these statistics. The numerical summaries are broken down according to the type of data with which you are dealing. For categorical data Frequencies will generate frequency distributions. Since there are many more statistics that you can compute for numerical data there are correspondingly more ways to get SPSS to compute them. We shall focus on Explore and Descriptives. For expository purposes we shall use the data in 100FILMS.SAV discussed in Laboratory Session 2 to obtain frequency distribution for the variable oscar.

FREQUENCIES

Frequencies generates a frequency distribution, a relative frequency distribution, and a cumulative frequency distribution of data.

- A frequency distribution simply lists each category along with the number of observations in each.
- A relative frequency distribution lists each category along with the percentage of the observations in each.
- A cumulative relative frequency distribution is often useful for data where the categories have some inherent ordering to them. Student evaluation responses, where you grade a question on a scale from 1 to 5, have such an ordering. A cumulative relative frequency distribution lists the categories in increasing order and shows the percentage of observations that are in categories that are equal to or below each category.

![Figure 1—Dialogue Box for Frequencies](image)

---

18
You invoke **Frequencies** by selecting **Analyze > Descriptive Statistics > Frequencies** to bring up the dialogue box shown in Figure 1. The variables for which frequency distributions are to be created are entered in the **Variable(s)** box in the usual manner. The dialogue box to get frequency distributions for **oscar** should look like Figure 1. Clicking **OK** will create in an Output window the results shown in Figure 2. In addition to the frequency distributions mentioned, you can also obtain a number of graphs and statistics. We will not go into them here. Click the **Statistics** and **Charts** buttons to see what is available.

**EXPLORE**

We saw in Laboratory Session 2 that **Explore** can be used to create stemplots and histograms. In addition to them, **Explore** can compute a wide variety of numerical summaries of data for a quantitative variable. Of the statistics that we have discussed in class, or will discuss, **Explore** computes and displays the following.

- The number of non-missing values (N).
- The number of missing values (N*).
- The mean.
- The median.
- The standard deviation (StDev).
- The standard error of the mean (SEMean). This is the standard deviation divided by the square root of the number of non-missing values.
- The minimum observation (Min).
- The maximum observation (Max).
- The first quartile (Q1). The first quartile is the number such that 25% of the data are less than or equal to it. If you are familiar with percentiles, it is the 25th percentile.
- The third quartile (Q3). The third quartile is the number such that 75% of the data are less than or equal to it. It is the 75th percentile.
- The interquartile range(IQR)
- The range.

You run **Explore** by selecting **Analyze > Descriptive Statistics > Explore** to bring up the dialogue box shown in Figure 3. Since we will not be splitting the file in this example you should leave the **Factor List** box empty. To compute statistics on the year in which the picture was made select and place **year** in the **Dependent List** box. The dialogue box should look like the one shown in Figure 3. In order to get the quartiles you must click **Statistics** to get the dialogue box shown in Figure 4. Make sure that **Descriptives and Percentiles** are checked. Click **Continue** to return to the main dialogue box. Finally clicking **OK** will cause results like those shown in Figure 5 to appear in an Output window.
Figure 3—Dialogue Box for Explore

**Explore: Statistics**
- Descriptives
- Confidence Interval for Mean: 95%
- M-estimators
- Outliers
- Percentiles

Continue Cancel Help

Figure 4—Dialogue Box for Statistics

**Case Processing Summary**

<table>
<thead>
<tr>
<th>YEAR</th>
<th>Valid</th>
<th>Percent</th>
<th>Missing</th>
<th>Percent</th>
<th>Total</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>100</td>
<td>100.0%</td>
<td>0</td>
<td>.0%</td>
<td>100</td>
<td>100.0%</td>
</tr>
</tbody>
</table>
### Descriptives

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Statistic</th>
<th>Std. Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>YEAR Mean</td>
<td>1959.46</td>
<td>1.851</td>
</tr>
<tr>
<td>95% Confidence Interval for Mean Lower Bound</td>
<td>1955.79</td>
<td></td>
</tr>
<tr>
<td>95% Confidence Interval for Mean Upper Bound</td>
<td>1963.13</td>
<td></td>
</tr>
<tr>
<td>5% Trimmed Mean</td>
<td>1959.43</td>
<td></td>
</tr>
<tr>
<td>Median</td>
<td>1959.50</td>
<td></td>
</tr>
<tr>
<td>Variance</td>
<td>342.554</td>
<td></td>
</tr>
<tr>
<td>Std. Deviation</td>
<td>18.508</td>
<td></td>
</tr>
<tr>
<td>Minimum</td>
<td>1915</td>
<td></td>
</tr>
<tr>
<td>Maximum</td>
<td>1996</td>
<td></td>
</tr>
<tr>
<td>Range</td>
<td>81</td>
<td></td>
</tr>
<tr>
<td>Interquartile Range</td>
<td>31.50</td>
<td></td>
</tr>
<tr>
<td>Skewness</td>
<td>-.004</td>
<td>.241</td>
</tr>
<tr>
<td>Kurtosis</td>
<td>-.751</td>
<td>.478</td>
</tr>
</tbody>
</table>

### Descriptives

**Descriptives**

*Descriptives* computes a limited number of statistics for quantitative data. Select **Analyze > Descriptive Statistics > Descriptives** to bring up the dialogue box shown in Figure 6.

### Figures

**Figure 5**—Example of **Explore**

**DESCRIPTIVES**

*Descriptives* computes a limited number of statistics for quantitative data. Select **Analyze > Descriptive Statistics > Descriptives** to bring up the dialogue box shown in Figure 6.
You enter the variable(s) whose descriptive statistics you wish to compute in the Variable(s) box in the usual manner. To select the statistics to be computed and stored click the Options button to bring up the dialogue box shown in Figure 7.

![Descriptives: Options](image)

Figure 7—Dialogue Box for Options

Check the statistics you wish to compute and click Continue to return to the main dialogue box. Clicking OK will cause the requested statistics to be displayed in an Output window.

**BOXPLOT**

Your text introduces the idea of a box plot in Topic 5 on page 86. There you were asked to use the five number summary to construct one. The “box” is a rectangle that goes from the first quartile to the third quartile. A line is drawn in the box at the location of the median. Then “whiskers” are drawn from the ends of the box to the maximum and minimum observed values. This method is different from the way SPSS constructs a box plot. SPSS draws a boxplot for a single sample in Explore. Refer back to Figure 2 in Laboratory Session 2. You will notice that there are radio buttons for drawing boxplots. Checking either of the top two will cause a boxplot for the variable(s) in the Dependent List box to be drawn.

A second method uses **Graphs > Interactive > Boxplot**. Selecting it opens the dialogue box shown in Figure 8. Drag the variable to be plotted to the box with the vertical arrow and click OK to produce the plot.

There are three ways to draw side-be-side boxplots for a variable divided into groups. The first uses Explore. Here you select the group identifying variable and place it in the Factor List box.

The second method uses **Graphs > Boxplot**. Selecting it brings up the dialogue box shown in Figure 9. Click Simple and then Define to bring up the dialogue box shown in Figure 10. Place the variable to be plotted in the Variable box and the group identifying variable in the Category Axis box. To boxplots for the year movies were produced for the Oscar and non-Oscar group, the dialogue box should look like the one shown in Figure 10. Clicking OK will draw the boxplots in an Output window.

The third method uses the interactive boxplot. Select **Graphs > Interactive > Boxplot** to bring up the dialogue box shown in Figure 8. Place the variable to be plotted in the box with the vertical arrow and the group identifying
variable in the box with the right point arrow. Then click OK to cause the boxplots to be drawn in an Output window.

Figure 8—Dialogue Box for **Interactive Boxplot**

The boxplot that SPSS produces is different from the one you learn to draw in Topic 5. In Topic 6 you learn to modify the box plot in the following way. In this modified boxplot the whiskers extend from the end of the box out
to the observation that is closest to, but not more than, a distance of $1.5(Q_3 - Q_1)$ from the end of the box.

Observations that are more distant than this are termed outliers and are shown individually on the plot with an asterisk (*). SPSS adds an additional modification. The whiskers are drawn in the same manner as in the modified boxplot. Observations that are more than $1.5(Q_3 - Q_1)$ but less than $3(Q_3 - Q_1)$ from the ends of the box are designated as being mild outliers and are drawn using a circle. Observations that are more than $3(Q_3 - Q_1)$ from the ends of the box are designated as being extreme outliers and are drawn using in asterisk (*).

Figure 10—Dialogue Box for Define Simple Boxplot
Laboratory Activities

Submit a copy of your SPSS Output windows along with answers to all the questions asked and explanations that are asked for.

1. Load the SPSS data file stored in the file LAB2.SAV.
2. Compute the mean, median, standard deviation, and interquartile range for the data in pop1 through pop10.
3. Recall from Laboratory Session 2, Activity 6, which columns you paired up. Do you feel that the statistics computed from the samples of size 100 are more or less accurate reflections of the true population quantities than those computed from the samples of size 1000? Explain your reasoning.
4. Draw a boxplot for pop6 through pop10. Briefly describe how their respective boxplots reflect the properties of each distribution.

For Activities 5 through 7 use only the statistics computed on pop6 through pop10

5. Recall from Laboratory Session 2, Activity 5, which columns you considered to be skewed and which columns you considered to be symmetric. For the skewed distributions compare the value of the mean with that of the median. Similarly, for the symmetric distributions compare the value of the mean with that of the median. Explain briefly what relationship, if any, you see between the values of the mean and median and whether or not the distribution is symmetric.
6. In each case in Activity 5 make a determination of whether the mean or median is more representative of a “typical” value of a member of the population. Explain your reasoning. You might want to plot the values of the mean and median on the histograms you created in Laboratory Session 2.
7. Recall from Laboratory Session 2, Activity 3, which columns you considered to have a large number of outliers. Make an assessment of whether the standard deviation or the interquartile range is a better measure the amount of variability in each population. Explain your reasoning.
8. You have seen in class that the median is more resistant to the effects of outliers in the data than is the mean. Later in the course you will see that the mean, and not the median, is used to make inferences about the population mean. It is logical to ask why this is so. It turns out that, in many instances, the mean is superior to median. In this activity we will illustrate one of these properties. That is, when there are no outliers it turns out that the mean will, on average, be closer to the population mean than is the median. To show this we provide the SPSS script COMP_M.SBS. It takes 1000 samples of size 10 from a distribution that is symmetric and mound shaped like the one shown on page 43 of your text. The actual mean and median of the distribution is 0. The mean and median of each sample are computed. Finally, the standard deviation of these 1000 sample means is computed, and the standard deviation of the 1000 sample medians is also computed. The script then displays these two quantities. Open an empty SPSS Data View and create two empty variables, mean and median. Run the script and comment on what you see regarding the differences between the results for the mean and for the median. Recall that standard deviation is a measure of variability in data.
LABORATORY SESSION 4

A FIRST LOOK AT SOME SOCIOLOGICAL DATA

Instructions:
1. Review what you did in Laboratory Sessions 2 and 3.

SPSS Commands

In this laboratory you will use numerical, graphical and tabular summaries to gain information about some of the variables in the NSFH data set. The variables in this data set are described in the appendix to this manual.

We shall begin with a description of the new commands for this session. This will be followed by some examples using the NSFH data set.

RECODE

The Recode command is used to change the values of a variable into a new set of values. There are two ways to implement it. Which one you use will depend on whether or not you wish to overwrite an existing variable with the new values or whether you want to store the new values in a new variable. It is usually good practice to create a new variable. That way, if you want to use the original values, they will not be lost. For this reason we will discuss only how to change the values and place them in a new variable.

Select Transform > Recode > Into Different Variables to bring up the dialogue box shown in Figure 1. As an example, we again work with the data in 100FILMS.SAV. Assume that we want to change the values in oscar. The new variable is to be called won. Select oscar and place it in the Input Variable->Output Variable box. Next type won in the Name box in the Output Variable area and click on Change. The dialogue box should look like the one shown in Figure 2. Click on Old and New Values to bring up the dialogue box shown in Figure 3. Suppose we want to change an entry of 0 to 1 and an entry of 1 to 2. This is indicated by typing each old value in the Value box in the Old Value area followed by typing the value it is to be changed to in the Value box in the New Value area and then clicking Add. Any values in the original variable that you do not give a new value will declared to be missing data. When you have completed this for all of the original values the Old and New Values dialogue box should look
like the one shown in Figure 4. Clicking *Continue* followed by *OK* will create the new variable.

![Recode into Different Variables](image1.png)

Figure 2—Filled in Dialogue Box for **Recode**

![Recode into Different Variables: Old and New Values](image2.png)

Figure 3—Dialogue Box for **Old and New Values**

If you want to recode a numerical variable into a string variable, check *Output variables are strings* in the **Old and New Values** dialogue box.
SELECT CASES
Sometimes you want to analyze only a subset of the data. SPSS provides a very flexible command, Select Cases, to do this. Select Data > Select Cases to bring up the dialogue box shown in Figure 5.

From the dialogue box you can see that you have the following options for selecting cases.
- Select all cases.
- Select those cases that satisfy user specified conditions.
Select a random sample of cases.  
Select cases whose case numbers are in a specified range  
Use a filter variable. A filter variable is a variable containing only 0’s and 1’s. The cases where the filter variable is 1 will be selected for analysis.

Suppose, for example, we want to analyze the year in which a picture was made for only those films that won the Oscar. There are two ways to do this.

Method 1:  
Check Use filter variable and place oscar in the box with the right pointing arrow.  
If you want to permanently delete unselected cases check Deleted in the UnselectedCases Are area. Otherwise check Filtered.  
Click OK to select the cases.

Method 2:  
Check If condition is satisfied to bring up the dialogue box shown in Figure 6.

EXAMPLES

We will now give some examples of using these commands to manipulate the data in the NSFH data. These data are stored in the file NSFH.SAV. In your activities you will need to do things similar to what will be shown. Your activities will NOT BE EXACTLY like the examples. It will be necessary for you to use those aspects that are directly relevant to your tasks.

1. Selecting Cases Belonging to a Specific Category

There are times when you wish to analyze data to compare values belonging to different categories. For example, you might wish to investigate the distribution of incomes according to marital status. Many SPSS commands allow you to do this easily as part of the command dialogue box. We have seen this in the interactive histograms and boxplots and in Boxplot. Other commands, such as Histogram, do not have such an option. To overcome this we can use Data > Split Cases which we introduced in Laboratory Session 2. Suppose we wish to draw a histogram for the incomes of only those respondents that are married.
In NSFH.SAV incomes are stored in `irearn`. Marital status is in `m2cp01`. To do what we want all that is needed is to use **Select Cases** with an *If condition satisfied of* `m2cp01 = 1`

Now, running **Histogram** on `irearn` will produce a histogram of the incomes of the married respondents.

2. **Selecting Cases that Belong to More than One Group**

There will be occasions when you will want to separate data into two or more groups. As an example, suppose we want to analyze earnings and ethnic background for those respondents who are married and those respondents who are divorced. The relevant variables are `irearn`, `m484` and `m2cp01`. To do this we use **Select Cases** with an *If condition satisfied of* `m2cp01 = 1 | m2cp01 = 3`. The vertical bar stands for *or*. 
Laboratory Activities

Submit a copy of your SPSS Output windows along with answers to all the questions asked and explanations that are asked for.

1. Load the SPSS data stored in the file NSFH.SAV. Save a copy of this data file.
2. Look at the description of the 36 variables in this data file that is given in the appendix of this manual on page 113. List three variables that contain numerical data. List three variables that contain categorical data where the categories have an inherent order. List three variables that contain categorical data where the categories have no inherent order. Note that caseid, region, sample, and weight are not data. They contain bookkeeping values used by the survey team. Do not use them.
3. Describe as best you can the shape of the distributions for the columns you chose that contain numerical data. Use the methods you learned in Laboratory Session 2 and Laboratory Session 3. Be sure to address the shape features listed on pages 43 and 44 of your text.
4. For the columns you chose that contain numerical data, compute the most appropriate measure of the center and the most appropriate measure of spread. Briefly explain your reasons for choosing the measure you did.
5. Draw a histogram of incomes for each sex, male and female. Discuss any differences you see in the shape of the two distributions. Do the data seem to indicate that men earn more than women?
6. Consider the variable E401. Create a new variable that has the following meaning. A value of 1 denotes that the husband wanted the divorce but the wife did not. A value of 2 denotes that the husband wanted the divorce more than the wife did. A value of 3 denotes that the husband and wife wanted the divorce equally. A value of 4 denotes that the wife wanted the divorce more than the husband did. A value of 5 denotes that the wife wanted the divorce but the husband did not. In other words the new variable will contain the responses made by male respondents the same as in E401, and the responses made by female respondents in reverse order from those in E401. To do this you can use Transform > Compute with the following Numeric Expression:

   \[ e401 * (m2dp01 = 1) + (6 - e401) * (m2dp01 = 2) \]

   Explain why this works. Name the target variable whatever you wish
7. Use this new variable to answer the following question. Does it appear that husbands are more likely than wives to want a divorce? Briefly explain your reasoning.
LABORATORY SESSION 5

EXPLORING RELATIONSHIPS

Instructions:
1. Read the discussion of SPSS commands before coming to lab.

SPSS Commands

In class you should have been exposed to the notion of constructing scatter plots and calculating correlation coefficients. In this laboratory you shall use these tools to analyze relationships in a variety of data.

SCATTERPLOT

The Scatterplot command can be used to generate scatter plots. It is found in the Graphs menu. Suppose you want a scatter plot a car’s highway miles per gallon as the response variable and its weight as the explanatory variable for the data in CARS99.SAV. Clicking on the Scatterplot item brings up the dialogue box shown in Figure 1.

![Figure 1—Dialogue Box for Scatterplot](image)

Check Simple and then click Define to bring up the dialogue box shown in Figure 2. Select hwy_mpg for the Y Axis and weight for the X Axis. The dialogue box should look like the one shown in Figure 2. Then click OK and the requested scatter plot will be drawn in an Output window. Each automobile is shown as a circle in the plot.

INTERACTIVE SCATTER PLOT

Suppose we wanted a scatter plot of highway miles per gallon versus weight but you wanted to use different symbols for each type of car. This can be done using an interactive scatter plot. Select Graphs > Interactive > Scatterplot to bring up the dialogue box shown in Figure 3. Drag the response variable (hwy_mpg) to the box with the vertical arrow through it. Drag the explanatory variable (weight) to the box with the right pointing arrow through it. Now to generate a plot with different symbols for each type of car drag the group identifying variable (type) to the Style box in the Legend Variables region. Clicking OK will cause the desired scatter plot to be drawn in an Output window. Next to the plot will be a legend showing which type of car goes with which symbol. The one we just drew is shown in Figure 4.

CORRELATE

The correlation coefficient is calculated in SPSS with the Correlate command. It is called by selecting Analyze > Correlate > Bivariate to bring up the dialogue box shown in Figure 5. Enter the two variables whose correlation is to be computed in the Variables box. In Figure 5 we are asking SPSS to compute the correlation between highway miles per gallon and weight. Make sure that Pearson is checked in the Correlation Coefficients area. The correlation coefficient you learned about in Topic 9 of the text is often referred to as the Pearson Product Moment Correlation. Click OK and the output shown in Figure 6 will appear in an Output window. The output shows that
the correlation between highway miles per gallon and itself is 1, the correlation between weight and itself is 1, and the correlation between weight and highway miles per gallon is -0.799. The number opposite \( N \) is the sample size. The value for \( \text{Sig} \) will not be covered in this course.

![Simple Scatterplot Dialogue Box](image)

Figure 2—Dialogue Box for *Simple Scatterplot*

If you enter more than two variables in the *Variables* box, then SPSS will compute the correlation between all possible pairs of variables. The output will also be a two way table with the correlation between a pair of variables being shown in the cell whose row is one variable and whose column is the other variable.
Figure 3—Dialogue Box for Interactive Scatterplot
Figure 4—Example of Interactive Scatterplot

Figure 5—Dialogue Box for Bivariate Correlations
Correlations

<table>
<thead>
<tr>
<th></th>
<th>Highway MPG</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pearson Correlation</td>
<td>1</td>
<td>-.799**</td>
</tr>
<tr>
<td>Sig. (2-tailed)</td>
<td>.000</td>
<td>N</td>
</tr>
<tr>
<td>N</td>
<td>106</td>
<td>106</td>
</tr>
<tr>
<td>Weight Pearson Correlation</td>
<td>-.799**</td>
<td>1</td>
</tr>
<tr>
<td>Sig. (2-tailed)</td>
<td>.000</td>
<td>N</td>
</tr>
<tr>
<td>N</td>
<td>106</td>
<td>109</td>
</tr>
</tbody>
</table>

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

Figure 5—Sample Output for Bivariate Correlations

COMPUTE

The Compute command allows you to perform arithmetic operations on variables to create new variables. You used it in class in Activity 2-3 to compute the Fan Cost Index and in Activity 2-4 to create a variable that assigned a 1 to a state where more than 25% of the students took SAT and a 0 if not. In this laboratory session we look at how to use Compute with other functions. You access the command by selecting Transform > Compute to bring up the dialogue box shown in Figure 6. This brings up the dialogue box shown in Figure 6.

The transformed values are stored in the variable designated in the Target Variable box. The equation that defines the transformation is entered in the Numeric Expression box. The buttons under the EXPRESSION box allow you to enter numbers, decimal points, arithmetic operations and relational operators. Most are what you would expect. Possible exceptions are that the asterisk (*) signifies multiplication and the double asterisks (**) are for exponentiation. The box on the right lists the functions that are available. They are in alphabetic order and are selected in the usual manner. Highlight the desired function and then click the up pointing arrow just above the Functions box. It is also possible to simply type your expression in the Numeric Expression box, but be careful to avoid making typographical errors.
As an example, we shall use **Compute** to compute and store the square root of the weights of the cars in CARS99.SAV. The target variable is **sqrtwt**. Find and select the square root function—**SQRT**. When you do this note that the name of the function is followed by (?). Replace the question mark with the variable or expression whose square root is to be taken. Figure 6 shows what the dialogue box should look like to compute the square root of the weights. Clicking on **OK** will perform the desired computation.

**Merging SPSS Data files**

Suppose you have two data files and that you wish to use variables from both of them in a single analysis. If you open one of the files and then open the second one, the result will be that only the second file will be in SPSS. SPSS, however, provides the option of opening one file and then merging the second with it. Suppose we wish to merge the data in MANUF.SAV with the data in LABOR.SAV. This is done by first opening one the two data files containing the columns to be merged and then selecting **Data > Merge files > Add Variables** to bring up the dialogue box shown in Figure 7.

![Add Variables: Read File](image)

**Figure 7—Dialogue Box for Merge files**

Double click LABOR to open the dialogue box shown in Figure 8. Since LABOR.SAV has two variables named **month** and **year** as does MANUF.SAV, they are excluded from the merge. If there are other variables that you do not want to be included in the combines data file, select them and click the left pointing arrow to enter them in the **Excluded Variables** box. Once you have the variables you want in the merged data file, click **OK** to perform the merge.
Figure 8—Dialogue Box for **Add Variables**
Laboratory Activities

Submit a copy of your SPSS Output windows along with answers to all the questions asked and explanations that are asked for.

1. Refer to the descriptions of the economic time series data sets given in the appendix. Find the data sets that contain the variables that correspond to unemployment and the prime interest rate. Merge these two data files. Create a scatter plot to investigate the relationship, if any, between unemployment and the prime interest rate. Write a description of what you observe. If there is a relationship, do you feel that one of these variables can be considered to be the “cause” of the other? If so, why? If not, why not?

2. Look at the description of the NSFH data. Find two numerical variables that you believe could be related to each other. Use scatter plots and correlations to determine whether your guess was correct. Explain your reasoning as to why you felt the two variables you chose could be related.

3. The SPSS data file HOUSTON.SAV contains census data for the city of Houston, Texas. Verify that a scatter plot shows that there is a relationship between time and population, but the relationship is not linear. The shape of the relationship looks somewhat like a parabola. If this is correct, then a scatter plot of the square root of population versus year should be a straight line. Use the square root function in Transform > Compute to compute and store Sqrt(pop) in a new variable. You now have the square root of population in the new column. Does there appear to be a linear relationship between time and the square root of population?

4. Other models of population growth assert that populations grow exponentially over time. This means that a formula like \( \text{Pop} = a \cdot b^{\text{year}} \) will fit the observed population growth. In this equation, \( a \) and \( b \) are constants. The accuracy of this model can be verified by comparing year with the logarithm of population to see if these two variables show a linear relationship. Use the LG10 function in Transform > Compute to compute and store LG10(pop) in a new variable. You now have the logarithm of population in the new column. Does there appear to be a linear relationship between time and the logarithm of population? Is it fairly strong or relatively weak?

5. Does the square root transformation or the logarithmic transformation do the better job?

6. Save a copy of your data file and output file from in this session for use in Laboratory Session 6.
LABORATORY SESSION 6

REGRESSIONS

Instructions:
1. Bring a copy of your results from Laboratory Session 5 to lab.
2. Read the discussion of the SPSS commands before coming to lab.

SPSS Commands

In this session you will be using the SPSS regression command. The command does a lot more than we have been discussing so far in this course.

REGRESSION

The Regression command performs least squares regressions. It uses the least squares criterion to determine the straight line that best fits the data. As an example we will use the data in CARS99.SAV to use an automobile’s weight to predict its highway mileage. To run it, select Analyze > Regression > Linear to bring up the dialogue box shown in Figure 1. Recall that the regression equation is of the form \( \hat{y} = a + bx \). Here, \( a \) is the y-intercept, and \( b \) is the slope. SPSS refers to the response (\( \hat{y} \)) variable as the dependent variable, and the explanatory variable (\( x \)) as the independent variable. In the dialogue box you select the response variable in the Dependent box. The explanatory variable goes in the Independent(s) box. Underneath the Independent(s) box there are a number of options that can be chosen. The only ones that might be of interest in this course are listed under Save. Clicking on that button opens the dialogue box shown in Figure 2. To store the fits (\( \hat{y} \)) check Unstandardized in the Predicted Values area. To store the residuals (\( y_i - \hat{y}_i \)) check Unstandardized in the Residuals area. Then click Continue.

When you run the regression the fits and residuals will be stored in new variables in the Data View. The fits will appear in a variable named pre_n, and the residuals will be stored in a variable named res_n. For the first regression \( n \) will be 1, for the second regression \( n \) will be 2, and so on. The dialogue box in Figure 1 will compute a regression equation to use automobile weight to predict highway mileage. Clicking on OK generates the following output in an Output window.

<table>
<thead>
<tr>
<th>Variables Entered/Removed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
</tr>
<tr>
<td>1</td>
</tr>
</tbody>
</table>

\( ^{\text{a}} \) All requested variables entered.

\( ^{\text{b}} \) Dependent Variable: HWY_MPG

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>.799 ( ^{\text{a}} )</td>
<td>.639</td>
<td>.636</td>
<td>1.991</td>
</tr>
</tbody>
</table>

\( ^{\text{a}} \) Predictors: (Constant), WEIGHT
This output contains much more information than you have been exposed to thus far in the course. Most of it will not be covered until the second semester. In terms of what has been covered in your text, the output gives the regression equation and the coefficient of determination, $r^2$. SPSS lists it as $R^2$. You probably noticed what appeared to be two different versions of $r^2$ with two different values. The one labeled $R$-sq is the coefficient of determination discussed in your text. The one labeled Adjusted $R$ Square is called the adjusted coefficient of determination. Its main purpose in life is to help compare different regression models with each other. For this course we will only consider the unadjusted value of $r^2$.

The ANOVA section will be left to the second course. You should note, however, that if you divide the regression sum of squares by the total sum of squares, you can obtain the coefficient of determination. Also the square root of the mean square for error is the same as the value shown in Std. Error of the Estimate.

In the Coefficients section, the only information that is relevant to this course are the values stored in the $B$ column of the Unstandardized Coefficients area. The value in the row labeled (Constant) is the value of the $y$-intercept, $a$. The value in the row labeled with the name of the explanatory variable is the slope coefficient, $b$. It should be noted that if the explanatory variable has an entry in the Label in the Variable view, then that label, rather than the variable name will be displayed in the output.
Figure 1—Dialogue Box for the REGRESS Command
Figure 2—Dialogue Box for Save
FITTED LINE PLOT

Once you have obtained a regression line, a valuable diagnostic tool is to draw this line in your scatter plot. SPSS allows you to do this using an interactive scatter plot. Select **Graphs > Interactive > Scatterplot** to bring up the dialogue box shown in Figure 3.

![Dialogue Box for Interactive Scatterplot](image)

Drag the response variable to the box with the vertical arrow, and drag the explanatory variable to the box with the right pointing arrow as shown in Figure 3. Click the *Fit* tab to bring up the dialogue box shown in Figure 4. In the *Method* area click the arrow on the right side of the box and select *Regression* as shown. This will cause the regression equation to be computed and drawn in the scatterplot along with the equation and the coefficient of determination. Clicking *OK* will draw the fitted line plot in an Output window as shown in Figure 5. If you do not like the location of the regression equation you can double click the graph to open the interactive graph editor. Now you can drag the regression equation and the coefficient of determination to the desired location. Clicking outside the editor window will close it.
Figure 4—Fit in Interactive Scatterplot
Linear Regression

\[ \text{hwy\_mpg} = 45.69 + (-0.01 \times \text{weight}) \]

R-Square = 0.64

Figure 5—Example of a Fitted Line Plot
Laboratory Activities

Submit a copy of your SPSS Output windows along with answers to all the questions asked and explanations that are asked for.

1. Refer to Activities 1 and 2 of Laboratory Session 5. For each situation where a straight-line relationship appeared reasonable do the following.
   a) Compute the regression line.
   b) Use your line to predict the value of $y$ for two values of $x$ that are not in the data.
   c) Assess how well your regression line describes the original data. Explain your reasoning.

2. Refer to Activities 3 through 5 of Laboratory Session 5. Either LG10(‘POP’) versus ‘Year’ or Sqrt(‘POP’) versus ‘Year’ displayed a better straight-line scatter plot. For the one that did, compute the regression line between these two variables. If you chose the square root of population, you would estimate the population in a given year as follows.
   a) Use your regression equation to predict the square root of the population.
   b) Square the value you obtained in Part (a) to get the estimate of population.

   If you chose the logarithm of the population, you would estimate the population in a given year as follows.
   a) Use your regression equation to predict the logarithm of the population.
   b) Raise 10 to the power of the value you obtained in Part (a) to get the estimate of population.

   Use your equation to predict the population of Houston in 2000 and 2020. Do you think this model does a good job of predicting future population? Explain your reasoning.

3. (Extra Credit) Thus far we have been using a single variable to predict the value of another. This process is referred to as simple regression. Many times you may wish to predict the value of a variable based on the values of two or more other variables. Using more than one predictor variable is called multiple regression. The idea is that different variables provide different information about the value of the response variable. The combined information can be used to make more accurate predictions. You perform a multiple regression in SPSS by specifying more than one variable in the Independent(s) box of the Analyze > Regression > Linear dialogue box. Scatter plots of each predictor variable plotted against the response variable are generally used to assess whether each predictor has a straight-line relationship with the response variable. It only accepts a single predictor variable. In this activity you are to investigate the ability to predict the Consumer Price Index (punew) by using the Index of Production (ip) and Cash, Checkable, Savings and Time Deposits (fm2) and Unemployment Rate (lhur). Note that these are in different data files. This means that you will have to merge them. Once they are merged we suggest you delete all variables except for the four you will be using. Not all of the variables have the same number of cases. Use individual scatter plots of punew with each of the other three variables to determine which of these show a reasonably strong linear relationship with Consumer Price Index. Only two of the three will show roughly linear scatter plots when compared with punew. Use those variables that show a linear relationship as predictors of Consumer Price Index. One way to assess your regression is to evaluate $r^2$. The interpretation is identical to that for simple regression. You can also draw a scatterplot of the fitted $y$ values versus the actual $y$ values. It should be roughly a straight line for a good fit. Save the fits and draw such a scatterplot. What do your plot and $r^2$ tell you about the quality of this regression?
LABORATORY SESSION 7
MORE ON RELATIONSHIPS

Instructions:
1. Read the description of the SPSS commands before coming to lab.

SPSS Commands

In this session you will be using the SPSS to create tables and segmented bar graphs. As an example we shall be trying to evaluate the relationship between a respondent’s level of education (compled) and his/her perception of health (e207) in the NSFH data.

CROSSTABS

The SPSS *Crosstabs* command generates two-way tables of categorical data. Select *Analyze > Descriptive Statistics > Crosstabs* to bring up the dialogue box shown in Figure 1.

![Figure 1—Dialogue Box for the TABLE Command](image)

Select *compled* for the row variable and *e207* for the column variable. Many times you also want to compute row and column percentages. To do this click *Cells* to bring up the dialogue box shown in Figure 2. Check the percentages that you wish to have computed and displayed in your table. The dialogue box shown displays and observed counts, the row percentages, the column percentages, and the percent each cell is of this entire table. Click *Continue* to return to main dialogue box and then click *OK* to generate the table. You will notice that SPSS prints the contents of the variable label. If you think this is too hard to read—which it is—you can edit the contents of the table. Double click in the table area to open an editor. Notice that the border of the table area has slanted lines. Now click on the label you want to change. Note that all entries that are the same become highlighted. Now simply
change the contents to whatever you want. We changed the row and column percent labels to read simply Row percent and Column percent. The final result is what is shown in Figure 3.

![Crosstabs: Cell Display](image)

**Figure 2—Dialogue Box for Cells**

<table>
<thead>
<tr>
<th>Respondent’s level of education completed</th>
<th>Compared with other people your age, how would you describe your health?</th>
<th>Crosstabs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-formal schooling</td>
<td>Very poor</td>
<td>Poor</td>
</tr>
<tr>
<td>Some elementary school</td>
<td>Count</td>
<td>25</td>
</tr>
<tr>
<td>Row percent</td>
<td>25.0%</td>
<td>37.5%</td>
</tr>
<tr>
<td>Column percent</td>
<td>2.9%</td>
<td>9%</td>
</tr>
<tr>
<td>% of Total</td>
<td>1%</td>
<td>3%</td>
</tr>
<tr>
<td>Some high school</td>
<td>Count</td>
<td>6</td>
</tr>
<tr>
<td>Row percent</td>
<td>6.3%</td>
<td>14.1%</td>
</tr>
<tr>
<td>Column percent</td>
<td>2.4%</td>
<td>5.3%</td>
</tr>
<tr>
<td>% of Total</td>
<td>6%</td>
<td>1.4%</td>
</tr>
<tr>
<td>High school diploma or GED</td>
<td>Count</td>
<td>1</td>
</tr>
<tr>
<td>Row percent</td>
<td>1.6%</td>
<td>25.0%</td>
</tr>
<tr>
<td>Column percent</td>
<td>2.4%</td>
<td>35.7%</td>
</tr>
<tr>
<td>% of Total</td>
<td>1%</td>
<td>1%</td>
</tr>
<tr>
<td>Some college</td>
<td>Count</td>
<td>4</td>
</tr>
<tr>
<td>Row percent</td>
<td>1.0%</td>
<td>5.3%</td>
</tr>
<tr>
<td>Column percent</td>
<td>1.1%</td>
<td>1.6%</td>
</tr>
<tr>
<td>% of Total</td>
<td>2%</td>
<td>3%</td>
</tr>
<tr>
<td>Bachelor’s degree</td>
<td>Count</td>
<td>2</td>
</tr>
<tr>
<td>Row percent</td>
<td>3%</td>
<td>5%</td>
</tr>
<tr>
<td>Column percent</td>
<td>5.7%</td>
<td>14.3%</td>
</tr>
<tr>
<td>% of Total</td>
<td>1%</td>
<td>1%</td>
</tr>
<tr>
<td>Some graduate study</td>
<td>Count</td>
<td>4</td>
</tr>
<tr>
<td>Row percent</td>
<td>1.0%</td>
<td>5.3%</td>
</tr>
<tr>
<td>Column percent</td>
<td>1.1%</td>
<td>1.6%</td>
</tr>
<tr>
<td>% of Total</td>
<td>2%</td>
<td>3%</td>
</tr>
<tr>
<td>Total</td>
<td>Count</td>
<td>35</td>
</tr>
<tr>
<td>Row percent</td>
<td>1.9%</td>
<td>3.8%</td>
</tr>
<tr>
<td>Column percent</td>
<td>15.6%</td>
<td>20.0%</td>
</tr>
<tr>
<td>% of Total</td>
<td>1.9%</td>
<td>3.8%</td>
</tr>
</tbody>
</table>

**Figure 3—Output for Crosstabs**

We now return to the dialogue box in Figure 1. The Statistics button allows you to compute a number of correlation like statistics for the table. These are not covered in this course, and we will discuss them no further.

**Constructing Marginal Distribution Bar Charts**

Suppose we want to draw a bar chart for completed. Bar charts are created in SPSS by selecting Graphs > Bar to bring up the dialogue box shown in Figure 4. Check Simple and click Define to bring up the dialogue box shown in Figure 5.
Constructing Segmented Bar Charts

Suppose we want a segmented bar chart to graphically display the relationship, if any, between \textit{compled} and \textit{e207}. Segmented bar charts can be drawn by using an interactive bar chart. Select \textbf{Graphs} $>$ \textbf{Interactive} $>$ \textbf{Bar} to bring up the dialogue box shown in Figure 6. Drag the explanatory variable, \textit{compled}, to the box with the right pointing arrow. Segments can be drawn by either using different colors or different patterns. We will use different patterns.
Drag the response variable, e207, to the box labeled Style in the Legend Variables area. To get stacks select Stack in the button to the right of the Style box. To get the bars to go to 100% make sure that 100% stacked is checked. Then clicking OK will draw the segmented bar chart shown in Figure 7. In this case we edited the legend box to fit the page.
Compared with other people your age, how would you describe your health?

- Very poor
- Poor
- Fair
- Good
- Excellent

Bars show counts

**Respondent's level of education completed**

Figure 7—Segmented Bar Chart
Laboratory Activities

Submit a copy of your SPSS Output windows along with answers to all the questions asked and explanations that are asked for.

1. Load the NSFH.SAV data file. You may need to free up some space in the data file in order to store new values you will create in the lab. Delete all columns except for m2cp01, m486 and m492a. Save this altered data file.

2. Analyze and discuss the relationship, if any, between religion (m486) and marital status (m2cp01). Sometimes the segmented bar charts that SPSS creates can be difficult to read. If that turns out to be the case here, then use Crosstabs and analyze the table you create. Use religion as the explanatory variable and marital status as the response variable.

3. Recode the variable on how often a person attends religious services, M492A, as follows. Let 1 = never, 2 = once, twice or three times a month, 3 = at least four times a month. Use Transform > Compute with a Numeric Expression of (m492a >= 0) + (m492a = 1) + (m492a = 2) + (m492a = 3) + (m492a >= 4)*2. Explain why this works. The values stored in m492a are the number of religious services attended per month. For the rest of the lab we shall refer to the recoded m492a as religiosity.

4. Analyze and discuss how the relationship you saw in Activity 2 is affected when you control for religiosity. By the term control we means that you want to see if the pattern of a relationship changes when you look at the relation between the same two variables when you separate the sample according the categories of a third variable known as the controlling variable. There are two ways to do this. You can create a two-way table using Analyze > Descriptive Statistics > Crosstabs making religion the column variable and marital status the row variable and select your recoded m486 in the Layer 1 of 1 box. This will generate separate tables for each different value of your recoded m486. The second way draws separate segmented bar charts for each level of religiosity. You can split the file according to your recoded m486 and then construct a segmented bar chart using religion as the explanatory variable and marital status as the response variable. Another way is to not split the file, but make your recoded m486 a Panel Variable in the interactive bar chart dialogue box. Determine if the form of the relationship between religion and marital status appears to change with a change in religiosity. Describe what you see.
LABORATORY SESSION 8

A FIRST LOOK AT

EXPERIMENTAL DESIGN

Instructions:
1. Read the description of the SPSS commands before coming to lab.

The preceding laboratory sessions have presented some SPSS tools for data analysis. You have been given the opportunity to apply them to various types of data from Sociology and Economics. The types of data you have been working with come from what are termed observational studies. This means that the investigator has a population with certain properties and a random sample is taken from that population.

A different type of study is the experimental design. Here, the investigator typically wishes to determine if there is a difference in the response of an observational unit when a unit is subjected to one set of conditions versus when a unit is subjected to a different set of conditions. A classic example of this type of study is the Salk Polio Vaccine trials in the 1950’s. A vaccine had been developed to prevent polio, and it was necessary to determine if it was effective. To do this, a so-called double blind experiment was conducted. Children were randomly separated into two groups. One group was given the vaccine, and the second group was given a placebo. No child knew to which group he or she belonged. This is a blind experiment. To make it a double blind experiment, the people who would examine the children to determine the effectiveness of the vaccine also did not know the group membership of any child that he or she examined. This was to prevent inadvertent bias on the part of the investigators. Such studies are now what the FDA requires for proper testing of new drugs.

The main point of the preceding description is that, in an experimental design, the investigator creates group membership by controlling whether or not a subject gets one treatment or another. These kinds of experiments are common in Psychology and Biology.

In this laboratory session we shall introduce two common ways of designing experiments for comparing two groups. One design, sometimes called a between groups design, is what was done in the Salk vaccine trials. Namely, a group of subjects is selected. Each member of the sample is then randomly assigned to one of the two groups. One group is then given one of the treatments, and the other group is given the other treatment. By the term treatment statisticians mean that a specified set of conditions is established before a measurement is taken. For example, a subject is given either the vaccine or the placebo. Receiving the vaccine would be one treatment, and receiving the placebo would be another treatment.

A second design is often referred to as a within subjects design. Another term is paired comparisons. Here each member of the group of subjects is given both treatments. Such a design is useful when the magnitude of the difference between the responses of each group is likely to be less than the variability among subjects. For example, suppose we wished to determine whether there is a difference in grip strength in a person’s dominant hand versus the non-dominant hand using a group of 50 people. Suppose we selected 25 people and measured the grip strength in the dominant hand and measured the grip strength in the non-dominant hand for the remaining 25. Overall strength varies considerably from person to person. Thus, the difference in grip strength between the dominant and non-dominant hand, if any, would be much less than the difference in overall grip strength between the subjects. Therefore, in order to isolate the difference, a within subjects design would be appropriate.

We will now describe two experimental designs from Psychology. Some members of the class may already be familiar with them from other courses. We will be doing some different things with the data than was done there.
A Between Groups Experiment

It is desired to determine if one method for memorizing items is more effective than another. One method is to repeat the item to be memorized as many times as possible (repetition). The other method is to create a mental picture (visualization) of the item to be memorized. In the experiment, 54 people were asked to memorize word pairs (e.g., DOG—CAT). Repetition was to be used by 27 of the subjects and the other 27 subjects used visualization. A total of 12 word pairs were to be memorized. Each subject was then shown one member of each pair and asked to give the other member. The data from a run of this experiment are stored in the SPSS data file MEMORY.SAV.

A Within Subjects Experiment

The question in this experiment is whether people respond more quickly to a visual stimulus or an audio stimulus. In the experiment, subjects were exposed to a stimulus and asked to push a button as soon as possible after noting the stimulus. The time between onset of the stimulus and pressing the button was measured. In the experiment, 58 subjects had their reaction times measured using the visual stimulus. The same 58 subjects also had their reaction times measured using the audio stimulus. The order in which this was done varied from subject to subject. The data from a run of this experiment are stored in the SPSS data file RESPONSE.SAV. A within subjects design is appropriate here due to the variability between subjects in overall reaction speed.
Laboratory Activities

Submit a copy of your SPSS Output windows along with answers to all the questions asked and explanations that are asked for.

1. Open the SPSS data file MEMORY.SAV.
2. Describe the shape of the distribution of the sample from the repetition group. Use the techniques covered in Laboratory Sessions 2 and 3.
3. Describe the shape of the distribution of the sample from the visualization group. Again, use the techniques covered in Laboratory Sessions 2 and 3.
4. Compare box plots for the groups. In particular, do the box plots indicate that there is a difference between the memorization methods?
5. Recall how you controlled for religiosity in Activity 4 of Laboratory Session 7. Now, separate the memorization results under each method according to sex. Then, answer the following question. Is there a difference in the memorization results when you control for sex?
6. Open the SPSS data file RESPONSE.SAV. Compute the difference between the visual and auditory response of each subject and store the values in an unused column. Remember that you can use Transform > Compute to do this easily.
7. Describe the shape of the distribution of these differences. Again, use the techniques covered in Laboratory Sessions 2 and 3.
8. Do these differences indicate that the type of stimulus matters? Explain your reasoning.
LABORATORY SESSION 9

CENTRAL LIMIT THEOREM

Instructions:
1. Read the description of the SPSS commands before coming to lab.
2. Review Topic 18 in your text.

This session will investigate the effect embodied in the Central Limit Theorem. Your text describes one version of this theorem that is applicable to sample proportions. It is introduced in Topic 16 in your statistics text. The theorem states that, if the sample size, \( n \), is large enough, then the distribution of the sample proportion, \( \hat{p} \), will be approximately normal. The distribution will have a mean of \( \theta \), the population proportion, and a standard deviation of \( \sqrt{\theta(1-\theta)/n} \). You were given the rule of thumb that the approximation is valid when sample sizes are such that \( n\theta \geq 10 \) and \( n(1-\theta) \geq 10 \).

There is also a version of the Central Limit Theorem that applies to sample means. It is introduced in Topic 17, and it states that, if the sample size, \( n \), is large enough, then the distribution of the sample mean, \( \bar{X} \), will be approximately normal. The distribution will have a mean of \( \mu \), the population mean, and a standard deviation of \( \sigma/\sqrt{n} \), the population standard deviation divided by the square root of the sample size. In this case, the rule of thumb that is usually given is that the approximation is valid if the sample size is 30 or more.

In this laboratory session we shall investigate these rules of thumb. In particular, we will be interested in seeing how large a sample is really necessary for the Central Limit Theorem to be useful. Sometimes a sample of size 30 is more than adequate, and sometimes much larger samples are needed.

SPSS Commands

Q-QPLOT

The Q-Q plot command is a useful diagnostic tool for investigating if sample data comes from a population that is normally distributed. Suppose we want to see if the data in the variable uniform in the data file LAB9.SAV follows a normal distribution. Select and place uniform in the Variables box. Make sure that Normal is selected in the Test Distribution area. Then click OK to have SPSS draw the plots shown in Figure 2 and Figure 3 in an Output window.

If the data follow a normal distribution, then they should follow the straight line that is drawn in the plot shown in Figure 2. The plot shown in Figure 3 should be a random scattering about the horizontal line that is drawn if the distribution is normal. The plots shown do not support the idea that distribution of the population is approximately normal.
Figure 1—Dialogue Box for Q-Q Plots
SAMPLE

The script SAMPLE.SBS has been written for this laboratory session. It draws a user specified number of samples of a size specified by the user from data stored in a variable chosen by the user. It computes and stores the sample mean for each sample in the variable mean. It then draws normal Q-Q plots of these sample means. It has a dialogue box shown in Figure 4.
Figure 4—Dialogue Box for Sample

Click the variable you want to sample, enter the number of samples and the size of each sample as shown in Figure 4. Click OK to run the script. The dialogue box shown above will take 100 samples of 10 each from the data stored in the variable uniform.
Laboratory Activities

Submit a copy of your SPSS Output windows along with answers to all the questions asked and explanations that are asked for.

1. Load the LAB9.SAV data file. This data file consists of four columns, each of which contains 2000 observations. The variable uniform contains observations from a population that is symmetric but not normal. The variable normal contains observations from a normal population. The variable exzpon contains observations from a population that is skewed to the right and, hence, not normal. The variable lognorm contains observations from a population that is very badly skewed to the right. Draw histograms of each of the variables to see what the shape of each distribution is.

2. For the data in each column try to determine how large a sample is needed for the sample means to have an approximate normal distribution. Use the script SAMPLE.SBS. We suggest that you begin with sample sizes of 10 and work your way up, as necessary, until an approximate straight line is obtained on your Q-Q plots. By approximate, we mean precisely that. The means you only need to have a graph that is roughly a straight line. For the columns containing the skewed data, we recommend that you use fairly large increments when increasing the sample size. This will save time.

3. What conclusions can you draw about a relationship between the shape of the distribution of the population and how large a sample is needed for the sample means to be normally distributed?

4. The rule of thumb for the Central Limit Theorem for sample proportions is that the theorem is valid if the sample size, n, is such that both quantities \( n\theta \) and \( n\left(1-\theta\right) \) are greater than or equal to 10. According to this rule of thumb what size sample is needed when \( \theta = 0.5 \)? What sample size would be needed when \( \theta = 0.2 \)? What sample size would be needed when \( \theta = 0.9 \)?

5. We wish to show how this new rule of thumb reflects the dependency on the skewness of the population. To do this, we shall draw the distributions of populations with \( \theta = 0.2 \), \( \theta = 0.5 \) and \( \theta = 0.9 \). Open the data file BINOMIAL.SAV. The entries in the variable x are the number of successes. The entries in the rows of the variables pop_2, pop_5, and pop_9 contain the probabilities of observing the number of successes, out of 20, shown in x. By plotting the entries in the variables pop_2, pop_5, and pop_9 you can get a plot that shows the shape of the population. Use the Graphs > Line command. Indicate that you want a Simple plot and click Define. In the Define dialogue box place the data to be plotted in the Category Axis box. The dialogue box shown below will draw a line plot for the pop_2 variable.
Draw a line plot for each of the variables. Now, based on your observations from Activity 3 and Activity 5 write a brief explanation how this rule of thumb reflects the connection between skewness and the size of sample needed for the Central Limit Theorem to be appropriate.
LABORATORY SESSION 10

BUILDING CONFIDENCE IN CONFIDENCE INTERVALS

Instructions:
1. Read the description of the SPSS commands before coming to lab.

This session will investigate the ideas underlying confidence. SPSS does not currently have the capability for computing confidence intervals for a population proportion. In that case, we provide a script that does this.

SPSS Commands

CIPROP

The script CIPROP.SBS computes a confidence interval for a population proportion using summary data. When you run the script the dialogue box shown in Figure 1 opens.

![Figure 1—Dialogue Box for CIPROP](image)

You enter either the sample proportion or the number of successes in your sample in the box labeled Phat or X. You enter the size of the sample in the box labeled N. You enter the desired confidence level, in percent, in the box labeled Confidence Level. The dialogue box shown in Figure 1 will construct a 95% confidence interval for a sample of 50 that contains 16 successes. Clicking OK will construct the confidence interval using the formula you learned about in Topic 19 and prints the results in an Output window. Figure 2 shows the output for the data entered in the dialogue box shown in Figure 1. It shows that the 95% confidence interval for $\theta$ runs from .1907 to .4493.

<table>
<thead>
<tr>
<th>N</th>
<th>X</th>
<th>Phat</th>
<th>Lower</th>
<th>Upper</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>16</td>
<td>.3200</td>
<td>.1907</td>
<td>.4493</td>
</tr>
</tbody>
</table>

![Figure 2—Sample Output for CIPROP](image)
Laboratory Activities

Submit a copy of your SPSS Output windows along with answers to all the questions asked and explanations that are asked for.

1. Start SPSS and load the data file LAB10.SAV.
2. The script CONFSIM.SBS demonstrates the idea behind confidence level. Begin by opening an empty Data View and create an empty variable with the name `phat`. The script has the following dialogue box.

![Dialogue box for simulate confidence intervals]

You enter the number of samples desired, the sample size, the actual value of θ and the confidence level, in percent. The dialogue box shown will take 200 samples of 25 each from a population with an actual proportion of 0.5 and compute 95% confidence intervals. Run the script three times, each time indicating that you want to take 200 samples of 25 each from a population with an actual proportion of 0.5 for confidence levels of 90%, 95% and 99%, respectively. Briefly explain what connection, if any, you see between the confidence level and the proportion returned by the macro. What did you expect to see? Discuss any difference between what occurred and what you expected.

3. Open the data file LAB10.SAV. It shows the results of 50 flips of a fair coin. A 0 is a tail, and a 1 is a head. The script CIDEMO.SBS further demonstrates the idea behind confidence level. Run Analyze > Descriptive Statistics > Frequencies to get either the number of successes or the sample proportion. Use CIPROP.SBS to compute a 95% confidence interval for the data in the variable `x`. Using this interval, consider the following situation. Suppose you were to take another 100 samples of the same size using the same coin and compute the proportion for each one. Answer the following question. How many of these proportions do you anticipate
will fall in the confidence interval you just computed? Now run CIDEMO.SBS. It has the following dialogue box.

Enter the number of additional samples to be taken, the lower confidence limit you obtained using CIPROP.SBS, and the upper confidence limit you obtained using CIPROP.SBS. Then click OK to run the script. It will take the additional samples, compute the sample proportion for each one and then show how many of these sample proportions fall in your confidence interval in an Output window. Compare your anticipated value with that returned by the macro. Discuss what you think is going on here.

4. Load the data file NSFH.SAV.

5. Compute a 95% confidence interval for the proportion of the respondents that are Roman Catholic. Use Analyze > Descriptive Statistics > Frequencies to get the sample size and the proportion of respondents that are Roman Catholic. Then run CIPROP.SBS to get the confidence interval. Does this sample support a claim that 25% of Americans are Roman Catholic? Explain your reasoning.
LABORATORY SESSION 11

CONCEPTS IN HYPOTHESIS TESTING

Instructions:
1. Read the description of the SPSS commands before coming to lab.

SPSS Commands

TESTPROP

SPSS does not have a built-in command to test hypotheses regarding a population proportion. We provide a script, TESTPROP.SBS that runs the tests discussed in Topic 21 of your text. Running the script brings up the dialogue box shown in Figure 1.

Enter either the number of successes or the sample proportion in the box labeled \( \text{Phat or X} \). Enter the sample size in the box labeled \( N \). Enter the value of \( \theta_0 \) in the box labeled \( \text{Proportion} \). Select the appropriate alternative hypothesis and then click \( \text{OK} \) to run the script. The dialogue box shown in Figure 1 will test \( H_0 : \theta = 0.6 \) versus \( H_a : \theta > 0.6 \) for a sample of 50 containing 41 successes. The results will be printed in an Output window. The Output window for the test shown in Figure 1 is shown in Figure 2.

![Figure 1—Dialogue Box for Testprop](image)

![Figure 2—Sample Output for Testprop](image)
Laboratory Activities

Submit a copy of your SPSS Output windows along with answers to all the questions asked and explanations that are asked for.

1. Start SPSS.
2. The script TDEMO.SBS demonstrates the meaning of the significance level of a test. The dialogue box looks like the following.

Enter the number of samples in the box labeled Number of Samples, the sample size in the box labeled N, and the desired significance level as a proportion (i.e. between 0 and 1) in the box labeled Significance Level. The script will take the requested number of samples of N observations from a population having a true proportion of 0.8. Each sample is then used to test the null hypothesis that $\theta = 0.8$ against the alternative hypothesis that $\theta \neq 0.8$. It returns the proportion the time that the test results are significant at the requested significance level. Open a new Data View and create an empty variable sig. Run TDEMO.SBS three times using 200 samples of 100 each at significance levels of 10%, 5% and 1%, respectively. Discuss what connection, if any, you see between the significance level and the proportion returned by TDEMO.SBS. What were you expecting to see? Discuss any differences between what you expected and what actually occurred.

3. The purpose of this activity is to investigate connections between tests and confidence intervals. Open the data file LAB10.SAV. Use CIPROP.SBS a 95% confidence interval for the data in variable x. Now use TESTPROP.SBA on the data in the variable x repeatedly using different values of $\theta$. Use an alternative hypothesis of $\theta \neq \theta_0$ in each case. Try to determine the range of all values of $\theta_0$ for which the null hypothesis would be accepted using a significance level of 5%. That is, determine, as best you can, the smallest value of $\theta_0$ for which the p-value is greater than or equal to 0.05 and also the largest value of $\theta_0$ for which the p-value is greater than or equal to 0.05. Recall that the null hypothesis is accepted at a given significance level if the p-value is greater than the significance level. Repeat the above process using a 90% confidence interval and a 10% significance level. Do you see any connection between the confidence intervals and the range of values for $\theta_0$? If so, briefly discuss what you think it is.

4. Open a new Data View and create an empty variable sig. Run TDEMO.SBS using a sample size of 5 and significance levels of 10%, 5% and 1%. Discuss what you see here compared to what you saw in Activity 2. What do you think is going on?

5. Load the data file NSFH.SAV.

6. Test the hypothesis that 20% of the population is either black or Hispanic. Use Analyze > Descriptive Statistics > Frequencies to obtain either the proportion of respondents who are black or Hispanics or the number of respondents who are black or Hispanic and use TESTPROP.SBS to run the text.. Explain what you find.
LABORATORY SESSION 12
TWO GROUP INference

Instructions:
1. Read the discussion of the SPSS commands before coming to lab.

This session deals with inference for comparing the proportions of two independent groups. You will get the opportunity to test hypotheses in more complex situations.

SPSS Commands

CI2PROP

SPSS does not provide built-in commands to construct confidence intervals for and test hypotheses regarding the difference between two population proportions. Therefore, we provide the SPSS script CI2PROP.SBS to construct confidence intervals for the difference between two population proportions. Running the script opens the dialogue box shown in Figure 1.

![Figure 1—Dialogue Box for CI2PROP](image)

Enter the sample proportion or the number of successes from group 1 in the Phat 1 or X 1 box, the sample size from group 1 in the N1 box, the sample proportion or the number of successes from group 2 in the Phat 2 or X 2 box, the sample size from group 2 in the N2 box, and the desired confidence level, in percent, in the Confidence Level box. Click OK to run the script. The dialogue box shown in Figure 1 will create the output shown in Figure 2.

TEST2PROP

In order to conduct tests of hypotheses regarding the difference between two population proportions, we provide the SPSS script TEST2PROP.SBS. Running the script opens the dialogue box shown in Figure 3.
Enter the sample proportion or the number of successes from group 1 in the \textit{Phat 1 or X 1} box, the sample size from group 1 in the \textit{N1} box, the sample proportion or the number of successes from group 2 in the \textit{Phat 2 or X 2} box, the sample size from group 2 in the \textit{N2} box, and select the desired alternative hypothesis in the \textit{Alternative} box. Click \textit{OK} to run the script. The dialogue box shown in Figure 3 will create the output shown in Figure 4.
Laboratory Activities

Submit a copy of your SPSS Output windows along with answers to all the questions asked and explanations that are asked for.

1. Start SPSS and load the data file NSFH.SAV.
2. This activity will walk you through a problem to give an indication of the processes involved. The question is whether blacks and Hispanics are more likely to have children and not be married than whites who are not married. By “not married” we mean people who are either divorced, widowed, or have never been married. The relevant variables are **m204**—the number of biological children, **m484**—race or ethnicity, and **m2cp01**—marital status. We need two new variables. One variable will contain 1’s for all the whites, blacks and Hispanics who are not married and have biological children, and 0’s for everyone else. The second variable will contain 1’s for all the blacks and Hispanics, 2’s for all whites, and missing values for everyone else. Proceed as follows.
   a) Use Transform > Compute to create a variable containing 1’s for all the whites, blacks and Hispanics who are not married and have biological children, 2’s for all whites, blacks and Hispanics who are not married and do not have biological children, and 0’s for everyone else. Use a Numeric Expression of $0 + ((m2cp01 > 2) & (m204 > 0) & (m484 < 4)) + ((m2cp01 > 2) & (m204 = 0) & (m484 < 4)) * 2$. Explain why this works.
   b) In Variable View indicate that 0 is a missing value for your new variable. Create value labels to make output readable.
   c) Use Transform > Recode to create a new variable **race** containing 1’s for all the blacks and Hispanics, 2’s for all whites, and missing values for everyone else. To do this recode **m484** so that 1 and 3 become 1, 2 becomes 2, and do not recode the others. Create value labels to make the output readable.
   d) Split the file using groups according to the values in **race** and run Analyze > Descriptive Statistics > Frequencies to get the information necessary to run TEST2PROP.SBS.
   e) Run TEST2PROP.SBS to test whether the proportion of unmarried blacks and Hispanics who have biological children is greater than the proportion of unmarried whites that have biological children. What do you conclude?
   f) Run CI2PROP on these summary statistics. Does your confidence interval confirm your conclusion? Explain your reasoning.
3. With Activity 2 as background, determine if the proportion of people who divorce and cohabit before marriage is lower than the proportion of people who divorce and do not cohabit before marriage. Briefly describe how you manipulated the original data in NSFH.SAV to get variables to answer this question. Explain your conclusion.
LABORATORY SESSION 13

POPULATION MEANS

Instructions:
1. Read the discussion of the SPSS commands before coming to lab.

This session looks at making inferences regarding population means. SPSS provides four commands for this. Two of them are confidence interval commands and two of them are hypothesis testing commands.

SPSS Commands

SPSS has two commands that construct confidence intervals for a population mean. Both are, in some way, based on the normal distribution.

EXPLORE

When the sample mean can be assumed to have a normal distribution, the formula for a confidence interval for the mean is given by

\[ \bar{X} \pm t^* \frac{s}{\sqrt{n}}, \]

where \( \bar{X} \) is the sample mean, \( s \) is the true population standard deviation, \( n \) is the sample size and \( t^* \) is the critical value from the \( t \) tables. The sample mean will be, at least approximately, normally distributed if the population has a normal distribution or the sample size is greater than or equal to 30. If the sample size is less than 30, you can use Q-Q PLOT to check if a sample appears to have come from a normal population. Recall that Q-Q PLOT was described in Laboratory Session 9. Select Analyze > Descriptive Statistics > Explore to bring up the dialogue box shown in Figure 1. Enter the variable whose confidence interval you want in the Dependent List box. Make sure that either Both or Statistics is checked in the Display area. Click Statistics to bring up the dialogue box shown in Figure 2. Enter the desired confidence level, in percent, in the box labeled Confidence Interval for Mean and click Continue. Finally clicking OK will display the desired results in an Output window.

EXAMPLE

We shall construct a 95% confidence interval for the mean year in which the American Film Institute’s top 100 films was made. The data are in the file 100FILMS.SAV. The dialogue boxes shown in Figure 1 and Figure 2 will do this. Clicking OK in the dialogue box shown in Figure 1 will generate the output shown in Figure 3. The confidence interval is (1955.79, 1963.13).
Figure 1—Dialogue Box for **Explore**

**Explore: Statistics**

- Descriptives
- Confidence Interval for Mean: 95%
- M estimators
- Outliers
- Percentiles

Figure 2—Dialogue Box for **Statistics**
Descriptives

<table>
<thead>
<tr>
<th>Year</th>
<th>Statistic</th>
<th>Std. Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>1959.46</td>
<td>1.851</td>
</tr>
<tr>
<td>95% Confidence Interval for Mean Lower Bound</td>
<td>1955.79</td>
<td></td>
</tr>
<tr>
<td>Upper Bound</td>
<td>1963.13</td>
<td></td>
</tr>
<tr>
<td>5% Trimmed Mean</td>
<td>1959.43</td>
<td></td>
</tr>
<tr>
<td>Median</td>
<td>1959.50</td>
<td></td>
</tr>
<tr>
<td>Variance</td>
<td>342.554</td>
<td></td>
</tr>
<tr>
<td>Std. Deviation</td>
<td>18.508</td>
<td></td>
</tr>
<tr>
<td>Minimum</td>
<td>1915</td>
<td></td>
</tr>
<tr>
<td>Maximum</td>
<td>1996</td>
<td></td>
</tr>
<tr>
<td>Range</td>
<td>81</td>
<td></td>
</tr>
<tr>
<td>Interquartile Range</td>
<td>31.50</td>
<td></td>
</tr>
<tr>
<td>Skewness</td>
<td>-.004</td>
<td>.241</td>
</tr>
<tr>
<td>Kurtosis</td>
<td>-.751</td>
<td>.478</td>
</tr>
</tbody>
</table>

Figure 3—Sample Output for Explore

ONE SAMPLE T TEST

The second command is One Sample T Test. It constructs a confidence interval for the population mean and also conducts a test of $H_0 : \mu = \mu_0$ versus $H_a : \mu \neq \mu_0$. Select Analyze > Compare Means > One-Sample T Test to bring up the dialogue box shown in Figure 4. Select the variable to be analyzed in the Test Variable(s) box. Enter the desired value for $\mu_0$ in the Test Value box. Click Options to bring up the dialogue box shown in Figure 5. Enter the desired confidence level, in percent in the box labeled Confidence Interval. Click Continue and then click OK to conduct the test and construct the confidence interval. The results will be shown in an Output window.

Figure 4—Dialogue Box for One Sample T Test
EXAMPLE
We shall construct a 95% confidence interval for the mean year in which the films in 100FILMS.SAV were made and test the hypothesis that the mean year is 1960 against the alternative hypothesis that the mean year is not 1960. The dialogue boxes shown in Figure 4 and Figure 5 do this. The resulting output is shown in Figure 6. It shows that the $p$-value for the test is .771 indicating that there is no evidence against the null hypothesis. The confidence interval results are for the difference between the sample mean and the value of $\mu_0$ used in your test. To get a confidence interval for the population mean add $\mu_0$ (in this case 1960) to the two confidence limits to get a 95% confidence interval of (1955.79, 1963.13) as before. Consult the SPSS Companion Manual for your text to see how to use SPSS output to conduct one-sided tests.

Example for One-Sample Test

PAIRED SAMPLES T TEST

You can conduct the matched pairs $t$ test on data by using Transform > Compute to calculate the differences between each pair and then use Analyze > Compare Means > One Sample T Test to conduct the test. However, SPSS provides a built-in command for the test. The data must be organized such that members of each pair appear in separate variables. Select Analyze > Compare Means > Paired Samples T Test to bring up the dialogue box shown in Figure 7. Select the two variables whose difference is to be computed for the test in the Paired Variables box. You click them both before clicking the right pointing arrow. Be aware that SPSS form differences by subtracting the variable that is higher alphabetically from the one that is lower. By default SPSS will construct a 95% confidence interval for the mean difference. If you want a different confidence level click Options to bring up the dialogue box shown in Figure 8. Enter the desired confidence level, in percent, in the Confidence Interval box. Then click Continue. Clicking OK will run the test and construct the confidence interval, and the results will be displayed in an Output Window.

Example

To illustrate we will test the hypothesis that the mean age for husbands at the time of marriage is different than that for wives and construct a 95% confidence interval for this mean difference. The data are stored in MARRIAGE.SAV. The entries in the dialogue boxes shown in Figure 7 and Figure 8 are set up to do this. Clicking OK produces the output shown in Figure 9. Again, if you need to conduct a one-sided test consult the
SPSS Companion manual for directions on how to interpret the output in this context. In this case if you wanted to test if husbands are older the $p$-value would be 0.0345.

Figure 7—Dialogue Box for **Paired Samples T Test**

Figure 8—Dialogue Box for **Options**

Figure 9—Sample Output for **Paired Samples T Test**
Laboratory Activities

Submit a copy of your SPSS Output windows along with answers to all the questions asked and explanations that are asked for.

1. Start SPSS and load the data file RESPONSE.SAV.
2. Recall that the experiment was to determine if response times are different when subjects are given a visual cue versus an audible cue. Each subject was asked to react to both cues. The reaction time for each was recorded. Theory suggests that reaction to the audible cue should be quicker than to the visual one. Answer the following questions.
   a) Is Paired Sample T Test appropriate for testing the differences you just computed? That is, do we need that the differences be normally distributed, and if so, are they? If it is not appropriate, stop here and go on to Activity 3.
   b) Test on the null hypothesis that there is no difference in mean reaction time. The appropriate alternative will depend on how you computed your differences.
   c) What conclusions do you draw from this test? Explain your reasoning.
   d) Construct a 99% confidence for the mean difference in reaction time.
   e) What conclusions can you draw from your confidence interval? Explain your reasoning.
3. Load the SPSS data file PRICES.SAV
4. The question here is whether or not there is an inflationary bias in the producer price index since 1960. The values of PW for 1960 and before are in rows 1 through 180. Run Data > Select Cases to select the data for 1960 and before. Compute the mean for these selected cases. This will be the mean for your null hypothesis. Now select cases for 1961 and beyond. Conduct a t test of whether the mean producer price index for the period 1961 and beyond is greater than the mean producer price index for 1960 and before. What do you conclude? Explain your reasoning.
5. Load the SPSS data file NSFH.SAV.
6. Extract those respondents who are divorced. Do the data indicate that the average divorced person attends religious services less than once a month? Explain how you reach your conclusion. The relevant variables are m2cp01 and m492A.
LABORATORY SESSION 14

COMPARING TWO POPULATION MEANS

Instructions:
1. Read the discussion of the SPSS commands before coming to lab.

This session looks at making inferences for comparing two population means. SPSS provides two commands for this. Both make inferences about the means of the groups. For the purposes of this discussion we assume that group 1 is generically labeled as \( X \), and the second group is generically labeled as \( Y \). When testing whether or not two means are equal, there are two forms for the \( t \)-test. Both are based on the normal distribution. Throughout the following discussion we assume that the two groups are generically labeled \( X \) and \( Y \). The relevant sample statistics are as shown in the following table.

<table>
<thead>
<tr>
<th>Sample Size</th>
<th>( X )</th>
<th>( Y )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample Mean</td>
<td>( \bar{X} )</td>
<td>( \bar{Y} )</td>
</tr>
<tr>
<td>Standard Deviation</td>
<td>( s_X )</td>
<td>( s_Y )</td>
</tr>
</tbody>
</table>

One case is based on the assumption that the two population standard deviations are equal, and the other does not require that this be the case.

When the population standard deviations are not necessarily equal, the appropriate version of the \( t \)-test is the following.

\[
t = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{s_X^2}{n_X} + \frac{s_Y^2}{n_Y}}}
\]

The corresponding confidence interval is be given by the following.

\[
\bar{X} - \bar{Y} \pm t^* \sqrt{\frac{s_X^2}{n_X} + \frac{s_Y^2}{n_Y}}
\]

Some texts, including yours, give a rule that the degrees of freedom associated with this test should be the minimum of \( n_X - 1 \) and \( n_Y - 1 \). SPSS, however, uses the following formula to compute the degrees of freedom for this test.

\[
d.f = \frac{\left(\frac{s_X^2}{n_X} + \frac{s_Y^2}{n_Y}\right)^2}{\frac{1}{n_X - 1}\left(\frac{s_X^2}{n_X}\right)^2 + \frac{1}{n_Y - 1}\left(\frac{s_Y^2}{n_Y}\right)^2}
\]

When the population standard deviations can be assumed to be equal, the appropriate form the \( t \)-test is the following.

\[
t = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{(n_X - 1)s_X^2 + (n_Y - 1)s_Y^2}{n_X + n_Y - 2}\left(\frac{1}{n_X} + \frac{1}{n_Y}\right)}}
\]

The corresponding confidence interval is be given by the following.
Here, the appropriate degrees of freedom are \( d.f. = n_x + n_y - 2 \). We discuss how to assess the validity of the assumption of equal standard deviations later.

**SPSS Commands**

SPSS has a single command that tests hypotheses and constructs confidence intervals for the difference between the means of two independent populations. It is based on the fact that the sample means have a normal distribution.

**INDEPENDENT SAMPLES T TEST**

The independent samples \( t \) test command requires that the data be stored such that the observed values are in one variable and a second variable identifies group membership. The procedure tests \( H_0 : \mu_1 = \mu_2 \) versus \( H_a : \mu_1 \neq \mu_2 \). Select **Analyze > Compare Means > Independent-Samples T Test** to bring up the dialogue box shown in Figure 1.

![Figure 1—Dialogue Box for Independent Samples T Test](image)

Enter the variable to be tested in the **Test Variable(s)** box. Enter the group identifying variable in the **Grouping Variable** box. When you do you will see the variable name followed by (?,?). This allows for the fact that there

![Figure 2—Dialogue Box for Define Groups](image)
may be more than two categories in your group identifying variable. So, you need to define which values of the group identifying variable identify group 1 and group 2. To do this click Define Groups to bring up the dialogue box shown in Figure 2. Enter the value that corresponds to observations from group 1 in the Group 1 box, and enter the value that corresponds to observations from group 2 in the Group 2 box. Click Continue to return to the main dialogue box. You will notice that the question marks have been replaced by the values you just entered. SPSS constructs a 95% confidence interval for the difference between group means by default. In order to change this click Options to bring up the dialogue box shown in Figure 3.

Figure 3—Dialogue Box for Options

Enter the desired confidence level, in percent, in the Confidence Interval box. Click Continue to return to the main dialogue box. Click OK to conduct the test and construct the confidence interval. The results will appear in an Output window.

EXAMPLE
Suppose we want to test whether the mean calories in roast beef sandwiches at Arby’s is different from the mean calories in their chicken sandwiches and construct a 95% confidence interval for the difference between these two means. The data are stored in ARBYS99.SAV. The dialogue boxes shown in Figure 1, Figure 2, and Figure 3 are set up to do this. When you click OK in the main dialogue box you get the output shown in Figure 4 in an Output window.

<table>
<thead>
<tr>
<th>Meat</th>
<th>N</th>
<th>Mean</th>
<th>Std. Deviation</th>
<th>Std. Error Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Calories</td>
<td>12</td>
<td>486.75</td>
<td>140.165</td>
<td>40.462</td>
</tr>
<tr>
<td>Chicken</td>
<td>8</td>
<td>458.50</td>
<td>107.549</td>
<td>38.024</td>
</tr>
</tbody>
</table>

Figure 4—Sample Output from Independent Samples T Test

In the output you notice an entry for Levene’s test for equality of variances. Variance is the square of standard deviation. The null hypothesis for this test is \( H_0 : \sigma_1 = \sigma_2 \), and the alternative hypothesis is \( H_a : \sigma_1 \neq \sigma_2 \). If
the $p$-value indicates that there is no evidence against the null hypothesis, then you should use the results in the Equal variances assumed row. If there is evidence against the null hypothesis then you should use the results in the Equal variances not assumed row. In this case we can assume that the standard deviations are equal. The test shows that there is no evidence against the hypothesis that the mean calories is the same for the two types of meat. The 95% confidence interval for the difference in population means is (-94.943, 151.443). For samples this small you should check for normal distributions by drawing Q-Q plots for the roast beef sandwiches and for the chicken sandwiches.
Laboratory Activities

Submit a copy of your SPSS Output windows along with answers to all the questions asked and explanations that are asked for.

1. Start SPSS and load the data file RESPONSE.SAV.
2. Recall that the experiment was to determine if response times are different when subjects are given a visual cue versus an audible cue. Each subject was asked to react to both cue and reaction time for each was recorded. Theory suggests that reaction to the audible cue should be quicker than to the visual one.
   a) Is Independent Samples T Test appropriate for testing for difference in the mean reaction time for a visual stimulus (VIS) and the mean reaction time for an audible stimulus (AUD)? If not, explain why not, and stop here and go on to Activity 3.
   b) Reorganize the data so that all response times are in a single variable and create a new variable type that contains a 1 for the audio stimulus and a 2 for the visual stimulus. Test for a difference in mean reaction time. The appropriate alternative will depend on which stimulus you designated as group 1 and group2.
   c) What conclusions do you draw from this test? Explain your reasoning.
3. Load the SPSS data file MEMORY.SAV.
4. We are interested in determining whether or not there is a difference between the two memorization methods. Conduct the appropriate two-sample *t* test. What conclusion do you reach? Explain your reasoning.
5. For this activity you will use the data in LABOR.SAV. Is there evidence that the unemployment rate for women differs from that for men? Explain how you reach your conclusion.
APPENDIX

DESCRIPTION

OF

THE

DATA SETS
SOCIOLOGY DATA SET
The National Survey of Families and Households (NSFH)
A Socio-Demographic Data Set

I. Brief Description of the NSFH

The NSFH is a cross-sectional national probability data set of 13,008 cases centered on family and household issues. The University of Wisconsin's Demography and Ecology Center collected the data in 1987-88. The survey includes a main sample of the non-institutionalized population of the US 19 years of age and older and oversamples of Blacks, Puerto Ricans, Chicanos, single-parents, persons with step-children, cohabiting persons and persons recently married. Within households, one adult was randomly selected as the primary respondent. There is a main face-to-face interview with the primary respondent as well as a self-administered questionnaire with the primary respondent covering sensitive topics. There is also a self-administered questionnaire completed by the primary respondent's spouse or partner when appropriate.

The NSFH data set described below is a simple random sample of 2000 cases drawn from the NSFH full sample of 13,008. It was extracted by Prof. Robert Kelly of the Sociology Department at Le Moyne College and placed in SPSS readable format by Prof. William C. Rinaman of the Department of Mathematics and Computer Science at Le Moyne College. They are stored in the SPSS file NSFH.SAV.

II. Variable List

<table>
<thead>
<tr>
<th>Column</th>
<th>Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>NCASEID</td>
<td>Identification number of case in the sample</td>
</tr>
<tr>
<td>2</td>
<td>REGION</td>
<td>Region of the US where interview was conducted</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1 = Northeast</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2 = Northcentral</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3 = South</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4 = West</td>
</tr>
<tr>
<td>3</td>
<td>SAMPLE</td>
<td>Whether case is part of main sample or over sample</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1 = Main</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2 = Over Sample</td>
</tr>
<tr>
<td>4</td>
<td>WEIGHT</td>
<td>Case weight to be used when the individual is the unit of analysis for the nation as a whole</td>
</tr>
<tr>
<td>5</td>
<td>MARCOHAB</td>
<td>Respondent’s current marital/cohabitation status.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1 = married, living with spouse</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2 = married, spouse absent</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3 = cohabiting—separated</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4 = cohabiting—divorced</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5 = cohabiting—widowed</td>
</tr>
<tr>
<td></td>
<td></td>
<td>6 = cohabiting—never married</td>
</tr>
<tr>
<td></td>
<td></td>
<td>7 = not cohabiting—separated</td>
</tr>
<tr>
<td></td>
<td></td>
<td>8 = not cohabiting—divorced</td>
</tr>
<tr>
<td></td>
<td></td>
<td>9 = not cohabiting—widowed</td>
</tr>
<tr>
<td></td>
<td></td>
<td>10 = not cohabiting—never married</td>
</tr>
<tr>
<td>6</td>
<td>K1</td>
<td>Respondent has biological child/children under 18 in his/her household.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1 = yes</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2 = no</td>
</tr>
<tr>
<td>7</td>
<td>K2</td>
<td>Respondent has non-biological child/children under 18 in his/her household</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1 = yes</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2 = no</td>
</tr>
<tr>
<td>Column</td>
<td>Name</td>
<td>Description</td>
</tr>
<tr>
<td>--------</td>
<td>--------</td>
<td>-------------</td>
</tr>
<tr>
<td>8</td>
<td>K3</td>
<td>Respondent has non-biological children under 18 in his/her household. Respondent has any step child under 18 of spouse/partner living in the household.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1 = yes</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2 = no</td>
</tr>
<tr>
<td>9</td>
<td>AYOC</td>
<td>Age of youngest child residing with respondent with relationship of biological, step, adopted, foster or child of partner/lover.</td>
</tr>
<tr>
<td>10</td>
<td>CURMARCO</td>
<td>Date of beginning of current marriage/cohabitation.</td>
</tr>
<tr>
<td>11</td>
<td>COMPLED</td>
<td>Respondent’s level of education completed.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0 = no formal schooling</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1 = some elementary school</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2 = some high school</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3 = high school diploma or GED</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4 = some college</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5 = Bachelor’s degree</td>
</tr>
<tr>
<td></td>
<td></td>
<td>6 = some graduate study</td>
</tr>
<tr>
<td>12</td>
<td>IREARN</td>
<td>Respondent’s total earnings from wages and salary and self-employed income.</td>
</tr>
<tr>
<td>13</td>
<td>E207</td>
<td>Compared with other people your age, how would you describe your health?</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1 = very poor</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2 = poor</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3 = fair</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4 = good</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5 = excellent</td>
</tr>
<tr>
<td>14</td>
<td>E401</td>
<td>If divorced since Jan. 1, 1977, who wanted the marriage to end more?</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1 = I wanted the marriage to end but my husband/wife did not</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2 = I wanted it to end more than my husband/wife did not</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3 = We both wanted it to end</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4 = My husband/wife wanted it to end more than I did</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5 = My husband/wife wanted it to end but I did not</td>
</tr>
<tr>
<td>15</td>
<td>E402</td>
<td>If divorced since Jan. 1, 1977, how would you describe your current relationship with your former husband/wife?</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1 = very unfriendly</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2 = somewhat unfriendly</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3 = neither friendly nor unfriendly</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4 = somewhat friendly</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5 = very friendly</td>
</tr>
<tr>
<td>16</td>
<td>E1295E</td>
<td>Marriage is a lifetime relationship and should never be ended except under extreme circumstances.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1 = strongly agree</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2 = agree</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3 = neither agree nor disagree</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4 = disagree</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5 = strongly disagree</td>
</tr>
<tr>
<td>17</td>
<td>M2NUM</td>
<td>Number of people in respondent’s household at time of interview.</td>
</tr>
<tr>
<td>18</td>
<td>M2BP01</td>
<td>Respondent’s age, in years, at time of interview.</td>
</tr>
<tr>
<td>19</td>
<td>M2CP01</td>
<td>Respondent’s marital status.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1 = married</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2 = separated due to marital problems</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3 = divorced</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4 = widowed</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5 = never married</td>
</tr>
<tr>
<td>Column</td>
<td>Name</td>
<td>Description</td>
</tr>
<tr>
<td>--------</td>
<td>---------</td>
<td>-----------------------------------------------------------------------------</td>
</tr>
<tr>
<td>20</td>
<td>M2DP01</td>
<td>Respondent’s sex.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1 = male</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2 = female</td>
</tr>
<tr>
<td>21</td>
<td>BKMK2</td>
<td>Respondent is currently cohabiting.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1 = cohabiting</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2 = not cohabiting</td>
</tr>
<tr>
<td>22</td>
<td>M95</td>
<td>Number of time respondent has been married.</td>
</tr>
<tr>
<td>23</td>
<td>M96m</td>
<td>Date of respondent’s first marriage.</td>
</tr>
<tr>
<td>24</td>
<td>M99</td>
<td>How did first marriage end?</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1 = divorce</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2 = separation</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3 = death of partner</td>
</tr>
<tr>
<td>25</td>
<td>M100M</td>
<td>Date of first divorce.</td>
</tr>
<tr>
<td>26</td>
<td>M113</td>
<td>Ever cohabited with first spouse?</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1 = yes</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2 = no</td>
</tr>
<tr>
<td>27</td>
<td>M204</td>
<td>Number of biological children.</td>
</tr>
<tr>
<td>28</td>
<td>M205P01M</td>
<td>Birth date of first child.</td>
</tr>
<tr>
<td>29</td>
<td>M484</td>
<td>Race/ethnicity.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1 = Black</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2 = White—not of Hispanic origin</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3 = Hispanic</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4 = American Indian</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5 = Asian</td>
</tr>
<tr>
<td></td>
<td></td>
<td>6 = Other</td>
</tr>
<tr>
<td>30</td>
<td>M486</td>
<td>Religious preference.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0 = No religion</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1 = Roman Catholic</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2 = Jewish</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3 = Protestant</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4 = Other</td>
</tr>
<tr>
<td>31</td>
<td>M492A</td>
<td>How often respondent attends religious services.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0 = never</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1 = once</td>
</tr>
<tr>
<td></td>
<td></td>
<td>90 = 90 or more</td>
</tr>
<tr>
<td>32</td>
<td>M501</td>
<td>Level of father’s schooling.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0 = no schooling</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1 = some elementary school</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2 = some high school</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3 = high school graduate</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4 = some college</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5 = Bachelor’s degree</td>
</tr>
<tr>
<td></td>
<td></td>
<td>6 = some graduate study</td>
</tr>
<tr>
<td></td>
<td></td>
<td>7 = GED</td>
</tr>
<tr>
<td>33</td>
<td>M502</td>
<td>Level of mother’s schooling.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Codes same as for M501</td>
</tr>
<tr>
<td>Column</td>
<td>Name</td>
<td>Description</td>
</tr>
<tr>
<td>--------</td>
<td>------</td>
<td>-------------</td>
</tr>
</tbody>
</table>
| 34     | M505 | Did respondent’s family ever receive public assistance while respondent was a child?  
1 = yes  
2 = no |
| 35     | M528 | Respondent currently working for pay at any job?  
1 = yes  
2 = no |
| 36     | M535 | If employed, hours worked last week. |
I. Brief Description of the Data

The Economics Department at Le Moyne College has provided six sets of economic data. They are time series for the period from January, 1946 through December, 1995. These data were obtained from the CITIBASE archive. The data were obtained by Prof. Dixie Blackley of the Economics Department at Le Moyne College and were placed in SPSS format by Prof. William C. Rinaman of the Department of Mathematics and Computer Science at Le Moyne College.

II. The Data Files

CONSTR.SAV

<table>
<thead>
<tr>
<th>Column</th>
<th>Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Year</td>
<td>Year</td>
</tr>
<tr>
<td>2</td>
<td>Month</td>
<td>Month — 1 = January, …, 12 = December</td>
</tr>
<tr>
<td>3</td>
<td>HSFR</td>
<td>New privately owned housing units started, in thousands, seasonally adjusted annual rate</td>
</tr>
</tbody>
</table>

FINANCE.SAV

<table>
<thead>
<tr>
<th>Column</th>
<th>Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Year</td>
<td>Year</td>
</tr>
<tr>
<td>2</td>
<td>Month</td>
<td>Month — 1 = January, …, 12 = December</td>
</tr>
<tr>
<td>3</td>
<td>FM1</td>
<td>Cash and checkable deposits, in billions, monthly average of daily values.</td>
</tr>
<tr>
<td>4</td>
<td>FM2</td>
<td>Cash, checkable, savings and time deposits, in billions, monthly average of daily values</td>
</tr>
<tr>
<td>5</td>
<td>FYGN3</td>
<td>Yield on 3 month Treasury bills, auction, in percent per annum</td>
</tr>
<tr>
<td>6</td>
<td>FYPR</td>
<td>Prime loan rate, in percent per annum</td>
</tr>
<tr>
<td>7</td>
<td>FYGM3</td>
<td>Yield on 3 month Treasury bills, secondary market, in percent per annum</td>
</tr>
<tr>
<td>8</td>
<td>FYAAAC</td>
<td>Moody’s AAA corporate bond rate, percent per annum</td>
</tr>
<tr>
<td>9</td>
<td>FYAAAI</td>
<td>Moody’s AAA industrial bond rate, percent per annum</td>
</tr>
<tr>
<td>10</td>
<td>FYAVG</td>
<td>Average yield on corporate bonds, Moody, percent per annum</td>
</tr>
<tr>
<td>11</td>
<td>FYIND</td>
<td>Average yield on industrial bonds, Moody, percent per annum</td>
</tr>
</tbody>
</table>

INDUST.SAV

<table>
<thead>
<tr>
<th>Column</th>
<th>Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Year</td>
<td>Year</td>
</tr>
<tr>
<td>2</td>
<td>Month</td>
<td>Month — 1 = January, …, 12 = December</td>
</tr>
<tr>
<td>3</td>
<td>IPC</td>
<td>Index of industrial production, consumer goods (1987 = 100)</td>
</tr>
<tr>
<td>4</td>
<td>IPCD</td>
<td>Index of industrial production, consumer durables (1987 = 100)</td>
</tr>
<tr>
<td>5</td>
<td>IPCN</td>
<td>Index of industrial production, consumer non-durables (1987 = 100)</td>
</tr>
<tr>
<td>6</td>
<td>IP</td>
<td>Index of industrial production, total (1987 = 100)</td>
</tr>
</tbody>
</table>
**LABOR.SAV**

<table>
<thead>
<tr>
<th>Column</th>
<th>Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Year</td>
<td>Year</td>
</tr>
<tr>
<td>2</td>
<td>Month</td>
<td>Month—1 = January, … , 12 = December</td>
</tr>
<tr>
<td>3</td>
<td>LHP16</td>
<td>Labor force participation rate, 16 and over, seasonally adjusted</td>
</tr>
<tr>
<td>4</td>
<td>LHMP16</td>
<td>Male labor force participation rate, 16 and over, seasonally adjusted</td>
</tr>
<tr>
<td>5</td>
<td>LHFPI6</td>
<td>Female labor force participation rate, 16 and over, seasonally adjusted</td>
</tr>
<tr>
<td>6</td>
<td>LHUMR</td>
<td>Male unemployment rate, 16 and over, seasonally adjusted</td>
</tr>
<tr>
<td>7</td>
<td>LHUFR</td>
<td>Female unemployment rate, 16 and over, seasonally adjusted</td>
</tr>
<tr>
<td>8</td>
<td>LHUR</td>
<td>Unemployment rate, 16 and over, seasonally adjusted</td>
</tr>
<tr>
<td>9</td>
<td>LHURM</td>
<td>Married men unemployment rate, 16 and over, seasonally adjusted</td>
</tr>
<tr>
<td>10</td>
<td>LHURMF</td>
<td>Married women unemployment rate, 16 and over, seasonally adjusted</td>
</tr>
<tr>
<td>11</td>
<td>LHURHF</td>
<td>Female head of household unemployment rate, 16 and over, seasonally adjusted</td>
</tr>
</tbody>
</table>

Note: Labor Force Participation Rate = \( \frac{\text{Labor Force}}{\text{Noninstitutional Population}} \times 100 \)

**MANUF.SAV**

<table>
<thead>
<tr>
<th>Column</th>
<th>Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Year</td>
<td>Year</td>
</tr>
<tr>
<td>2</td>
<td>Month</td>
<td>Month—1 = January, … , 12 = December</td>
</tr>
<tr>
<td>3</td>
<td>MFGS</td>
<td>Manufacturing shipments, in millions of dollars, seasonally adjusted</td>
</tr>
<tr>
<td>4</td>
<td>IVM</td>
<td>Manufacturing inventories, in millions of dollars, seasonally adjusted</td>
</tr>
<tr>
<td>5</td>
<td>MO</td>
<td>Manufacturing new orders, in millions of dollars, seasonally adjusted</td>
</tr>
<tr>
<td>6</td>
<td>MDS</td>
<td>Manufacturing shipments, durables, in millions of dollars, seasonally adjusted</td>
</tr>
<tr>
<td>7</td>
<td>IVMD</td>
<td>Manufacturing inventories, durables, in millions of dollars, seasonally adjusted</td>
</tr>
<tr>
<td>8</td>
<td>MDO</td>
<td>Manufacturing new orders, durables, in millions of dollars, seasonally adjusted</td>
</tr>
<tr>
<td>9</td>
<td>MNS</td>
<td>Manufacturing shipments, non-durables, in millions of dollars, seasonally adjusted</td>
</tr>
<tr>
<td>10</td>
<td>IVMN</td>
<td>Manufacturing inventories, non-durables, in millions of dollars, seasonally adjusted</td>
</tr>
<tr>
<td>11</td>
<td>MNO</td>
<td>Manufacturing new orders, non-durables, in millions of dollars, seasonally adjusted</td>
</tr>
</tbody>
</table>

**PRICES.SAV**

<table>
<thead>
<tr>
<th>Column</th>
<th>Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Year</td>
<td>Year</td>
</tr>
<tr>
<td>2</td>
<td>Month</td>
<td>Month—1 = January, … , 12 = December</td>
</tr>
<tr>
<td>3</td>
<td>PZUNEW</td>
<td>Consumer price index for urban consumers, all items, all urban consumers, not seasonally adjusted</td>
</tr>
<tr>
<td>4</td>
<td>PUNEW</td>
<td>Consumer price index for urban consumers, all items, all urban consumers, seasonally adjusted</td>
</tr>
<tr>
<td>5</td>
<td>PW</td>
<td>Producer price index, all commodities, not seasonally adjusted</td>
</tr>
</tbody>
</table>
PSYCHOLOGY DATA SETS

I. Brief Description

The Psychological data sets represent the result of experiments conducted either in courses within the Department of Psychology at Le Moyne College or conducted by members of the faculty of that department. The contents of the files will be listed after a brief description of the experiment is given.

II. The Data Files

MEMORY.SAV

It is desired to determine if one method for memorizing items is more effective than another. One method is to repeat the thing to be memorized as many times as possible (repetition). The other method is to create a mental picture (visualization) of the item to be memorized. In the experiment 54 people were asked to memorize word pairs (e.g. DOG—CAT). Repetition was to be used by 27 of the subjects and the other 27 subjects used visualization. There were 12 word pairs to be memorized. Each subject was then shown one member of each pair and asked to give the other member.

<table>
<thead>
<tr>
<th>Column</th>
<th>Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Sex</td>
<td>Sex of subject, M = male, F = female</td>
</tr>
<tr>
<td>2</td>
<td>Method</td>
<td>Memorization method, R = repetition, I = visualization</td>
</tr>
<tr>
<td>3</td>
<td>Score</td>
<td>Number of correct responses, out of 12</td>
</tr>
</tbody>
</table>

RESPONSE.SAV

The issue in this experiment is whether people respond more quickly to a visual stimulus or an audio stimulus. In the experiment, subjects were exposed to a stimulus and asked to push a button as soon as possible after noting the stimulus. The time between onset of the stimulus and pressing the button was measured. In the experiment 58 subjects had their reaction time measured using the visual stimulus. The same 58 subjects also had their reaction time measured using the audio stimulus. The order in which this was done differed from subject to subject.

<table>
<thead>
<tr>
<th>Column</th>
<th>Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Sex</td>
<td>Sex of subject, M = male, F = female</td>
</tr>
<tr>
<td>2</td>
<td>DH</td>
<td>Dominant hand, L = left, R = right</td>
</tr>
<tr>
<td>3</td>
<td>Stim</td>
<td>Which stimulus was done first, V = visual, A = audio</td>
</tr>
<tr>
<td>4</td>
<td>Vis</td>
<td>Reaction time to visual stimulus, in msec.</td>
</tr>
<tr>
<td>5</td>
<td>Aud</td>
<td>Reaction time to audio stimulus, in msec.</td>
</tr>
</tbody>
</table>
INDEX

Bar charts, 49
  Segment, 50
between groups design, 54
blind experiment, 54
Boxplot, 22
Central Limit Theorem, 57
Ciprop, 63
coefficient of determination, 41
comparing the proportions of two independent
groups, 68
Comparing two population means, 77
Compute, 36
certainty intervals for a population mean, 71
Correlate, 32
Correlation, 32
Crosstabs, 48
data file, 5
Data View, 5
Descriptives, 21
double blind experiment, 54
Editing and manipulating data, 11
Entering and saving data, 9
Entering data, 9
experimental design, 54
Explore, 13, 19, 71
first quartile, 19
Fitted line plot, 44
Frequencies, 18
help, 8
Histogram, 14
  Interactive, 15
Independent samples t test, 78
Interactive histogram, 15
Interquartile range, 19
Keeping a record of an SPSS session, 11
least squares regressions, 40
maximum, 19
mean, 19
median, 19
Merging SPSS files, 37
minimum, 19
missing values, 19
nominal variable, 7
non-missing values, 19
NSFH data, 26
One Sample test, 73
ordinal variable, 7
Output window, 5
paired comparisons, 54
Paired samples t test, 74
placebo, 54
PTWO, 68
Q-Q plot, 57
r^2, 41
Range, 19
Recode, 26
Regression, 40
Retrieving previously saved data, 9
Sample, 59
Saving data, 10
scale variable, 7
Scatter plot, 32
  Interactive, 32
Select cases, 28
SPSS help, 8
standard deviation, 19
standard error of the mean, 19
Stemplots, 13
Test2prop, 68
Testprop, 66
third quartile, 19
treatment, 54
two-way tables, 48
Variable view, 5, 6
within subjects design, 54