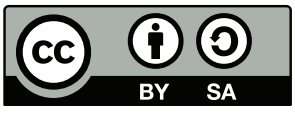
LAB MANUAL

PSPP

GEORGE SELF

August 2017 – Edition 1.0

George Self: *Lab Manual,* PSPP, August 2017



FORWARD

I have taught BASV 316, *Introductory Methods of Analysis*, online for the University of Arizona in Sierra Vista since 2010 and enjoy working with students on research methodology. From the start, I wanted students to work with statistics that are commonly found in research. It is my belief that the best way to understand what statistics can, and cannot, prove is to calculate values using a known dataset. As I evaluated statistical software for this class I had three criteria:

- **Open Educational Resource (OER)**. It is important to me that students use software that is available free of charge and is supported by the entire web community.
- **Platform**. While most of my students use a Windows-based system, some use Macintosh and it was important to me to use software that is available for all of those platforms. As a bonus, most OER software is also available for the Linux system, though I'm not aware of any of my students who are using Linux.
- **Longevity**. I wanted a system that could be used in other college classes or in a business setting after graduation. That way, any time a student spends learning the software in my class will be an investment that can yield results for many years.

I originally wrote a series of six lab exercises (later expanded to nine) using R-Project since that software met these three criteria. Moreover, R-Project is a recognized standard for statistical analysis and could be easily used for even peer-reviewed published papers. Unfortunately, I found R-Project to be confusing to students since it is text-based with rather complex commands. I found that I spent a lot of time just teaching students how to set up a single test with R-Project instead of analyzing the result. In the spring of 2017 I changed to `SOFA` (*Statistics Open For All*) because it is much easier to use and still met my criteria. However, SOFA is a stand-alone product that students would not likely be able to use beyond this class so in the fall of 2017 I changed to `PSPP`. `PSPP` looks and works like SPSS, which is the leading statistical analysis software package used in research around the world. `PSPP` is a simplified version of SPSS and has only a few statistical analysis options available, but those are enough for many undergraduate projects. Moreover, I believe that students who learn `PSPP` in this class and later need the power of SPSS will find the transition to be very smooth since the two programs have a similar menu structure.

This lab manual explores many aspects of PSPP but does not attempt to dig into every corner of this software. It is my hope that students will find the labs instructive and will then be able to use PSPP, or smoothly transition into SPSS, for other classes. This lab manual is published under a Creative Commons license with a goal that other instructors will modify it to meet their own needs. I always welcome comments and will improve this manual as I receive feedback.

—George Self

Contents

ACRONYMS

GUI Graphic User Interface

PSPP This is not an acronym though it looks like it should be.

SPSS Statistical Package for the Social Sciences

Part I

INTRODUCTION

This part contains only one lab exercise and that provides instructions for downloading and installing PSPP. Also included is an introduction to the different types of data and some simple data transformations available in PSPP.

Part II

DESCRIPTIVE

Descriptive statistics attempt to describe data and include tools like frequency counts, means, and crosstabs. This part contains five labs that explore several different descriptive statistics available in PSPP.

1 INTRODUCTION

1.1 INTRODUCTION

Statistical analysis is the core of nearly all research projects and researchers have a wide variety of statistical tools that they can use, like *SPSS*, *SAS*, and *R*. Unfortunately, these analysis tools are expensive or difficult to master so this lab manual introduces PSPP, an open source statistical analysis program that is free of charge and easy to use. Even though PSPP looks like an acronym it is not. Those letters are intended to imply that it is the opposite of SPSS in the sense that it is provided free of charge but is compatible with the SPSS language and datasets.

Before downloading and diving into PSPP there are two important background fundamentals that must be considered: hypothesis and data.

1.2 DISCUSSION

1.2.1 *Hypothesis*

A hypothesis is an attempted explanation for some observation and is often used as a starting point for further investigation. For example, imagine that a physician notices that babies born of women who smoke seem to be lighter in weight than for women who do not smoke. That could lead to a hypothesis like "smoking during pregnancy is linked to light birth-weights." As another example, imagine that a restaurant owner notices that tipping seems to be higher on weekends than through the week. That might lead to a hypothesis that "the size of tips is higher on weekends than weekdays." After creating a hypothesis a researcher would gather data and then statistically analyze that data to determine *if* the hypothesis is true but additional investigation may be needed to explain *why* it is true.

In a research project there are usually two related competing hypotheses: the *Null Hypothesis* and the *Alternate Hypothesis*.

- Null Hypothesis (abbreviated H_0). This is sometimes described as the "skeptical" view; that is, the explanation for some observed phenomenon is mistaken. For example, the null hypothesis for the smoking mother observation mentioned above would be "smoking has no effect on a baby's weight" and for the tipping observation would be "there is no difference in tipping on the weekend."

- Alternate Hypothesis (abbreviated H_a). This is the suggested explanation for an observed phenomenon. In the case of the smoking mother mentioned above the alternative hypothesis would be that "smoking causes a decrease in birth weight." This is called the "alternate" because it is different from the status quo which is encapsulated in the null hypothesis.

One commonly used example of the difference between the null and alternate hypothesis comes from the trial court system. When a jury deliberates about the guilt of a defendant they start from a position of "innocent until proven guilty," which would be the null hypothesis. The prosecutor is asking the jury to accept the alternate hypothesis, or "the defendant committed the crime."

For the most part, researchers will never conclude that the alternate hypothesis is true. There are always confounding variables that are not considered but could be the cause of some observation. For example, in the smoking mothers example mentioned above, even if the evidence indicates that babies born to smokers are lighter in weight the researcher could not state conclusively that smoking caused that observation. Perhaps non-smoking mothers had better health care, perhaps they had better diets, perhaps they exercised more, or any of a number of other reasonable explanations not related to smoking.

For that reason, the result of a research project is normally reported with one of two phrases similar to these:

- *The null hypothesis is rejected.* If the evidence indicates that there is a significant difference between the status quo and whatever was observed then the null hypothesis would be rejected. For the "tipping" example above, if the researcher found a significant difference in the amount of money tipped on weekends compared to weekdays then the null hypothesis (that is, tipping is the same on weekdays and weekends) would be rejected.

- *The null hypothesis cannot be rejected.* If the evidence indicates that there is no significant difference between the status quo and whatever was observed then the researcher would report that the null hypothesis could not be rejected. For example, if there was no significant difference in the birth weights of babies born to smokers and non-smokers then the researcher failed to reject the null hypothesis.

Often a research hypothesis is based on a prediction rather than an observation and that hypothesis can be tested to see if there is any significant difference between it and the null hypothesis. Imagine a hypothesis like "walking one mile a day for one month decreases blood pressure." A researcher could easily test this by measuring the blood pressure of a group of volunteers, have them walk a mile every day for a month, and then measure their blood pressure at the end of the experiment to see if there was any significant difference.

1.2.2 *Data*

1.2.2.1 *Types of Data*

There are four types of data, divided into two main groups, and it is important to understand the difference between them since that determines appropriate statistical tests to be used in data analysis.[1]

- **Qualitative**. Qualitative data groups observations into a limited number of categories; for example, type of pet (cat, dog, bird, etc.) or place of residence (Arizona, California, etc.). Because qualitative data do not have characteristics like means or standard deviations, they are analyzed using non-parametric tests, as described in Lab **??** on page **??**. Qualitative data can be further divided into two sub-types, nominal and ordinal.
  - **Nominal**. Nominal data are categories that do not overlap and have no meaningful order, they are merely labels for attributes. Examples of nominal data include occupations (custodial, accounting, sales, etc.) and blood type (A, B, AB, O). A special subcategory of nominal data is binary, or dichotomous, where there are only two possible responses, like "yes" and "no". Nominal data are sometimes stored in a database using numbers but they cannot be treated like numeric data. For example, binary data, like "Do you rent or own your home?" can be stored as "1 = rent, 2 = own" but the numbers in this case have no numeric significance and could be replaced by words like "Rent" and "Own."
  - **Ordinal**. Ordinal data, like nominal, are categorical data but, unlike nominal, the categories imply some sort of order (which is why it is called "ordinal" data). One example of ordinal data is the "star" rating system for movies. It is clear that a five-star movie is somehow better than a four-star movie but there is no way to quantify the difference between those two categories. As another example, it is common for hospital staff members to ask patients to rate their pain level on a scale of one to ten. If a patient reports a pain level of "seven" but after some sort of treatment later reports a pain level of "five" then the pain has clearly decreased but it would be impossible to somehow quantify the exact difference in those two levels. Ordinal scales are most commonly used for Likert-type survey questions where the responses are selections like "Strongly Agree", "Agree", "Neutral", "Disagree", "Strongly Disagree". Ordinal data are also used when numeric data are grouped. For

1 Appendix A: Datasets, on page 63, lists all of the datasets used in this lab manual and specifies the type of data each contains.

example, if a dataset included respondents' ages then those numbers could be grouped into categories like "20 – 29" and "30 – 39." Those groups would typically be stored in the dataset as a single number so maybe "2" would represent the ages "20 – 29," which would be ordinal data.

- **Quantitative**. Quantitative data are numbers, typically counts or measures, like a person's age, a tree's height, or the number of widgets sold. Quantitative data are measured with scales that have equal divisions so the difference between any two values can be calculated. Quantitative data are discrete if they are represented by integers, like the count of words in a document, or continuous if they are represented by fractional numbers, like a person's height. Because quantative data has characteristics like means and standard deviations, they are analyzed using parametric tests, as described in Lab **??** on page **??**. Quantitative data can be further divided into two sub-types, interval and ratio[2].
  - **Interval**. Interval data use numbers to represent quantities where the distance between any two quantities can be calculated but there is no true zero point on the scale. One example is a temperature scale where the difference between 80°and 90°is calculated to be the same as the difference between 60°and 70°. It is important to note that interval data do not include any sort of true zero point, thus zero degrees Celsius does not mean "no temperature," and without a zero point it is not reasonable to make a statement like 20°is twice as hot as 10°.
  - **Ratio**. Ratio data, like interval data, use numbers to describe a specific measurable distance between two quantities; however, unlike interval data, ratio data have a true zero point. A good example of the use of ratio data is the sales report for an automobile dealership. Because the data are a simple count of the number of automobiles sold it is possible to compare one month to another. Also, since the scale has a true zero point (it is possible to have zero sales) it is possible to state that one month had twice the sales of another.

1.2.2.2 *Shape of Data*

ABOUT THE NORMAL DISTRIBUTION (BELL CURVE) When the quantitative data gathered from some statistical project are plotted on a graph they often form a "normal distribution" (sometimes called a "bell curve" due to its shape). As an example, consider the Scholastic Aptitude Test (SAT) which is administered to more than 1.5 million

2 PSPP lumps both Interval and Ratio data into a single type called "Scale."

high school students every year. Figure 1 was created with fake data but illustrates the results expected of a typical SAT administration.

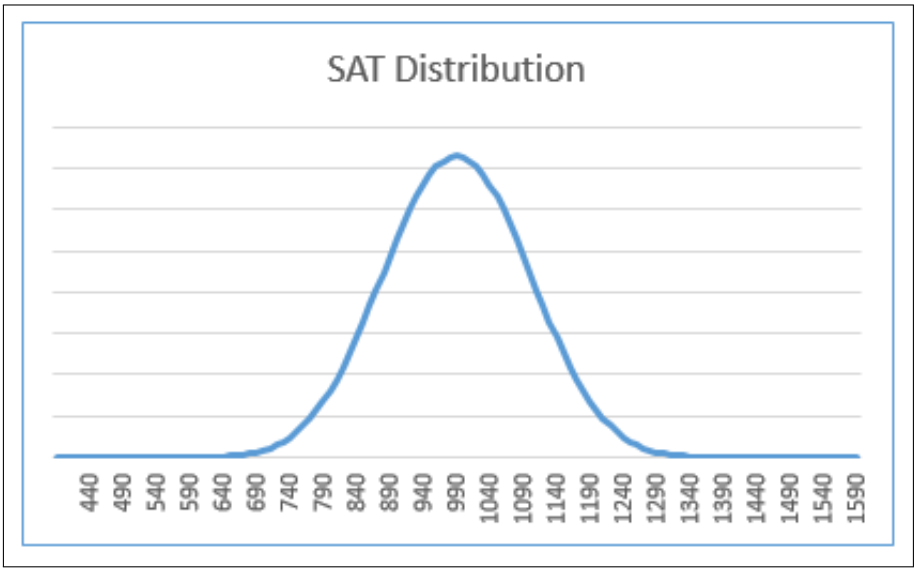


Figure 1: Normal Distribution

SAT scores lie between 400 and 1600 as listed across the X-Axis and the number of students who earn each score is plotted. Since the most common score is 1000 that score is at the peak of the curve. Very few students scored above 1300 or below 650 and the curve is near the lower bound beyond those points. This illustrates a normal distribution where most scores are bunched near the center of the graph with only a few at either extreme.

The normal distribution is important because it permits researchers to test hypothesis about the sample. For example, perhaps a researcher hypothesized that the students in university "A" had a higher graduation rate than at university "B" because their SAT scores were higher. Because SAT scores have a normal distribution the researcher could use specific tests, like a t-test, to try to support the hypothesis. However, if the data were not normally distributed then the researcher would need to know that and select a different group of statistical tests.

EXCESS KURTOSIS One way to mathematically describe a normal distribution is to calculate the length of the tails of a bell curve, and that is called its *excess kurtosis*. For a normal distribution the excess kurtosis is 0.00, a positive excess kurtosis would indicate longer tails while a negative excess kurtosis would indicate shorter tails. Intuitively, many people believe the excess kurtosis represents the "peaked-ness" of the curve since longer tails would tend to lead to a more peaked graph; however, excess kurtosis is a measure of the

data outliers, which would be only present in the tails of the graph; so excess kurtosis is not directly indicative of the the "sharpness" of the peak. It is difficult to categorically state that some level of excess kurtosis is good or bad. In some cases, data that form a graph with longer tails are desired but in other cases they would be a problem.

Following are three examples of excess kurtosis. Notice that as the excess kurtosis increases the tails become longer.

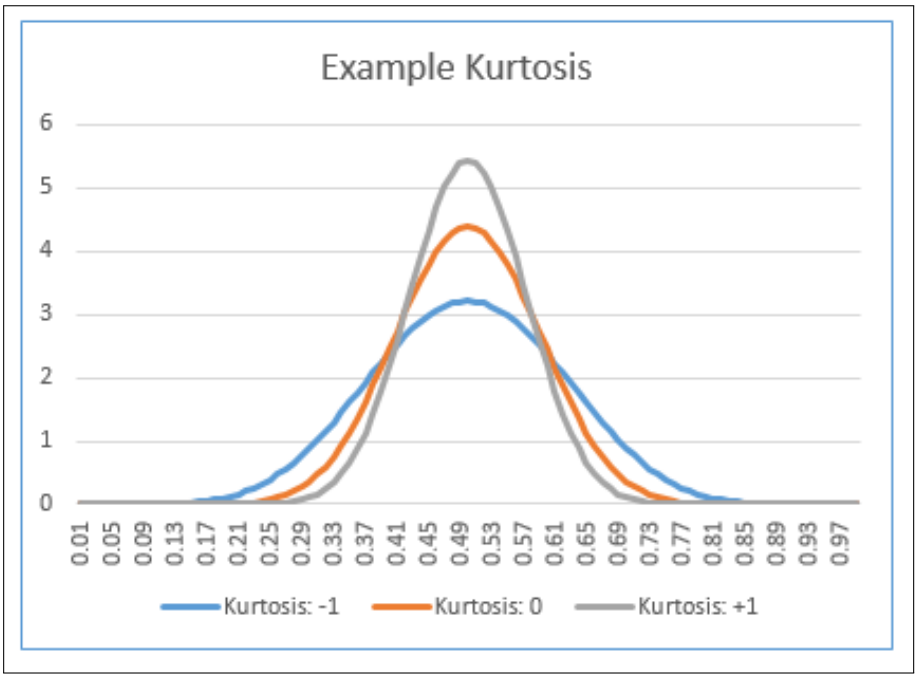


Figure 2: Kurtosis in a Normal Distribution

SKEW The second numerical measure of a normal distribution that is frequently reported is its *skew*, which is a measure of the symmetry of the curve about the mean of the data. The normal distribution in Figure 1 has a skew of 0.00. A positive skew indicates that the tail on the right side is longer, which means that there are several data points on the far right side of the graph "pulling" the tail out that direction. A negative skew indicates that the tail on the left side of the graph is longer. Following are three examples of skew:

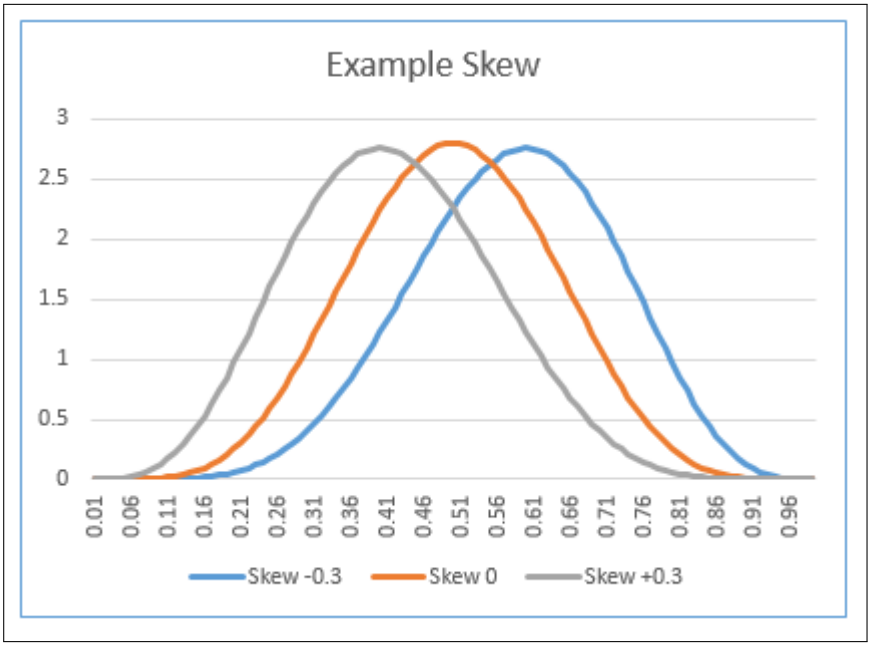


Figure 3: Skew in a Normal Distribution

1.3 PROCEDURE

1.3.1 *Installing and Starting PSPP*

There are versions of PSPP available for Windows, MacOS, and Linux; so whatever operating system is being used there is a version that will work. The PSPP downloads can be found at:

`https://www.gnu.org/software/pspp/get.html`.

The download and installation process is fairly straightforward so there is no additional information about that here. Students should contact their instructor if they have trouble downloading or installing PSPP.

1.3.2 *Importing Data*

For simplicity, data files that are already prepared and useable "out of the box" are made available with this lab manual[3]. Therefore, the "Import" feature will not be considered in this lab, but students who want to know more about importing data can find more information online.

3 PSPP can read and save SPSS data files so students can load datasets in that format directly into PSPP without further processing.

Students should create a single folder for all labs and store the data files in that folder. The datasets[4] for all of the activities in this manual are available in a ZIP file located at:

https://goo.gl/hA04Gg

Download the "Data Files" for Edition 1.0, August 2017, and extract all of the **.SAV** files to a folder and then open them as needed.

The following datasets should be available to complete the labs in this manual:

- bdims
- births
- cafe
- cars
- email
- gifted
- rivers

The description for all datasets can be found in Appendix 6.1.

1.3.3 *Using PSPP*

To open a dataset, click File → Open. For example, to open the *bdims* dataset:

1. Start PSPP and Click File → Open.
2. On the "Select File" screen, click *bdims.sav* and then click "Open" at the bottom of the screen.

When a dataset is first opened in PSPP it is presented in a spreadsheet-like view. For example, Figure 4 shows the top 10 rows of the cars dataset after it was first opened.

4 Appendix A: Datasets, on page 63, details the structure and contents of all datasets used in this manual.

Case	driveTrain	mpgCity	passenge	price	type	weight	
1	front	25	5	15.90	small	2705	
2	front	18	5	33.90	midsize	3560	
3	front	19	6	37.70	midsize	3405	
4	rear	22	4	30.00	midsize	3640	
5	front	22	6	15.70	midsize	2880	
6	front	19	6	20.80	large	3470	
7	rear	16	6	23.70	large	4105	
8	front	19	5	26.30	midsize	3495	
9	front	16	6	34.70	large	3620	
10	front	16	5	40.10	midsize	3935	
11	front	21	6	15.90	midsize	3195	

Figure 4: The Top Of The Cars Data Table

1.3.3.1 *Activity 1*

Start PSPP and open the *cafe* dataset. Take a screen capture of the first ten rows of that data table. It should look something like Figure 4, which shows the top 10 rows of the *cars* dataset.

1.3.4 *Variables*

At the bottom of the data window are two buttons used to change the window from Data View to Variable View.

Figure 5: Data/Variable View Buttons

The Variable View lists all of the variables in the dataset along with the attributes for those variables.

Variab	Name	Type	Width	Decimal	Label
1	fage	Numeric ...	2	0	Father's Age
2	gained	Numeric ...	3	0	Mother's Weight Gain
3	gender	String ...	6		Baby's Gender
4	habit	String ...	9		Mother's Smoking Habit
5	lowbirthweight	String ...	14		Low Birth Weight?
6	mage	Numeric ...	2	0	Mother's Age
7	marital	String ...	11		Mother's Marital Status
8	mature	String ...	11		Mother's Maturity Level
9	premie	String ...	9		Premature Birth?
10	visits	Numeric ...	2	0	Mother's Hospital Visits
11	weeks	Numeric ...	2	0	Gestation Period
12	weight	Numeric ...	3	2	Birth Weight
13	whitemom	String ...	9		White Mother?

Figure 6: Data/Variable View Buttons

Figure 6 shows the first six attributes for the variables in the *births* dataset, but all of the attributes are described below.

1. **Variable** is simply a one-up number assigned by PSPP when the dataset is created.
2. **Name** is the name of the variable. In Figure 6, Variable one is named "fage."
3. **Type** is the type of data the variable contains. PSPP permits the following types of data: *Numeric*, *Comma*, *Dot*, *Scientific Notation*, *Date*, *Dollar*, *Custom Currency*, and *String*[5].
4. **Width** and **Decimal** is how much data is stored. For numeric data the Width is the size of the integer portion of the number while the Decimal attribute is the size of number's decimal portion. For PSPP the size indicates how many places are displayed, not the size of the number. For example, the number 999 has a width of three since there are three places. For string data the width is the number of characters the variable can contain.
5. **Label** is a label displayed in reports to make it easier to understand. For example, variable one is named "fage" but in a report it would be labeled as "Father's Age" to make it easier to understand that variable.
6. **Value Labels** are labels for values, normally only used for nominal data. For example, it is common to store "yes/no" type of

5 The labs in this manual use only *numeric*, *date*, and *string* types of data.

data where 0 is "no" and 1 is "yes." By specifing those values as Value Labels that would be displayed in reports rather something like 1, which would be difficult to understand.

7. **Missing Values** are how missing values should be displayed. By default PSPP displays missing values with a dot but some sort of code could be used instead and that code is entered here.
8. **Column** sets the width of the column displayed in the Data View.
9. **Align** determines if a column is left, center, or right aligned.
10. **Measure** determines how the variable is used by PSPP and can be Scale, Ordinal, or Nominal.
11. **Role** is the role the variable plays in the dataset. The possible roles are Input, Output, Both, None, Partition, and Split. For all of the labs in this manual the role for all variables are always "Input."

Any of the above attributes can be changed but the researcher is responsible to not "break" the dataset by changing an attribute to an inappropriate value.

1.3.4.1 *Activity 2*

Start PSPP and open the *cafe* dataset. Switch to the *Variable View* and take a screen capture of all 13 rows and the first seven columns (From *Variable* to *Value Labels*) of the variable view. It should look something like Figure 6, which shows the first six columns for the variable view of the *births* dataset.

1.3.5 *Menus*

The top of the PSPP screen has a menu bar with important selections. Most of the labs in this manual will focus on the *Analyze* menu item. Students may also want to consider Appendix B, page 68, and work through some of the *Transform* menu items.

1.3.6 *Syntax Files*

PSPP is, technically, a command line program with scores of commands and options that are designed to be entered in a text box. The labs in this manual, though, teach students how to use a Graphic User Interface (GUI) rather than the command line. It may be easiest to think about the GUI as a "front end" to the command line. Unfortunately, students who download and read the PSPP User's Guide can quickly become confused since it discusses commands like:

```
REGRESSION /VARIABLES={age, tip} /STATISTICS={R, ANOVA}
```

with no clear relationship between a command like that and the GUI that most students use. Technically, the GUI is called "PPSPIRE" and is just an adjunct to PSPP.

It is possible to open a command line box and manually enter commands. To access a command line box, Click File → New → Syntax.

There are many commands and options available via the command line that are not available using menus in the GUI, but those are not encountered in the labs in this manual (with only a couple of exceptions).

1.4 DELIVERABLE

Complete the following activities in this lab:

Number	Name	Page
1.3.3.1	Activity 1	13
1.3.4.1	Activity 2	15

Consolidate the responses for all activities into a single document and submit that document for grading.

2

LAB 02: FREQUENCIES

2.1 INTRODUCTION

Nominal and Ordinal data items are normally reported in frequency tables where the counts for a particular item are displayed and this lab explores frequency tables and visualization techniques that are used to make frequencies easier to comprehend.

2.2 DISCUSSION

2.2.1 *Frequency Tables*

A frequency table simply lists a count of the number of times that some nominal or ordinal data item appears in a dataset. These types of tables are common around election time when polls report the number of people who voted for or against some proposition. As an example, here is a frequency table for the passenger rating in the *cars* dataset.

passengers

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
	4	10	18.52	18.52	18.52
	5	28	51.85	51.85	70.37
	6	16	29.63	29.63	100.00
	Total	54	100.0	100.0	

Figure 7: Number of Passengers Per Car

The above table shows that 10 cars in the dataset were rated for four passengers, 28 for five passengers, and 16 for six passengers, for a total of 54 rated cars. The table also shows the various row percentages so the researcher could report that 18.5% of the cars were rated for four passengers. The "Valid Percent" column indicates the percentage of cases when missing cases are removed. In this dataset there were no missing cases so the Valid Percent column is the same as the "Percent" column.

A second example of a frequency table was created from the *email* dataset. This frequency table shows the number of images that were attached to messages.

image

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
	0	3076	97.53	97.53	97.53
	1	59	1.87	1.87	99.40
	2	13	.41	.41	99.81
	3	5	.16	.16	99.97
	5	1	.03	.03	100.00
	Total	3154	100.0	100.0	

Figure 8: Images Per Message

Figure 8 shows that 97.5% of 3154 email messages contained no images while a small number of messages contained one or more images.

Frequency tables are only useful for nominal or ordinal data-type items. To illustrate why this is true, imagine creating a survey for all of the students at the University of Arizona and including "age" (interval-type data) as one of the survey questions. Attempting to create a frequency table for the ages of the respondents would have, potentially, more than 65 rows since student ages would range from about 15 to more than 80 and each row would report the number of students for that age. While a frequency table that large could be created it would have so many rows that it would be virtually unusable.

2.2.2 *Visualizing Frequency*

There are many ways to visualize frequency data and people often find that graphs aid in comprehension. This lab introduces the visualization tools available in PSPP.

2.2.2.1 *Histogram*

A histogram is a graph that shows how often various responses were selected on a survey. These are often presented as a graphic representation of the statistical data found in a frequency table in order to aid in understanding. Histograms are only used for data that are interval or ratio in nature, for example, age or height.

As an example of a histogram, Figure 9 shows the mother's age from the *births* dataset.

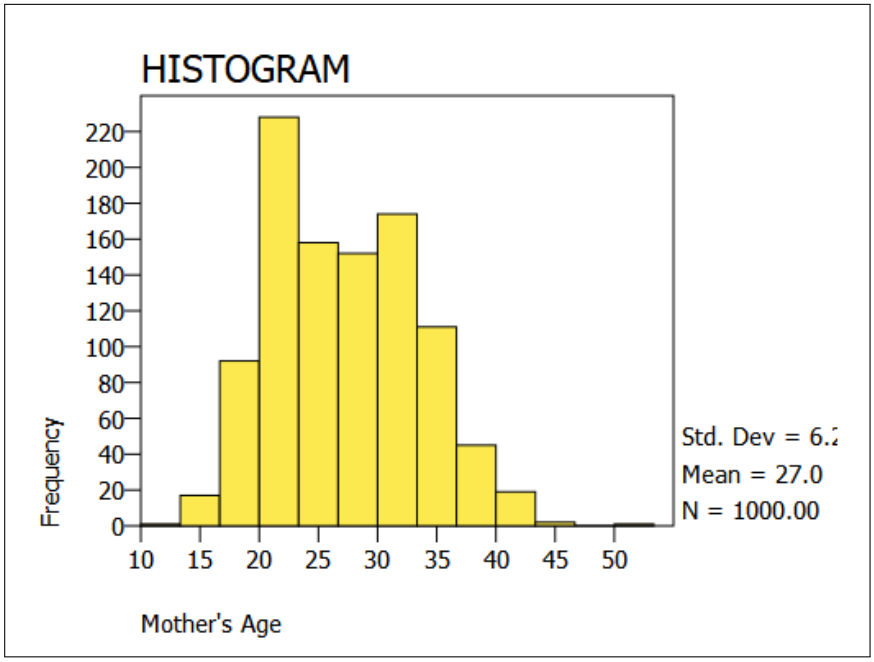


Figure 9: Histogram of Mother's Age

Notice that there is not a separate bar for each age; rather, PSPP has clustered three years into the same bar. Thus, there is a bar that combines 20-22 and not separate bars for 20, 21, and 22.

As another example, Figure 10 shows a histogram for baby's weight from the *births* dataset.[1]

1 Using a histogram aids a researcher in determining if a dataset is normally distributed and skewed. Figure 10 shows a normally distributed dataset since there is a clear peak in the middle trailing off on both sides. It also shows a negative skew since the tails on the left side of the peak are longer. Lab 1.2.2.2 on page 8 discusses the shape of a normal distribution.

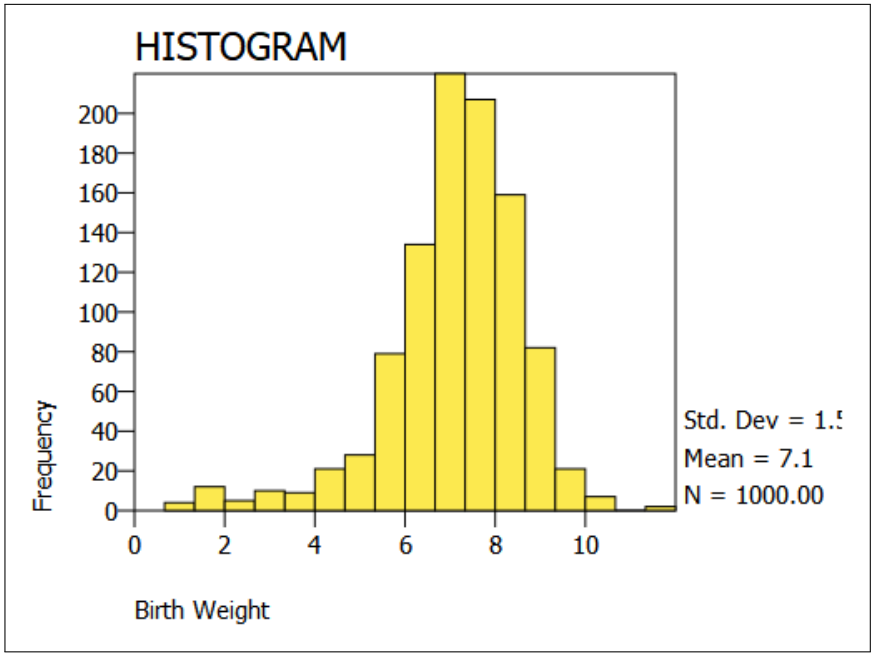


Figure 10: Histogram of Baby's Weight

As in Figure 9, each bar represents a range of weights so any weight between 6 and 6.33 pounds is clustered in a single bar.

2.2.2.2 *Bar Chart*

A bar chart is used to display the frequency count for ordinal or nominal data. There are technical differences between a bar chart and a histogram but for the purposes of this lab manual they can be considered identical displays for different types of data. Figure 11 is a bar chart showing the prevalence of various drive trains in the *cars* dataset.

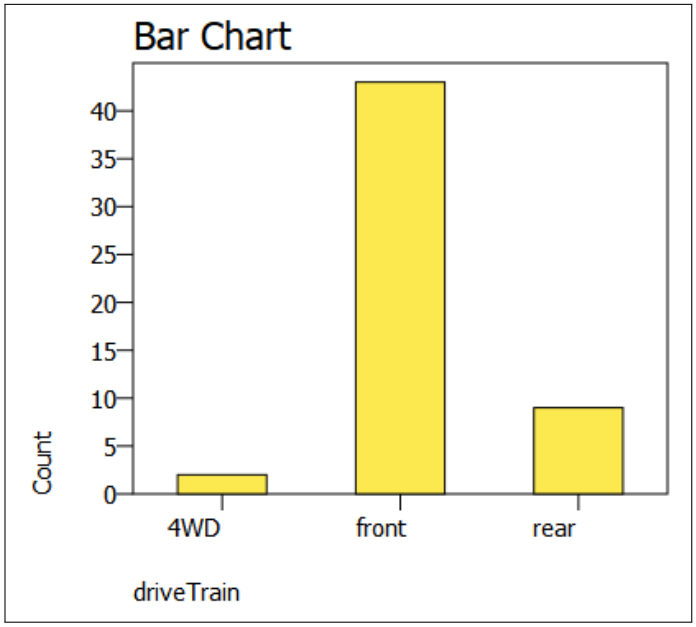


Figure 11: Prevelance of Types of Drive Trains

Figure 12 shows the maturity level of the mothers in the *births* dataset and unsurprisingly indicates that most mothers are younger.

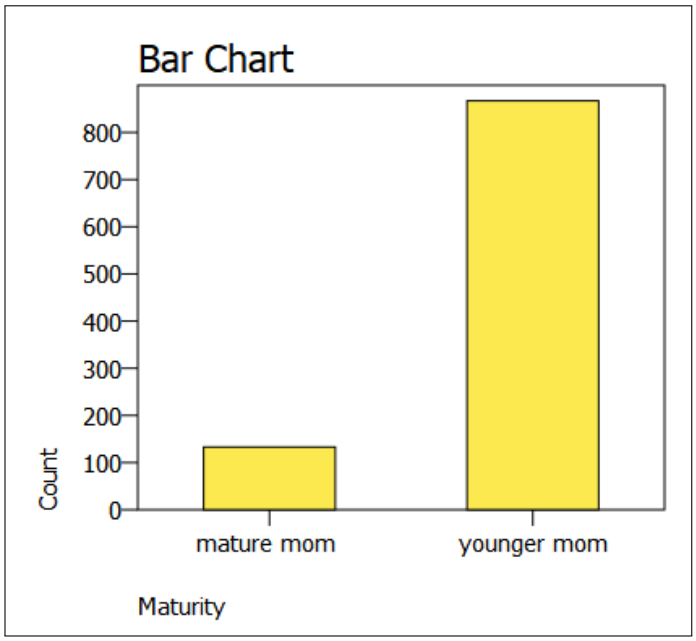


Figure 12: Maturity of Mothers

2.2.2.3 *Clustered Bar Chart*

A clustered bar chart displays two or more variables and is used to display ordinal or nominal data. In general, clustered bar charts can

be difficult to interpret and should be avoided. Figure 13 is a clustered bar chart that shows the incidence of premature births by the mother's smoking habit in the *births* dataset.

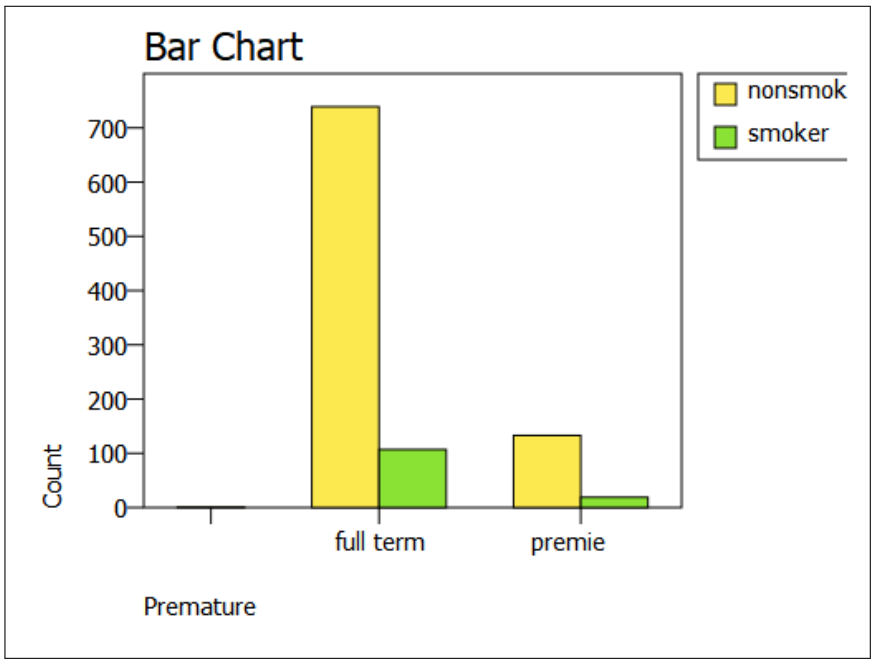


Figure 13: Permature Births By Smoking Habit

Figure 14 illustrates the problem with a clustered bar chart. This is a chart that shows the number of passengers for each type of car in the *cars* dataset. Notice that no large cars have four or five passengers and no small cars have six passengers so those bars are missing and that can make the chart difficult to interpret.

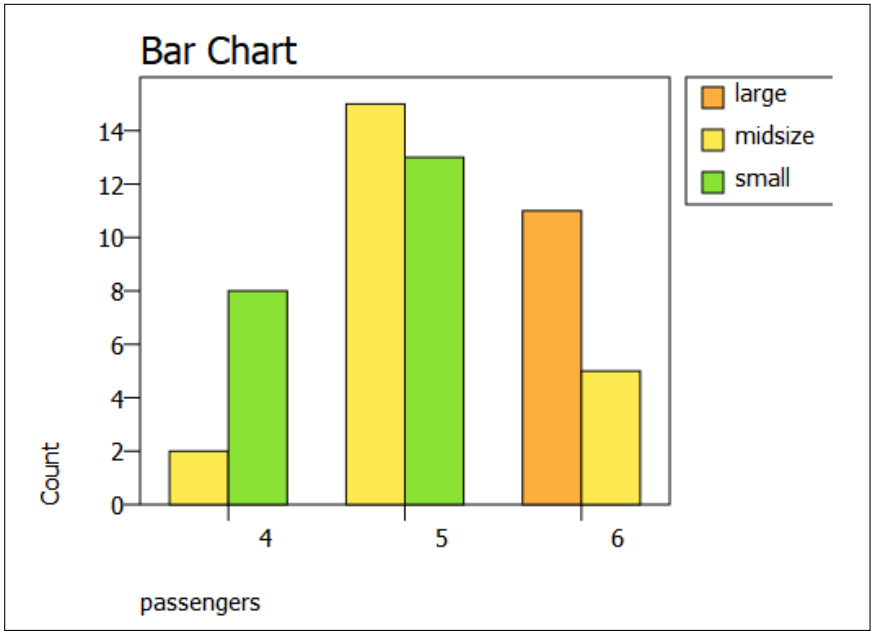


Figure 14: Number of Passengers By Car Type

2.2.2.4 *Pie Chart*

A pie chart is commonly used to display nominal or ordinal data; however, pie charts are notoriously difficult to understand, especially if the writer uses some sort of 3-D effect or "exploded" slices. The human brain seems able to easily compare the *heights* of two or more bars, as in histograms and bar charts, but the *areas* of two or more slices of a pie chart are difficult to compare. For this reason, pie charts should be avoided in research reports. If they are used at all, they should only illustrate one slice's relationship to the whole, not comparing one slice to another; and no more than four or five slices should be presented on one chart.

Figure 15 shows the types of numbers found in messages in the *email* dataset. This pie chart is easy to interpret since there are only three slices and each slice is easy to compare to the whole.

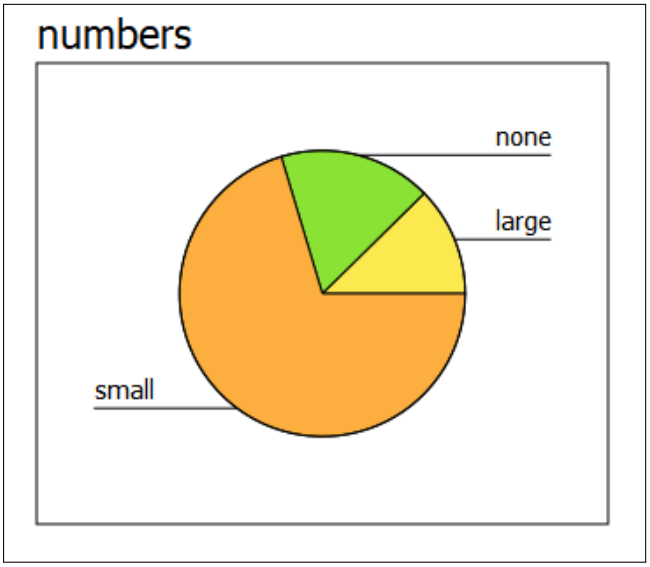


Figure 15: Types of Numbers in Email Messages

As an extreme example of a poorly used pie chart, consider Figure 16. Even ignoring the problem of the numbered labels overlapping, making them impossible to read, the slices are so numerous and small that it is impossible to differentiate between them. For example, the "one" and "two" slices are impossible to compare. For this pie chart, about all that can be stated is that most email messages have zero dollar signs.

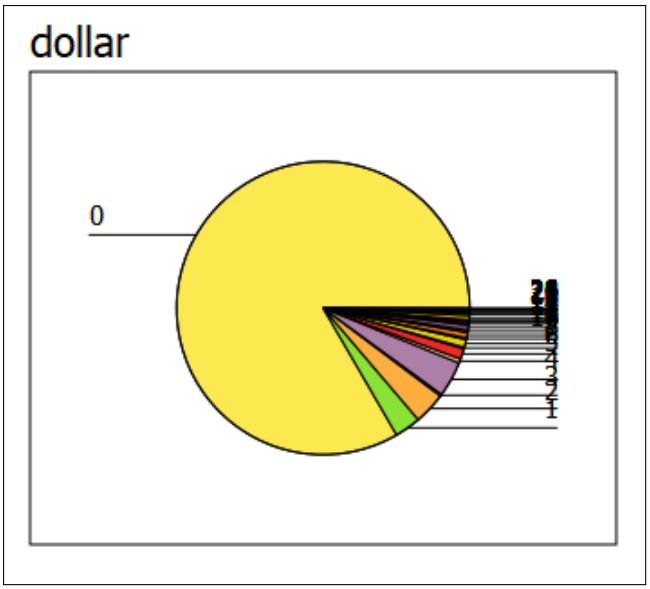


Figure 16: Number of Times a Dollar Sign Used in Email Messages

2.3 PROCEDURE

2.3.1 *Frequency Table*

Start PSPP and open the *email* dataset, then:

1. Click Analyze → Descriptive Statistics → Frequencies
2. Click the word *format* in the left column and then click the right-arrow button near the center of the window to move *format* to the "Variables" box on the right side of the window. (Alternatively, double-click the word *format* in the left column to move it to the "Variables" box.)
3. Uncheck all "Statistics" options in the lower-right box.
4. It is safe to explore the "Charts" and "Frequency Tables" options but do not select any of those options, they will be demonstrated later.
5. Click OK to generate the frequency table.

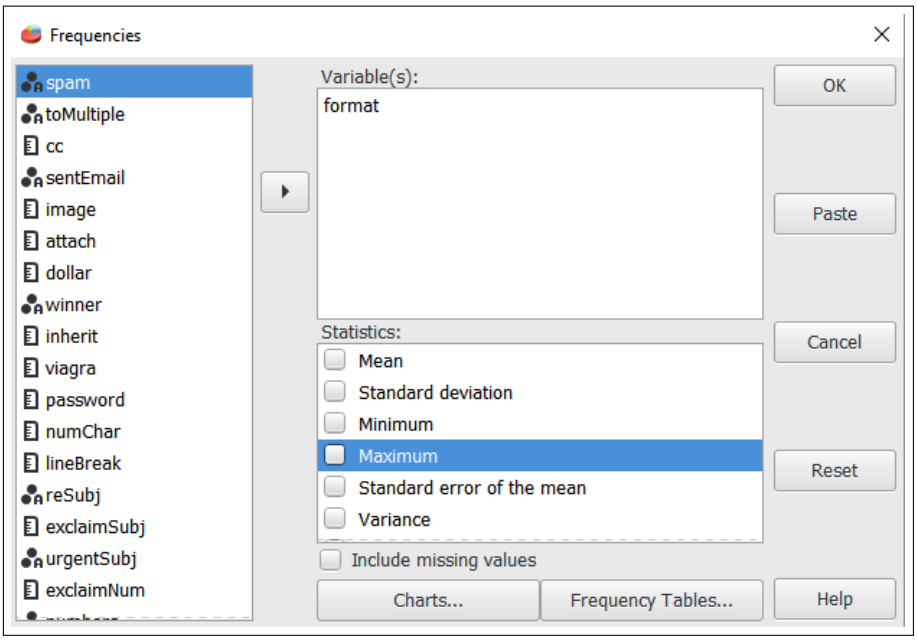


Figure 17: Generating a Frequency Table

format

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
	html	2011	63.76	63.76	63.76
	text	1143	36.24	36.24	100.00
	Total	3154	100.0	100.0	

Figure 18: Email Format Frequency Table

2.3.1.1 *Activity 1: Frequency Table*

Using the *births* dataset, produce a frequency table for *gender*.

2.3.1.2 *Activity 2: Frequency Table*

Using the *cafe* dataset, produce a frequency table for *meal*.

2.3.2 *Histogram*

Start PSPP and open the *bdims* dataset, then:

1. Click GRAPHS → HISTOGRAM [2]
2. Click the word *age* in the left column and then click the right-arrow button near the center of the window to move *age* to the "Variable" box on the right side of the window. (Alternatively, double-click the word *age* in the left column to move it to the "Variable" box.)
3. Click OK to generate the histogram.

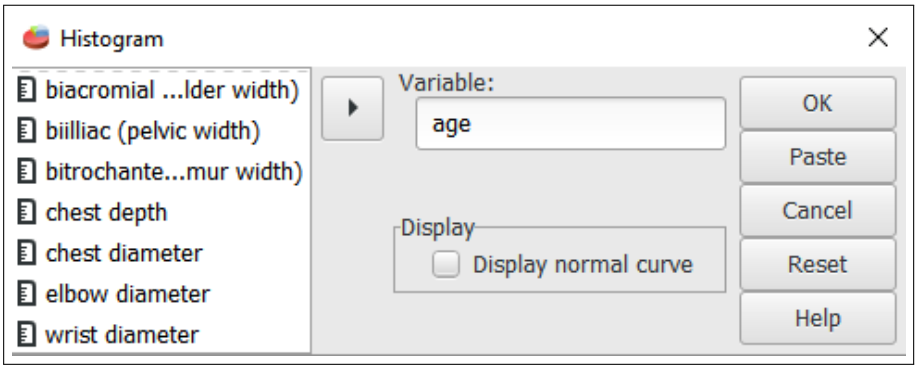


Figure 19: Generating a Histogram

2 Histograms can also be specified as an optional chart when creating a Frequency Table.

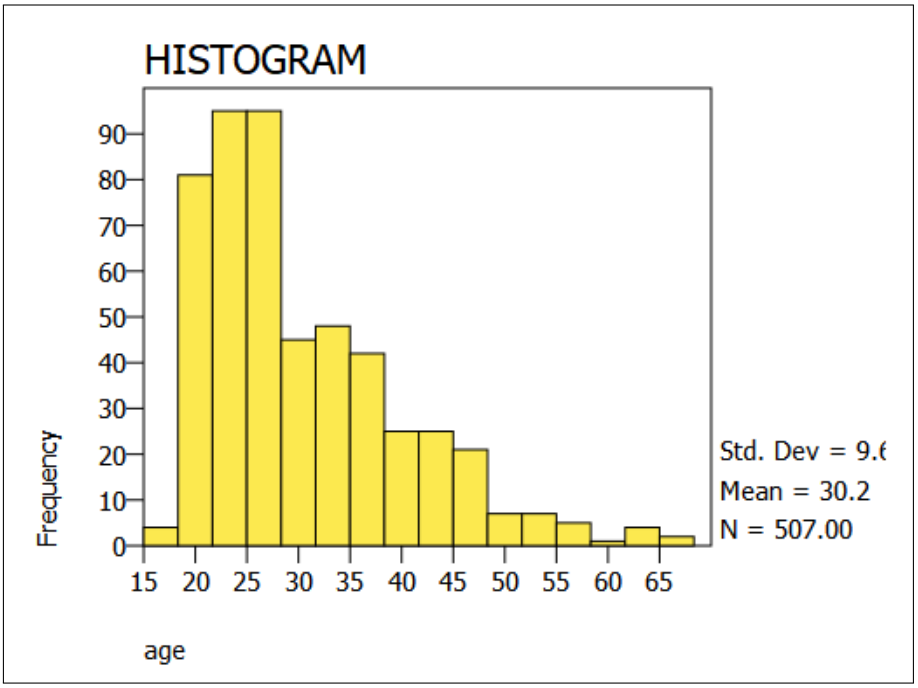


Figure 20: Age Histogram

2.3.2.1 *Activity 3: Histogram*

Using the *email* dataset, produce a histogram for *numChar* (the number of characters, in thousands, in the email message).

2.3.2.2 *Activity 4: Histogram*

Using the *cafe* dataset, produce a histogram for *age*.

2.3.3 *Bar Chart*

Start PSPP and open the *cars* dataset, then:

1. Click GRAPHS → BARCHART[3]
2. Click the word *passengers* in the left column and then click the right-arrow button near the center of the window to move *passengers* to the "Category Axis" box on the right side of the window.
3. Click OK to generate the bar chart.

3 Bar charts can also be specified as an optional chart when creating a Frequency Table.

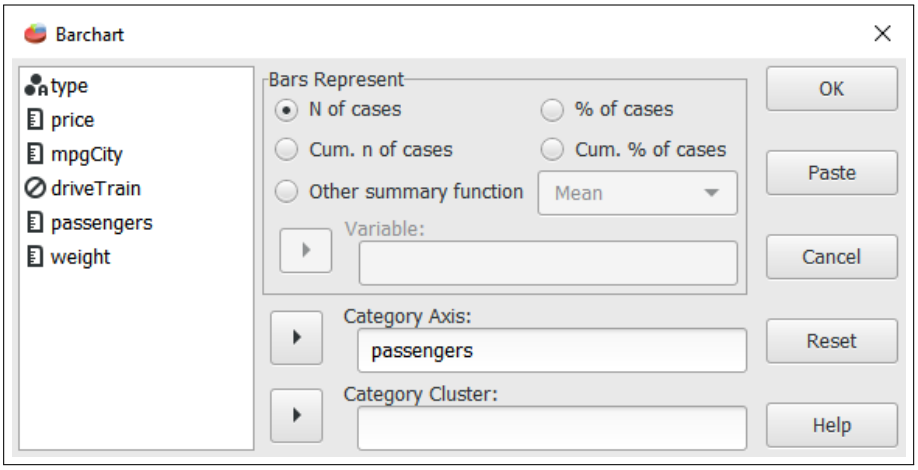


Figure 21: Generating a Bar Chart

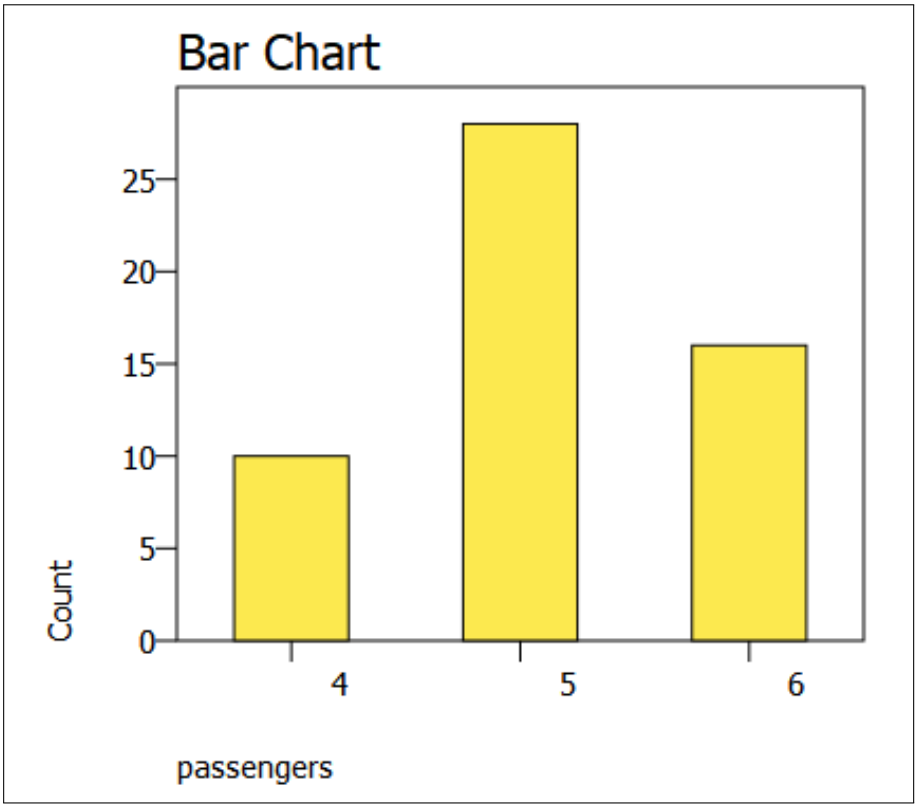


Figure 22: Passengers Bar Chart

2.3.3.1 *Activity 5: Bar Chart*

Using the *email* dataset, produce a bar chart for *numbers*.

2.3.3.2 *Activity 6: Bar Chart*

Using the *cafe* dataset, produce a barchart for *meal*.

2.3.4 *Clustered Bar Chart*

Start PSPP and open the *births* dataset, then:

1. Click GRAPHS → BARCHART
2. Click the word *premie* in the left column and then click the right-arrow button near the center of the window to move *premie* to the "Category Axis" box on the right side of the window.
3. Click the word *mature* in the left column and then click the right-arrow button near the bottom of the window to move *mature* to the "Category Cluster" box on the right side of the window.
4. Click OK to generate the bar chart.

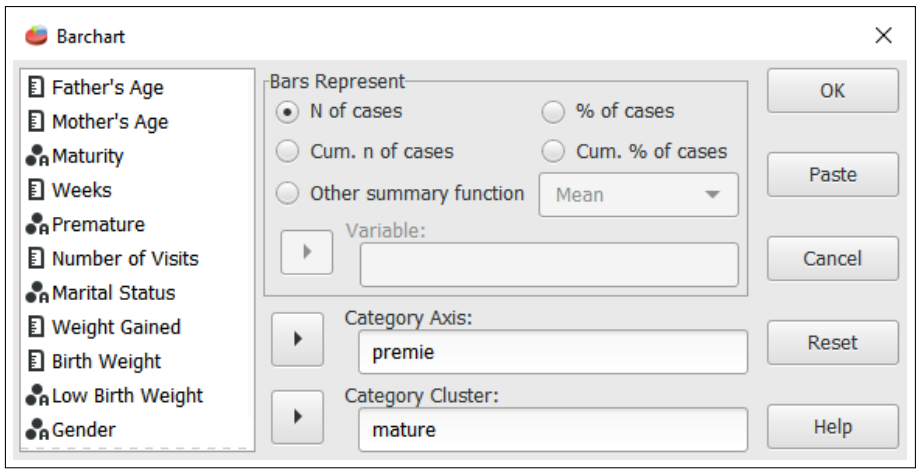


Figure 23: Generating a Clustered Bar Chart

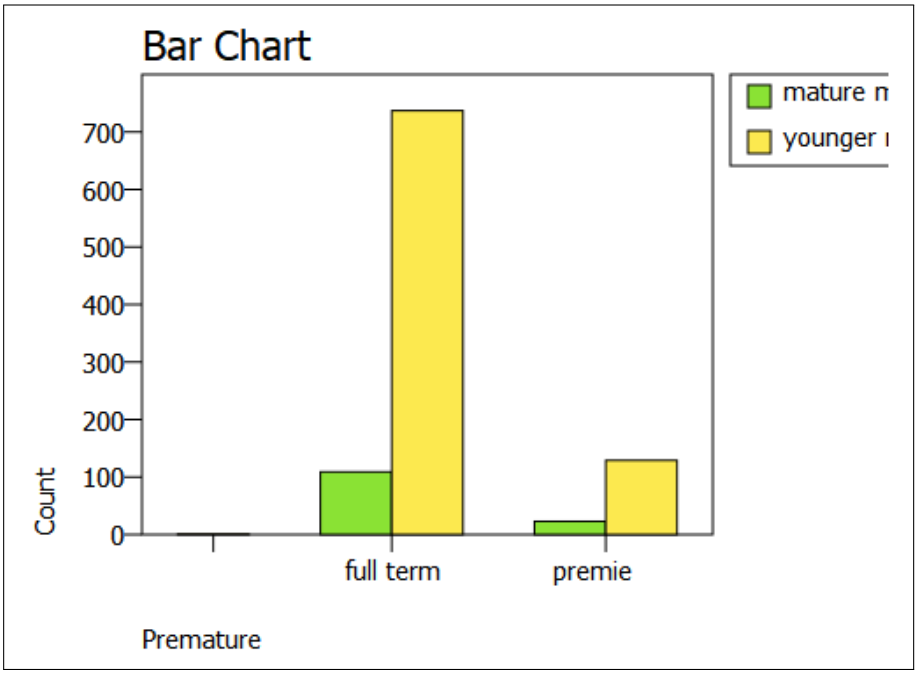


Figure 24: Clustered Bar Chart

2.3.4.1 *Activity 7: Clustered Bar Chart*

Using the *cars* dataset, produce a clustered bar chart using *drive train* by *type*.

2.3.4.2 *Activity 8: Clustered Bar Chart*

Using the *cafe* dataset, produce a clustered bar chart for *meal* by *sex*.

2.3.5 *Pie Chart*

Start PSPP and open the *cars* dataset, then:

1. Click Analyze → Descriptive Statistics → Frequencies
2. Click the word *driveTrain* in the left column and then click the right-arrow button near the center of the window to move *driveTrain* to the "Variable(s)" box on the right side of the window.
3. Uncheck all of the "Statistics" boxes in the lower-right text box.
4. Click Charts to open the charts options window.
5. Select "Draw pie charts"
6. Click Continue
7. Click OK to generate the pie chart.

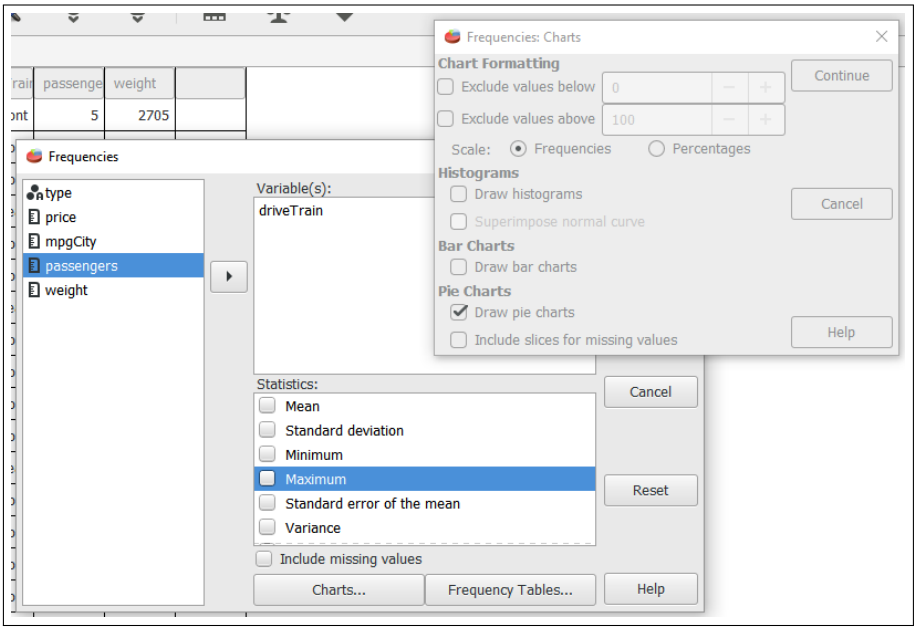


Figure 25: Generating a Pie Chart

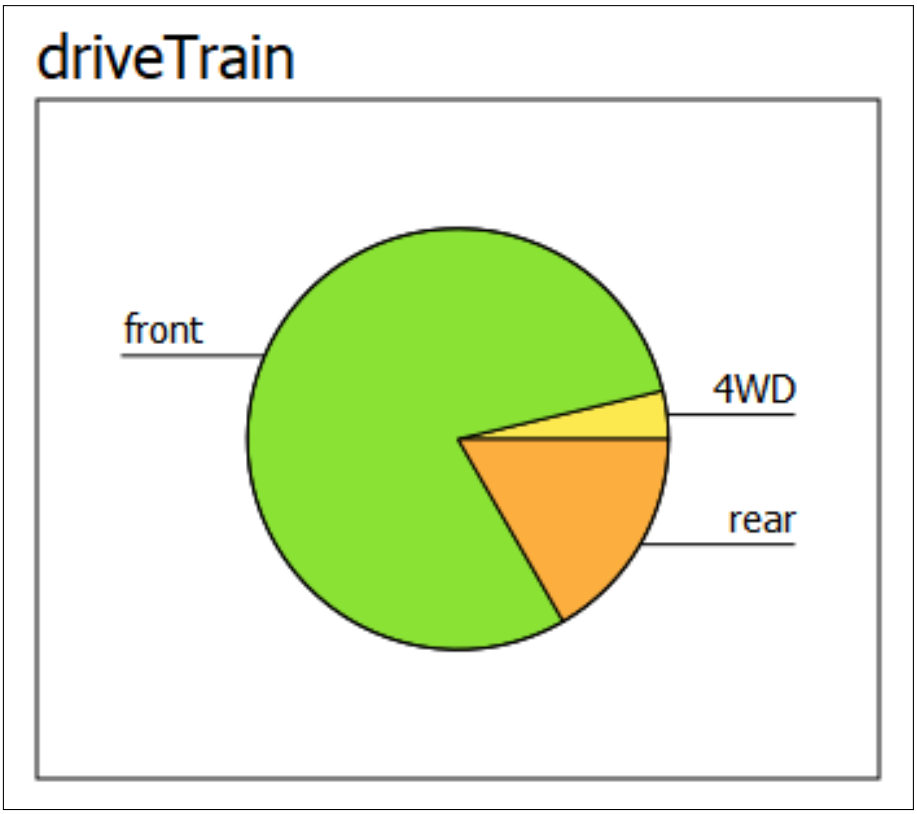


Figure 26: Pie Chart

2.3.5.1 *Activity 9: Pie Chart*

Using the *cars* dataset, produce a pie using *passengers*.

2.3.5.2 *Activity 10: Pie Chart*

Using the *cafe* dataset, produce a pie chart for *meal* by *sex*.

2.4 DELIVERABLE

Complete the following activities in this lab:

Consolidate the responses for all activities into a single document and submit that document for grading.

3

LAB 03: COMMON DESCRIPTIVE MEASURES

3.1 INTRODUCTION

One of the goals of descriptive statistics is to summarize and characterize the data so data scientists can determine its value and whether further research is needed. Often a simple measure like comparing the mean and median for a dataset can indicate a problem with outliers or skew. There are many statistics that are reported as descriptives and this lab both defines and demonstrates how those statistics are generated with PSPP.[1]

3.2 DISCUSSION

Each of the following sections describe one commonly-used statistical measure.

3.2.1 *Hinges*

Occasionally, the phrase "Tukey's Hinges" appears in statistical literature. The two hinges for a dataset are the medians for the lower half and the upper half of the data, but those halves also include the dataset median. For the simple dataset above, the lower hinge is the median of 5, 7, 10, and 13, or 8.5. The upper hinge is the median of 13, 17, 19, and 23, or 18. Quartiles and hinges usually have about the same accuracy but quartiles are more commonly used.

3.2.2 *Inter-Quartile Range*

Another measure that is occasionally used is the Inter-Quartile Range (*IQR*); that is, the difference between *Q1* and *Q3*. This is used to produce a better range for an element in a dataset that includes extreme outliers.

3.2.3 *Kurtosis*

One way to mathematically describe a normal distribution is to calculate the length of the tails of a bell curve, and that is called its *excess kurtosis*. For a normal distribution the excess kurtosis is 0.00, a

1 The information about calculated statistics in this lab is dependent on the datatypes discussed in Lab 1.

positive excess kurtosis would indicate longer tails while a negative excess kurtosis would indicate shorter tails. Intuitively, many people believe the excess kurtosis represents the "peaked-ness" of the curve since longer tails would tend to lead to a more peaked graph; however, excess kurtosis is a measure of the data outliers, which would be only present in the tails of the graph; so excess kurtosis is not directly indicative of the the "sharpness" of the peak. It is difficult to categorically state that some level of excess kurtosis is good or bad. In some cases, data that form a graph with longer tails are desired but in other cases they would be a problem.

Following are three examples of excess kurtosis. Notice that as the excess kurtosis increases the tails become longer.

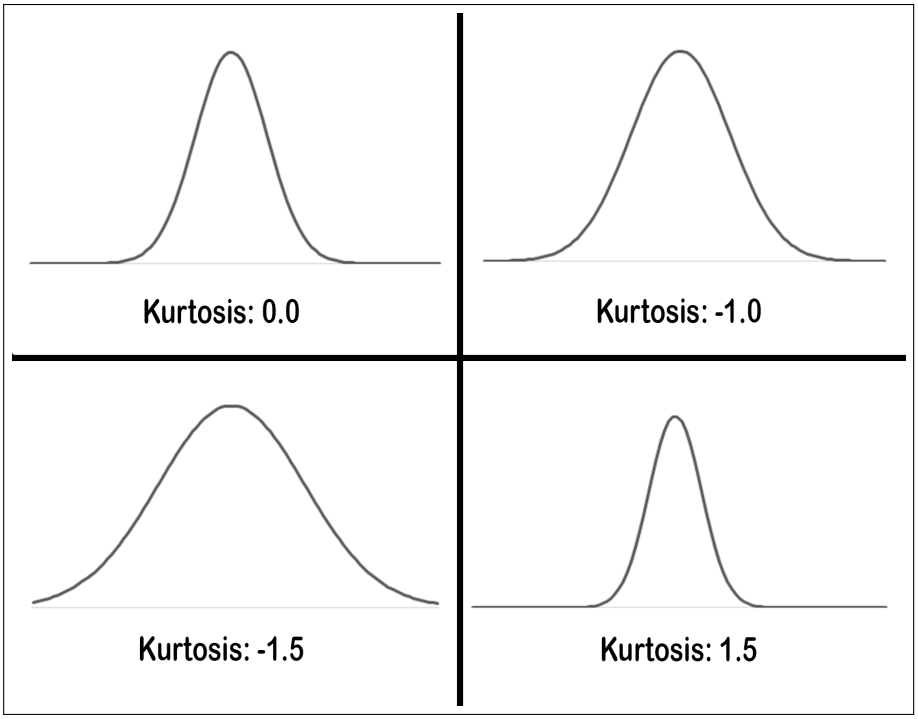


Figure 27: Kurtosis in a Normal Distribution

3.2.4 *Mean*

The mean is calculated by adding all of the data items together and then dividing that sum by the number of items, which is taught in elementary school as the *average*. For example, given the dataset: $6, 8, 9$, the total is 23 and that divided by 3 (the number of items) is 7.66; so the mean of $6, 8, 9$ is 7.66.

If a dataset has outliers, or values that are unusually large or small, then the mean is often skewed such that it no longer represents the "average" value. As an example, the length (in miles) of the 141 longest rivers in North America ranges from 135 to 3710 and the mean

of these values is 591.18 miles[2]. Unfortunately, because the lengths of the top few rivers are disproportionately higher than the rest of the values in the dataset (their lengths are *outliers*), the mean is skewed upward. One way compensate for outliers is to use a *trimmed mean* (sometimes called a *truncated mean*). A trimmed mean is calculated by removing a specified number of values from both the top and bottom of the dataset and then finding the mean of the remaining values. In the case of the rivers dataset, if 5% of the values are trimmed from the data (or 2.5% from both the top and bottom) then the remaining items create a "trimmed" mean of 518.79. Trimming the dataset effectively removes both upper and lower outliers and produces a much more reasonable central value for this dataset. In actual practice, a trimmed mean is not commonly used since it is difficult to know how much to trim from the dataset and the resulting mean may be just as skewed as if no values were trimmed; thus, when outliers are suspected, the best "middle" term to report is the median.

3.2.5 *Median*

The median is found by listing all of the data items in numeric order and then mechanically finding the middle item. For example, using the dataset $6, 8, 9$, the middle item (or median) is 8. If the dataset has an even number of items, then the median is calculated as the mean between the two middle items. For example, in the dataset $6, 8, 9, 13$ the median is 8.5, which is the mean of 8 and 9, the two middle terms.

The median is very useful in cases where the dataset has outliers. As an example of using a median rather than a mean, consider the dataset $5, 6, 7, 8, 30$. The mean is $(5 + 6 + 7 + 8 + 30)/5 = 56/5 = 11.2$. However, 11.2 is clearly much higher than most of the other numbers in that dataset since one outlier, 30, is significantly driving up the mean. A much better representation of the central term for this dataset would be 7, which is the median. To re-visit the river lengths introduced above, the median of the dataset is 425, which is much more representative of the "middle" length than using either the mean or the trimmed mean.

As another example, suppose a newspaper reporter wanted to find the "average" wage for a group of factory workers. The ten workers in that factory all have an annual salary of $\$25,000$; however, the supervisor has a salary of $\$125,000$. In the newspaper article, the supervisor is quoted as saying that the employees in his company have an average salary of $\$34,090$. That is correct if the mean of all those salaries is reported, but that number is clearly higher than any sort of reasonable "average" salary for workers in the factory due to the one outlier (the supervisor's salary). In this case, the median of $\$25,000$ would be much more representative of the "average" salary.

2 These data are found in the *rivers* dataset.

The median is typically reported for salaries, home values, and other datasets where one or two outliers would significantly distort the reported "middle" value.

If the dataset contains no outliers and is normally distributed[3], then the mean and median are the same; but if there are outliers then these two measures become separated, often by a large amount. Consider the rivers dataset mentioned in the *Mean* section above. That dataset has a mean of 591.18 and a median of 425. This difference, 166.18, is about 28% of the mean and is significant. The size of this difference would tell a researcher that there are outliers or other influences that are skewing the data.

3.2.6 *Minimum/Maximum*

The minimum and maximum values of an element in a dataset are, as the name implies, the smallest and largest values. As an example, the dataset $6, 7, 8, 9, 10$ has a minimum of 6 and a maximum of 10. The *rivers* dataset has a minimum of 135 and a maximum of 3710.

3.2.7 *Mode*

The mode is used to describe the center of nominal or ordinal data and is nothing more than the value that is most often found in the dataset. For example, if a question asked respondents to select their zip code from a list of five local codes and "12345" was selected more often than any other then that would be the mode for that item. Calculating the mode is no more difficult than counting the number of times the various values are found in the dataset and reporting the value found most frequently.

As an example, the *cars* dataset includes the following types of drive trains:

Type	Frequency
4WD	2
Front	43
Rear	9

Since the most common type of drive train is "Front" that would be the mode for this data item.

It does not make much sense to calculate the mean or median for nominal or ordinal data since those are categories; however, reports frequently contain the mean for Likert-style questions (ordinal data) by equating each level of response to a number and then calculating the mean of those numbers. For example, imagine that a stu-

3 The Normal Data Distribution, along with the terms skew and kurtosis, is covered in Lab 1.

dent housing survey asked respondents to select among "Strongly Disagree, Disagree, Neutral, Agree, Strongly Agree" for a statement like "I like the food in the cafeteria." That is clearly ordinal data and while "Agree" is somehow better than "Disagree" it would be wrong to try to quantify that difference as "one point better." Sometimes, though, researchers will assign a point value to those responses like "Strongly Disagree" is one point, "Disagree" is two points, and so forth. Then they will calculate the mean for the responses on a survey item and report something like "The question about the food in the cafeteria had a mean of 3.24." It would be impossible to know what that means. Are students 0.24 units above "Neutral" on liking the cafeteria food? Thus, the mean or median should not be reported for nominal or ordinal data.

3.2.8 *N*

One of the simplest of measures is nothing more than the number of items in a dataset. For example, the dataset $5, 7, 13, 22$ contains 4 items. In statistics, the number of items in a dataset is usually represented by the letter N, therefore, in the simple dataset in this paragraph, $N = 4$.

3.2.9 *Quartiles*

A measure that is closely related to the median is the first and third quartile. The first quartile (Q_1) is the score that splits the lowest 25% of the values from the rest and the third quartile (Q_3) splits the highest 25% of the values from the rest. The second quartile (Q_2) is the same as the median and, normally, the term "median" is used rather than Q_2. For example, consider this dataset:

$$5, 7, 10, 13, 17, 19, 23$$

The median of this dataset is 13 because three values are smaller and three are larger. The first quartile is 7, which is the median for the lower half of the values (not including 13, the median of the dataset); or the score that splits the lowest 25% from the rest of the data. The third quartile is 19, which is the median for the upper half of the scores; or the score that splits the highest 25% from the rest of the data.

3.2.10 *Range*

The range of an element in a dataset is the maximum value minus the minimum value. As an example, the dataset $6, 7, 8, 9, 10$ has a range of $10 - 6$ or 4. The *rivers* dataset has a range of 3575.

3.2.11 *Skew*

The second numerical measure of a normal distribution that is frequently reported is its *skew*, which is a measure of the symmetry of the curve about the mean of the data. The normal distribution in Figure **??** has a skew of 0.00. A positive skew indicates that the tail on the right side is longer, which means that there are several data points on the far right side of the graph "pulling" the tail out that direction. A negative skew indicates that the tail on the left side of the graph is longer. Following are three examples of skew:

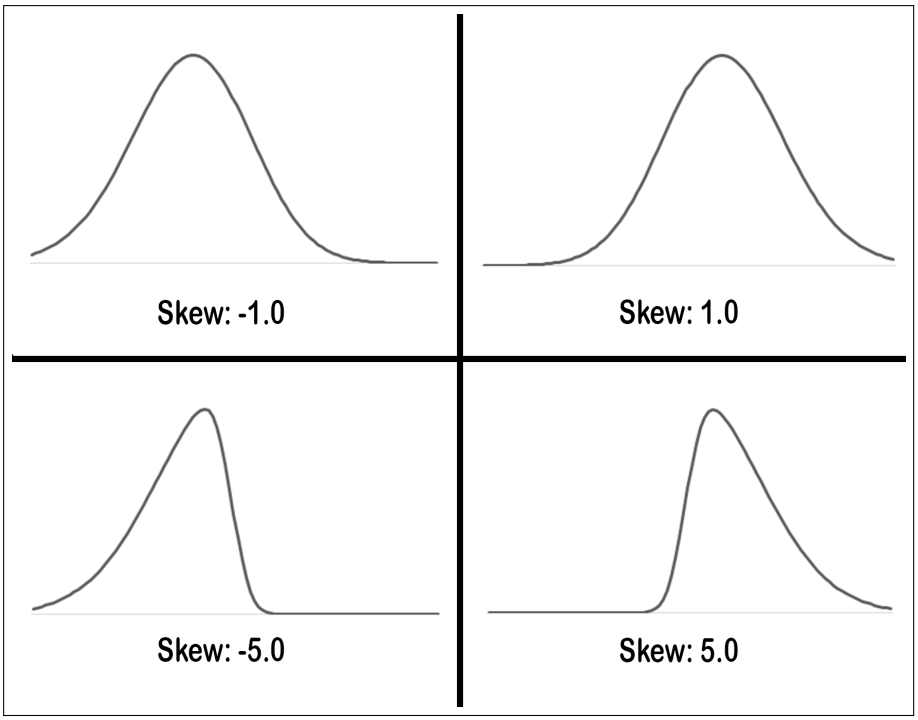


Figure 28: Skew in a Normal Distribution

3.2.12 *Standard Deviation*

The standard deviation of a dataset is a number that indicates how much variation there is in the data; or how "scattered" the data are from the mean. In general, the larger the standard deviation then the more variation there is in the data. A dataset with a small standard deviation would create a sharply peaked normal distribution curve while a large standard deviation would create a flatter curve.[4]

Once a standard deviation is calculated, then about 68.2% of the samples will lie closer to the mean than that number. To put it another way, one standard deviation explains about 68.2% of the vari-

4 The concept of the normal distribution curve was presented in Lab 01.

ance from the mean. To show this concept graphically, consider the following graph of the scores on an examination:

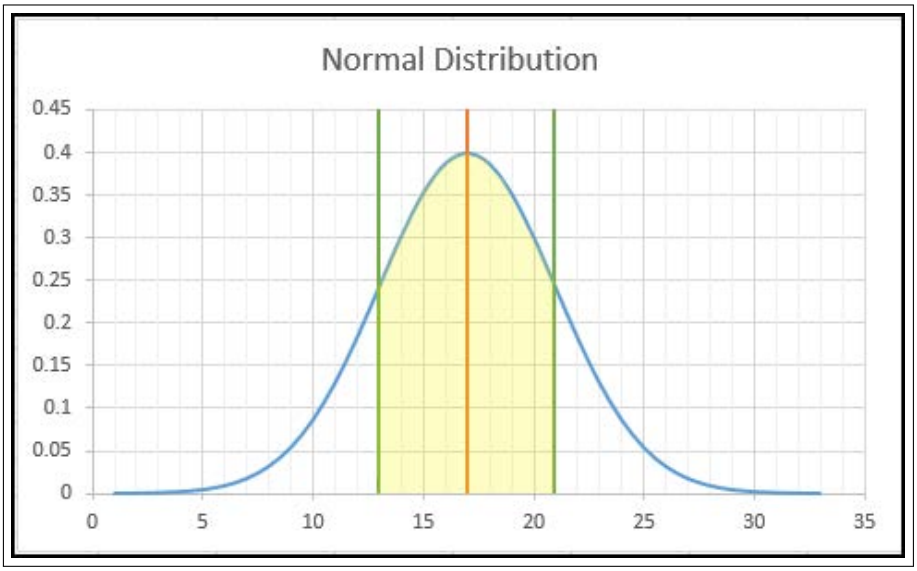


Figure 29: Illustration Of Standard Deviation

The mean of this distribution is marked with a vertical line in the center of the bell curve. One standard deviation up and one standard deviation down are marked by two other vertical lines. The shaded area under the curve would include about 68.2% of all scores for this dataset. In the same way, two standard deviations from the mean would include about 95.4% of the data points; and three standard deviations would include more than 99.7% of the data points (the second and third standard deviation are not indicated on the graph).

As one last example, imagine a class with 500 students where the professor administered an examination worth 100 points. If the mean score for that examination was 80 and the standard deviation was 5, then the professor would know that the scores were fairly tightly grouped (341 scores of the 500 (68.2%) were between $75 - 85$, within 5 points of the mean), and this would probably be good news for the professor. On the other hand, if the mean score was 60 and the standard deviation was 15, then the scores were "all over the place' (more precisely, 341 scores of the 500 were between 45-75), and that may mean that the professor would have to re-think how the lesson was taught or that the examination itself was flawed.

It is difficult to categorically state whether a specific standard deviation is good or bad; it is simply a measure of how concentrated the data are around the mean. For something like a manufacturing process where the required tolerance for the parts being produced is tight then the standard deviation for the weights of random samples pulled off of the line must be very small; that is, the parts must be as

nearly identical as possible. However, in another context, the standard deviation may be quite large. Imagine measuring the time it takes a group of high school students to run 100 yards. Some would be very fast but others would be much slower and the standard deviation for that data would likely be large.

3.2.13 *Standard Error*

If multiple samples are taken of a large population the mean of each of those samples will be slightly different. The "standard error of the mean" is the same as the standard deviation for a single sample. As an example, imagine that the following dataset are all of the test scores for 100 students who took the same departmental final examination last fall: 34, 44, 46, 46, 50, 52, 52, 53, 54, 54, 55, 56, 57, 58, 58, 58, 59, 60, 60, 62, 62, 62, 62, 64, 64, 64, 65, 65, 65, 66, 67, 68, 68, 68, 68, 69, 69, 69, 69, 70, 70, 70, 70, 70, 70, 71, 71, 71, 72, 72, 73, 73, 73, 73, 74, 75, 75, 75, 75, 75, 76, 77, 77, 77, 78, 78, 78, 78, 79, 79, 79, 79, 79, 79, 79, 80, 80, 80, 80, 81, 81, 82, 85, 86, 86, 86, 86, 86, 87, 87, 87, 90, 92, 93, 95, 95, 97, 100, 100, 100

Of course, with only 100 data points it would be very easy to calculate various statistics with the entire population, but that would not be possible if the sample were tens-of-thousands so a sample would be drawn at random from the entire population and it would be assumed that the sample would represent the entire population. Imagine that, at random, the following samples of ten were drawn from the population above:

Sample	Sum	Mean
50, 56, 64, 72, 72, 75, 79, 80, 80, 93	721	72.1
58, 69, 70, 75, 78, 79, 80, 81, 86, 95	771	77.1
46, 58, 66, 70, 73, 75, 79, 85, 95, 100	747	74.7
54, 60, 62, 66, 68, 70, 70, 71, 71, 73	665	66.5
69, 72, 73, 76, 77, 77, 79, 79, 79, 81	762	76.2

The mean of the randomly-drawn sample should be the same as the mean for the entire population. However, it would be possible that the sample included only the high scores and the mean for that sample would be far different from the mean for a sample of low scores. The table above lists five different means depending on the samples drawn.

The actual mean for this dataset is 72.14. The standard error of the mean is 4.12 so the means of samples of ten values drawn at random would be expected to fall between 59.78 and 84.50, which is three standard errors below and above the mean.

The standard error is a useful way to estimate how accurately the mean of a sample reflects the mean of the entire population.

3.2.14 *Standard Error of the Kurtosis*

Kurtosis is defined earlier in this lab. Like the mean of a randomly-drawn sample would be expected to differ from the mean of the entire population (expressed as the Standard Error), the kurtosis of a randomly-drawn sample would be expected to differ from the kurtosis of the entire population. That expected difference would be expressed as the Standard Error of the Kurtosis.

3.2.15 *Standard Error of the Skew*

Skew is defined earlier in this lab. Like the mean of a randomly-drawn sample would be expected to differ from the mean of the entire population (expressed as the Standard Error), the skew of a randomly-drawn sample would be expected to differ from the skew of the entire population. That expected difference would be expressed as the Standard Error of the Skew.

3.2.16 *Sum*

Another descriptor of a dataset that is occasionally reported is the sum, which is nothing more than the values of all of the items in an element added together. As an example, the dataset $6, 7, 8$ has a sum of 21. The *rivers* dataset has a sum of 83357. It should be rather obvious that the sum by itself does not offer much information without knowing the number of items in the dataset and the range of the values.

3.2.17 *Variance*

Statistically, the variance is a measure of how far a sample is spread out from the mean. A dataset with a wide variance would have values that are spread out "all over the place" while a dataset with a smaller variance would be more uniform. This is about the same definition that was used for Standard Deviation (page 38). In fact, there is a statistical relationship between variance and standard deviation: the standard deviation of a dataset is the square root of the variance. It is natural to wonder, then, which should be used. A standard deviation is measured with the same units as the dataset. So, for example, if some element of a dataset were measuring the height of all incoming students in inches then the standard deviation would also be in inches. Variance, on the other hand, is measured in square units so

it is more difficult to compare the variance with the raw data. To be blunt, the variance has great utility in advanced statistics classes where students are considering theory but for most undergraduate classes where statistics are being used as a tool rather than the object of the study then the standard deviation is all that is needed.

3.3 PROCEDURE

Labs Two through Six will generate most of the above measures so for this lab only a Frequency Table, as found in Lab Two, is used to generate a few simple statistics in this lab.

3.3.1 *Frequency Table with Statistics*

Start PSPP and open the *email* dataset, then:

1. Click Analyze → Descriptive Statistics → Frequencies
2. Click the word *image* in the left column and then click the right-arrow button near the center of the window to move *image* to the "Variables" box on the right side of the window. (Alternatively, double-click the word *image* in the left column to move it to the "Variables" box.)
3. Check *Mean*, *Minimum*, and *Maximum* "Statistics" options in the lower-right box.
4. Click OK to generate the frequency table.

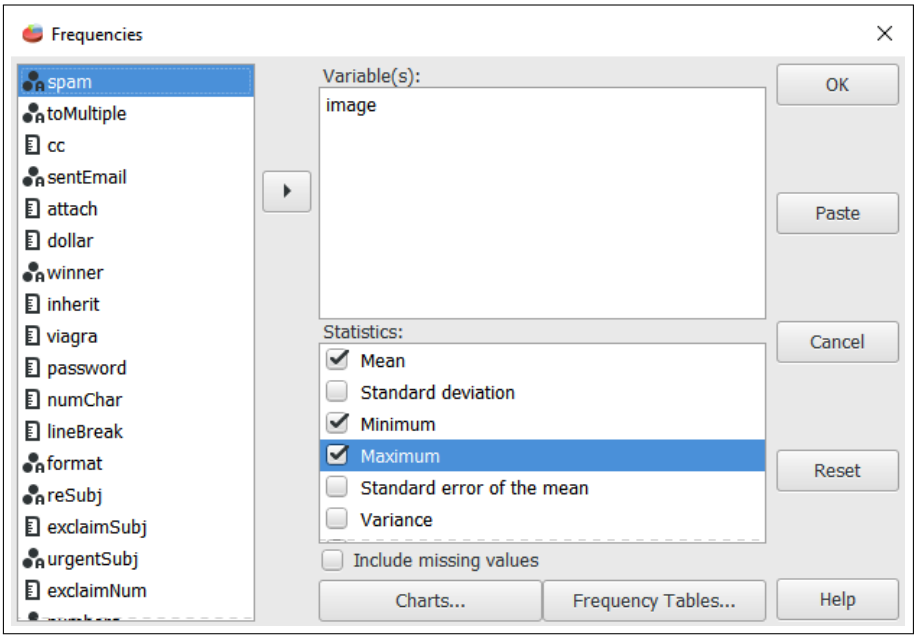
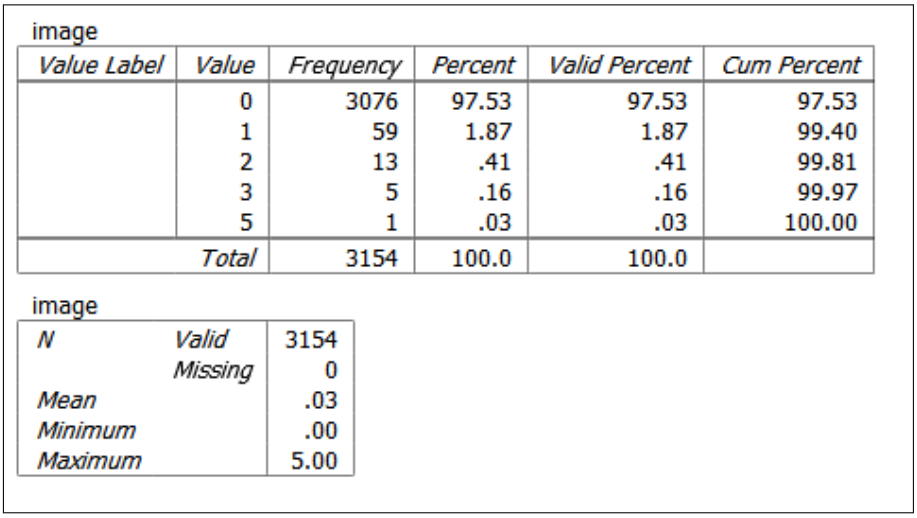


Figure 30: Generating a Frequency Table With Statistics



image

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
	0	3076	97.53	97.53	97.53
	1	59	1.87	1.87	99.40
	2	13	.41	.41	99.81
	3	5	.16	.16	99.97
	5	1	.03	.03	100.00
	Total	3154	100.0	100.0	

image

N	*Valid*	3154
	Missing	0
Mean		.03
Minimum		.00
Maximum		5.00

Figure 31: Email Image Frequency Table With Statistics

Figure 31 shows some simple statistics for the *image* data element.

Statics are only generated for Interval or Ratio data.

3.3.1.1 *Activity 1: Frequency Table With Statistics*

Using the *gifted* dataset, produce a frequency table for *eduTV* (the number of hours children spend watching educational television programs each week). Include the *mean*, *minimum*, and *maximum* values with the table.

3.3.1.2 *Activity 2: Frequency Table With Statistics*

Using the *cafe* dataset, produce a frequency table for *ptysize* (the size of the dining party). Include the *mean*, *minimum*, and *maximum* values with the table.

3.4 DELIVERABLE

Complete the following activities in this lab:

Number	Name	Page
3.3.1.1	Activity 1: Frequency Table With Statistics	43
3.3.1.2	Activity 2: Frequency Table With Statistics	44

Consolidate the responses for all activities into a single document and submit that document for grading.

4

LAB 04: DESCRIPTIVES]

4.1 INTRODUCTION

It is typical for a researcher to report a number of descriptive attributes of a dataset, like its mean or standard deviation and this lab describes how to generate that sort of statistical analysis.

4.2 DISCUSSION

4.2.1 *Descriptive Measures*

The descriptive measures commonly reported are[1]:

- Kurtosis
- Mean
- Maximum Value
- Minimum Value
- Range
- Skewness
- Standard Deviation
- Standard Error
- Sum
- Variance

4.2.2 *Z-Scores*

A Z-Score is how many standard deviations[2] a data point lies above or below the mean. In other words, a Z-Score re-scales, or standardizes, the data points so they can be more easily compared. Here are a few facts about Z-Scores:

- A z-score equal to 0 is a data point that is equal to the mean.

1 These measures are described in LAB 03: COMMON DESCRIPTIVE MEASURES on page 3.
2 Standard Deviation is described on page 38

- A z-score less than 0 is a data point that is less than the mean.
- A z-score greater than 0 is a data point that is greater than the mean.
- A z-score equal to +1 is a data point that is 1 standard deviation greater than the mean.
- A z-score equal to −1 is a data point that is 1 standard deviation less than the mean.

For example, imagine that a student took an exam and scored 15. Is the a good score? It would depend on the mean and standard deviation for that exam. The easiest way to check is to convert the student's score of 15 to a Z-Score, which accounts for both mean and standard deviation. If a score of 15 is converted to a Z-Score of +1 then the student scored higher than about 68% of the other students so that would be a good score.

Z-Scores can also be used to compare the results of two tests where the mean and standard deviation may be different. For example, imagine that a state changes the annual test that is given to fourth, sixth, and ninth graders. A particular student scored 40 on the math portion in the fourth grade and then scored 30 on the new text in the sixth grade. That would not necessary mean that the student is doing worse because the test may be harder. In order to compare those two test results a researcher would have to know the mean and standard deviation of the two tests. Again, converting the student's scores to Z-Scores would help determine whether the student is making progress. Imagine that for the fourth grade exam the student's Z-Score is −0.3, or slightly below the mean, but on the sixth grade test the student's Z-Score is +0.3, or slightly above the mean. Even though the raw scores seem to show a decrease in score the Z-Score reveals that the student is making above normal progress.

Calculating Z-Scores is a simple formula:

$$Z = \frac{X - mean}{SD}$$

Where X is a specific data point and SD is the standard deviation of the dataset. Thus, if a dataset has a mean of 50 and a standard deviation of 10 then the data point 55 has a Z-Score of +0.5:

$$+0.5 = \frac{55 - 50}{10}$$

4.3 PROCEDURE

4.3.1 *Descriptives*

Start PSPP and open the *bdims* dataset, then:

1. Click Analyze → Descriptive Statistics → Descriptives
2. Click the phrase *navel (abdominal) girth* in the left column and then click the right-arrow button near the center of the window to move it to the "Variables" box on the right side of the window. (Alternatively, double-click the phrase *navel (abdominal) girth* in the left column to move it to the "Variables" box.) NOTE: This is displayed as "nav.gi" in the "Variables" box since that is the actual name of the variable.
3. Check the following "Statistics" options in the lower-right box: Mean, Standard Deviation, Minimum, and Maximum.
4. Uncheck all options at the bottom of the window.
5. Click OK to generate the descriptives.

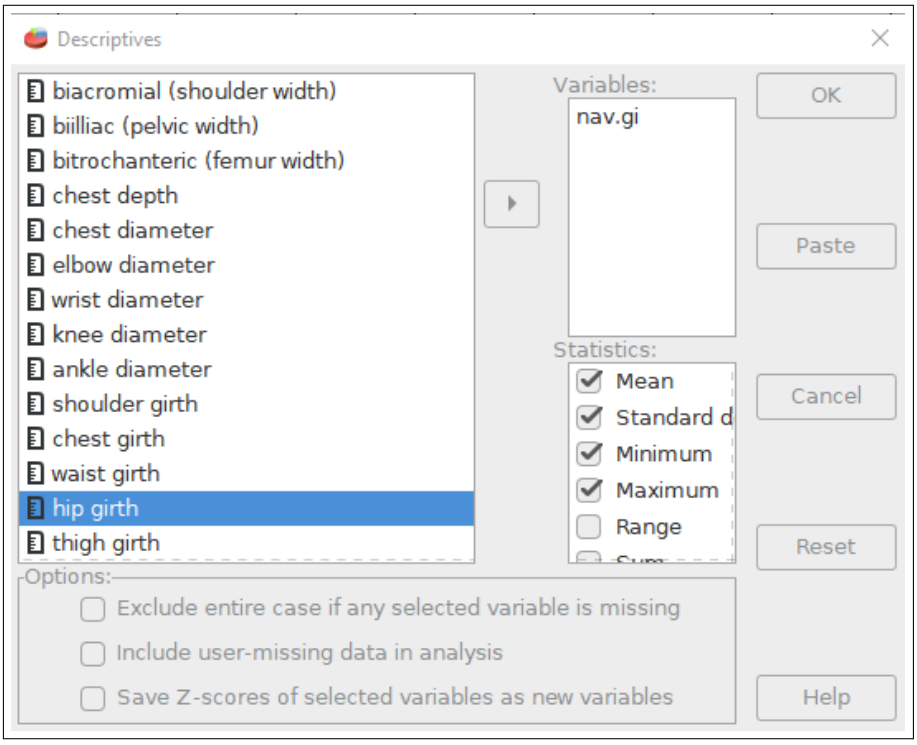


Figure 32: Generating Descriptives

```
DESCRIPTIVES
  /VARIABLES= nav.gi.

Valid cases = 507; cases with missing value(s) = 0.
```

Variable	*N*	*Mean*	*Std Dev*	*Minimum*	*Maximum*
navel (abdominal) girth	507	85.65	9.42	64.00	121.10

Figure 33: Navel Girth Descriptives

4.3.1.1 *Activity 1: Descriptives*

Using the *bdims* dataset, produce the descriptives for *age*. Include the mean, standard deviation, minimum, and maximum statistics.

4.3.1.2 *Activity 2: Descriptives*

Using the *cafe* dataset, produce the descriptives for *age*. Include the mean, standard deviation, minimum, and maximum statistics.

4.3.2 *Z-Scores*

Start PSPP and open the *gifted* dataset, then:

1. Click Analyze → Descriptive Statistics → Descriptives
2. Click the word *motherIQ* in the left column and then click the right-arrow button near the center of the window to move it to the "Variables" box on the right side of the window. (Alternatively, double-click the word *motherIQ* in the left column to move it to the "Variables" box.)
3. Check the following "Statistics" options in the lower-right box: Mean, Standard Deviation, Minimum, and Maximum.
4. Check the "Save Z-scores of selected variables as new variables" option at the bottom of the window.
5. Click OK to generate the descriptives.

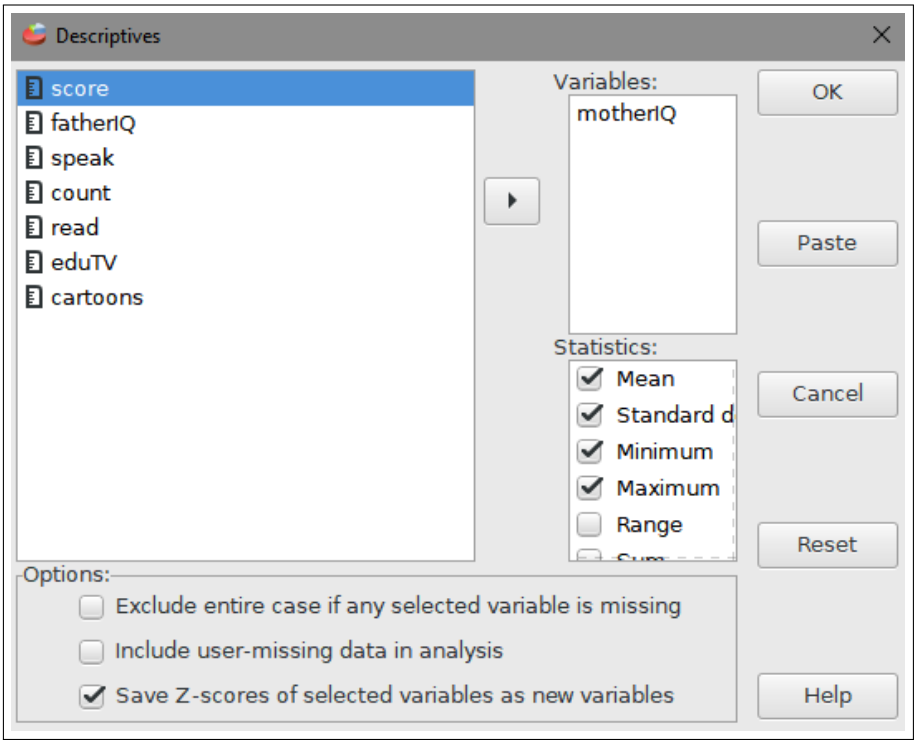


Figure 34: Generating Descriptives

```
DESCRIPTIVES
  /VARIABLES= motherIQ
  /SAVE.
```

Mapping of variables to corresponding Z-scores.

Source	*Target*
motherIQ	ZmotherIQ

Valid cases = 36; cases with missing value(s) = 0.

Variable	*N*	*Mean*	*Std Dev*	*Minimum*	*Maximum*
motherIQ	36	118.17	6.50	101.00	131.00

Figure 35: Descriptives for Mother's IQ

After the Descriptives are generated the Z-Scores for Mother's IQ are calculated and added to a new field in the dataset.

Case	score	fatherIQ	motherIQ	speak	count	read	eduTV	cartoons	ZmotherI
1	159	115	117	18	26	2	3.00	2.00	-.18
2	164	117	113	20	37	3	1.75	3.25	-.79
3	154	115	118	20	32	2	2.75	2.50	-.03
4	157	113	131	12	24	2	2.75	2.25	1.97
5	156	110	109	17	34	2	2.25	2.50	-1.41
6	150	113	109	13	28	2	1.25	3.75	-1.41
7	155	118	119	19	24	2	2.00	3.00	.13

Figure 36: Z-Scores for Mother's IQ

The Z-Scores are in the last column on the right. For example, case one has Mother's IQ of 117 and Z-Score of -0.18 which is slightly below the mean for this dataset. Case four, on the other hand, has Mother's IQ of 131 and Z-Score of $+1.97$, which is significantly above the mean for this dataset.

4.3.2.1 *Activity 3: Z-Scores*

Using the *gifted* dataset, produce the Z-Scores for *speak*. Report the Z-Score for any case where the "speak" raw score is 18.

4.3.2.2 *Activity 4: Z-Scores*

Using the *cafe* dataset, produce the Z-Scores for *age*. Report the Z-Score for any case where the "age" raw score is 30.

4.4 DELIVERABLE

Complete the following activities in this lab:

Number	Name	Page
4.3.1.1	Activity 1: Descriptives	48
4.3.1.2	Activity 2: Descriptives	48
4.3.2.1	Activity 3: Z-Scores	50
4.3.2.2	Activity 4: Z-Scores	50

Consolidate the responses for all activities into a single document and submit that document for grading.

5

LAB 05: EXPLORE

5.1 INTRODUCTION

Preparing and "cleaning" data before it is used in statistical analysis is an important and time-consuming process. It should make sense that to find a mean of a data element is rather pointless if that element contains bad data. While the process of cleaning data is beyond the scope of this manual there are a few simple steps researchers can take to attempt to verify the integrity of the data. This lab examines the PSPP *Explore* feature that permits that sort of examination.

5.2 DISCUSSION

5.2.1 *Data Problems*

There are a number of simple tests that researchers can perform on a data element before they start working with it to determine if there are problems that may foul further statistical analysis. While some data analysis programs include tools to help clean and prepare data for analysis it is probably better to use software that has been specifically created for that purpose so the data can be ready before it is ever imported into a program like PSPP. For students who want to explore this aspect of data analysis, Open Refine is an excellent place to start. This was formerly *Google Refine* and is an easy-to-use, but powerful, data preparation tool with a many free online tutorials and other help.

5.2.1.1 *Data Element Type*

When data are first imported into the data base it is possible for the automatic process to misidentify the type of data being imported. For example, a numeric data element may be imported as text and that would limit the types of analysis that could be completed. This type of error is normally very easy to detect and correct by simply changing the data element's attribute settings.

5.2.1.2 *Duplicate Data*

Often records in a dataset are duplicated and this would create a problem when analyzing the data. Some data analysis software includes built-in procedures to detect and correct duplicate data, but this type

of error can more efficiently be corrected with data preparation software made to correct these errors.

5.2.1.3 *Missing Data*

Often, survey data will have missing values because the respondents did not complete one or more questions for some reason. This creates missing data and researchers need to account for those missing items. There are several techniques used to allow for missing data.

- Ignore records. It is possible to simply exclude an entire record from analysis if there are any missing elements within that record. This should only be done if the number of records being excluded is less then 5% of the total number of records. Excluding records also does not work well for data from matched trials ("before/after" types of experiments) or time values (where something is being measured over a long period of time).
- Setting missing values to a fixed value. It is possible to simply find-replace all missing values with some fixed number, like 0. This, though, leads to a number of additional problems and should not be used.
- Estimating values. It is common to replace missing values with an estimate for the missing data. For example, missing values could be replaced by the mean of the entire dataset or by the mean that is calculated from nearby values.

5.2.1.4 *Outliers*

It is common for interval and ratio data to include outliers, or values that are far outside the range of "normal" values. For example, a survey of home prices in a neighborhood may reveal that the mean value of a home is $150K but there is one mobile home in the neighborhood that is valued at only $75K. That low value would be an outlier that would concern researchers.

There are several methods used to deal with outliers. If there is an obvious coding error, like a decimal point in the wrong place, then it can simply be corrected. Outliers can also be systematically changed to a value that is three standard deviations from the mean. They can also be excluded from analysis by using a method like trimmed means. Finally, if the outliers seem to contain data that is important for further analysis, then the dataset can be split such that the outliers end up in their own subset where they can receive more scrutiny.

5.2.1.5 *Data Distribution*

Most statistical tests require data to be normally distributed and if the dataset has some other distribution, or is badly skewed or shows

an abnormal kurtosis[1], then the statistical analysis may fail. To try to correct the distribution it is possible to apply some sort of transformation to the data to see if a better fit can be achieved. As one example, the common logarithm can be taken of the data and that may fit a normal distribution curve better than the raw data.

5.3 PROCEDURE

5.3.1 *Data Element Type*

The simplest way to check and correct an improperly imported data error is to examine the attributes for the data elemets to see if they seem to be correct. Within PSPP, click VARIABLE VIEW at the bottom of the data window and look carefully at each variable's attributes.

Variab	Name	Type	Width	Decimal	Label	Value Labels	Missing Values	Column	Align	Measure	Role
1	spam	String	4			None	None	4	Left	Nominal	Input
2	toMultiple	String	11			None	None	11	Left	Nominal	Input
3	cc	Numeric	2	0		None	None	8	Right	Scale	Input
4	sentEmail	String	10			None	None	10	Left	Nominal	Input
5	image	Numeric	5	0		None	None	8	Right	Scale	Input
6	attach	Numeric	6	0		None	None	8	Right	Scale	Input
7	dollar	Numeric	6	0		None	None	8	Right	Scale	Input
8	winner	String	6			None	None	6	Left	Nominal	Input
9	inherit	Numeric	7	0		None	None	8	Right	Scale	Input
10	viagra	Numeric	6	0		None	None	8	Right	Scale	Input
11	password	Numeric	8	0		None	None	8	Right	Scale	Input
12	numChar	Numeric	9	2		None	None	8	Right	Scale	Input
13	lineBreak	Numeric	11	0		None	None	8	Right	Scale	Input
14	format	String	6			None	None	6	Left	Nominal	Input
15	reSubj	String	7			None	None	7	Left	Nominal	Input
16	exclaimSubj	Numeric	12	0		None	None	8	Right	Scale	Input
17	urgentSubj	String	11			None	None	11	Left	Nominal	Input
18	exclaimNum	Numeric	12	0		None	None	8	Right	Scale	Input
19	numbers	String	6			None	None	6	Left	Nominal	Input

Figure 37: Attributes for the Email Dataset Elements

In Figure 37, the type of data is found in column three and that can be changed if necessary. The data width is found in columns four and five where string data (that is, text data) is restricted to only the number of characters in column four. For numeric data the integer part is restricted to the width specified in column four and the decimal to that specified in column five. The other column of interest is 11, labeled "Measure." That indicates how the data are being used by PSPP and that can be easily changed between Nominal, Ordinal, and Scale (that is both ratio and interval).

1 Lab 1 contains information on data distributions, skew, and kurtosis.

5.3.2 *Duplicate Data*

PSPP does not include a method to detect duplicate data; therefore, a pre-processing program, like Open Refine should be used to prepare the dataset, include removing duplicate records, before it is imported.

5.3.3 *Missing Data*

Start PSPP and open the *births* dataset, then:

1. Click Analyze → Descriptive Statistics → Explore
2. Click the word *visits* in the left column and then click the right-arrow button near the center of the window to move *visits* to the "Dependent List" box on the right side of the window. (Alternatively, double-click the word *visits* in the left column to move it to the "Dependent List" box.)
3. Click Statistics at the bottom of the screen.
4. Select all three statistics functions: Descriptives, Extremes, and Percentiles.
5. Click Continue to close the Statistics window.
6. Click OK to generate the requested information.

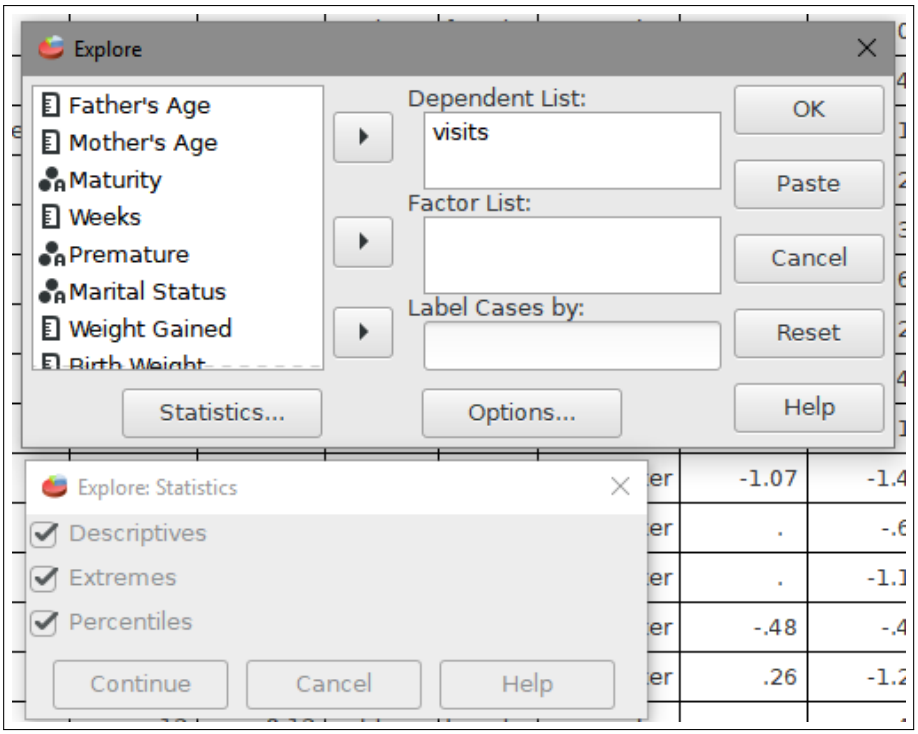


Figure 38: Generating Explore Information

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Number of Visits	991	99.1%	9	0.9%	1000	100%

Extreme Values

			Case Number	Value
Number of Visits	Highest	1	1000	30
		2	999	30
		3	998	30
		4	997	30
		5	996	30
	Lowest	1	10	0
		2	11	0
		3	12	0
		4	13	0
		5	14	0

Percentiles

		Percentiles						
		5	10	25	50	75	90	95
Number of Visits	HAverage	5.00	7.00	10.00	12.00	15.00	16.00	18.00
	Tukey's Hinges			10.00	12.00	15.00		

Descriptives

			Statistic	Std. Error
Number of Visits	Mean		12.10	.13
	95% Confidence Interval for Mean	Lower Bound	11.86	
		Upper Bound	12.35	
	5% Trimmed Mean		12.10	
	Median		12.00	
	Variance		15.64	
	Std. Deviation		3.95	
	Minimum		.00	
	Maximum		30.00	
	Range		30.00	
	Interquartile Range		5.00	
	Skewness		.17	.08
	Kurtosis		2.16	.16

Figure 39: Explore Information

The "Case Processing Summary" box at the top of Figure 39 indicates that the *age* element is missing nine cases, which is only about 0.9% of the total number of cases in the dataset. Becuase this is well under 5% of the total number of cases, it would be reasonable to simply ignore the missing cases if the *visits* data were being analyzed.

5.3.4 *Activity 1: Missing Data*

Using the *births* dataset, determine the number and percent of missing cases for *Father's Age* and *Weight Gained*.

5.3.5 *Outliers*

There is no rigid statistical definition of "outlier" so there is no one way to detect them. However, there are several simple techniques can be used to determine if a data element has suspected outliers.

"Tukey's Test" is to take the Interquartile Range (IQR) and multiply that by 1.5 then subtract that number from the first quartile and add

it to the third quartile. Any values that lie outside those boundaries should be investigated as potential outliers. As an example, consider Figure 39.

1. The IQR of 5.00 is found in the "Descriptives" table.
2. Multiply the IQR by 1.5 to get 7.5.
3. The values of the first (25%) and third (75%) quadrants are found in the "Percentiles" box.
4. Subtract $10.00 - 7.5$ to get the lower bound of 2.5.
5. Add $15.00 + 7.5$ to get the upper bound of 22.5.
6. Use the "Extreme Values" box to see if any values lie below the lower bound or above the upper bound and there are many.

Another approach is to look for values that are more than three standard deviations from the mean of the dataset. Again, using Figure 39:

1. The mean of 12.10 is found in the "Descriptives" table.
2. The standard deviation of 3.95 is found in the "Descriptives" table.
3. Multiply the standard deviation by 3 to get 11.85.
4. Subtract $12.10 - 11.85$ to get the lower bound of 0.25.
5. Add $12.10 + 11.85$ to get the upper bound of 23.95.
6. Use the "Extreme Values" box to see if any values lie below the lower bound or above the upper bound and there are many.

It is perhaps easiest to use the Z-Score to look for outliers. Since the Z-Score is the number of standard deviations a data point lies from the mean, researchers can look for any Z-Scores above $+3.0$ or below -3.0 to detect outliers. The method for calculating Z-Scores with PSPP was described on page 45. The Z-Scores for *Visits* in the *Births* dataset was calculated and Figure 40 shows the top 19 and bottom 11 cases for that data element.

Case	visits	Zvisits
1	.	.
2	.	.
3	.	.
4	.	.
5	.	.
6	.	.
7	.	.
8	.	.
9	.	.
10	0	-3.06
11	0	-3.06
12	0	-3.06
13	0	-3.06
14	0	-3.06
15	0	-3.06
16	0	-3.06
17	2	-2.56
18	2	-2.56
19	2	-2.56

989	22	2.50
990	22	2.50
991	22	2.50
992	23	2.75
993	24	3.01
994	25	3.26
995	26	3.51
996	30	4.52
997	30	4.52
998	30	4.52
999	30	4.52
1000	30	4.52

Figure 40: Z-Scores for Visits

Notice that the top nine cases have missing data, indicated by a dot, but cases $10-16$ have a Z-Score of -3.06, which are more than three standard deviations below the mean and would be likely outliers. In the same way, cases $993-1000$ are all more than three standard deviations above the mean and would be likely outliers.

5.3.6 *Activity 2: Outliers*

Using the *births* dataset, determine the outliers for the *weight* data element. For this activity, assume that any weight that is more than three standard deviations below or above the mean is an outlier. Report the weight, the Z-Score, and number of times that weight appears in the dataset. The response should be similar to this:

Weight	Z-Score	Number
1.20	-3.55	3
1.18	-3.54	1
11.75	+3.10	2

5.3.7 *Activity 3: Outliers*

Using the *cafe* dataset, determine the outliers for the *miles* data element. For this activity, assume that any values more than three standard deviations below or above the mean is an outlier. Report the miles, the Z-Score, and number of times that value appears in the dataset. The response should be similar to the table in Activity 2.

5.4 DELIVERABLE

Complete the following activities in this lab:

Number	Name	Page
5.3.4	Activity 1: Missing Data	55
5.3.6	Activity 2: Outliers	57
5.3.7	Activity 3: Outliers	58

Consolidate the responses for all activities into a single document and submit that document for grading.

Part III

INFERENTIAL

Inferential statistics attempt to infer some characteristic of an entire population from probabilities calculated in a sample and include tools like correlation, regression, and comparing means. This part contains eight labs that explore several different inferential statistics available in PSPP.

Part IV

APPENDIX

6 APPENDIX

6.1 APPENDIX A: DATASETS

There are a number of datasets used in the lab exercises and this appendix lists the elements in each dataset.

6.1.1 *bdims*

This is a dataset of the body girth measurements and skeletal diameter measurements, as well as age, weight, height and gender, for 507 physically active individuals, 247 men and 260 women.

- **age**. (Scale) The patient's age in years.
- **ank.di**. (Scale) The patient's ankle diameter in centimeters, measured as sum of two ankles.
- **ank.gi**. (Scale) The patient's ankle minimum girth in centimeters, measured as average of right and left girths.
- **bia.di**. (Scale) The patient's biacromial (shoulder width) in centimeters.
- **bic.gi**. (Scale) The patient's bicep girth in centimeters, measured when flexed as the average of right and left girths.
- **bii.di**. (Scale) The patient's biiliac (pelvic width) in centimeters.
- **bit.di**. (Scale) The patient's bitrochanteric (femur width) in centimeters.
- **cal.gi**. (Scale) The patient's calf maximum girth in centimeters, measured as average of right and left girths.
- **che.de**. (Scale) The patient's chest depth in centimeters, measured between spine and sternum at nipple level, mid-expiration.
- **che.di**. (Scale) The patient's chest diameter in centimeters, measured at nipple level, mid-expiration.
- **che.gi**. (Scale) The patient's chest girth in centimeters, measured at nipple line in males and just above breast tissue in females, mid-expiration.
- **elb.di**. (Scale) The patient's elbow diameter in centimeters, measured as sum of two elbows.

- **for.gi**. (Scale) The patient's forearm girth in centimeters, measured when extended, palm up as the average of right and left girths.
- **hgt**. (Scale) The patient's height in centimeters.
- **hip.gi**. (Scale) The patient's hip girth in centimeters, measured at at level of bitrochanteric (femur width) diameter.
- **kne.di**. (Scale) The patient's knee diameter in centimeters, measured as sum of two knees.
- **kne.gi**. (Scale) The patient's knee diameter in centimeters, measured as sum of two knees.
- **nav.gi**. (Scale) The patient's navel (abdominal) girth in centimeters, measured at umbilicus and iliac crest using iliac crest as a landmark.
- **sex**. (Nominal) The patient's sex coded as 0 for female and 1 for male.
- **sho.gi**. (Scale) The patient's shoulder girth in centimeters, measured over deltoid muscles.
- **thi.gi**. (Scale) The patient's thigh girth in centimeters, measured below gluteal fold as the average of right and left girths.
- **wai.gi**. (Scale) The patient's waist girth in centimeters, measured at the narrowest part of torso below the rib cage as average of contracted and relaxed position.
- **wgt**. (Scale) The patient's weight in kilograms.
- **wri.di**. (Scale) The patient's wrist diameter in centimeters, measured as sum of two wrists.
- **wri.gi**. (Scale) The patient's wrist minimum girth in centimeters, measured as average of right and left girths.

6.1.2 *births*

This is a random sample of 1000 births in North Carolina in 2004.

- **fage**. (scale) The father's age.
- **gained**. (scale) The mother's weight gain, in pounds.
- **gender**. (Nominal) The gender of the baby.
- **habit**. (Nominal) Whether the mother was a smoker.

- **lowbirthweight**. (Nominal) Whether the baby had a low birth weight.
- **mage**. (Scale) The mother's age.
- **marital**. (Nominal) Whether the mother was married.
- **mature**. (Nominal) The mother's maturity level.
- **premie**. (Nominal) Whether the baby was premature.
- **visits**. (Scale) The number of hospital visits made by the mother.
- **weeks**. (Scale) Length of pregnancy, in weeks.
- **weight**. (Scale) Birth weight of the baby, in pounds.
- **whitemom**. (Nominal) Whether the mother was white.

6.1.3 *cars*

This is a random sample for 1993 model cars that were in both *Consumer Reports* and *PACE Buying Guide*. Only vehicles of type "small," "midsize," and "large" were included. The dataset has 54 rows and these data elements for each row:

- **driveTrain**. (Nominal) Vehicle drive train with levels *4WD*, *front*, and *rear*.
- **mpgCity**. (Scale) City mileage (miles per gallon).
- **passengers**. (Ordinal) The vehicle passenger capacity.
- **price**. (Scale) Vehicle price in U.S. dollars.
- **type**. (Ordinal) The vehicle type with levels *large*, *midsize*, and *small*.
- **weight**. (Scale) Vehicle weight in pounds.

6.1.4 *email*

This dataset contains 3921 observations from the email received by one person over the first three months of 2012.

- **attach**. (Scale) The number of attached files.
- **cc**. (Scale) The number of people who were CCed on the message.
- **date**. (Scale) The date and time the email was sent.

- **dollar**. (Scale) The number of times a dollar sign or the word "dollar" appeared in the email.
- **exclaim_mess**. (Scale) The number of exclamation points in the email message.
- **exclaim_subj**. (Nominal) 0 if the email subject did not have an exclamation point, otherwise 1.
- **format**. (Nominal) 0 if the message was sent in text format and 1 if it used HTML format.
- **image**. (Scale) The number of images attached.
- **inherit**. (Scale) The number of times "inherit" (or an extension, such as "inheritance") appeared in the email.
- **line_breaks**. (Scale) The number of line breaks in the email (does not count text wrapping). This could be used as a surrogate for the length of the message.
- **num_char**. (Scale) The number of characters in the email, in thousands. This could be used as a surrogate for the length of the message.
- **number** (Ordinal) *None* if the message included no numbers, *small* if it included only numbers less than one million, or *large* if it included one or more big numbers.
- **password**. (Scale) The number of times "password" appeared in the email.
- **re_subj**. (Nominal) 1 if the subject started with any of these: "Re:", "RE:", "re:", or "rE:", otherwise 0.
- **sent_email**. (Nominal) 1 if email had been sent to the sender in the last 30 days, otherwise 0.
- **spam**. (Nominal) 1 if the message is spam, otherwise 0.
- **to_multiple**. (Nominal) 1 for a message that was sent to more than one person, otherwise 0.
- **urgent_subj**. (Nominal) 1 if the subject included the word "urgent," otherwise 0.
- **viagra**. (Scale) The number of times "viagra" appeared in the email.
- **winner**. (Nominal) 1 if the word "winner" appeared in the email, otherwise 0.

6.1.5 *gifted*

An investigator is interested in understanding the relationship, if any, between the analytical skills of young gifted children and the variables listed below. The analytical skills are evaluated using a standard testing procedure and the score on that test is included in the dataset. Data were collected from schools in a large city on a set of 36 children who were identified as gifted children soon after they reached the age of four.

- **Cartoons**. (Scale) Average number of hours per week the child watched cartoons on TV during the past three months.
- **Count**. (Scale) Age in months when the child first counted to ten successfully.
- **Edutv**. (Scale) Average number of hours per week the child watched an educational program on TV during the past three months.
- **Fatheriq**. (Scale) Father's IQ.
- **Motheriq**. (Scale) Mother's IQ.
- **Read**. (Scale) Average number of hours per week the child's mother or father reads to the child.
- **Score**. (Scale) The score earned on the test of analytical skills.
- **Speak**. (Scale) Age in months when the child first said "mommy" or "daddy."

6.1.6 *cafe*

This is simulated data. Customers of the Main Street Café completed surveys over a one week period.

- **age**. (Scale) The age in years of the person completing the survey.
- **bill**. (Scale) The bill for the meal.
- **day**. (Nominal) The day of the week the person visited the cafe. The levels are 1 for "Sunday", 2 for "Monday," etc.
- **food**. (Ordinal) A rating for the food, from one to five "stars."
- **length**. (Scale) The length of the visit in minutes.
- **meal**. (Nominal) The meal eaten stored as 0 for Breakfast, 1 for Lunch, 2 for Dinner, and 3 for Other.

- **miles**. (Scale) The number of miles between the visitor's home and the café.
- **pref**. (Nominal) A binary item for seating preference where 0 is for "booth" and 1 is for "table."
- **ptysize**. (Scale) The size of the dining party.
- **recmd**. (Nominal) A binary item for whether the customer would recommend the café to other people, stored as 0 for "no" and 1 for "yes."
- **sex**. (Nominal) The Sex of the person completing the survey. The levels are: 0 for male, 1 for female, and 2 for other.
- **svc**. (Ordinal) A rating for the service, from one to five "stars."
- **tip**. (Scale) The amount of tip left.

6.1.7 *rivers*

The Rivers dataset is a list of the lengths, in miles, of the longest 141 rivers in the United States.

> 135, 202, 210, 210, 215, 217, 230, 230, 233, 237, 246, 250, 250, 250, 255, 259, 260, 260, 265, 268, 270, 276, 280, 280, 280, 281, 286, 290, 291, 300, 300, 300, 301, 306, 310, 310, 314, 315, 320, 325, 327, 329, 330, 332, 336, 338, 340, 350, 350, 350, 350, 352, 360, 360, 360, 360, 375, 377, 380, 380, 383, 390, 390, 392, 407, 410, 411, 420, 420, 424, 425, 430, 431, 435, 444, 445, 450, 460, 460, 465, 470, 490, 500, 500, 505, 524, 525, 525, 529, 538, 540, 545, 560, 570, 600, 600, 600, 605, 610, 618, 620, 625, 630, 652, 671, 680, 696, 710, 720, 720, 730, 735, 735, 760, 780, 800, 840, 850, 870, 890, 900, 900, 906, 981, 1000, 1038, 1054, 1100, 1171, 1205, 1243, 1270, 1306, 1450, 1459, 1770, 1885, 2315, 2348, 2533

6.2 APPENDIX B: RECODING VARIABLES

6.2.1 *Background*

For efficiency, data are often stored in a database in a format that does not lend itself to easy analysis. For example, nominative values, like "no" and "yes" are frequently stored as 0 and 1. While that is efficient for storage it makes using a table or chart more difficult because the various data elements will be presented as something like 0 instead of "no" and it is incumbent upon the researcher to remember what the various codes mean; however, values in a dataset can be recoded to make them easier to use. As examples, "0/1" values can be recoded

to "no/yes" or a variable containing ages can be recoded so the ages are grouped, like ages $20 - 29$ can be recoded to 2.

6.2.2 *Recoding Variables With SOFA*

In SOFA, a data field is recoded into a new field so the dataset ends up with two fields that contain the same data but in different formats. As an example, imagine that a researcher is using the "spam" field of the *email* dataset and desires to use 0 instead of "no" and 1 instead of "yes," then that field would need to be recoded.

1. Start SOFA and select "Enter/Edit Data."
2. Data Tables: email
3. Click "Design"
4. Click the "Recode" button on the Data Table screen
5. Fill in the "Recode" screen as illustrated below. (Note: since each row is saved as it is entered, the cursor must be moved into Row 3, as illustrated, in order to save the changes made on Row 2).

Figure 41: Recoding Spam Field

6. Click "Recode"
7. The following message will be displayed

Figure 42: Recode Warning on Save

8. Click OK to dismiss the warning, click OK to complete the recode, and then click "Update" to file the results of the recode process.

When this process is completed the *email* dataset will contain a new field named "spamnum" that contains a 0 where "spam" is equal to "no" and a 1 where "spam" is "yes."

6.3 APPENDIX C: SOFA EXPORTS

SOFA creates several different types of exports and each are easy to generate and use. Export specifications are set in the area across the

center of the various pop-up windows (Report Tables, Charts, and Statistics) and generating reports is the same for all of the windows.

6.3.1 *Styles*

SOFA comes with seven built-in styles that can be selected from the box on the bottom-right of the window:

Figure 43: SOFA Styles

As each style is selected the display in the lower-left corner of the window is immediately updated to reflect the selected style.

6.3.2 *Exporting a File*

The output in the lower-left corner of the window can be exported in a file that can be opened by a program like Excel. In the Export drop-down box at the right-center of the window select "Current Output" and then click the "Export" button.

Figure 44: Selecting Export Type

Select the specific type of output desired in the pop-up window.

Figure 45: Specifying Desired Export

- Export as PDF. This option will generate a PDF file containing the current output. Note that a number of different outputs can be combined into a single PDF file by using the "Report" feature, described below.
- Export to spreadsheet (report tables only). This generates an Excel spreadsheet without any formatting. This is an excellent option if the data produced by SOFA needs further manipulation.
- Export as Images. SOFA will export the output as .PNG images that can be used in other programs or emailed. The quality of the image can be selected using the drop-down box. (NOTE: for most work the "Print Quality (300 dpi)" setting is adequate.)

Click the "Export" button to generate the desired export file.

6.3.3 *Copy/Paste Output*

In the dropdown "Export" box, select "Copy current output ready to paste" to copy the output displayed in the lower left corner of the window so it can be pasted directly into Word or some other program.

6.3.4 *Reports*

SOFA can combine the outputs for several operations into a single PDF report or series of .PNG files. This is a two-step process, first the various outputs are saved into a report and, second, the final report is exported.

To save the outputs into a single report, start by specifying the location and name for the report. By default, SOFA saves reports in the *sofastats/reports* folder and that is appropriate since SOFA can generate a number of files when producing a report. To specify the name for a new report, click the "Browse" button and enter the report's name (it is the file name). SOFA saves reports in .HTM format but a different format can be specified when the report is later exported.

Then, as outputs are produced, click the "Also add to report" button to add the current output, displayed in the lower-left corner of the window, to the report. Note: every time the "Also add to report" button is clicked the current output is added to the report so avoid clicking that button multiple times unless multiple copies of the current output are desired.

To export the report, select "Entire Report" in the export dropdown box. Select the format for the report (PDF, images, or Excel Spreadsheet) and click "Export."

Note: SOFA exports PDF files such that each saved output screen is on a different PDF page, which makes the PDF file rather long with a lot of blank space between pages. As an alternative, it may be possible to click the "View Report" button to open the report in a browser and then use the browser's print feature.