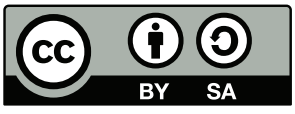
LAB MANUAL

SOFA: Statistics Open For All

GEORGE SELF

June 2017 – Edition 2.0

George Self: *Lab Manual,* SOFA: Statistics Open For All, June 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 as part of our studies and carry out the types of calculations that are discussed in the text. As I evaluated statistical software 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.

This lab manual explores every aspect of `SOFA` in a series of ten lab exercises plus a final. It is my hope that students will find the labs instructive and will then be able to use `SOFA` 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

Part I

LAB EXERCISES

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 *Stastics Open For All* (SOFA), an open source statistical analysis program that is free of charge and easy to use. Before downloading and diving into a statistics package there are two important background fundamentals that must be covered: hypothesis and data.

1.2 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 accurate. Additional investigation may be needed to explain *why* that observation 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 that was proffered for some observed phenomena was 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 hypothesis that is being suggested as an explanation for the observed phenomenon. In the case of the smoking mother mentioned above the alternative hypothesis would be that smoking causes a de-

crease 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:

- *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.

1.3 DATA

1.3.1 *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 9 on page 93. 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",

1 Appendix A: Datasets, on page 123, lists all of the datasets used in this lab manual and specifies the type of data each contains.

"Agree", "Neutral", "Disagree", "Strongly Disagree". Ordinal data are also used when numeric data are grouped. For 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 a truck's weight. 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 includes characteristics like means and standard deviations, they are analyzed using parametric tests, as described in Lab 10 on page 107. Quantitative data can be further divided into two sub-types, interval and ratio.
  - **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 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 on month with 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.

1.3.2 *Shape of Data*

1.3.2.1 *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" (some-

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

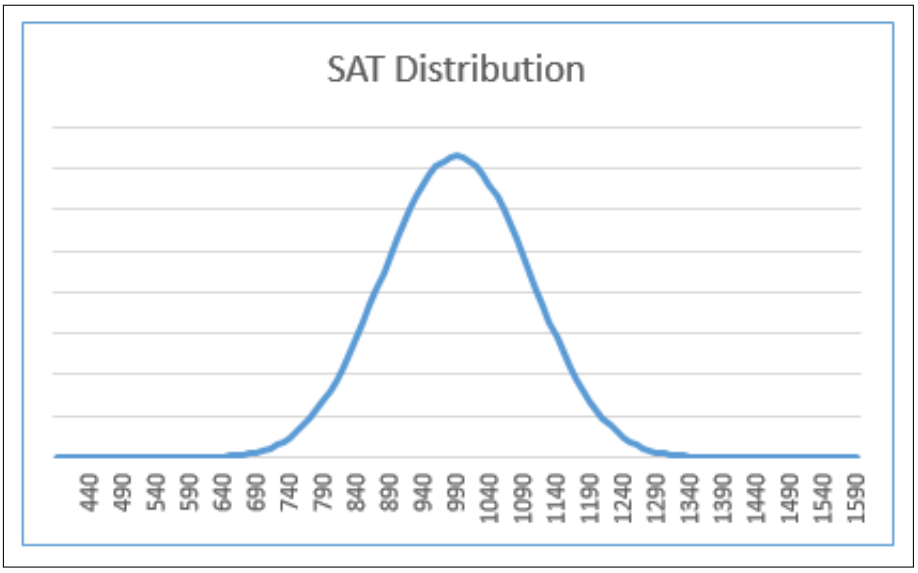


Figure 1: Normal Distribution

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 use a different group of tests.

1.3.2.2 *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.

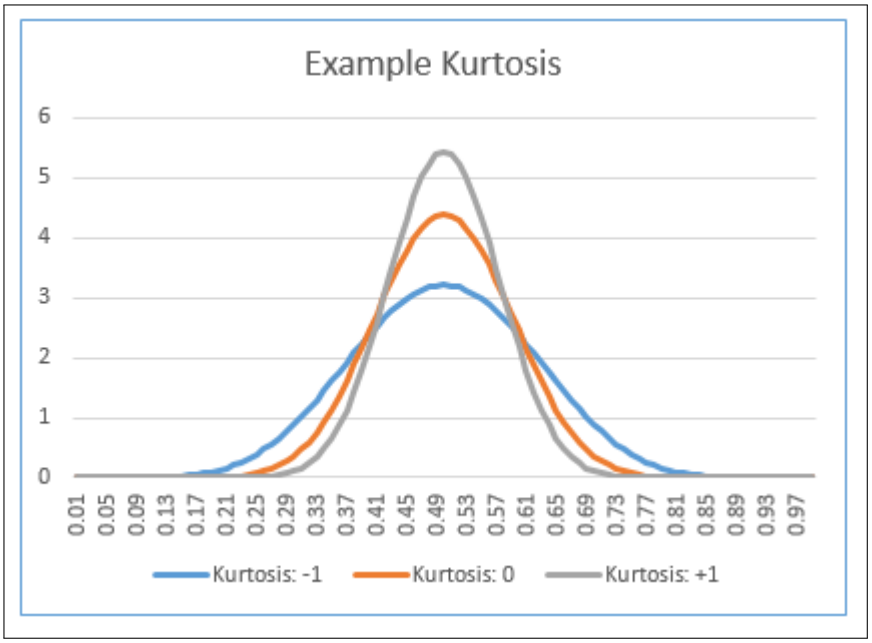


Figure 2: Kurtosis in a Normal Distribution

1.3.2.3 *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:

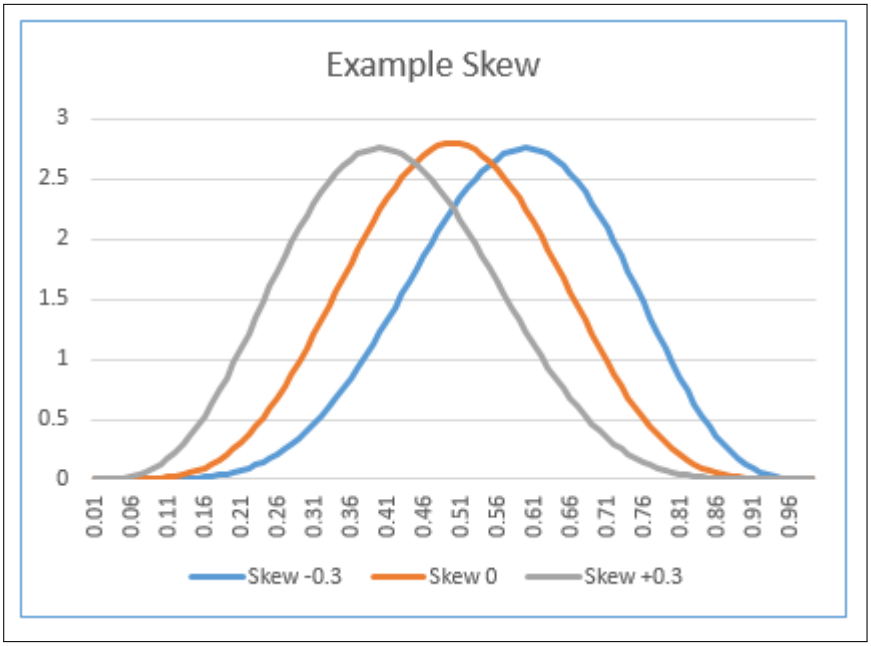


Figure 3: Skew in a Normal Distribution

1.4 INSTALLING AND STARTING SOFA

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

http://www.sofastatistics.com/downloads.php.

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

1.5 IMPORTING DATA

In order to work with the statistical analysis in SOFA the data must first be imported. SOFA makes it easy to import data, then those data are always available in SOFA's internal database until they are intentionally deleted. The datasets[2] for all of the activities in this manual are available in a ZIP file located at:

https://goo.gl/hA04Gg

Download the latest version of the Zipped "Data Files" and extract all of the .CSV files to a folder and then import each dataset into SOFA. As a start, to import the *bdims* dataset:

1. Open SOFA and click the "Import Data" button.

2 Appendix A: Datasets, on page 123, details the structure and contents of all datasets used in this manual.

2. On the "Select File" screen, find the bdims.csv file that was extracted from the dataset ZIP archive. The `SOFA` Table Name will be automatically filled in based upon the name of the .CSV file. While the table name can be changed, it is best to leave it at the default value so it matches the activities in this manual.

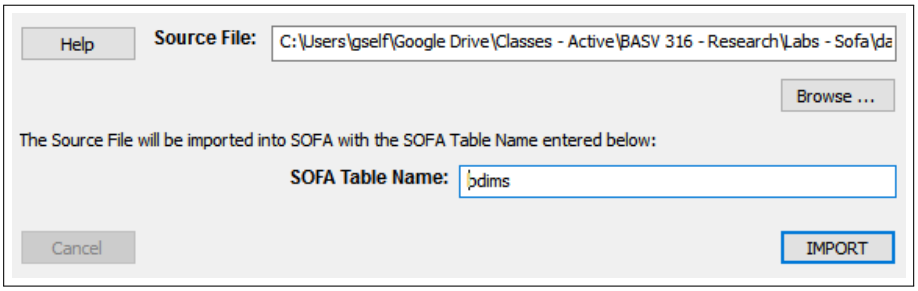
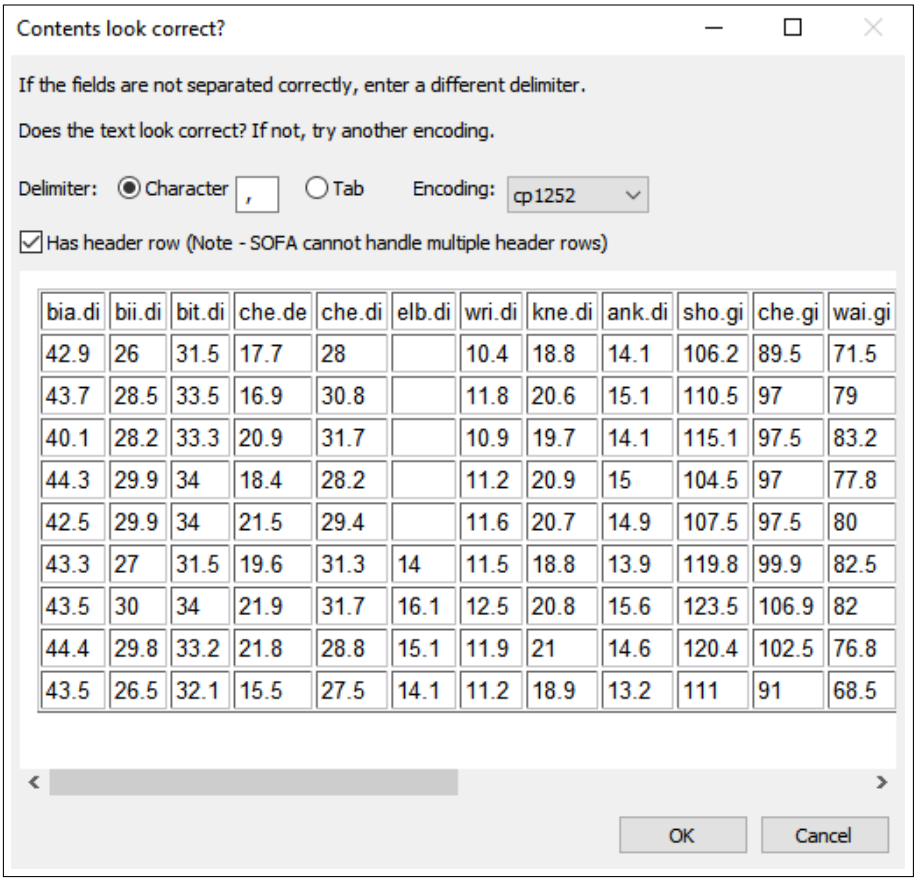


Figure 4: Finding CSV File To Import

3. Click the "Import" button.

4. A window with the first few lines of data from the CSV file will open. Be certain that the data looks correct. If it looks like there are run-on lines (that is, a long string of random characters) or the header line was not found, then adjust the various settings at the top of the check window.



bia.di	bii.di	bit.di	che.de	che.di	elb.di	wri.di	kne.di	ank.di	sho.gi	che.gi	wai.gi
42.9	26	31.5	17.7	28		10.4	18.8	14.1	106.2	89.5	71.5
43.7	28.5	33.5	16.9	30.8		11.8	20.6	15.1	110.5	97	79
40.1	28.2	33.3	20.9	31.7		10.9	19.7	14.1	115.1	97.5	83.2
44.3	29.9	34	18.4	28.2		11.2	20.9	15	104.5	97	77.8
42.5	29.9	34	21.5	29.4		11.6	20.7	14.9	107.5	97.5	80
43.3	27	31.5	19.6	31.3	14	11.5	18.8	13.9	119.8	99.9	82.5
43.5	30	34	21.9	31.7	16.1	12.5	20.8	15.6	123.5	106.9	82
44.4	29.8	33.2	21.8	28.8	15.1	11.9	21	14.6	120.4	102.5	76.8
43.5	26.5	32.1	15.5	27.5	14.1	11.2	18.9	13.2	111	91	68.5

Figure 5: Checking Import Data

5. Notice in Figure 5 that the column labeled "elb.di" has some blanks at the top. It is common for datasets to have missing data, but SOFA can easily work around that problem.

6. Click "OK."

7. Next, SOFA warns that there was a problem with Row 7.

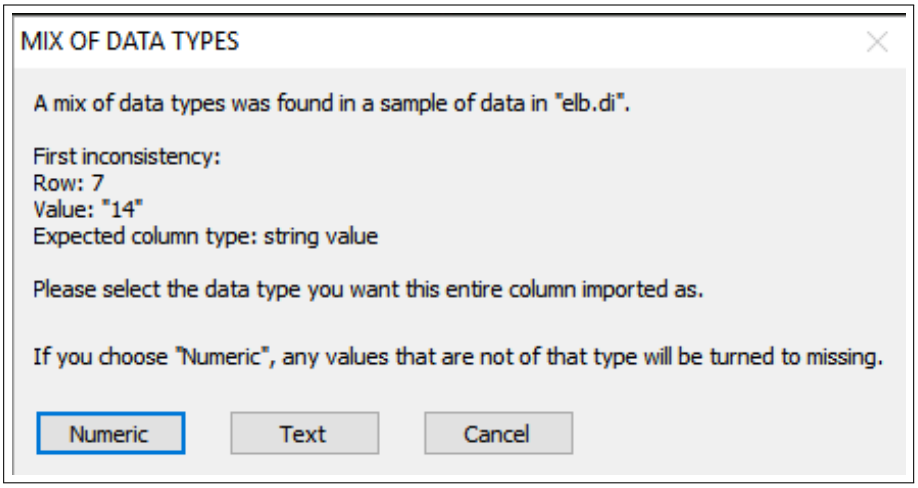


Figure 6: Mixed Data Warning

8. The column for "elb.di" had some missing values in its first few lines so SOFA assumed that the column contained text. Then, when SOFA found a number in row seven it was not sure if the column contained text or numbers.

9. Click "Numeric" to let SOFA know that the column contains numbers rather than text.

10. SOFA imports the data and then displays a "Success" screen.

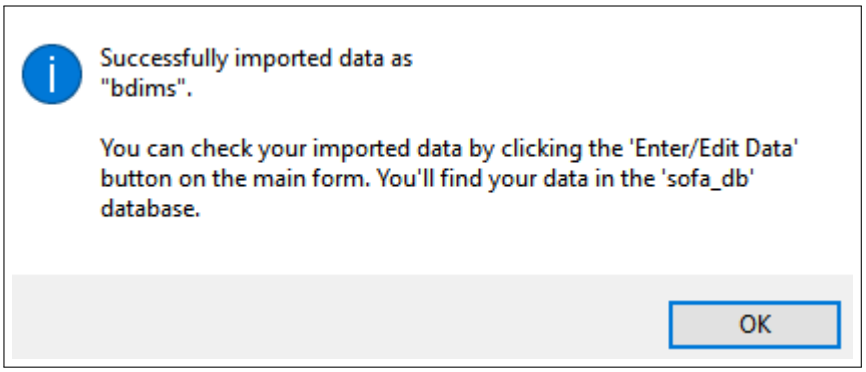


Figure 7: Import Success

11. Click "Close" to finish the import process.

12. All other datasets should now be imported. There should not be any additional problems with missing data or other warnings.

    - births
    - cars
    - doorsurvey

- email
- gifted
- maincafe
- rivers (Note: this dataset has only a single column of numbers.)
- tutoring

13. To check and be certain that all datasets were successfully loaded, go to the main SOFA screen and click the "Enter/Edit Data" button.

14. In the "Choose an existing data table..." screen, click the "Data tables" field to open that select box. The names of all of the datasets that were just loaded should be listed. (Note: *demo_tbl* is a default dataset that comes with SOFA.) If any dataset is missing it should be loaded before proceeding with the lab exercise so it will be available for future lessons.

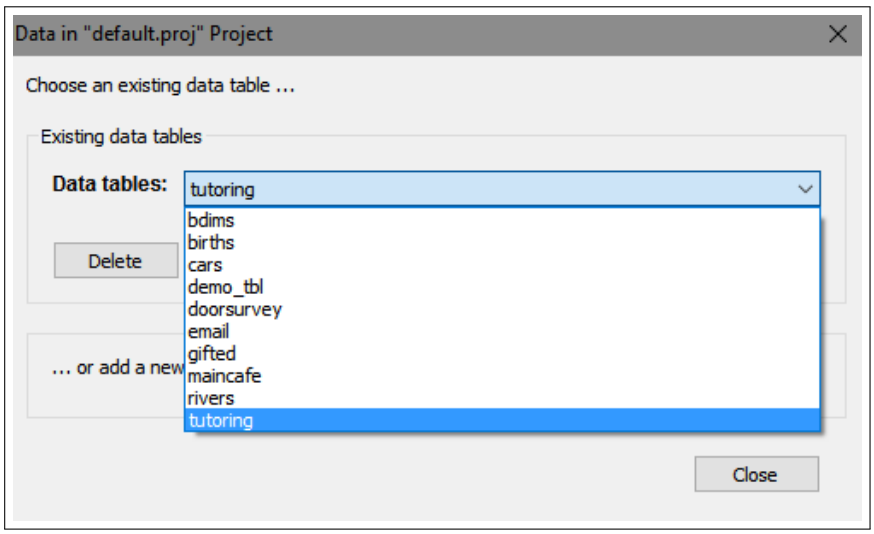


Figure 8: Checking The Datasets

1.5.1 *Activity 1*

Start SOFA and click the "Enter/Edit Data" button. Select the *gifted* data table and click the "Open" button. Take a screen capture of the first ten rows of that data table. It should look something like the following image (which shows the top of the *cars* data table).

Data from sofa_db.cars

	Sofa_Id	Type	Price	Mpgcity	Drivetrain	Passengers	Weight
1	1	small	15.9	25.0	front	5.0	2705.0
2	2	midsize	33.9	18.0	front	5.0	3560.0
3	3	midsize	37.7	19.0	front	6.0	3405.0
4	4	midsize	30.0	22.0	rear	4.0	3640.0
5	5	midsize	15.7	22.0	front	6.0	2880.0
6	6	large	20.8	19.0	front	6.0	3470.0
7	7	large	23.7	16.0	rear	6.0	4105.0
8	8	midsize	26.3	19.0	front	5.0	3495.0
9	9	large	34.7	16.0	front	6.0	3620.0
10	10	midsize	40.1	16.0	front	5.0	3935.0
11	11	midsize	15.9	21.0	front	6.0	3195.0

Figure 9: The Top Of The Cars Data Table

1.6 DELIVERABLE

Complete the following activity in this lab:

Number	Name	Page
1.5.1	Activity 1	13

Save the screen capture in a Word document and submit that document for grading.

2 CENTRAL MEASURES

2.1 INTRODUCTION

It is often desirable to characterize an entire dataset with a single number, and the number that is "in the middle" of the dataset would seem most logical to use. Students in elementary school are taught how to find the average of a group of numbers and they learn that the average is the best representation for that entire group. In statistics, though, there are several different numbers that are often used to represent an entire dataset, and these numbers are collectively known as the *Central Measure*, or numbers that are the "middle" of the dataset.

2.2 CENTRAL MEASURES

2.2.1 N

One of the simplest of measures is nothing more than the number of items in a dataset. For example, for the dataset $5, 7, 13, 22$ the number of items is 4. In statistics, the number of items in a dataset is usually represented by the letter N, therefore, in the simple dataset in this paragraph, $N = 4$. Technically, N does not identify the middle of a dataset but it is an important measure that is often reported and is included here for completeness.

2.2.2 *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 miles[1]. 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*

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

(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 removed from both the top and bottom (7 values from each end of the dataset, for 10% total) then 127 values remain with a range from 230 to 1450 and the trimmed mean for that dataset is 519. 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.

2.2.3 *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 where the median is the best central measure, 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 his workers 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. 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, 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 and a median of 425. This difference, 166, is about 28% of the mean and is significant. The size of this difference would tell a researcher that there are outliers in the dataset that may be skewing the mean.

2.2.4 *Mode*

The mode is used to describe the center of nominal or ordinal data and is nothing more than the item that was most commonly 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 student 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" or something. 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?

2.2.5 *Sum*

One last measure of a dataset that is occasionally reported is the sum, which is nothing more than the values of all of the items 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.

2.3 PROCEDURE

Start `SOFA` and select "Report Tables." Then:

2.3.1 *Calculating Mean, Median, and N*

1. Data Source Table:: rivers
2. Table Type: Row Stats

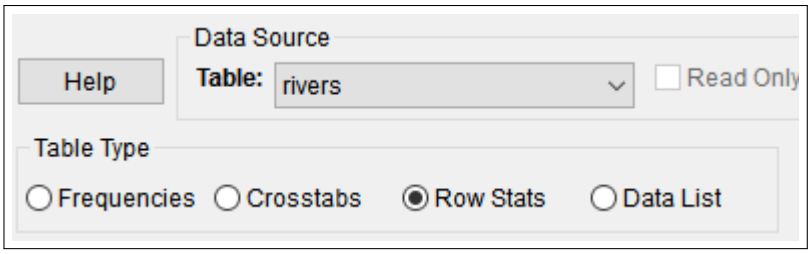


Figure 10: Central Measures for Rivers: Steps 1-2

3. Columns: Add -> Length (Length)

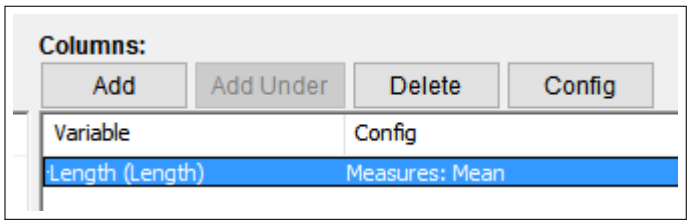


Figure 11: Central Measures for Rivers: Step 3

4. Just above the "Columns" window, click the "Config" button and select: Mean, Median, and N.

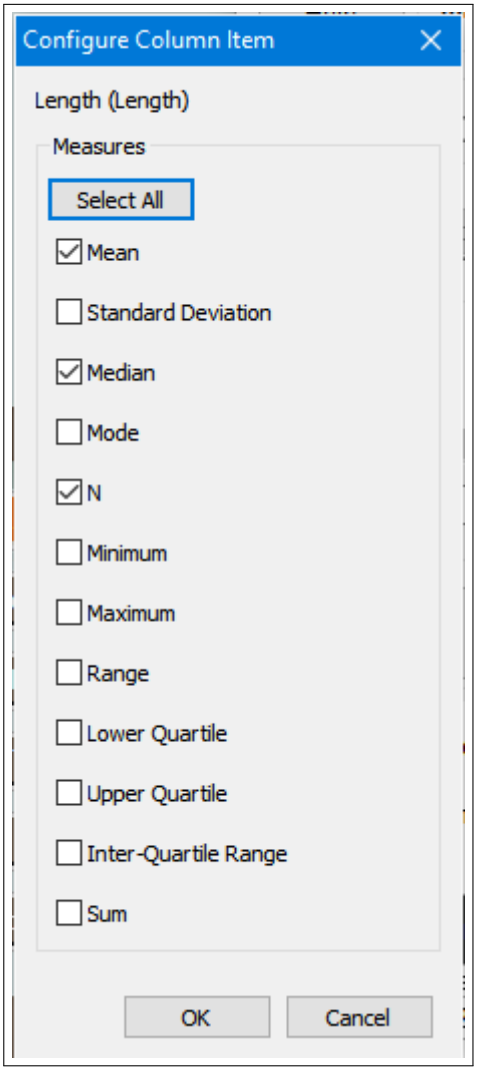
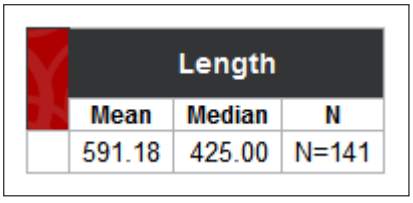


Figure 12: Central Measures for Rivers: Step 4

5. Read those values in the lower left corner of the "Make Report Table" window.



Length		
Mean	Median	N
591.18	425.00	N=141

Figure 13: Central Measures for Rivers: Step 5

6. Thus, there are 141 river lengths in the dataset, the mean length of those rivers is 591.18 miles and the median length is 425.00 miles.

7. To create a version of the output that can be saved and pasted into Word or some other program refer to the instructions in Appendix B: Recoding Variables (page 131).

2.3.2 *Activity 1: Central Measures*

Using the *maincafe* dataset in SOFA produce a table that contains the mean and median for Age, Bill, Length, and Miles data elements. The table should have a title of "Central Measure, Activity 1" and a subtitle of "Main Street Cafe Central Measures".

2.3.3 *Grouping*

It is frequently desirable to group data so means can be compared. For example, it may be useful to group the mean of some variable by gender or some other category. To create groups in SOFA:

Start SOFA and select "Report Tables." Then:

1. Data Source Table: births
2. Table Type: Row Stats
3. Select "Habit" for the row
4. Select "Weeks," and "Weight" for Columns
5. Click "Config" for each variable in Columns and select "mean" and "median" for those variables.

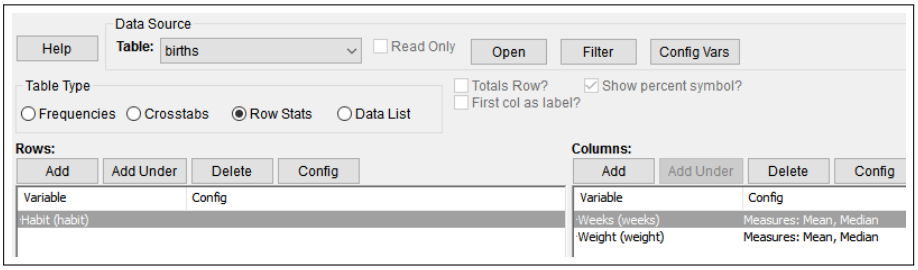


Figure 14: Setting Up Groups of Central Measures

Now, read the mean and median for weeks and weight grouped by smoking habit.

		Weeks		Weight	
		Mean	Median	Mean	Median
Habit	nonsmoker	38.32	39.00	7.14	7.31
	smoker	38.44	39.00	6.83	7.06

Figure 15: Central Measures for Weeks and Weight by Smoking Habit

In Figure 15 the mean length of pregnancy, in weeks, for non-smokers is 38.32 and for smokers is 38.44. The mean and median for weight can also be easily read from the chart.

Sub-groups can also be created with `SOFA`. As an example, start `SOFA` and select "Report Tables." Then:

1. Data Source: births
2. Table Type: Row Stats
3. Click Add and then select "Habit" for the row
4. For Rows, click "Add Under" and then select "Gender"
5. Select "Weight" for Columns
6. Click "Config" for the columns and select "mean" and "median"

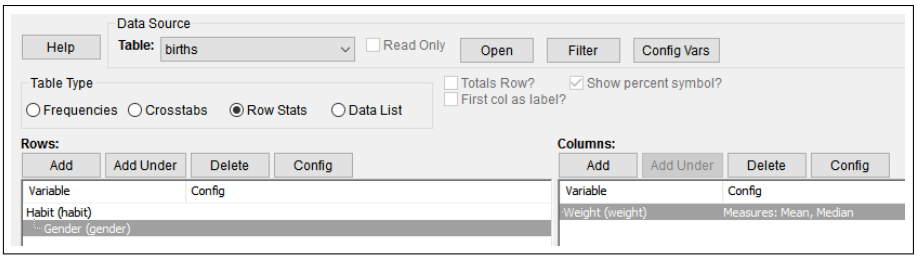


Figure 16: Setting Up Subgroups

				Weight	
				Mean	Median
Habit	nonsmoker	Gender	female	6.94	7.13
			male	7.36	7.50
	smoker	Gender	female	6.68	6.88
			male	6.96	7.31

Figure 17: Mean and Median for Subgroups

In Figure 17 female babies born to non-smokers have a mean weight of 6.94 pounds and females born to smokers have a mean weight of 6.68 pounds.

To create meaningful statistics, the "Row" variables, which are used for grouping, should be either nominal or ordinal and the "Column" variables should be interval or ratio.

2.3.4 *Activity 2: Grouping*

Using the *maincafe* dataset in SOFA produce a table that contains the mean and median for Bill when grouped by Meal. The table should have a title of "Central Measure, Activity 2" and a subtitle of "Main Street Cafe Grouped Central Measures".

2.3.5 *Mode*

The mode is useful for dataset items that are nominal or ordinal in nature rather than interval or ratio. To find the mode for numeric data with SOFA:

Start SOFA and select "Report Tables." Then:

1. Data Source: births
2. Table Type: Row Stats
3. Columns: Add -> Visits
4. Just above the "Columns" window, click the "Config" button and select: Mode.
5. Read those values in the lower left corner of the "Make Report Table" window.

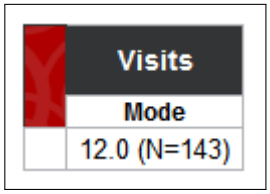


Figure 18: Mode for Visits

143 mothers visited the hospital 12 times, which is the mode for this data item and would be the best "average" to describe the number of hospital visits made by mothers.

If the data being analyzed is text rather than numeric then the mode must be found in a slightly different way. For example, the gender of the baby is listed in the dataset as "male" and "female" and SOFA will not calculate this mode since these are not numeric values. To find the mode for this type of data:

Start SOFA and select "Report Tables." Then:

1. Data Source: births
2. Table Type: Row Stats
3. Rows: Add -> Gender
4. Columns: Add -> Sofa_Id (Note: this is just a one-up number added by SOFA to each row of data as it is imported.)
5. Just above the "Columns" window, click the "Config" button and select: N.
6. The output table shows the frequency that each value appears in the dataset and the mode would be the largest of those frequencies.

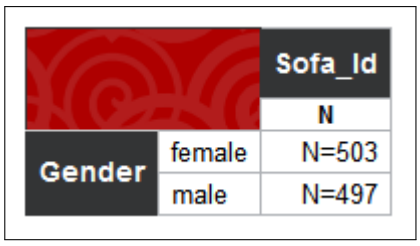


Figure 19: Mode for Baby Gender

For example, the information in Figure 19 shows that there were more females than males in the dataset so that is the mode for Gender.

2.3.6 *Activity 3: Mode for Numeric Data*

Using the *maincafe* dataset in `SOFA` produce a table that contains the mode for both Food and Svc (these are ordinal data). The table should have a title of "Central Measure, Activity 3" and a subtitle of "Main Street Cafe Food and Service Modes".

2.3.7 *Activity 4: Mode for Text Data*

Using the *maincafe* dataset in `SOFA` produce a table that contains the mode for Day. The table should have a title of "Central Measure, Activity 4" and a subtitle of "Main Street Cafe Day Mode".

2.4 EXAMPLES

The following examples were created from the *births* data and are provided for practice.

	Fage		Gained		Mage		Weight	
	Mean	Median	Mean	Median	Mean	Median	Mean	Median
	30.26	30.00	30.33	30.00	27.00	27.00	7.10	7.31

Figure 20: Various Means and Medians from the Births Data

		Sofa_Id
		N
Gender	female	N=503
	male	N=497
Habit	nonsmoker	N=873
	smoker	N=126
Premie	full term	N=846
	premie	N=152

Figure 21: Various Modes from the Births Data

		Weeks		Weight	
		Mean	Median	Mean	Median
Habit	nonsmoker	38.32	39.00	7.14	7.31
	smoker	38.44	39.00	6.83	7.06

Figure 22: Length of Term and Baby's Weight Grouped by Smoker

2.5 DELIVERABLE

Complete the following activities in this lab:

Number	Name	Page
2.3.2	Activity 1: Central Measures	20
2.3.4	Activity 2: Grouping	22
2.3.6	Activity 3: Mode for Numeric Data	24
2.3.7	Activity 4: Mode for Text Data	24

Screen capture the tables generated by each of these activities, add all of them into a single document, and submit that document for grading.

3

DATA DISPERSION

3.1 INTRODUCTION

One way to describe a dataset is to report its dispersion, or spread. For example, if a professor administered a test to 100 students and the scores were between $90-100$ that would be a fairly tight group but if another class had scores between $60-100$ that would indicate something completely different. This lab explores the concept of data dispersion and the methods used to describe that value.

3.2 MEASURES OF DATA DISPERSION

3.2.1 *Range*

The maximum and minimum values are those at the extreme ends of the dataset and the range is nothing more than the maximum minus the minimum values. For the 2016 version of the Scholastic Aptitude Test (SAT) the maximum score is 1600 and the minimum score is 400, so the range is $1600-400$, or 1200.

3.2.2 *Quartiles*

A measure that is closely related to the median[1] is the first and third quartile. The first quartile ($Q1$) is the score that splits the lowest 25% of the values from the rest and the third quartile ($Q3$) splits the highest 25% of the values from the rest. The second quartile ($Q2$) is the same as the median and, normally, the term "median" is used rather than $Q2$. 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.

Occasionally, the word "hinges" appears in statistical literature. The two hinges for a dataset are the medians for the lower half and the

1 The median was described in Lab 2.2.3, page 16.

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.

Another measure of dispersion that is occasionally used is the Inter-Quartile Range (*IQR*); that is, the difference between $Q1$ and $Q3$. This is used to counter the skew introduced by a dataset with extreme outliers.

3.2.3 *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.[2]

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

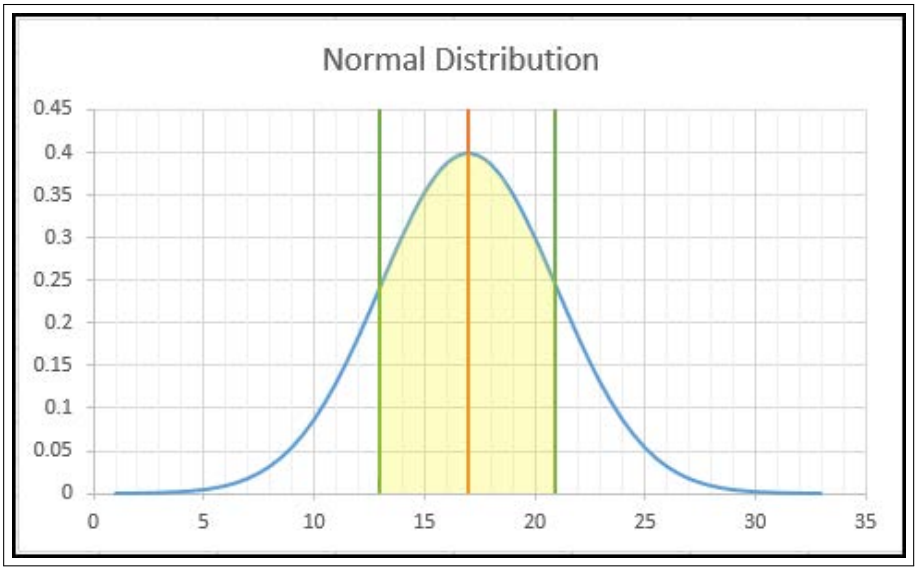


Figure 23: Illustration Of Standard Deviation

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

2 The concept of the normal distribution curve was presented in Lab 1.3.2.1 on page 6.

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 (these larger values 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 that 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.

3.3 PROCEDURE

3.3.1 *Statistical Calculations*

Start `SOFA` and select "Report Tables." Then:

1. Data Source Table: dbims
2. Table Type: Row Stats
3. Columns: Age (age)
4. Title: Age Statistics

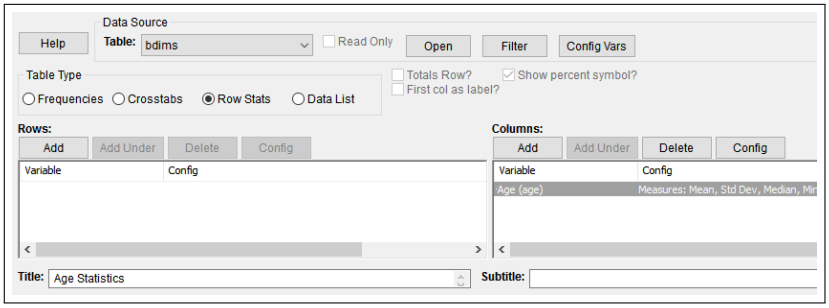


Figure 24: Setting Up Age

5. Just above the "Columns" window, click the "Config" button and select: Mean, Standard Deviation, Median, Minimum, and Maximum. Of course, any of the available measures, such as range or quartiles, can be selected depending upon what needs to be reported.

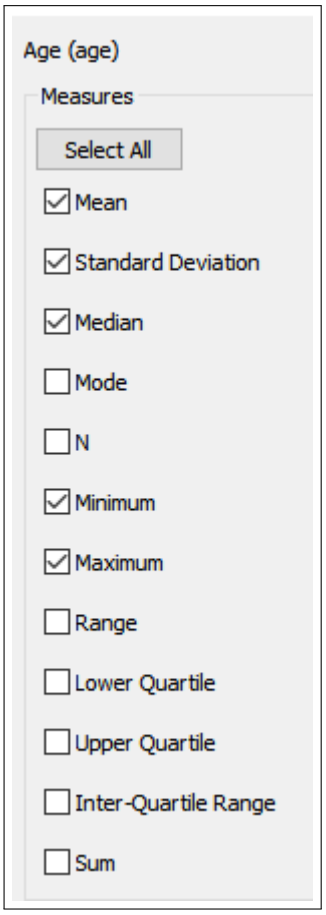
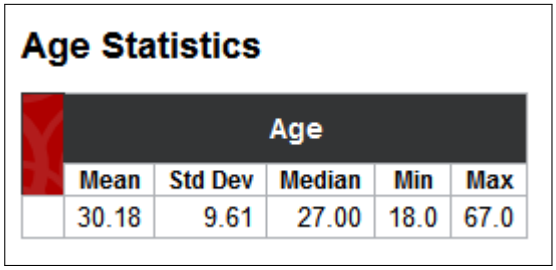


Figure 25: Configure Options for Age

6. Read those values in the lower left corner of the "Make Report Table" window.



Age Statistics

	Age				
	Mean	Std Dev	Median	Min	Max
	30.18	9.61	27.00	18.0	67.0

Figure 26: Statistics for Age

3.3.2 *Activity 1: Simple Statistics*

Using the *maincafe* dataset in `SOFA`, produce a table that contains the standard deviation, minimum, maximum, range, lower quartile, upper quartile, and inter-quartile range for Age. The table should have a title of "Data Dispersion, Activity 1" and a subtitle of "Measures of Dispersion".

3.3.3 *Grouping Variables*

To produce the statistics for grouped variables, like "the age statistics by sex," add the "sex" variable to the "Rows:" window and modify the title of the chart.

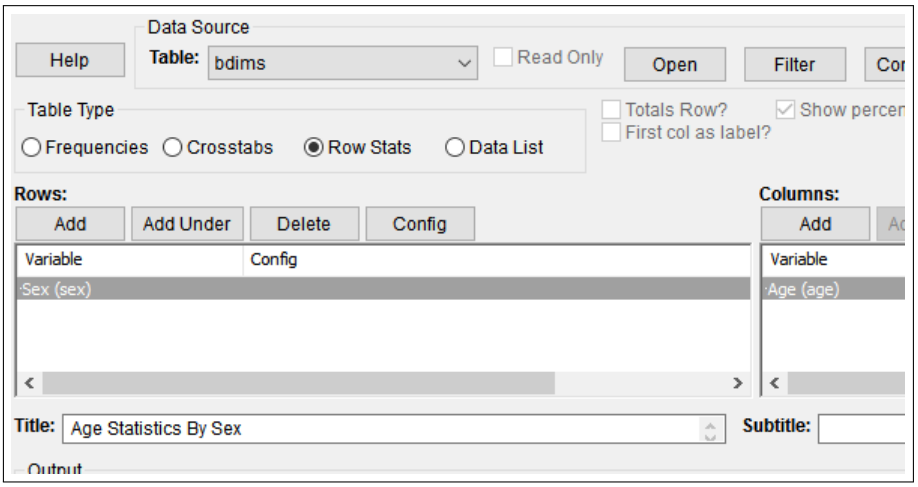


Figure 27: Grouping Age Statistics by Sex

Then, the output will display the selected Age statistics by sex.

Age Statistics By Sex

		Age				
		Mean	Std Dev	Median	Min	Max
Sex	female	28.77	8.85	26.00	18.0	67.0
	male	31.67	10.15	29.00	18.0	65.0

Figure 28: Age Statistics by Sex

3.3.4 *Activity 2: Grouped Statistics*

Using the *maincafe* dataset in SOFA, produce a table that contains the mean, standard deviation, and N for Age when grouped by Sex. The table should have a title of "Data Dispersion, Activity 2" and a subtitle of "Grouped Measures of Dispersion".

3.3.5 *Filtering*

SOFA provides researchers a way to filter the data such that only a specified subset is used in calculations. As an example, to analyze only the males in the dataset:

Start SOFA and select "Report Tables." Then:

1. Data Source Table: dbims
2. Table Type: Row Stats
3. Columns: Age (age)
4. Configure the Age column to display the Mean, Median, and N
5. Title: Age Statistics For Males
6. Click the "Filter" button near the top of the window
7. Select "Sex" in the dropdown list for the "Quick" button
8. Select the "=" comparator
9. Enter "male" as the match term (Note: do not use quote marks)
10. Add the optional label: "Analyze Age For Males Only"

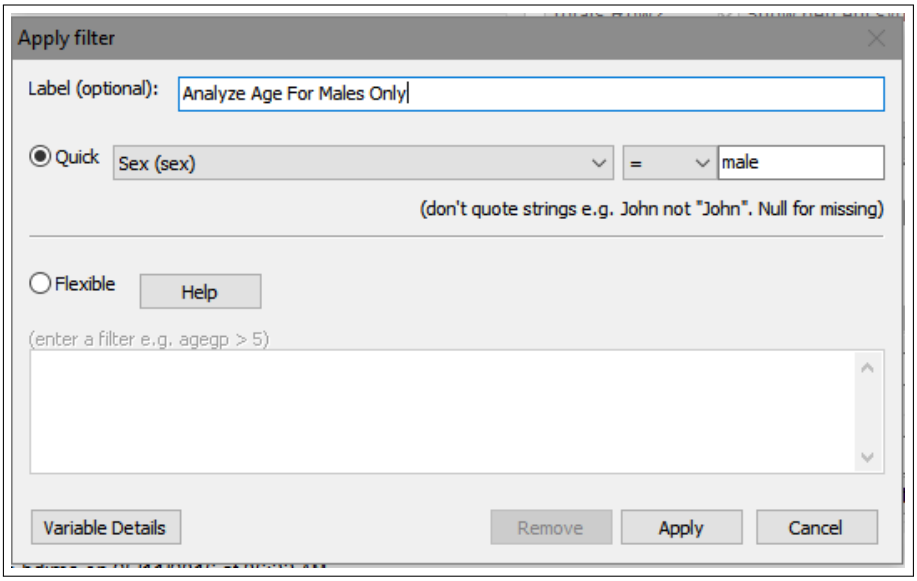


Figure 29: Setting Up a Filter

Then, the output will display the selected Age statistics for males only.

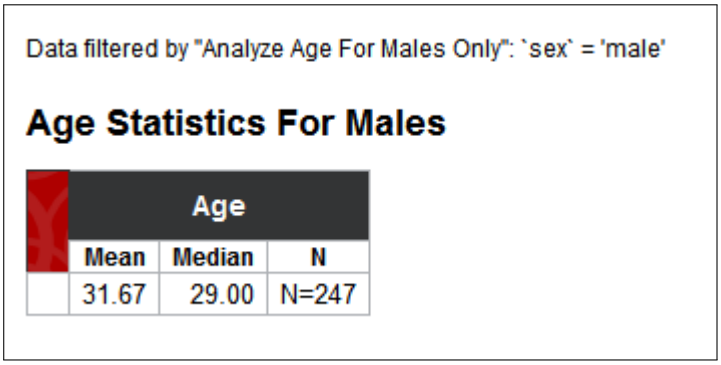


Figure 30: Age Statistics for Males

Important Note: Once a filter is applied it will remain until it is either manually removed or the SOFA session ends. To remove a filter, open the filter dialog box and click the "remove" button.

3.3.6 *Activity 3: Filtering*

Using the *maincafe* dataset in SOFA, produce a table that contains the mean, standard deviation, and N for Age for only males. The table should have a title of "Data Dispersion, Activity 3" and a subtitle of "Filtered Statistics".

3.4 DELIVERABLE

Complete the following activities in this lab:

Number	Name	Page
3.3.2	Activity 1: Simple Statistics	31
3.3.4	Activity 2: Grouped Statistics	32
3.3.6	Activity 3: Filtering	33

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

4 VISUALIZING DISPERSION

4.1 INTRODUCTION

SOFA makes it easy to calculate various measures of dispersion, as covered in Lab 3; however, most people find it easier to understand the dispersion of data when that is presented graphically. Fortunately, SOFA has a great graphic tool for visualizing data dispersion: Box Plot (sometimes called "Box and Whisker" plot). A Box Plot graphically illustrates Q1, the median, Q3, and outliers (if any are present).

The *bdims* dataset is used to illustrate a Box Plot. Following is the statistical data for *Age* from the *bdims* dataset along with the box plot for that same data.

Age (years)						
Min	**Q1**	**Median**	**Q3**	**Max**	**Range**	**IQR**
18	23	27	36	67	49	13

Following is the box plot that illustrates the above statistics.

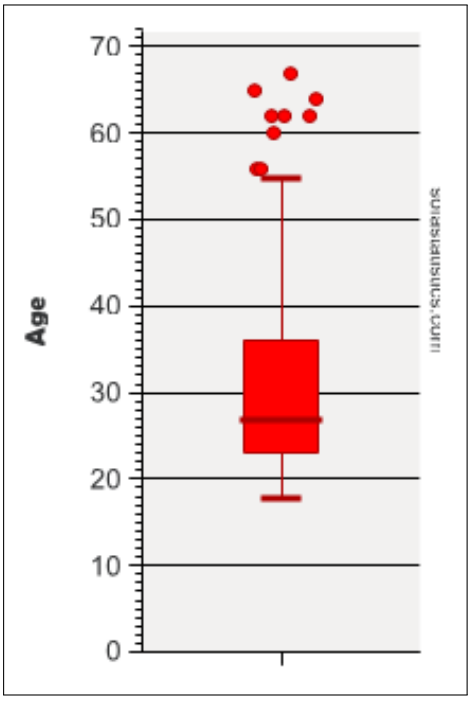


Figure 31: Box Plot for Bdims Age

In the above Box Plot, the median is indicated by a dark line at 27, Q1 is 23 (the lower edge of the box) and Q3 is 36 (the upper edge of

the box). That makes the Inter-Quartile Range (IQR) equal to 13 (the size of the box).

The "whiskers" are placed at 1.5XIQR above and below each quadrant. Thus:

1. **Upper**. Q3 + (1.5XIQR), or 36 + (1.5X13) = 55.5
2. **Lower**. Q1–(1.5XIQR), or 23 – (1.5X13) = 19.5

Note: `SOFA` makes it possible to set the whiskers at the maximum and minimum values, but that hides the outliers and is rarely done.

The circles above the box plot represent outliers. In this case there are nine outliers. If the data are a normal distribution, then the whiskers will enclose most of the values in the dataset and outliers will be rare.

Here is a second example from the same dataset:

Height (cm)						
Min	**Q1**	**Median**	**Q3**	**Max**	**Range**	**IQR**
147.20	163.80	170.30	177.80	198.10	50.90	14.00

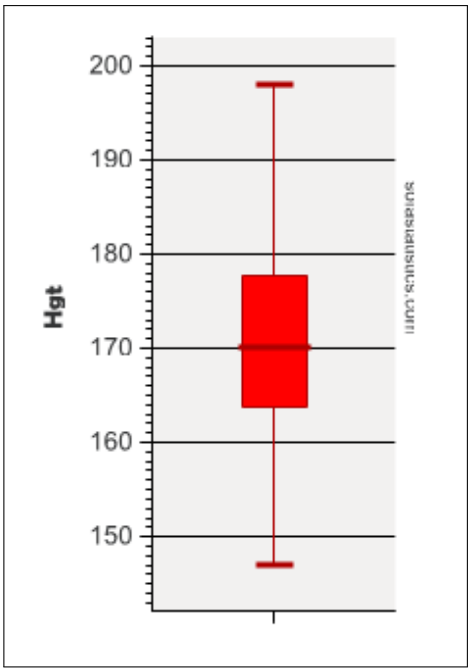


Figure 32: Box Plot for Bdims Height

In this case, there are no outliers so there are no circles above or below the whiskers. Also, the data are very "bunched up" and all values lie inside 1.5 times the minimum and maximum values, so the whiskers lie at the minimum and maximum.

Box plots become much more useful when more than one data item is plotted side-by-side for comparison. For example, the following

box plot is helpful in determining if there is a difference in height by sex.

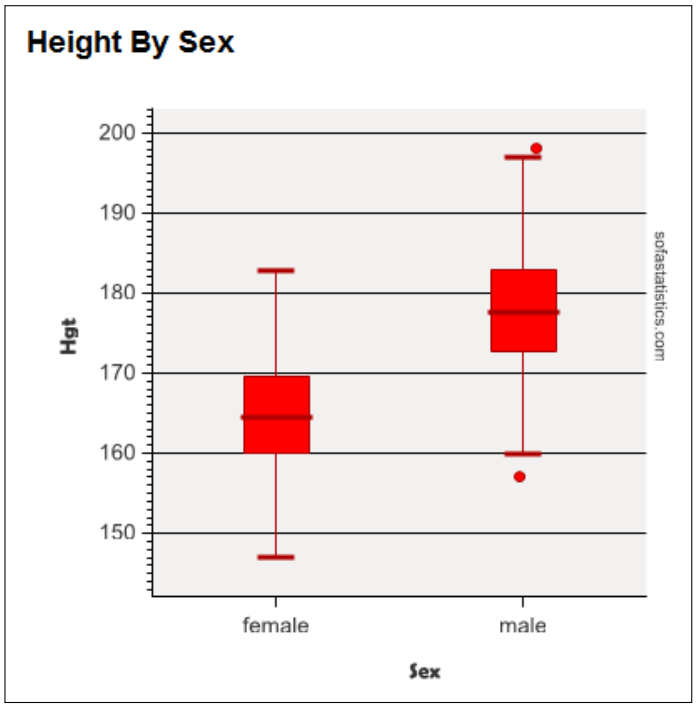


Figure 33: Comparing Height By Sex

By comparing the two box plots it is very easy to see that males are generally taller than females since that box is higher on the graph. Also notice that the "males" plot includes outliers at both the minimum and maximum values which indicates a greater variation in male heights than female.

As a final example of box plots, consider the *cars* dataset. The following box plot shows the price of a new automobile by the number of passengers it can carry.

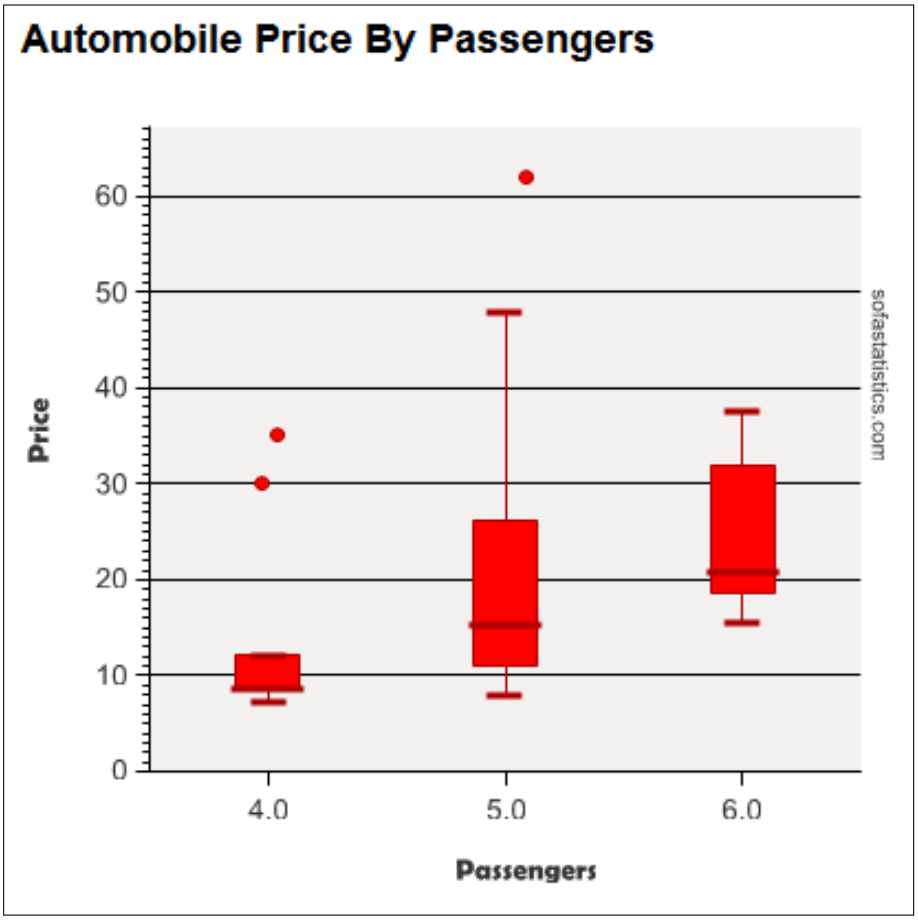


Figure 34: Car Prices By Passenger Capacity

An analysis of Figure 34 leads to the following:

- Generally, four-passenger cars are less expensive than five-passenger cars and they, in turn, are less expensive than six-passenger cars since the boxes for those vehicles are progressively higher on the graph.
- Five-passenger automobiles have a wider spread of prices than the other two types.
- Notice that for five-passenger automobiles the upper whisker is a long way from the top of the box. That indicates that the dataset are skewed such that most values are between about $12K and $27K but there are a number of values that are higher along with one outlier above $60K.
- The box for the four-passenger cars is small and the whiskers are nearly the same as the ends of the box which means that there is very little variation in the prices, though there are two extreme outliers.

4.2 PROCEDURE

4.2.1 *Boxplots*

Start SOFA and select "Charts." Then:

1. Data Source Table: births
2. Chart Types: Make Box and Whisker Plot (the last button on the right)
3. Variables Described: Gained (gained)
4. Title: Weight Gained

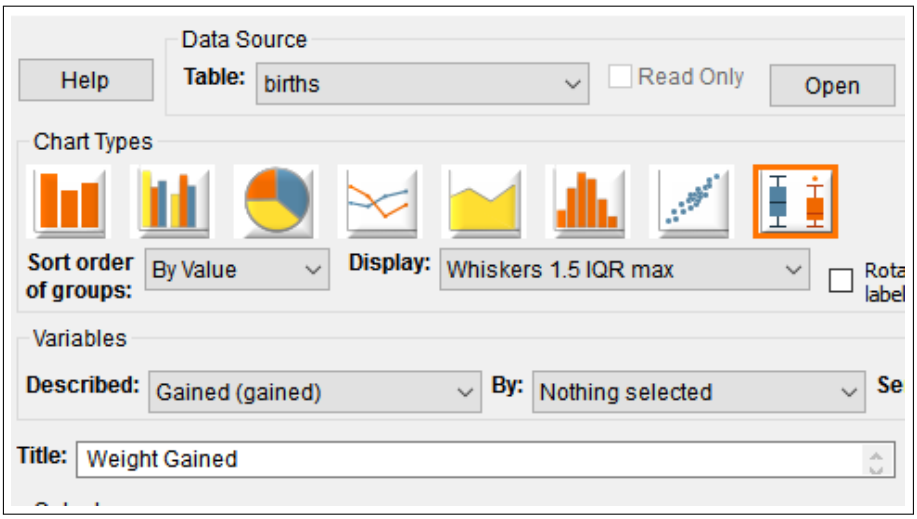


Figure 35: Set Up Boxplot for Weight Gained

Following is the plot generated by the above.

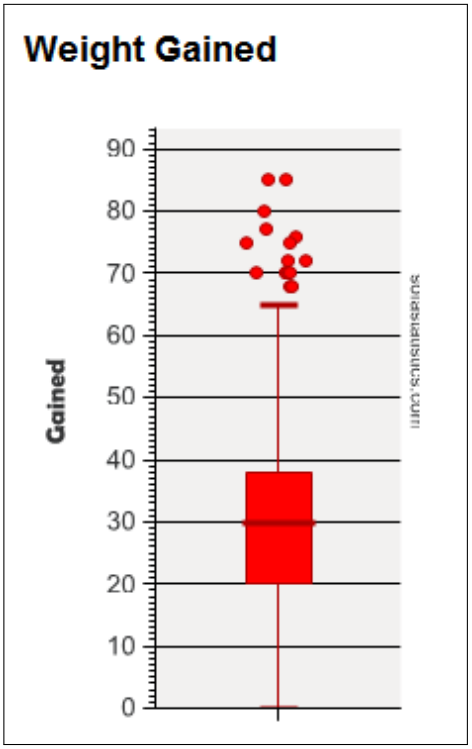


Figure 36: Box Plot of Weight Gained

4.2.2 *Activity 1: Simple Boxplot*

Using the *maincafe* dataset in `SOFA`, produce a boxplot for Age. The boxplot should have a title of "Visualizing Dispersion, Activity 1" and a subtitle of "Boxplot For Age".

4.2.3 *Grouped Boxplot*

To group two or more boxplots, select a grouping variable to create grouped plots. For example, to group the weight gain box plots by whether the mother was a smoker, select:

1. Data Source Table: births
2. Chart Types: Make Box and Whisker Plot (the last button on the right)
3. Variables Described: Gained
4. By: Habit
5. Title: Weight Gained by Smoking Habit

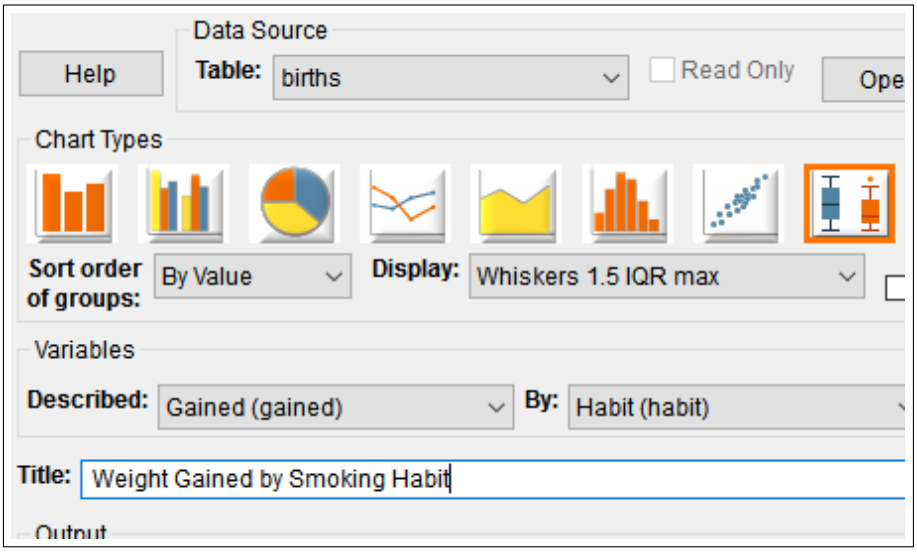


Figure 37: Set Up Boxplot for Weight Gained by Habit

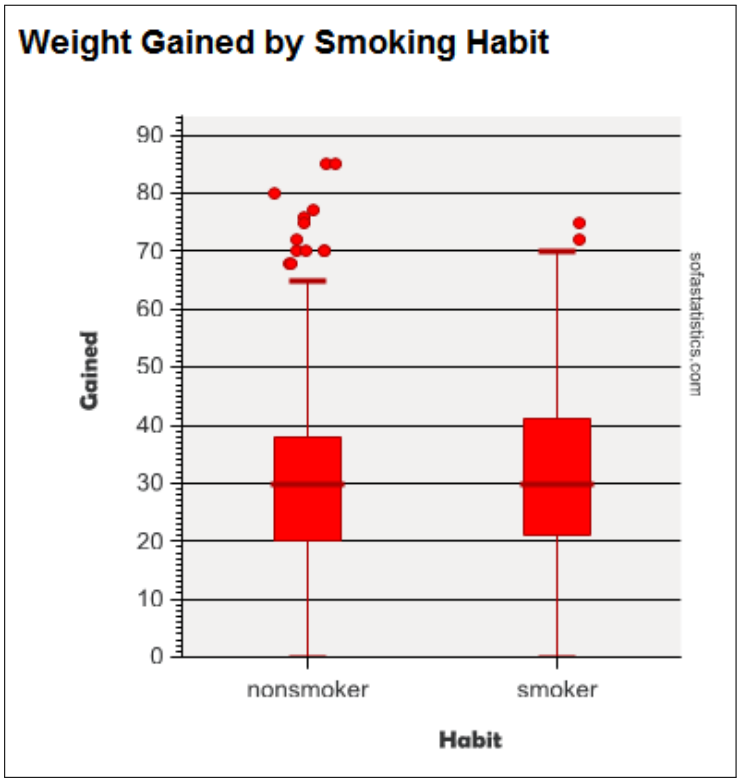


Figure 38: Box Plot of Weight Gained by Habit

4.2.4 *Activity 2: Grouped Boxplots*

Using the *maincafe* dataset in `SOFA`, produce a boxplot for Age grouped by Meal. The boxplot should have a title of "Visualizing Dispersion, Activity 2" and a subtitle of "Boxplot of Age Grouped By Meal".

4.2.5 *Grouped Boxplots By Series*

Finally, the data can be grouped by series. For example, to group the above box plots by the sex of the baby, select "Gender" as the "Series By" variable and then change the title and subtitle of the box plot.

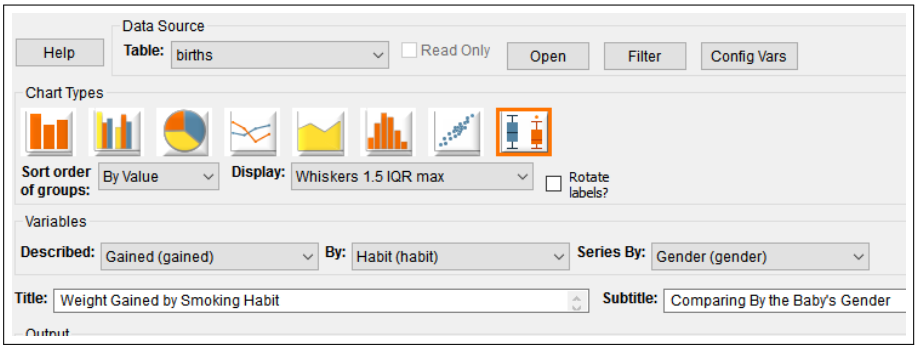


Figure 39: Set Up Boxplot for Weight Gained by Habit and Baby's Sex

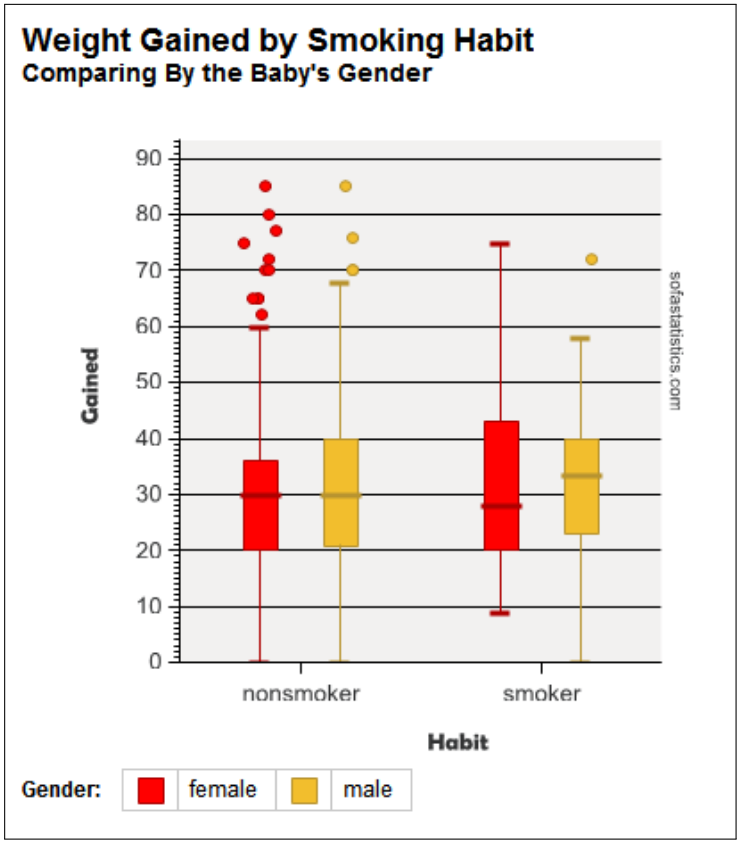


Figure 40: Box Plot of Weight Gained by Habit and Baby's Sex

In box plot 40 it is clear that for non-smokers, male baby's weights are more variable than females since that box is larger. Interestingly, the weight of female babies born to smokers has a much larger variability (the box is larger) and the mean of the weights for female babies is somewhat lower than for males.

As a second example of a box plot, the *email* dataset was selected and the size of the email message (number of characters) was plotted as a function of whether it was spam and contains large numbers.

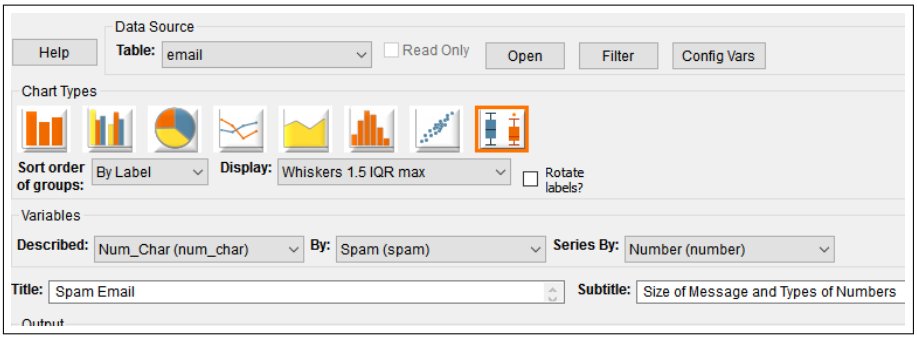


Figure 41: Set Up Boxplot for Email Data Set

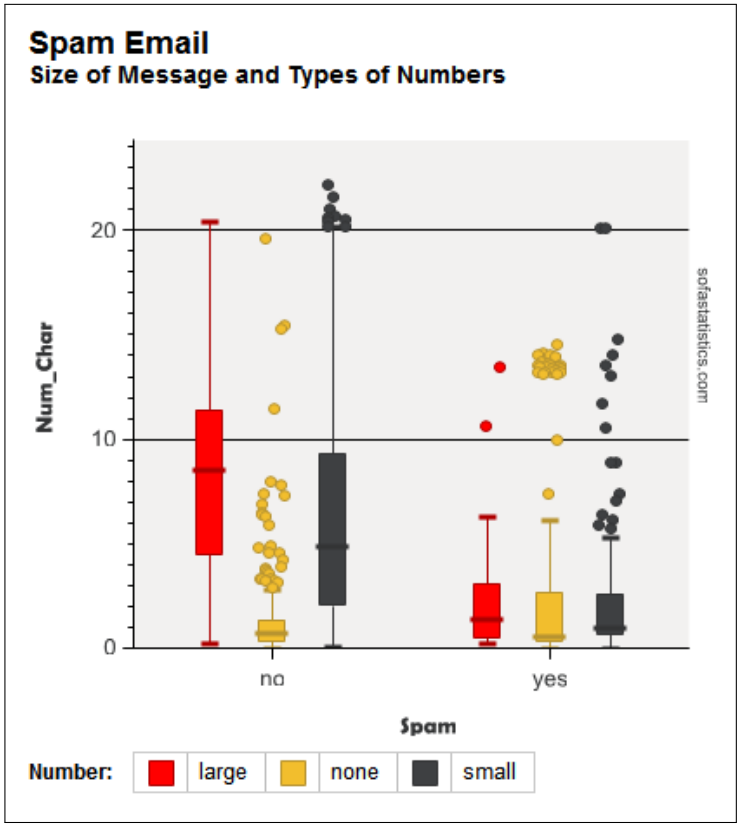


Figure 42: Example Box Plot for Email Data Set

To interpret box plot 42, notice that, generally, the spam messages contain fewer characters than non-spam messages (those plots are lower), indicating that spam messages are generally shorter than non-spam messages. Then, looking at only the non-spam plots (the three on the left), notice that if the message contains large numbers (the first plot from the left) it will generally contain more characters than messages with only small numbers (the third plot). Finally, notice that non-spam messages with no numbers (the second plot) are generally short (fewer characters, indicated by a box that is lower on the scale). There are a number of other characteristics that could be drawn from this one chart, such as a discussion about the outliers or the means for each type of message.

The last example (the email box plots) illustrates that a box plot contains a lot of information and is a valuable tool for research reports.

4.2.6 *Activity 3: Grouped Boxplots by Series*

Using the *maincafe* dataset in `SOFA`, produce a boxplot for Age grouped by Sex and using Meal as a series. The boxplot should have a title of

"Visualizing Dispersion, Activity 3" and a subtitle of "Boxplot of Age by Sex and Meal".

4.3 DELIVERABLE

Complete the following activities in this lab:

Number	Name	Page
4.2.2	Activity 1: Simple Boxplot	40
4.2.4	Activity 2: Grouped Boxplots	42
4.2.6	Activity 3: Grouped Boxplots by Series	44

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

5 FREQUENCY TABLES

5.1 INTRODUCTION

Nominal and Ordinal data items are normally reported in frequency tables where the counts for a particular item are displayed. Crosstabs are a type of frequency table used to compare the counts of items that have been grouped in some way. This lab explores both frequency tables and crosstabs.

5.2 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.

Passenger Ratings Per Car

		Freq	Col %
Passengers	4.0	10	18.5%
	5.0	28	51.9%
	6.0	16	29.6%
	TOTAL	54	100.0%

Figure 43: Passenger Ratings 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.

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.

Images Per Email

		Freq	Col %
Image	0.0	3076	97.5%
	1.0	59	1.9%
	2.0	13	0.4%
	3.0	5	0.2%
	5.0	1	0.0%
	TOTAL	3154	100.0%

Figure 44: Images Per Message

Figure 44 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.

5.3 CROSSTABS

A crosstab (sometimes called a contingency table or pivot table), is a table of frequencies used to display the relationship between two nominal or ordinal variables. These are commonly used around election time when pollsters create tables that show things like the number of people who voted for or against some proposition counted by gender, race, or some other factor. As an example of a crosstab, consider the following from the *email* dataset.

Spam By Number of Addressees

		Spam	
		no	yes
		Freq	Freq
To_Multiple	no	2288	333
	yes	521	12

Figure 45: Addressees As A Function of Spam

In Figure 45 notice that 2288 messages were sent to a single addressee and were identified as not spam while 521 messages were sent to multiple addressees and identified as not spam. By using a crosstab, a researcher can determine the frequency of some incident (email messages) by two different criteria (spam and multiple addressees).

Here is a second example from the *email* dataset:

Images in Spam

		Spam	
		no	yes
		Freq	Freq
Image	0.0	2733	343
	1.0	57	2
	2.0	13	0
	3.0	5	0
	5.0	1	0

Figure 46: Images As A Function of Spam

In Figure 46 notice that spam in general has no images. In fact, only two email messages out of 345 had one image, and no messages had more than one.

5.3.1 *Complex Crosstabs*

It is possible to create crosstabs that are quite complex, with multiple subcategories for both rows and columns. However, these tables are often too complex to be easy to interpret. Consider the table in Figure 47.

Images in Spam By Multiple Addressees

				Spam	
				no	yes
				Freq	Freq
Image	0.0	To_Multiple	no	2249	331
			yes	484	12
	1.0	To_Multiple	no	26	2
			yes	31	0
	2.0	To_Multiple	no	8	0
			yes	5	0
	3.0	To_Multiple	no	4	0
			yes	1	0
	5.0	To_Multiple	no	1	0

Figure 47: Images and Addressees As A Function of Spam

In the crosstab presented, 2249 email messages were identified as not spam, had zero images, and went to a single addressee. While this crosstab presents a lot of data in a compact form it is difficult to read and make sense of any one data cell. In general, it is preferable to have only one variable for both the rows and columns in a crosstab.

5.4 PROCEDURE

5.4.1 *Frequency Table*

Start `SOFA` and select "Report Tables." Then:

1. Data Source Table: email
2. Table Type: Frequencies
3. Rows: Format
4. Title: Types Of Email Messages
5. Row Config: Check "Total"
6. Column Config: Check "Frequency" and "Column %"

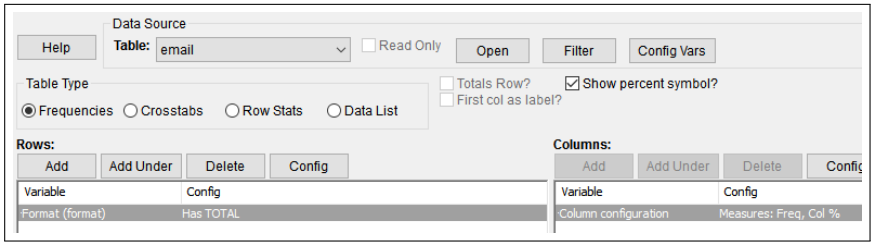


Figure 48: Setting Up Email Frequency Table

Types Of Email Messages

		Freq	Col %
Format	html	2011	63.8%
	text	1143	36.2%
	TOTAL	3154	100.0%

Figure 49: Email Frequency Table

5.4.2 *Activity 1: Frequency Table*

Using the *maincafe* dataset in SOFA, produce a frequency table for Meal. The table should include both frequency and percentage for each of the four types of meals. The table should have a title of "Frequency Tables, Activity 1" and a subtitle of "Frequency Table for Types of Meals".

5.4.3 *Crosstabs*

Start SOFA and select "Report Tables." Then:

1. Data Source Table: births
2. Table Type: Crosstabs
3. Rows: Habit
4. Columns: Premie
5. Title: Premature Births By Habit

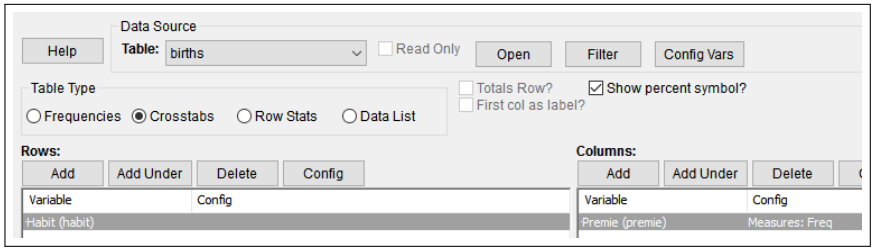


Figure 50: Setting Up Premature Births By Habit

Premature Births By Habit

		Premie	
		full term	premie
		Freq	Freq
Habit	nonsmoker	739	133
	smoker	107	19

Figure 51: Premature Births By Habit

To create a more complex crosstab, follow the instructions above for a simple crosstab but then click "Add Under" for Rows and select "Mature."

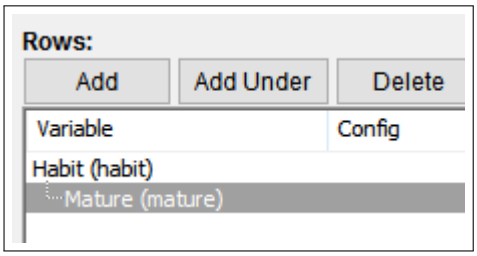


Figure 52: Setting Up Premature Births By Habit and Maturity

Premature Births By Habit and Maturity

				Premie	
				full term	premie
				Freq	Freq
Habit	nonsmoker	Mature	mature mom	100	21
			younger mom	639	112
	smoker	Mature	mature mom	9	2
			younger mom	98	17

Figure 53: Premature Births By Habit and Maturity

5.4.4 *Activity 2: Crosstabs*

Using the *maincafe* dataset in `SOFA`, produce a crosstab with the rows being Food and the columns being Svc. The crosstab should include only frequency and have a title of "Frequency Tables, Activity 2" and a subtitle of "Crosstab for Food and Service".

5.4.5 *Activity 3: Complex Crosstabs*

Using the *maincafe* dataset in `SOFA`, produce a crosstab with the rows being Pref with a sub-group of Sex and the columns being Meal. The crosstab should include only frequency and should have a title of "Frequency Tables, Activity 3" and a subtitle of "Crosstab of Preference by Sex and Meal".

5.5 DELIVERABLE

Complete the following activities in this lab:

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

6 VISUALIZING FREQUENCY

6.1 INTRODUCTION

Nominal and Ordinal data items are normally reported in frequency tables where the number of times a particular survey item was selected by respondents is displayed. However, there are many ways to visualize frequency data and many people find various charts and graphs to be useful. This lab introduces the visualization tools available in SOFA.

6.2 VISUALIZING DATA

6.2.1 *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. Histograms are especially useful for interval or ratio data since SOFA will automatically cluster the data into "bins."

As an example of a histogram, Figure 54 shows the mother's age from the *births* dataset.

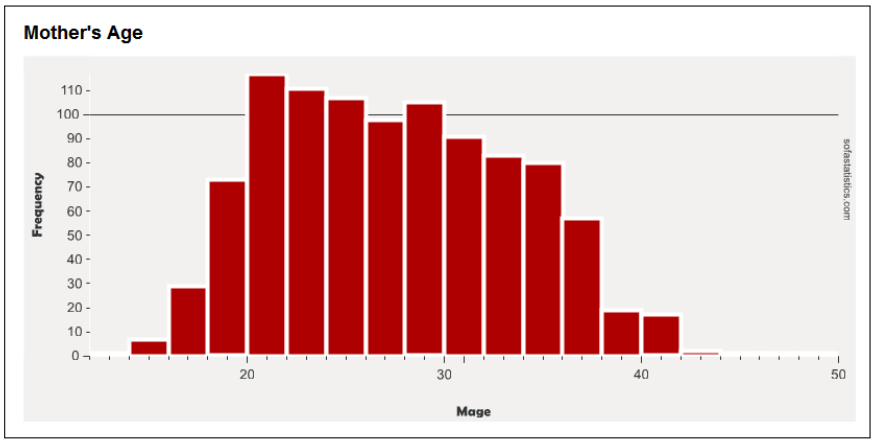


Figure 54: Histogram of Mother's Age

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

As another example, Figure 55 shows a histogram for baby's weight from the *births* dataset.[1]

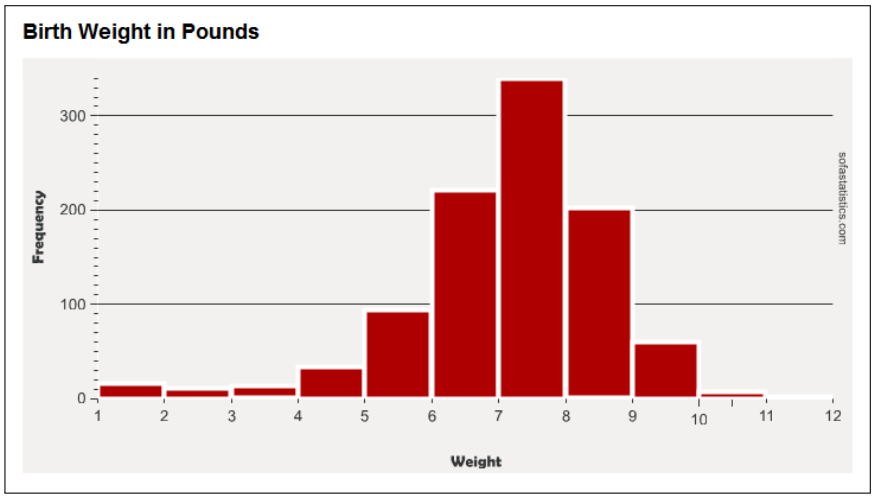


Figure 55: Histogram of Baby's Weight

As in Figure 54, any weight between 6 and 6.99 pounds is clustered in a single bar between 6 and 7.

6.2.2 *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. Figure 56 is a bar chart showing the prevalence of various drive trains in the *cars* dataset.

1 Using a histogram aids a researcher in determining if a dataset is normally distributed and skewed. Figure 55 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 1.3.2.1 on page 6 discusses the shape of a normal distribution.

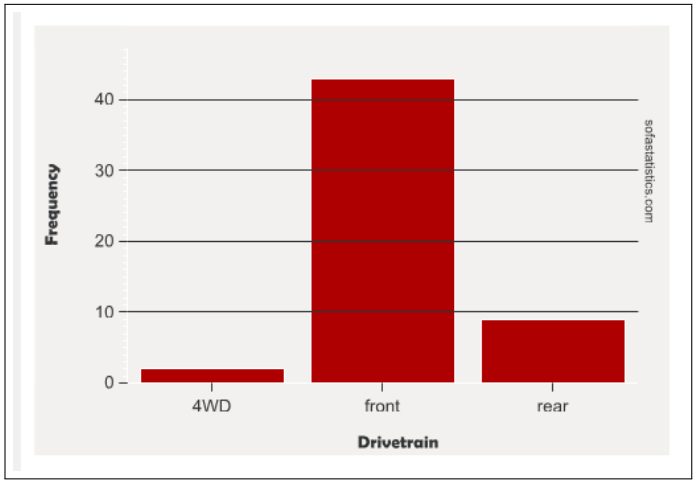


Figure 56: Prevelance of Types of Drive Trains

Figure 57 shows the maturity level of the mothers in the *births* dataset and unsurprisingly indicates that most mothers are younger.

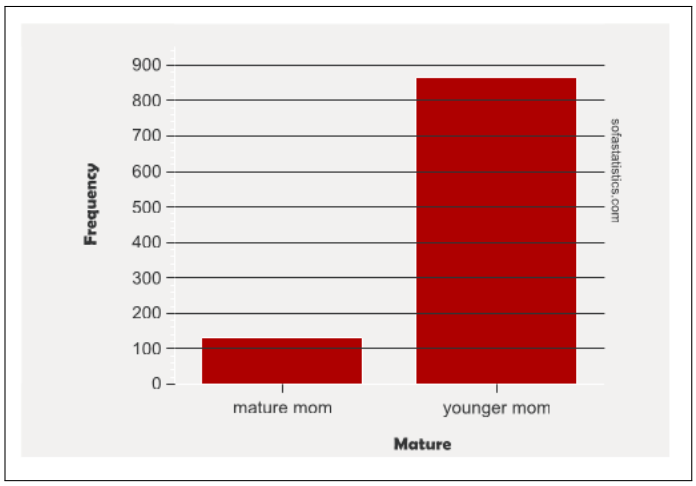


Figure 57: Maturity of Mothers

6.2.3 *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 58 is a clustered bar chart that shows the incidence of premature births by the mother's smoking habit in the *births* dataset.

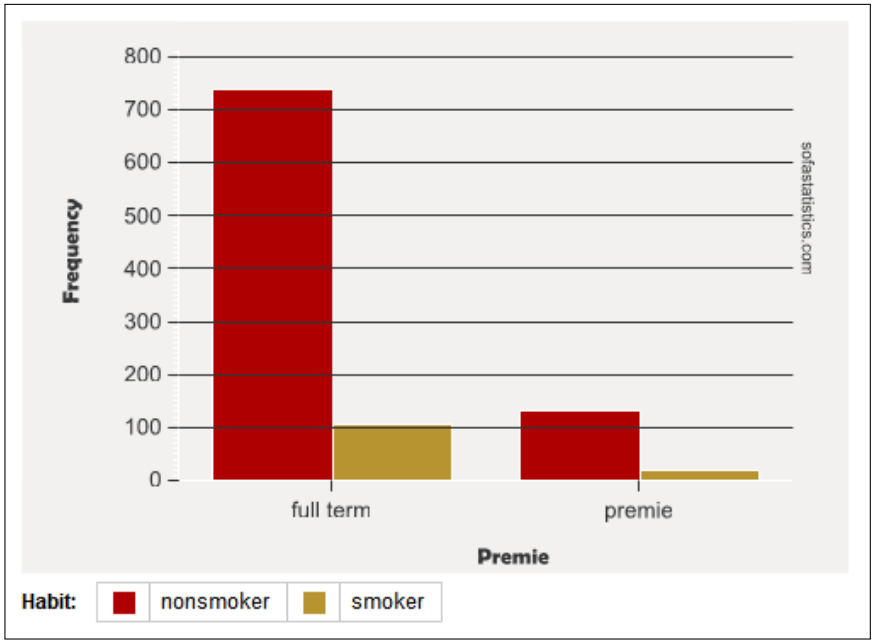


Figure 58: Permature Births By Smoking Habit

Figure 59 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.

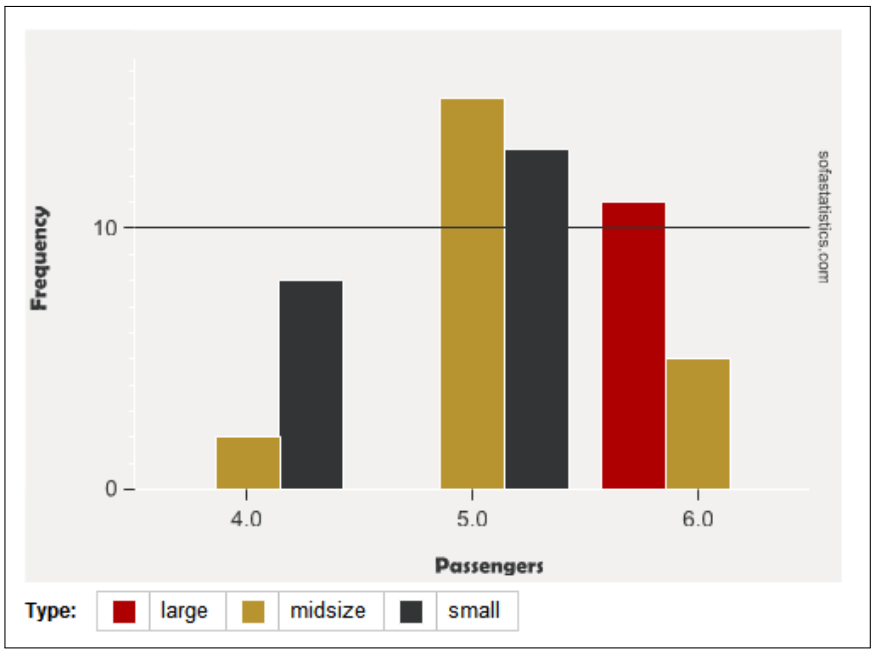


Figure 59: Number of Passengers By Car Type

6.2.4 *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 60 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.

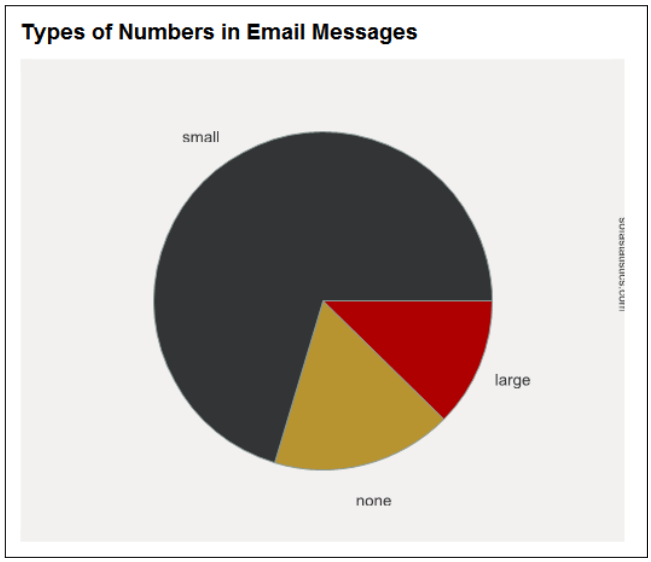


Figure 60: Types of Numbers in Email Messages

As an extreme example of a poorly used pie chart, consider Figure 61. Even ignoring the problem of the numbers 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.

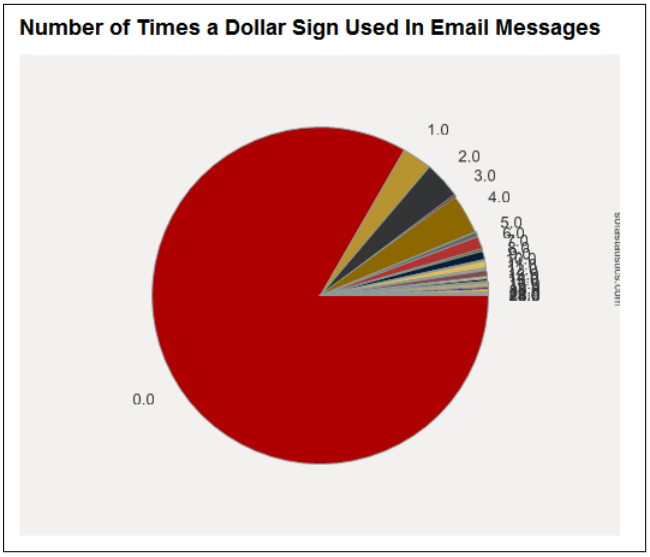


Figure 61: Number of Times a Dollar Sign Used in Email Messages

6.2.5 *Line Charts*

Line charts display the frequency of some value in a linear form that makes trend detection easier. As an example, consider the following from the *gifted* dataset which charts the number of hours that parents spent reading to their children.

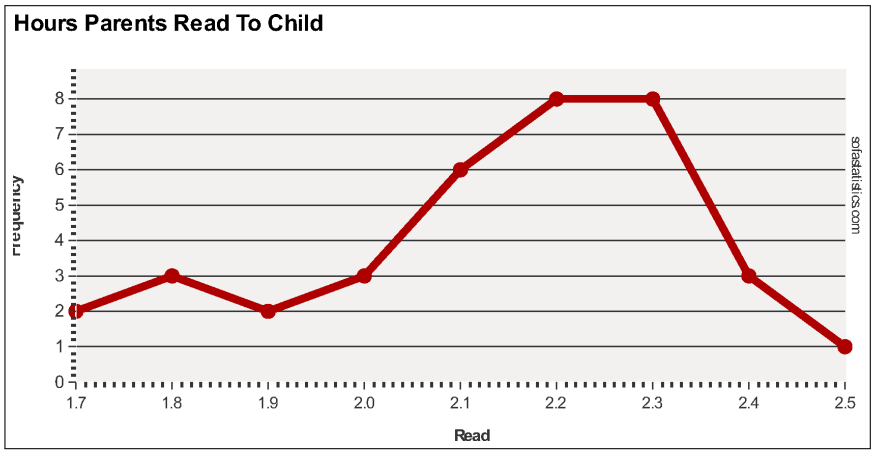


Figure 62: Hours Spent Reading

This chart makes it easy to see that most parents in this survey read to their children about 2.2 – 2.3 hours per week.

6.3 PROCEDURE

Start SOFA and select "Charts." Then:

6.3.1 *Histogram*

1. Data Source Table: bdims
2. Table Type: Histogram (sixth button)
3. Values: Age
4. Title: Age Statistics

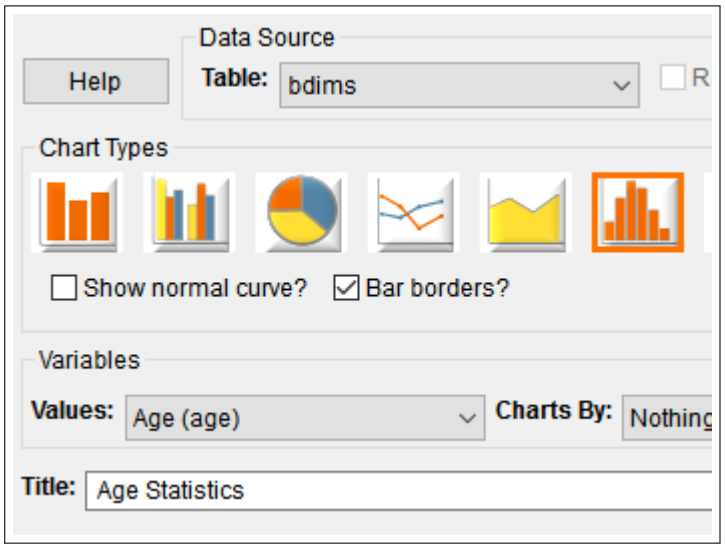


Figure 63: Setting Up Age Statistics

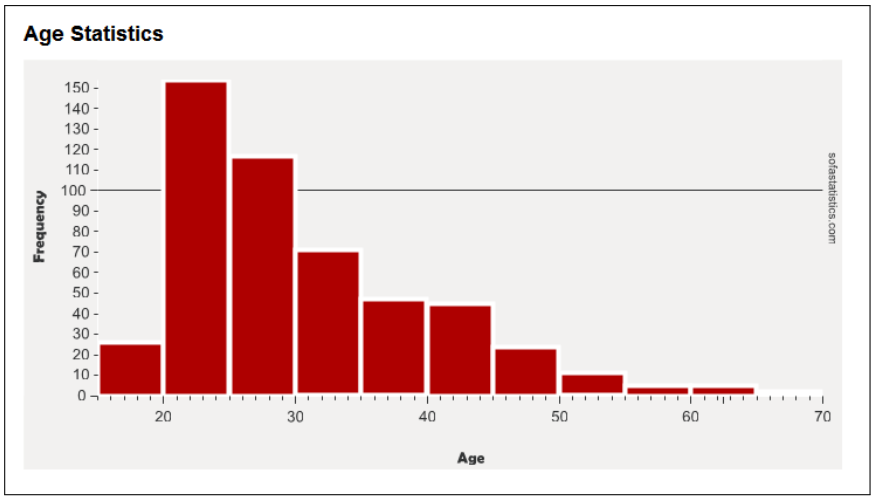


Figure 64: Age Statistics

6.3.2 *Activity 1: Histogram*

Using the *maincafe* dataset in SOFA, produce an histogram of Age. The histogram should have a title of "Visualizing Frequency, Activity 1" and a subtitle of "Histogram of Age".

6.3.3 *Line Charts*

To create a line chart:
Start SOFA and select "Charts" then:

1. Data Source Table: gifted
2. Table Type: Line Chart (fourth button)
3. Values: Read
4. Title: Hours Parents Read To Child

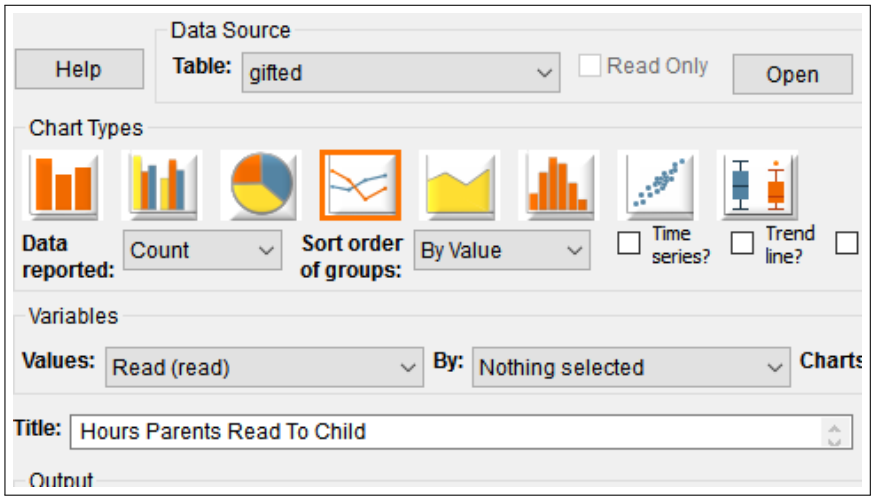


Figure 65: Setting Up Line Chart

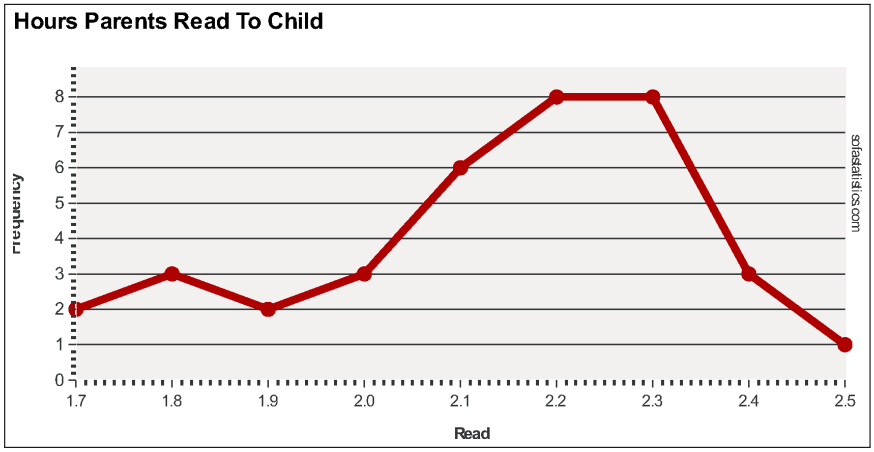


Figure 66: Line Chart For Time Parents Spend Reading To Children

If a "By" value is selected then a new line will be generated for each of the levels in the "By" variable. For example:

1. Data Source Table: births
2. Table Type: Line Chart (fourth button)
3. Values: Mage
4. By: Gender
5. Title: Mother's Age and Baby Gender

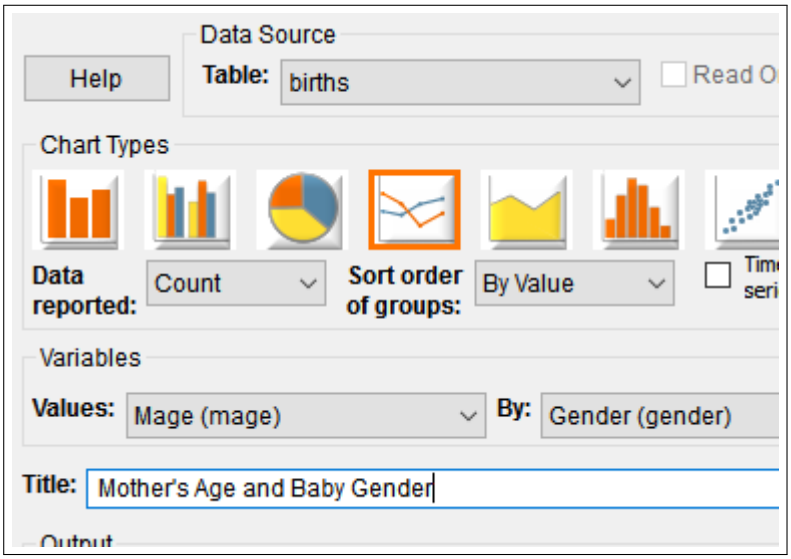


Figure 67: Setting Up Line Chart

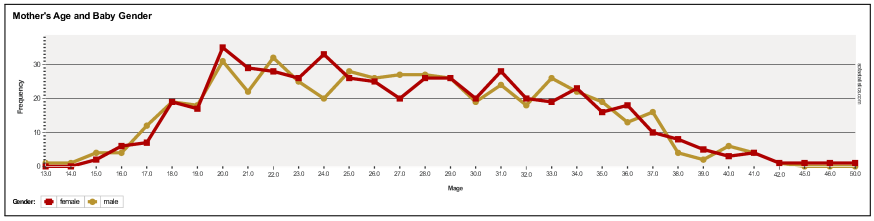


Figure 68: Line Chart For Mother's Age and Baby Gender

An "Area Chart" (button five) is the same as a line chart but the area under the line is colored which may make it easier to see trends.

1. Data Source Table: births
2. Table Type: Area Chart (fifth button)
3. Values: Mage
4. Title: Mother's Age

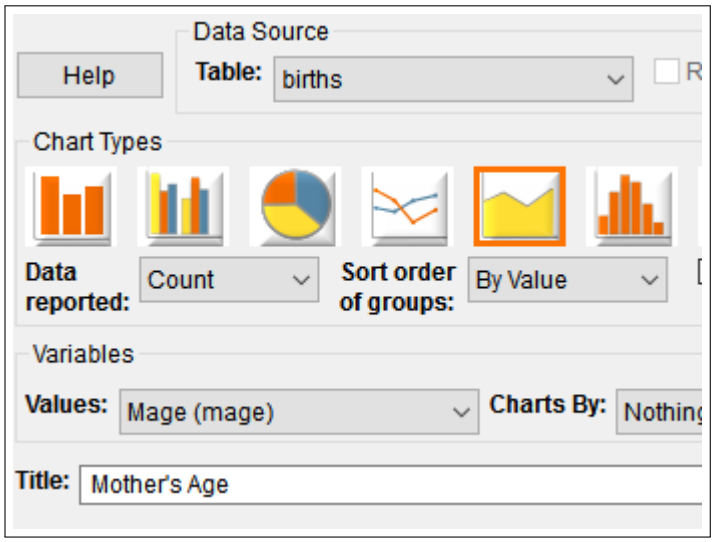


Figure 69: Setting Up Area Chart

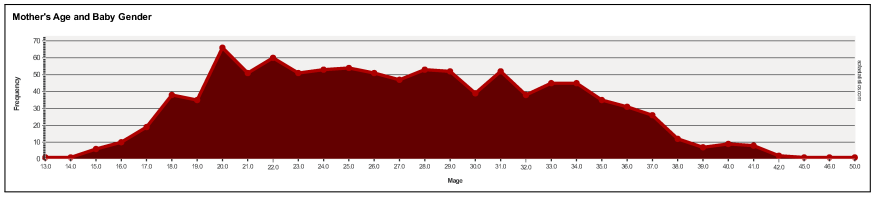


Figure 70: Area Chart For Mother's Age

Occasionally, data are in a "time series" where some observation was made over a period of time and SOFA is able to create line charts that display time series data.

1. Data Source Table: doorsurvey
2. Table Type: Line Chart (forth button)
3. Options: Check the "Time Series?" button
4. Dates/Times: Day
5. Title: Customer Counts

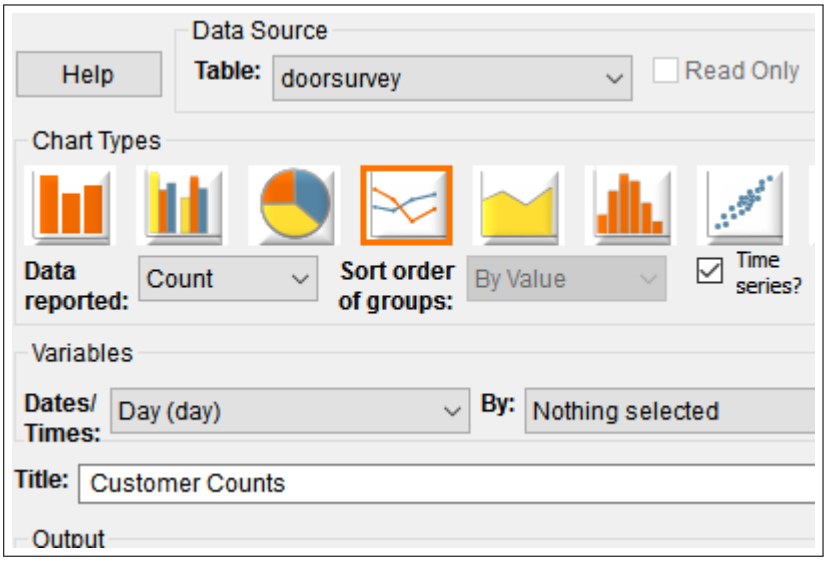


Figure 71: Setting Up Time Series Chart

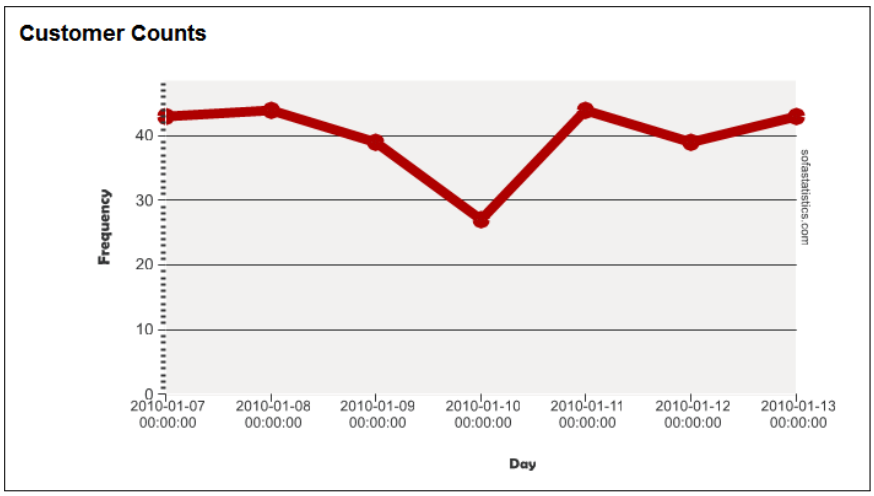


Figure 72: Time Series Chart For Customer Counts

SOFA is also able to display more than one time series on the same chart. For example, to display the number of customers by sex:

1. Data Source Table: doorsurvey
2. Table Type: Line Chart (forth button)
3. Options: Check the "Time Series?" button
4. Dates/Times: Day
5. By: Gender
6. Title: Customer Counts By Gender

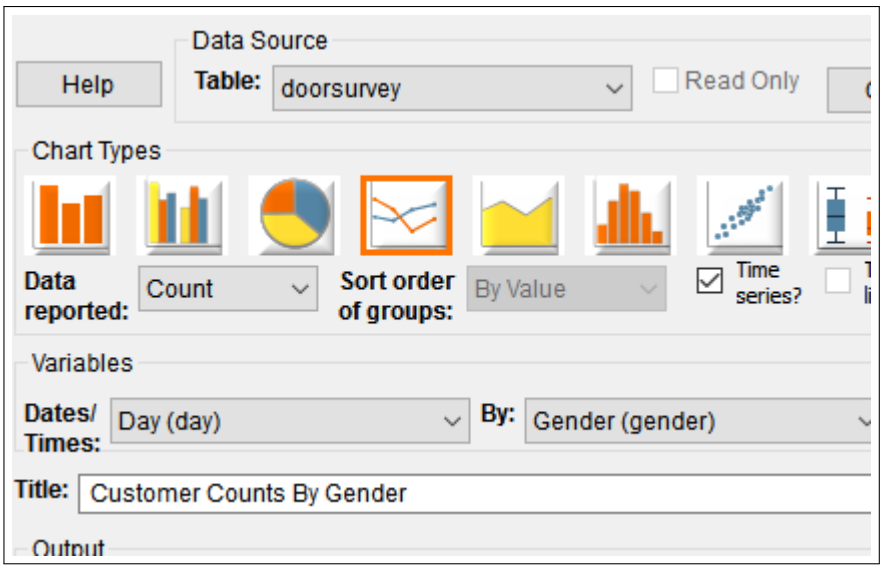


Figure 73: Setting Up Time Series Chart By Sex

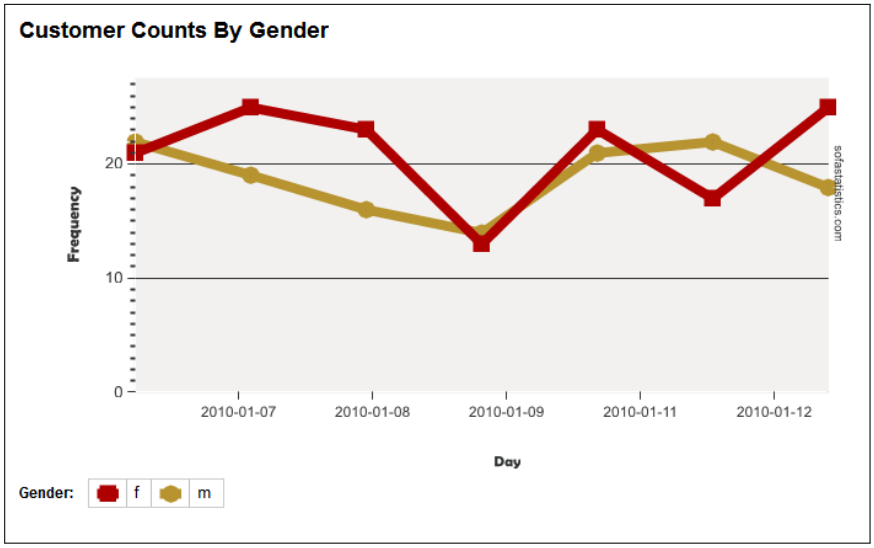


Figure 74: Time Series Chart For Customer Counts By Gender

SOFA includes a number of options for line charts that can be selected with the check boxes under the Chart Types buttons:

- **Trend Line?** is a straight line added to a time series chart to help visualize a variable's trend over time.
- **Smooth Line?** produces a smoothed-out line rather than jagged.
- **Rotate Labels?** makes longer labels fit the chart space better.
- **Hide Markers?** removes the "dots" from the line.
- **Major Labels Only?** condenses the chart by only showing the major time divisions.

6.3.4 *Activity 2: Line Chart*

Using the *maincafe* dataset in SOFA, produce a line chart of Ptysize by Sex. The line chart should have a title of "Visualizing Frequency, Activity 2" and a subtitle of "Line Chart of Party Size by Sex".

6.3.5 *Bar Chart*

1. Data Source Table: cars
2. Table Type: Bar Chart (first button)
3. Values: Passengers
4. Title: Passengers

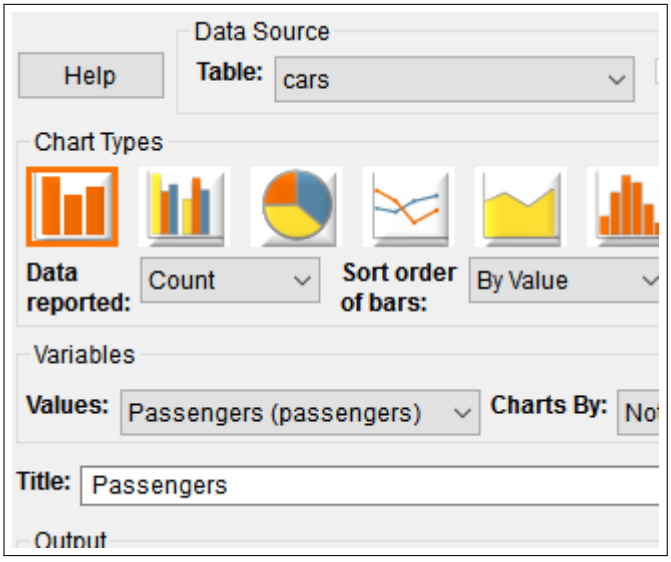


Figure 75: Setting Up Passengers

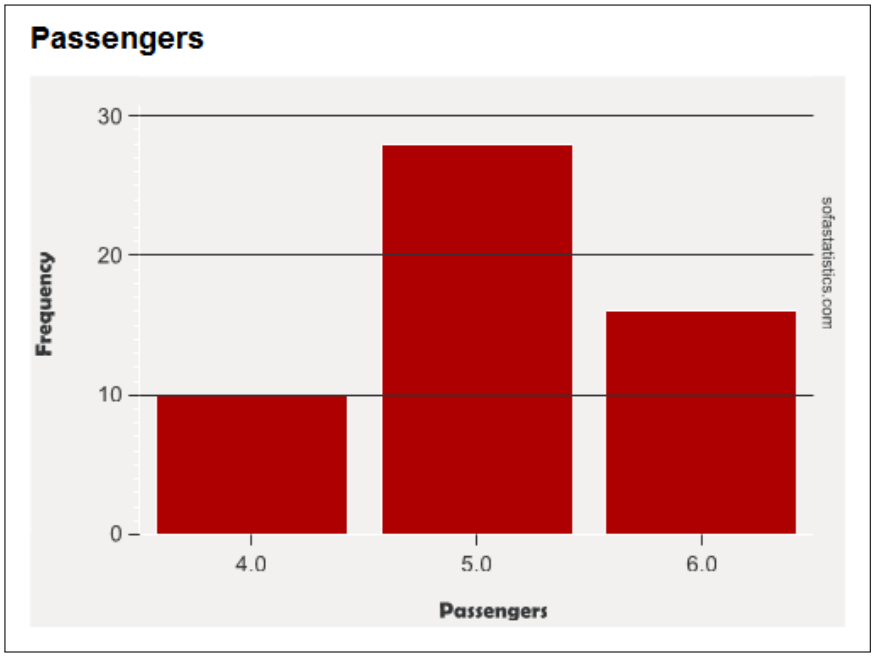


Figure 76: Passengers

6.3.6 *Activity 3: Bar Chart*

Using the *maincafe* dataset in SOFA, produce a bar chart of Meal. The bar chart should have a title of "Visualizing Frequency, Activity 3" and a subtitle of "Bar Chart of Meal".

6.3.7 *Clustered Bar Chart*

1. Data Source Table: cars
2. Table Type: Clustered Bar Chart (second button)
3. Values: Passengers
4. By: Drivetrain
5. Title: Passengers By Drive Train

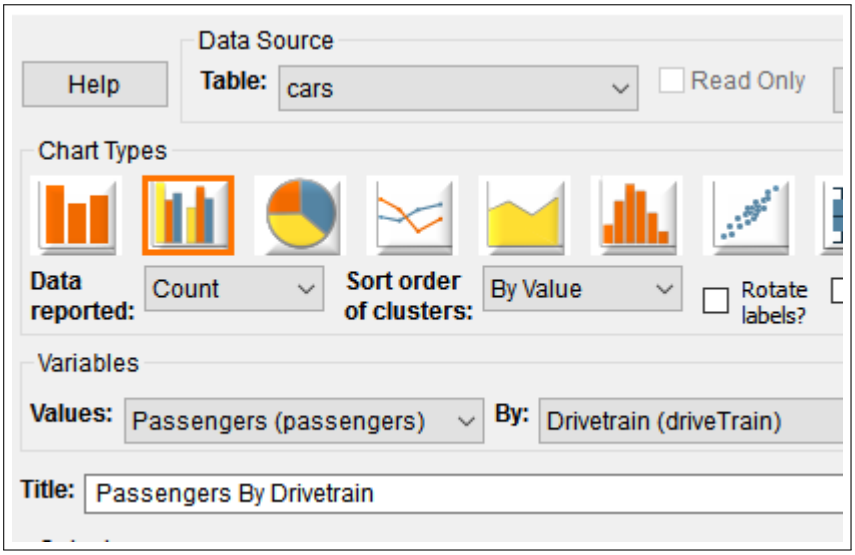


Figure 77: Setting Up Passengers By Drivetrain

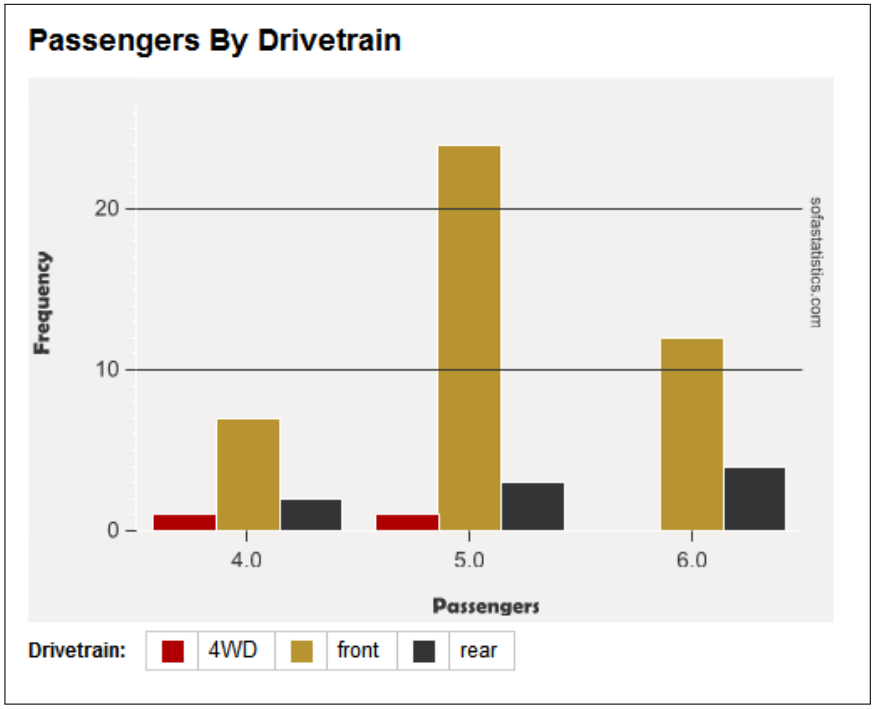


Figure 78: Passengers By Drivetrain

6.3.8 *Activity 4: Clustered Bar Chart*

Using the *maincafe* dataset in `SOFA`, produce a clustered bar chart of Meal by Sex. The bar chart should have a title of "Visualizing Frequency, Activity 4" and a subtitle of "Bar Chart of Meal by Sex".

6.3.9 *Pie Chart*

1. Data Source Table: cars
2. Table Type: Pie Chart (third button)
3. Values: Passengers
4. Title: Passengers

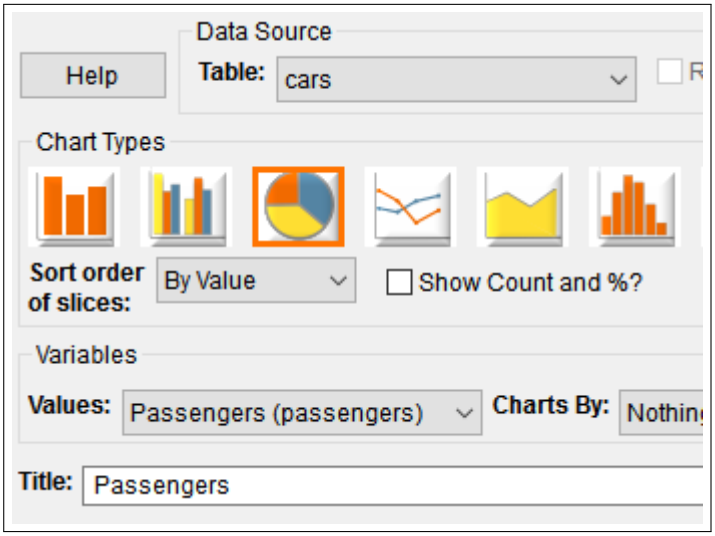


Figure 79: Setting Up Passengers

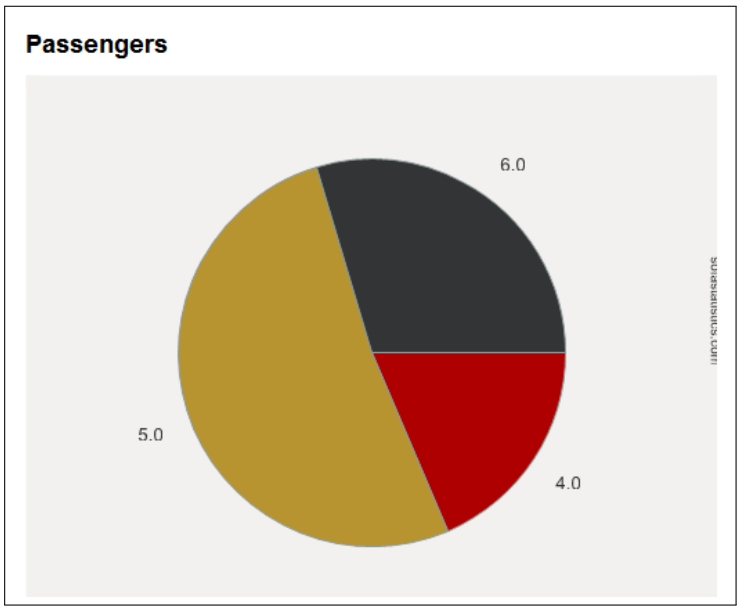


Figure 80: Passengers

6.3.10 *Activity 5: Pie Chart*

Using the *maincafe* dataset in SOFA, produce a pie chart of Food. The pie chart should have a title of "Visualizing Frequency, Activity 5" and a subtitle of "Pie Chart of Food".

6.4 DELIVERABLE

Complete the following activities in this lab:

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

7

CORRELATION

7.1 INTRODUCTION

Correlation is a method used to describe relationships between two variables. For example, if some research project plotted the ages of people who started smoking and the family income of those people then a correlation would attempt to determine if there is some relationship between those two factors.

This lab explores both correlation and scatter plots, which are graphic tools designed to make correlations easier to understand.

7.2 CORRELATION AND CAUSATION

From the outset of this lab, it is important to remember that there is a huge difference between correlation and causation. Just because two factors are correlated in some way does not lead to a conclusion that one is causing the other. As an example, if a research project found that students who spend more hours studying tend to get higher grades this would be an interesting correlation. However, that research, by itself, could not prove that longer studying hours causes higher grades. There could be other intervening factors that are not accounted for in this simple correlation (like the type of final examination used). As an egregious example to prove this point, consider that the mean age in the United States is rising (that is, people are living longer; thus, there are more elderly people in the population) and that human trafficking crime is increasing. While these two facts may be correlated, it would not follow that old people are responsible for human trafficking! Instead, there are numerous social forces in play that are not accounted for in this simple correlation. It is important to keep in mind that correlation does not equal causation as you read research.

7.2.1 *Pearson's R*

Pearson's Product-Moment Correlation Coefficient (normally called *Pearson's r*) is a measure of the strength of the relationship between two variables. Pearson's r is a number between -1.0 and $+1.0$, where 0.0 means there is no correlation between the two variables and either $+1.0$ or -1.0 means there is a perfect correlation. A positive correlation means that as one variable increases the other also increases. For example, as people age they tend to weigh more so a positive corre-

lation would be expected between age and weight. A negative correlation, on the other hand, means that as one variable increases the other decreases. For example, as people age they tend to run slower so a negative correlation would be expected between age and running speed. In general, both the strength and direction of a correlation is indicated by the value of "r":

Correlation	Description
+.70 or higher	Very strong positive
+.40 to +.69	Stong positive
+.30 to +.39	Moderate positive
+.20 to +.29	Weak positive
+.19 to −.19	No or negligible
−.20 to −.29	Weak negative
−.30 to −.39	Moderate negative
−.40 to −.69	Strong negative
−.70 or less	Very strong negative

As an example from the *bdims* dataset, the correlation between weight and waist girth (Wgt - Wai.Gi) is +0.904 so there is a very strong positive correlation between these two factors. This would be expected since people who weigh more would be expected to have larger waists. Here are a few other correlations from the *dbims* dataset:

Variables	Correlation	Description
Weight—Waist Girth	0.904	Very Strong Positive
Weight—Ankle Diameter	0.726	Very Strong Positive
Height—Chest Depth	0.553	Strong Positive
Height—Hip Girth	0.339	Moderate Positive
Height—Thigh Girth	0.116	No Correlation

7.2.2 *Spearman's Rho*

Pearson's r is only useful if both data elements being correlated are interval or ratio in nature. When the one or both data elements are ordinal or nominal then a statistically different process must be used to calculate a correlation, and that process is *Spearman's Rho*. Other than the process used to calculate Spearman's Rho, the concept is exactly the same as for Pearson's r and the result is a correlation between −1 and +1 where the strength and direction of the correlation is determined by its value.

For example, imagine that a dataset included information about the age of people who purchased various makes of automobiles. Since the "makes" would be selected from a list (Ford, Chevrolet, Honda,

etc.), Spearman's Rho would be used to calculate the correlation between the customers' preference for the make of an automobile and their age. Perhaps the correlation would come out to +0.534 (this is just a made-up number). This would indicate that there was a strong positive correlation between these two variables; that is, people tend to prefer a specific make based upon their age; or, to put it another way, as people age their preference for automobile make changes in a predictable way.

Pearson's r and Spearman's Rho both calculate correlation and it is reasonable to wonder which method should be used in any given situation. A good rule of thumb is to use Pearson's r if both data items being correlated are interval or ratio and use Spearman's rho if one or both are ordinal or nominal. Imagine a series of survey questions that permitted people to select from only a small group of possible answers. As an example, perhaps respondents are asked to select one of five responses ranging from "Strongly Agree" to "Strongly Disagree" for statements like "I enjoyed the movie." Restricting responses to only one of five options creates ordinal data and to determine how well the responses to these questions correlate with something like the respondents' ages, Spearman's Rho would be an appropriate choice.

As an example from the *email* dataset, the correlation between Image and CC is +0.808 so there is a very strong positive correlation between these two factors. Here are a few other correlations from the *email* dataset, all using Spearman's Rho:

Variables	Correlation	Description
Image—Exclaim_Mess	0.522	Strong Positive
Image—Line_Breaks	0.491	Strong Positive
Image—Attach	0.927	Very Strong Positive
Line_Breaks—Dollar	0.453	Strong Positive
Exclaim_Mess—CC	0.408	Strong Positive

7.3 SIGNIFICANCE

Most people use the word "significant" to mean "important" but researchers and statisticians have a much different meaning for "significant" and it is vital to keep that difference in mind.

In statistics and research, "significance" means that the experimental results were such that they would not likely have been produced by mere chance. For example, if a coin is flipped 100 times, heads should come up 50 times. Of course, by pure chance, it would be possible for heads to come up 55 or even 60 times. However, if heads came up 100 times, researchers would suspect that something unusual was happening (and they would be right!). To a researcher, the

central question of significance is "How many times can heads come up and still be considered just pure chance?" That number is the statistical significance level.

In general, researchers use one of three significance levels: 1%, 5%, or 10%. A researcher conducting The Great Coin-Tossing Experiment may start by simply stating "This result will be significant at the 5% level." That would mean that if the coin were tossed 100 times, then anything between $47.5 - 52.5$ (a 5% spread) "heads" tosses would be merely chance. However, 47 or 53 "heads" would be outside that 5% spread and would be significant.

It must seem somewhat subjective for a researcher to simply select the desired significance level, but most researchers in business and the social and behavioral sciences (like education, sociology, and psychology) tend to choose a significance level of 5%.[1] There is no real reason for choosing that level today other than it is just the way things have traditionally been done for many years. Therefore, if a researcher selected something other than 5%, peer researchers would want some explanation concerning the "weird" significance level.

Keep in mind, though, that statistical significance is not the same as practical significance. *Wikipedia*, that great repository of knowledge, includes this interesting example[2]:

> As used in statistics, significant does not mean important or meaningful, as it does in everyday speech. For example, a study that included tens of thousands of participants might be able to say with great confidence that residents of one city were more intelligent than people of another city by 1/20 of an IQ point. This result would be statistically significant, but the difference is small enough to be utterly unimportant.

7.3.1 *Chi-Square*

One of the most common measurements of statistical significance is the chi-square test, but this test should only be used if both variables are nominal. The basic idea behind this test is to determine what values would be expected by chance and compare that to the values actually obtained by experiment. The formula for a chi-square test is somewhat complex but `SOFA` handles chi-square calculations without any problem at all. The following chi-square values were obtained from the *email* dataset:

1 The reason that 5% is usually selected as a significance level can be traced back to R.A. Fisher who published a book in 1925 that included statistical tables for researchers. His work was profoundly influential and values in the 5% table became the most commonly cited levels in published research for many decades.

2 http://en.wikipedia.org/wiki/Statistical_significance

Variables	Chi-Square
Spam—Exclaim_Subj	0.256
Spam—Image	5.927
Spam—Urgent_Subj	15.374
Spam—Attach	104.825

While it is difficult to decide if any give chi-square statistic is "good" or "bad," in general the larger that number then the stronger the relationship between the two variables. In the above table, for example, there would seem to be virtually no relationship between Spam and Exclaim_Subj but there is a strong relationship between Spam and Attach.

As an aid to determining whether a calculated chi-square is significant, a *p-value* (that stands for "probability value") is usually also calculated. A p-value answers the question, "what is the probability that an observed phenomenon is due to chance?" For example, the p-value for each of the above chi-square statistics is:

Variables	Chi-Square	P-Value
Spam—Exclaim_Subj	0.256	0.6129
Spam—Image	5.927	0.2047
Spam—Urgent_Subj	15.374	8.820×10^{-5}
Spam—Attach	104.825	0.00

Notice that as the chi-square statistic gets larger the p-value gets smaller since there is an inverse relationship between those values. In the Significance section above, page 75, the concept of a 5% significance level was developed so any p-value smaller than 0.05 is considered significant; that is, it is not likely due to chance. If the researcher for the "email" project was using a 5% (0.05) significance level then there is no significance in the correlation between spam and exclaim_subj or spam and image since the p-level for those two correlations is greater than 0.05. However, the relationship between spam and urgent_subj and spam and attach is significant since the p-level for those two correlations is smaller than 0.05.

In short, a significant correlation would have a large chi-square statistic and a small p-level value (less than 0.05, normally).

7.3.2 *Degrees of Freedom*

One other statistic is often evoked when discussing significance: *Degrees of Freedom*. While this statistic can be confusing to understand[3],

3 There is an excellent discussion about Degrees of Freedom at http://blog.minitab.com/blog/statistics-and-quality-data-analysis/what-are-degrees-of-freedom-in-statistics

it is fairly easy to calculate. Keep in mind that degrees of freedom are only meaningful for nominal or ordinal data and is calculated by determining the number of levels for each of the variables under consideration, subtract one from each of those levels, and multiply the result. As an example, when calculating the "spam—Attach" chi-square, there were two levels for "spam" (no/yes) and eight levels for "attach" $(0, 1, 2, 3, 4, 5, 6, 7)$. Subtracting one from each of those leaves one and seven and then multiplying those two numbers ends with seven degrees of freedom. `SOFA` displays the degrees of freedom with the chi-square calculation.

Here are a few degrees of freedom from the *email* dataset to help explain this concept. In each case, the number of levels for each variable are in parenthesis following the variable name.

Variables	Degrees of Freedom
Spam (2)—Exclaim_Subj (2)	$(2-1)(2-1) = 1$
Spam (2)—Image (5)	$(2-1)(5-1) = 4$
Number (3)—Image (5)	$(3-1)(5-1) = 8$
Number (3)—Attach (8)	$(3-1)(8-1) = 14$

7.4 SCATTER PLOTS

A scatter plot is a two-dimensional graph of ordered pairs that indicates the relationship between two variables. While a scatter plot could be created from nominal or ordinal data, it is normally only used for interval and ratio data. For example, the correlation between weight and waist girth in the *dbims* dataset was calculated at 0.904 (see the table on page 74). Figure 81 is the scatter plot for that correlation:

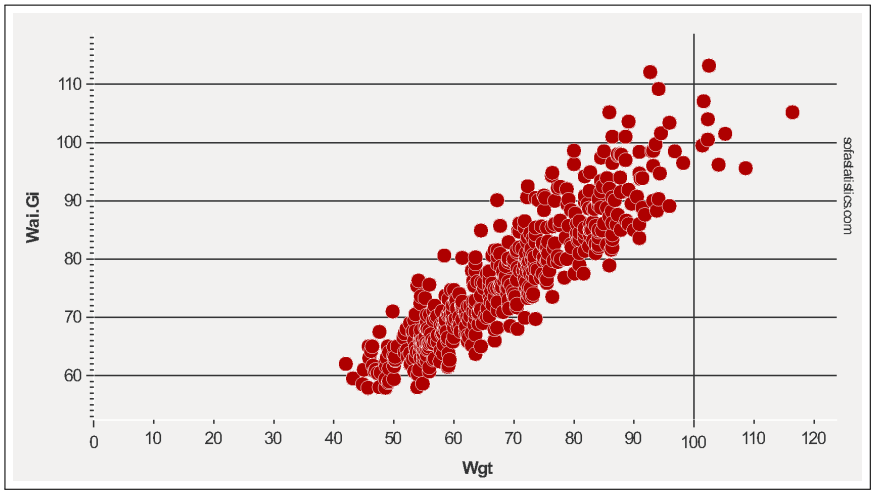


Figure 81: Weight-Waist Girth

Each of the red dots is a single point in the dataset. For example, the farthest dot to the right is for a weight of 116.4 and a waist girth of 105.2 Notice that, generally, as the weight increases (X-Axis) the waist girth also increases (Y-Axis), which indicates a positive correlation. Also notice that the dots are very close together which indicates a strong correlation. Compare this to the next illustration:

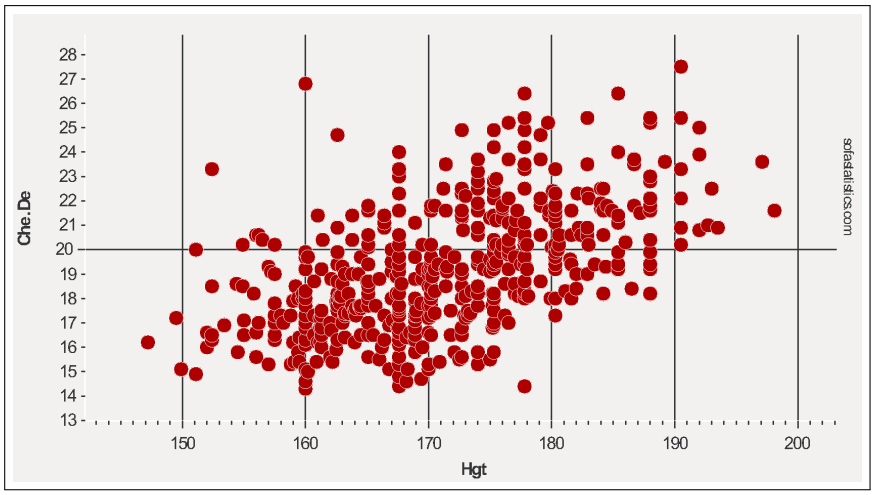


Figure 82: Height-Chest Depth

Figure 82 is the scatter plot for the correlation between height and chest depth, which was calculated at 0.553 (see the table on page 74). While there is still a clear upward trend to these dots (a positive correlation), they are more scattered out (a weak correlation). As one final example, consider the scatter plot for height and thigh girth.

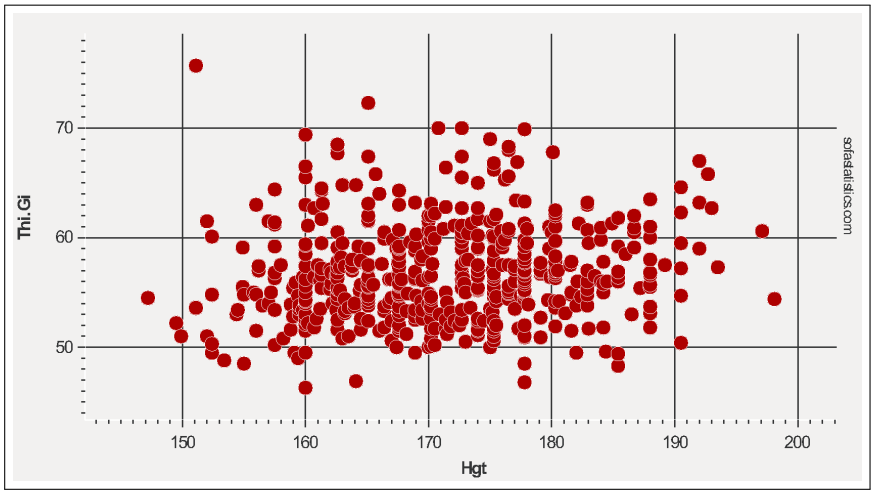


Figure 83: Height-Thigh Girth

The correlation 0.116 so there is no clear direction for the plot and the dots are very scattered.

7.5 PROCEDURE

7.5.1 *Pearson's R*

Start SOFA and select "Statistics" then:

1. Select "Select a Statistical Test Here"
2. Select "Correlation - Pearson's"
3. Click "Configure Test"

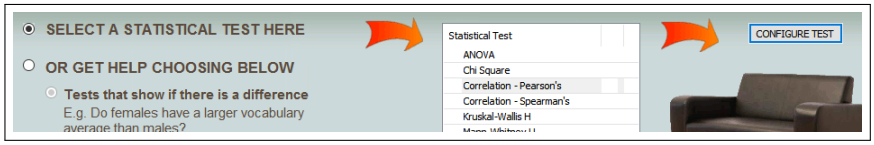


Figure 84: Starting Pearson's R

4. Data Source Table: gifted (It is desired to see if there is a correlation between the age when children first learn to count and the time spent watching cartoons.)
5. Group A: Count (this is the "X-Axis" or independent variable)
6. Group B: Cartoons (this is the "Y-Axis" or dependent variable)

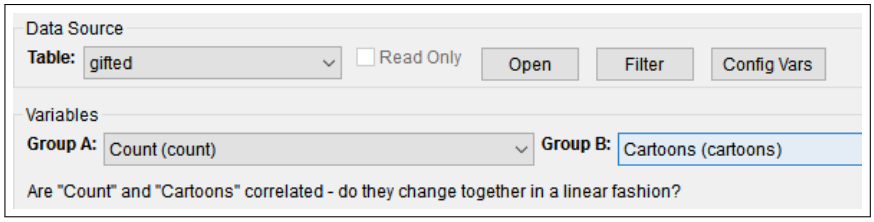


Figure 85: Calculating Pearson's R for Cartoons-Count

7. Click "Show Results" and read the results at the bottom of the window. SOFA reports the "Two-tailed p value" (described as simply "p-value" earlier in this lab document) of 0.3670, Pearson's R of 0.155, with 34 degrees of freedom.

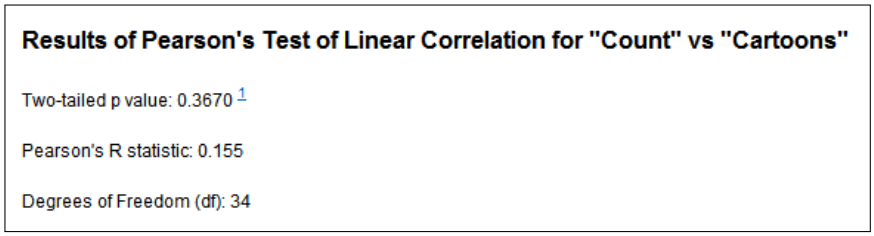


Figure 86: Pearson's R for Cartoons-Count

8. SOFA also displays a scatter plot with a regression line and reports its slope (0.023) and Y-Intercept (2.371)[4].

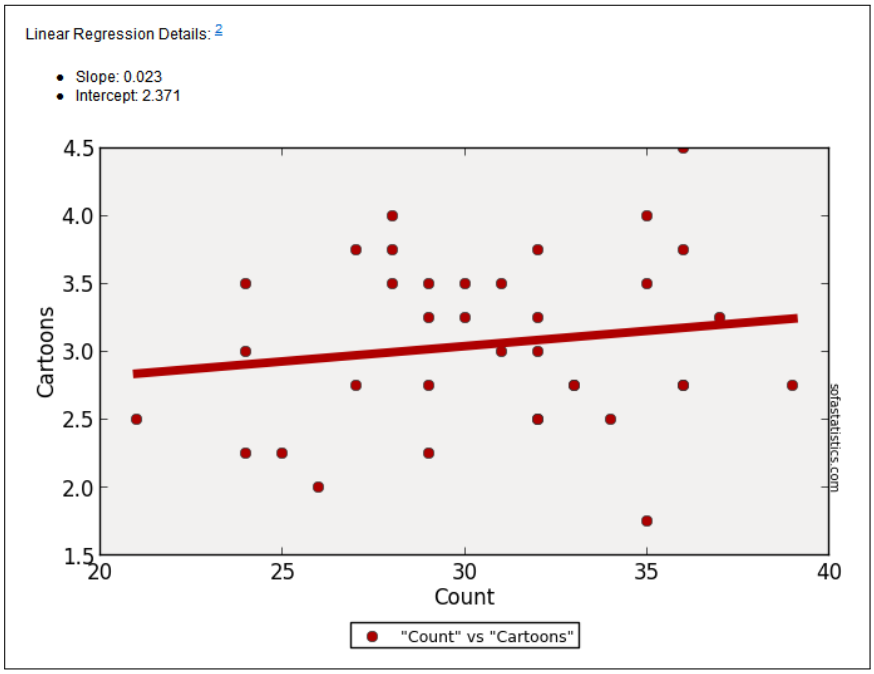


Figure 87: Scatterplot and Regression Line for Cartoons-Count

The following table lists the calculated statistics for several Pearson's R correlations from the *gifted* dataset. This could be used for practice in obtaining Pearson's R.

Variables	P	R	DF
Count-Edutv	0.2065	−0.216	34
Read-Score	1.006×10^{-3}	0.525	34
Motheriq-Read	0.8032	−0.043	34

7.5.2 *Activity 1: Pearson's R*

Using the *maincafe* dataset in SOFA, determine Pearson's R for Length (Group A) and Bill (Group B). Submit a screen capture from SOFA that shows the Two-tailed p value, Pearson's R statistic, Degrees of Freedom, and Linear Regression Details (Slope and Intercept).

7.5.3 *Spearman's Rho*

Start SOFA and select "Statistics" then:

4 Regression and the use of the slope/intercept information is covered in Lab 8, page 87

1. Select "Select a Statistical Test Here"
2. Select "Correlation - Spearmans's"
3. Click "Configure Test"

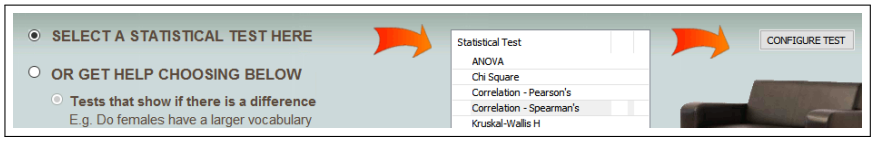


Figure 88: Starting Spearman's Rho

4. Data Source Table: email (It is desired to see if there is a correlation between the number of attachments to a message and its being identified as spam.)
5. Group A: Attach (this is the "X-Axis" or independent variable)
6. Group B: Spamnum (this is the "Y-Axis" or dependent variable)

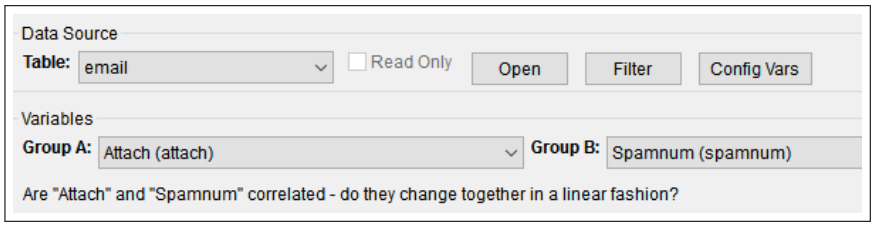


Figure 89: Setting Up Spearman's Rho

7. Read the results of calculating Spearman's Rho at the bottom of the window.

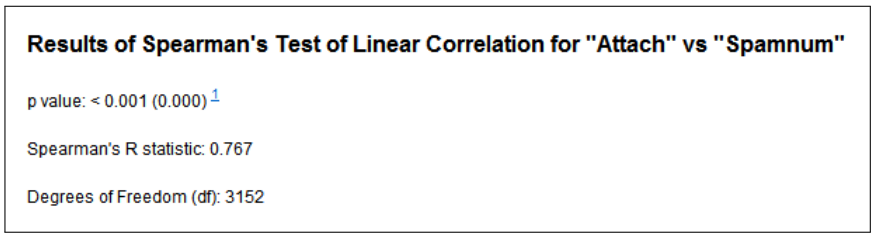


Figure 90: Spearman's Rho Results

8. SOFA also creates a scatter plot for Spearman's Rho but that is of limited value. Since the "spam" variable can only be 0 or 1 the scatter plot is basically two rows of dots and does not offer much information.

The following table lists the calculated statistics for several Spearman's Rho correlations from the *births* dataset. This could be used for practice in obtaining Spearman's Rho.

Variables	P	Rho	DF
Fage-Mage	1.716×10^{-181}	0.795	827
Mage-Weight	0.01443	0.077	998
Weeks-Weight	2.359×10^{-53}	0.46	996

7.5.4 *Activity 2: Spearman's Rho*

Using the *maincafe* dataset in SOFA, determine Spearman's Rho using Length for Group A and Ptysize for Group B. Submit a screen capture from SOFA that shows the p value, Spearman's R statistic, Degrees of Freedom, and Linear Regression Details (Slope and Intercept).

7.5.5 *Chi Square*

Start SOFA and select "Statistics" then:

1. Select "Select a Statistical Test Here"
2. Select "Chi Square"
3. Click "Configure Test"

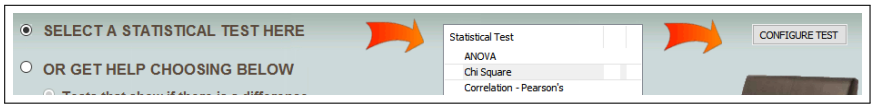


Figure 91: Starting Chi Square

4. Data Source Table: email (It is desired to see if the correlation between the types of numbers in a message and the number of attachments is significant.)
5. Group A: Number
6. Group B: Image
7. SOFA calculates the p-value of 0.01199, chi square of 19.593 and the degrees of freedom of 8.

Results of Pearson's Chi Square Test of Association Between "Number" and "Image"

p value: 0.01199 [1]

Pearson's Chi Square statistic: 19.593

Degrees of Freedom (df): 8

Figure 92: Chi Square Results

8. Chi square results are normally printed in a matrix where the expected and actual values for each of the correlated variables are compared.

		Image											
		0.0		1.0		2.0		3.0		5.0		TOTAL	
		Obs	Exp	Obs	Exp	Obs	Exp	Obs	Exp	Obs	Exp	Obs	Exp
Number	large	373	380.4	14	7.3	2	1.6	0	0.6	1	0.1	390	390.0
	none	538	531.5	7	10.2	0	2.2	0	0.9	0	0.2	545	545.0
	small	2165	2164.1	38	41.5	11	9.1	5	3.5	0	0.7	2219	2219.0
	TOTAL	3076	3076.0	59	59.0	13	13.0	5	5.0	1	1.0	3154	3154.0

Minimum expected cell count: 0.124

% cells with expected count < 5: 53.3

Figure 93: Chi Square Matrix

9. In the chi square matrix each level for both variables is calculated and compared. For example, in the matrix in Figure 93, "Large" numbers were observed 373 times and were expected 380.4 times. As long as the two values are close to each other then there is not much statistical significance in the values, in other words, the observed values are about what was expected. Under the matrix SOFA notes that the minimum expected cell count for the entire matrix is 0.124 and 53.3% of the expected values are less than 5.

7.5.6 *Activity 3: Chi Square*

Using the *maincafe* dataset in SOFA, determine Chi Square where Group A is Meal and Group B is Ptysize. Submit a screen capture from SOFA that shows the p value, Pearson's Chi Square statistic, Degrees of Freedom, the Ptysize Matrix, the Minimum expected cell count, and the percentage of cells with the expected count.

7.6 DELIVERABLE

Complete the following activities in this lab:

Number	Name	Page
7.5.2	Activity 1: Pearson's R	81
7.5.4	Activity 2: Spearman's Rho	83
7.5.6	Activity 3: Chi Square	84

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

8

REGRESSION

8.1 INTRODUCTION

Regression is a method used to describe relationships between two variables. Regression is similar to correlation, but it is a measure of the relationship between the *mean* of one variable and the corresponding values of another variable. Regression analysis is used to predict an unknown value for one variable when the mean of the related variable is known. For example, if some research project plotted the ages of people who started smoking and the family income of those people then a correlation would attempt to determine if there is some relationship between those two factors while a regression would attempt to predict the age someone would start smoking given that person's family income.

This lab explores the statistics of regression and demonstrates how to use regression to make predictions.

8.2 REGRESSION

A regression line can be drawn on a scatter plot to graphically show the relationship between two variables (this is sometimes called a "trend line" and a "line of best fit"). Moreover, if the data points in the scatter plot are all close to the regression line, then it is a strong correlation.

As an example, in the *bdims* dataset the calculated Pearson's r for weight and ankle diameter was 0.726. The regression line in Figure 94 makes it clear that the correlation is positive (the line slopes up and right) and since the dots are all fairly close to the regression line the correlation is strong.

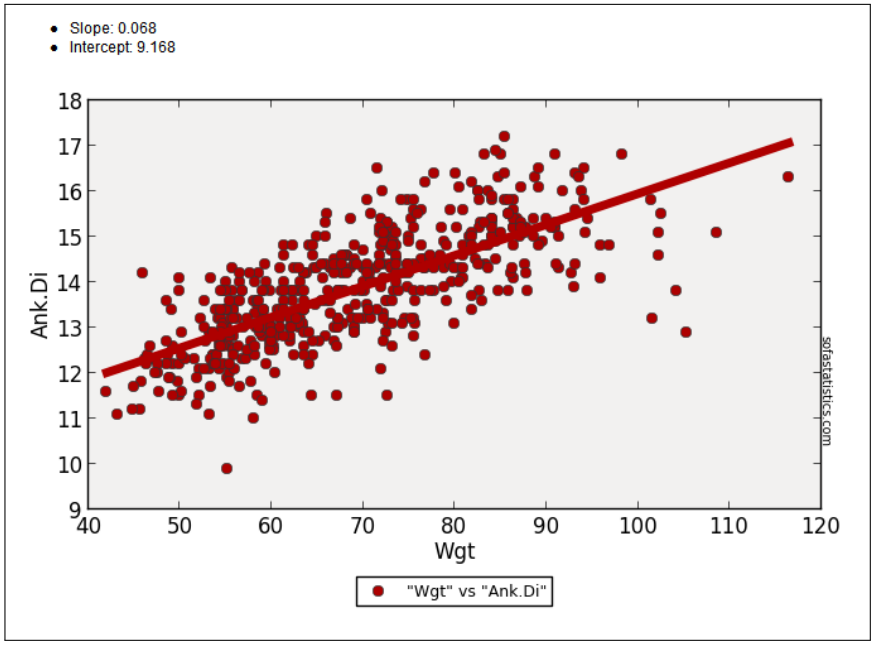


Figure 94: Weight-Ankle Diameter

As another scatter plot example, here is one for height and hip girth.

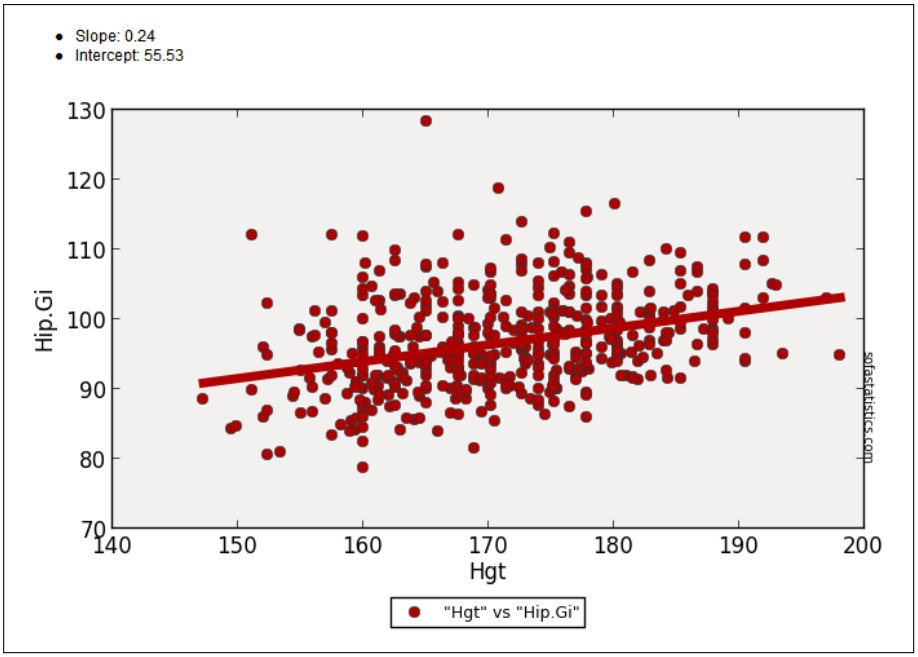


Figure 95: Height-Hip Girth

The correlation for height and hip girth is 0.339. While this is a positive number, it is weakly positive so the slope of the regression line is small and the dots are fairly scattered.

SOFA includes the slope and Y-Intercept data at the top left corner of a regression plot. That information can be used to predict the Y-Value for a given X-Value by using a simple "Slope-Intercept" equation:

$$y = mx + b$$

Where m is the slope of the regression line and b is the y-intercept. By plugging in a value for x a simple calculation will determine the corresponding value for y. For example, in Figure 95 the slope is 0.241 and the y-intercept is 55.427. To calculate the Hip Girth (the "y" value) for a height of 165 centimeters:

$$y = (0.241)(165) + 55.427$$
$$y = 95.192$$

It is important to keep in mind that Pearson's r and the slope of the regression line measure different aspects of a correlation, and even though they could coincidentally be nearly the same they should not be confused. It is possible to have a slope, for example, of -2.0, but never a correlation that size. In the case of Figure 95, the slope is 0.241 while the correlation is 0.339.

Regression analysis becomes less certain if the selected X-Value is at the edge or outside the main body of the scatter plot. For example, in Figure 95 it is mathematically possible to predict the value of hip girth (Y-Value) for a height of 140 centimeters (X-Value):

$$y = (0.241)(140) + 55.427$$
$$y = 89.167$$

However, since the selected X-Value is outside the main body of the scatter plot then the calculated Y-Value is suspect.

8.3 PROCEDURE

Start SOFA and select "Statistics" then:

1. Select "Select a Statistical Test Here"
2. Select "Correlation - Pearson's"
3. Click "Configure Test"
4. Data Source Table: gifted (It is desired to predict a child's score on a test of analytical skills when given the mother's IQ.)
5. Group A: Motheriq (this is the "X-Axis" or independent variable)

6. Group B: Score (this is the "Y-Axis" or dependent variable)

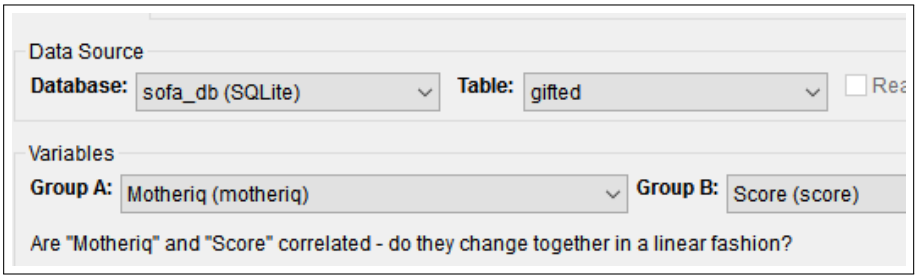


Figure 96: Calculating Pearson's R for Motheriq-Score

7. Click "Show Results" and scroll down to the scatter plot.

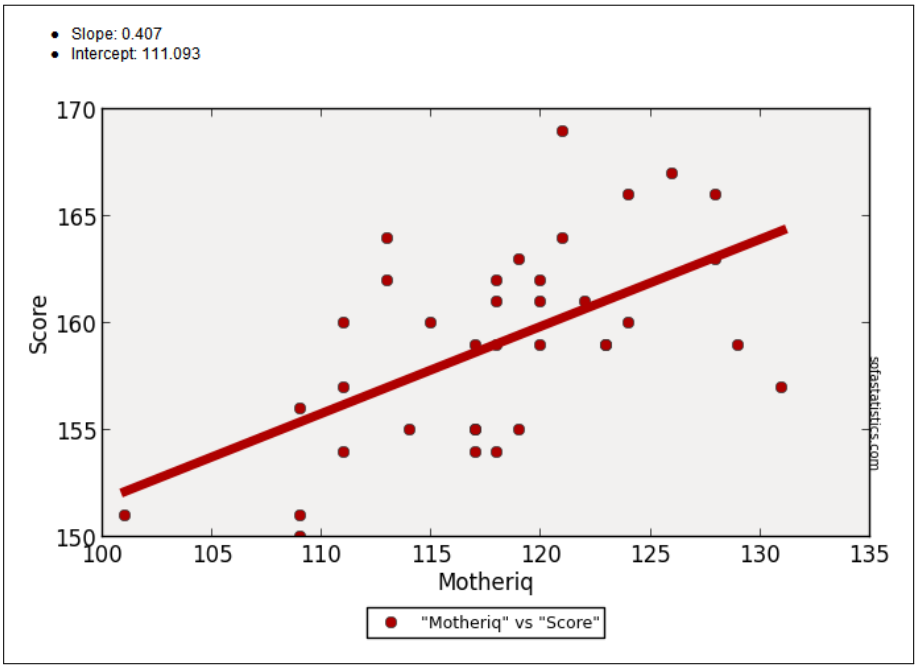


Figure 97: Scatter Plot for Motheriq-Score

8. Using the slope and y-intercept, calculate the expected test score for a mother's IQ of 118.

$$y = (0.407)(118) + 111.093$$
$$y = 159.119$$

9. Using the slope and y-intercept, calculate the expected test score for a mother's IQ of 122.

$$y = (0.407)(122) + 111.093$$
$$y = 160.747$$

The following table lists the predicted student's test score for several different variables in the *gifted* dataset. This could be used for practice in calculating regression predictions. In each case, Pearsons r was calculated and the "Group B" used is "Score."

Group A	Slope	Intercept	X-Value	Predicted Score
Fatheriq	0.25	130.429	117	159.679
Read	11.813	133.905	2.35	161.67
Speak	0.385	152.216	17	158.761

8.3.1 *Activity 1: Predictive Regression 1*

Using the *maincafe* dataset in SOFA, determine the slope and intercept for Length (Group A) and Bill (Group B). Use those numbers to predict the bill for a meal lasting 42 minutes. Round the bill to the nearest penny.

8.3.2 *Activity 2: Predictive Regression 2*

Using the *maincafe* dataset in SOFA, determine the slope and intercept for Age (Group A) and Tip (Group B). Use those numbers to predict the tip left by a customer who is 48 years old. Round the tip to the nearest penny.

8.4 DELIVERABLE

Complete the following activities in this lab:

Number	Name	Page
8.3.1	Activity 1: Predictive Regression 1	91
8.3.2	Activity 2: Predictive Regression 2	91

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

9 HYPOTHESIS TESTING: NONPARAMETRIC TESTS

9.1 INTRODUCTION

An important function for statistical analysis is to test a hypothesis to see if it adequately explains some observed phenomenon. The statistical processes for qualitative data are introduced in this lab and the processes for quantitative data are in HYPOTHESIS TESTING: PARAMETRIC TESTS on 107.[1]

9.2 KRUSKAL-WALLIS H

More Than 2 Groups : Ordinal/Nominal Data : Not Paired

This test is used to determine if there are any significant differences in three or more groups of ordinal or nominal data. Imagine that a researcher wanted to determine if there was a significant difference in the birth weight for babies born to women who were divided into three groups: "Heavy Smokers" (more than one pack per day), "Light Smokers" (one pack or less per day), and "Nonsmokers." The researcher would record the birth weights and mothers' smoking habits for all babies born in a single hospital for six months and then use a Kruskal-Wallis H test to see if there was a significant difference in those three groups.

9.3 WILCOXON SIGNED RANKS

2 Groups : Ordinal/Nominal Data : Paired

This test is used to determine if there are any significant differences in two groups of paired ordinal or nominal data. "Paired" data means that a single data point is observed two different times and those two observations are compared. Imagine that a researcher wanted to know if a speech from the company president could change workers' opinions. The researcher would ask a group of volunteers their opinion about a controversial policy both before and after that speech. The researcher would use a 1-to-5 point scale with questions like "I would vote for this policy." Later, the researcher would compare each person's before/after rating with a Wilcoxon Signed Ranks test to see if there was a significant change in opinion. In this case, the data are paired such that the before/after opinion for each volunteer is compared.

1 The definitions of "hypothesis" and the various data types are found in the INTRODUCTION, beginning on page 3.

9.4 MANN-WHITNEY U

2 Groups : Ordinal/Nominal Data : Not Paired

This test is used to determine if there are any significant differences in two groups of unpaired ordinal or nominal data. Imagine that a movie producer wanted to know if there was a difference in the way the audience in two different cities responded to a movie. As the audience members left the theater they would be asked to rate the movie on a scale of one to five stars. The ratings for the two cities would be collected and then a Mann-Whitney test would be used to determine if the difference in ratings between the cities was significant.

9.5 PROCEDURE

SOFA makes it easy to complete any of the statistical tests listed in this lab exercise. A Wizard provides help in selecting an appropriate test but users can also manually select and configure whatever test they need.

9.5.1 *Statistics Wizard*

If the type of data and the question being asked is known then SOFA includes a Wizard to help select an appropriate statistics test. To access the wizard, open SOFA and select "Statistics." Then click "Or Get Help Choosing Below."

- **Tests that show if there is a difference.**
  - **Groups**. Select the number of groups that are being analyzed: 2 or 3 *or more*. For help in determining how many groups are in the data click the "Groups" button near the bottom of the window to open the frequency table function[2] Also, the "Help" button to the right of the Groups selection provides good information.
  - **Normal**. Select whether the data are in a normal distribution[3]. For help in determining if the data are normal, click the "Normality" button near the bottom of the window. Also, the "Help" button to the right of the Normal selection provides good information.
  - **Independence**. Select whether the data are independent observations or paired.
- **Tests that show if there is a relationship.**

2 See Lab 5 on page 47 for more information about frequency tables.
3 See Section 1.3.2.1 on page 6 for more information about the normal distribution.

- **Data Type**. Select "Names" if the data are nominal and "Ordered" if Ordinal. For help in determing the data type, click the "Data Type" button near the bottom of the window. Also, the "Help" button to the right of the Data Type selection provides good information.
- **Normal**. (Note: this selection is only active for "Ordered" data) Select whether the data are in a normal distribution[4]. For help in determining if the data are normal, click the "Normality" button near the bottom of the window. Also, the "Help" button to the right of the Normal selection provides good information.

As an example, the *births* dataset contains information about 1000 births in North Carolina in 2004. Imagine that a researcher proposed this hypothesis: "The weight of babies born to smokers is less than the weight of babies born to nonsmokers." The null hypothesis would be "Smoking makes no difference in the birth weight of babies." To test that hypothesis, the researcher would:

Start `SOFA` and select "Statistics." Then:

1. Click "Or Get Help Choosing Below"
2. Click "Tests that show if there is a difference" since the hypothesis states that smoking causes a difference in birth weight.
3. The researcher proposes comparing two groups: smokers and nonsmokers. However, to be certain that the dataset does not include any other types of groups (like "former smokers"), the researcher clicks the "Groups" button and the "Make Report Table" window opens.[5] The researcher creates a frequency table for "Habit" and finds that there are only two groups: nonsmoker and smoker.

4 See Section 1.3.2.1 on page 6 for more information about the normal distribution.

5 The creation and use of Frequency Tables are covered in FREQUENCY TABLES on page 47.

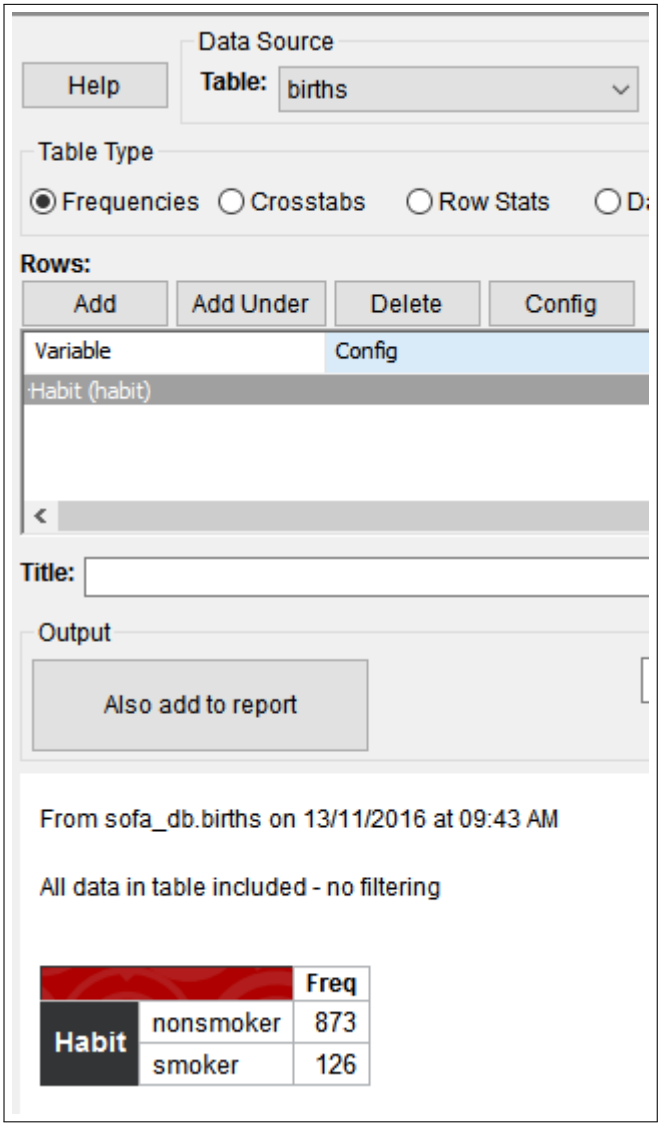


Figure 98: Verifying Groups

4. Having verified that there are two groups, the researcher must next determine if the birth weight is normally distributed. To do that, the researcher clicks then "Normality" button at the bottom of the window. In the "Normal Data?" window that pops up, the variable "weight" is selected and the following is returned by SOFA:

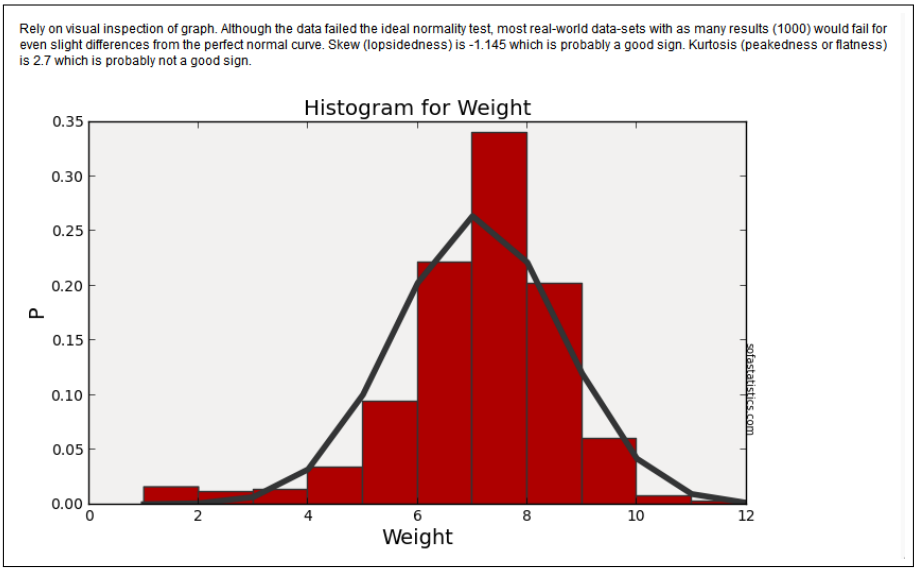


Figure 99: Verifying Normality

5. The text above the histogram reports "Rely on visual inspection of graph. Although the data failed the ideal normality test, most real-world data-sets with as many results (1000) would fail for even slight differences from the perfect normal curve. Skew (lopsidedness) is -1.145 which is probably a good sign. Kurtosis (peakedness or flatness) is 2.7 which is probably not a good sign." In this case, there are 1000 birth weights recorded so the skew and kurtosis numbers need to tempered by the visual appearance of the histogram. Since the histogram shows a large center peak with "shoulders" on both sides this data can be treated as normally distributed.

6. Next, the researcher must determine if the data are independent observations or paired. In this case there is no pairing indicated so the researcher chooses "Independent."

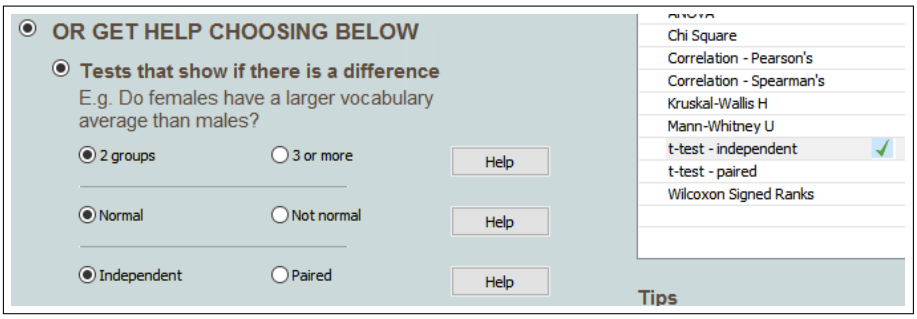


Figure 100: Choosing The Correct Statistical Test

7. SOFA indicates that "t-test - independent" is the correct test for the data being analyzed, so the researcher clicks the "Configure Test" button at the top right corner of the window and proceeds to execute that test. All of the various statistical tests in SOFA are described on the following pages of this lab.

9.5.2 *Activity 1: Wizard*

Using the various settings for the Wizard, what type of test is recommended for the following types of data? Note: in the Word document submitted for this lab, Activity 1 should have a simple listing, something like this (Note: these are not the correct answers to the listed tests):

1 Mann-Whitney U
2 t-test - paired
3 Kruskal-Wallis H

Here are the three types of data being analyzed:

No.	Test Type	Groups	Distribution	Indep	Cat
1	Difference	4	Normal	N/A	N/A
2	Difference	2	Not Normal	Paired	N/A
3	Relationship	N/A	Normal	N/A	Ordered

9.5.3 *Chi Square*

The Chi Square statistic is covered in Lab 7.3.1, Chi-Square, page 76. The SOFA activity related to Chi Square is on page 83.

9.5.4 *Correlation - Spearman's*

The Spearman's Rho statistic is covered in Lab 7.2.2, Spearman's Rho, page 74. The SOFA activity related to Spearman's Rho is on page 81.

9.5.5 *Kruskal-Wallis H*

1. Start SOFA and click the "Statistics" button.
2. Select "Kruskal-Wallis H" from the Statistical Test list at the top of the window and then click the "Configure Test" button.
3. Data Source Table: email
4. Averaged Variable: Dollar. This is the key difference between a Kruskal-Wallis H test and an ANOVA. The Kruskal-Wallis H

test expects the "averaged" variable to be a non-normal distribution, like ordinal or nominal data, while the ANOVA test expects the "averaged" variable to be a normal distribution, like ratio or interval.

5. Group By Variable: "Spam".
6. From Group: No
7. To: Yes

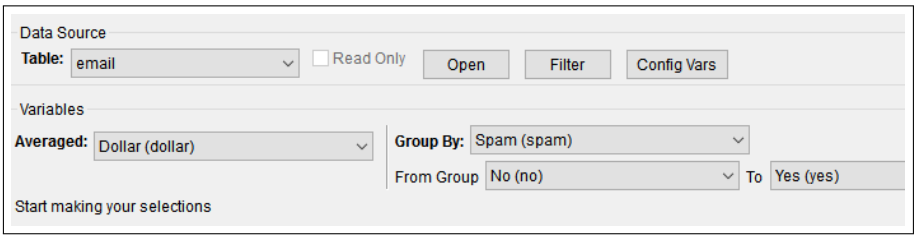


Figure 101: Setting Up a Kruskal-Wallis H Test

8. The results of that test are found in the results window:

Results of Kruskal-Wallis H test of average Dollar for Spam groups from "No" to "Yes"

p value: 0.1849 [1]

Kruskal-Wallis H statistic: 1.758

Degrees of Freedom (df): 1

Group	N	Median	Min	Max
No	2809	0.0	0.0	28.0
Yes	345	0.0	0.0	8.0

Figure 102: Results of a Kruskal-Wallis H Test

9. **p Value**. The most important statistic in the results window is the p-value. As always, when this number is encountered it is desired for it to have a value less than 0.05 (5%). In the example calculated for this exercise, the p-value is 0.1849, which is greater than 0.05, so there is no significant relationship between Dollar and Spam in this dataset.
10. **Kruskal-Wallis H statistic**. If the Kruskal-Wallis H statistic is greater than the Chi Square calculated value for the same input variables then the null hypothesis can be rejected. In a separate operation, the Chi Square statistic was calculated as 112.033 for Dollar and Spam. Since the Kruskal-Wallis H statistic was less

than the Chi Square then the null hypothesis could not be rejected. This agrees with the p-value calculated and, in general, if a p-value is available it should be used rather than rely on calculating and comparing to the Chi Square statistic.

11. **Others**. All other statistics displayed in the Kruskal-Wallis H results window have been explained elsewhere and will not be further detailed here.

12. **Another Example**. As another example, following is a Kruskal-Wallis H test for "Attach" and "Spam." Notice that the p-value is 4.587×10^{-5}, which is much less than the threshold of 0.05, so there is a significant relationship between the number of attachments to an email message and whether that message is spam.

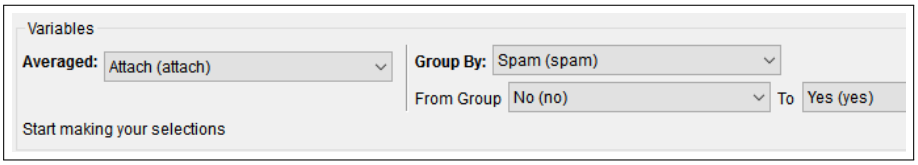


Figure 103: Setting Up a Kruskal-Wallis H Test

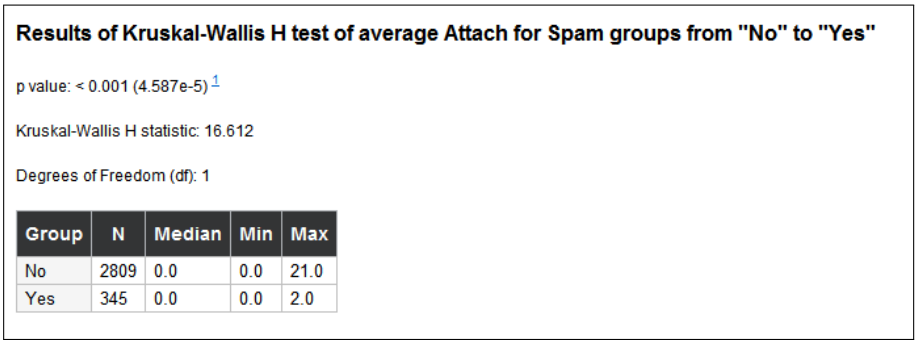


Figure 104: Results of a Kruskal-Wallis H Test

9.5.6 *Activity 2: Kruskal-Wallis H*

Using the *maincafe* dataset in `SOFA` conduct a Kruskal-Wallis H test and report the *p-value* for the following variables. Note: in the Word document submitted for this lab, Activity 2 should have a simple listing, something like this. (Notes: these are not the correct answers to the listed tests. To indicate tiny values use scientific notation in a form like $1.6e-5$ since that is easier to type.)

1 0.35
2 $1.27e-8$
3 237.4
4 0.73

Here are the variables to test:

Num	Averaged	Group By	From	To
1	Ptysize	Recmd	No	Yes
2	Food	Meal	Breakfast	Other
3	Bill	Day	Friday	Wednesday
4	Tip	Sex	Female	Other

9.5.7 *Mann-Whitney U*

1. Start `SOFA` and click the "Statistics" button.
2. Select "Mann-Whitney U" from the Statistical Test list at the top of the window and then click the "Configure Test" button.
3. Data Source Table: email
4. Ranked Variable: Attach
5. Group By Variable: Spam
6. Group A: No
7. Group B: Yes

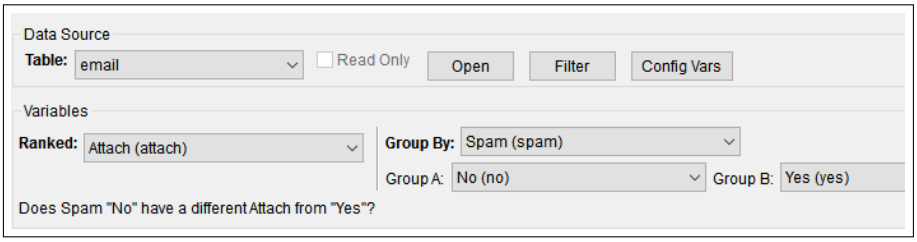


Figure 105: Setting Up a Mann-Whitney U Test

8. The results are shown in the figure below.

Results of Mann Whitney U Test of "Attach" for Spam "No" vs "Yes"

Two-tailed p value: 5.771e-3 [1]

U statistic: 454708.0 [2]

z: 2.761

Group	N	Median	Avg Rank	Min	Max
No	2809	0.0	1566.875	0.0	21.0
Yes	345	0.0	1664.006	0.0	2.0

Figure 106: Results of a Mann-Whitney U Test

9. **Two-tailed p value**. This is the p-value calculated by the test. The designation of "two-tailed" indicates that there were only two groups being analyzed, but the p-value calculated is interpreted exactly the same as for any other test. In this case, since 5.771×10^{-3} is less than 0.05 then there is a significant difference in the "No" and "Yes" spam groups.

10. **U Statistic**. This number is of little value except, perhaps, to compare the results of one Mann-Whitney U test to another. The most important result of a Mann-Whitney test is the p-value.

11. **z**. A Z-score is a representation of a score as a standard deviation above or below zero. For example, a Z-score of 1.0 indicates that the score is one standard deviation above zero. In the example calculated for this exercise, the Z-score is 2.761 which is more than two standard deviations higher than the mean and would indicate that the test results are significant.

12. **Other Statistics**. The other statistics presented in a Mann-Whitney U test are routine calculations explained elsewhere.

9.5.8 *Activity 3: Mann-Whitney U*

Using the *maincafe* dataset in `SOFA` conduct a Mann-Whitney U test and report the *p-value* for the following variables. Note: in the Word document submitted for this lab, Activity 3 should have a simple listing, something like this. (Notes: these are not the correct answers to the listed tests. To indicate tiny values use scientific notation in a form like $1.6e-5$ since that is easier to type.)

1 0.35
2 $1.27e-8$
3 237.4
4 0.73

Here are the variables to test:

Num	Ranked	Group By	From	To
1	Ptysize	Meal	Breakfast	Other
2	Ptysize	Pref	Booth	Table
3	Svc	Sex	Female	Other
4	Food	Day	Friday	Wednesday

9.5.9 *Wilcoxon Signed Ranks*

1. Start SOFA and click the "Statistics" button.
2. Select "Wilcoxon Signed Ranks" from the Statistical Test list at the top of the window and then click the "Configure Test" button.
3. Data Source Table: tutoring
4. Group A Variable: Expgrapre
5. Group B Variable: Expgra01

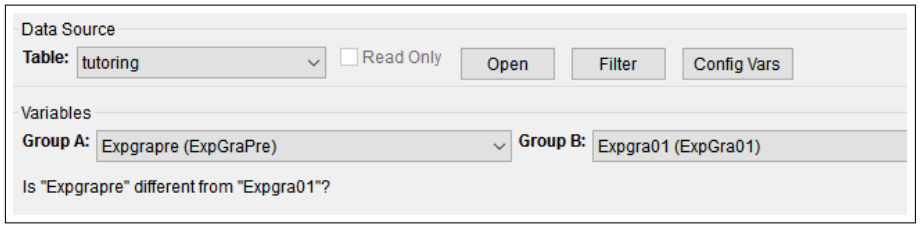


Figure 107: Setting Up a Wilcoxon Signed Ranks Test

6. SOFA calculates the following result.

Results of Wilcoxon Signed Ranks Test of "Expgrapre" vs "Expgra01"

Two-tailed p value: 2.789e-3 [1]

Wilcoxon Signed Ranks statistic: 344.0 [2]

Variable	N	Median	Min	Max
Expgrapre	100	3.0	1.0	4.0
Expgra01	100	3.0	1.0	4.0

Figure 108: Results of a Wilcoxon Signed Ranks Test

7. **p-Value**. As in all other statistical tests, the goal is a p-value less than 0.05 (5%). In the example calculated for this exercise, the p-value is 2.789×10^{-3}, which is well below the 0.05 threshold, so the null hypothesis would be rejected.

8. **t Statistic**. This number is of little value except, perhaps, to compare the results of one paired t-Test to another. In general, the greater the t-statistic the more likely the null hypothesis can be rejected, but it is challenging to find an appropriate "cutoff" score. Therefore, the most important result of an paired t-Test is the p-value.

9. **Degrees of Freedom**. This is calculated as the number of observations minus one. Thus, 99 degrees of freedom indicate 100 observations in the dataset.

10. **Other Statistics**. SOFA also calculates a number of statistics for each of the two groups, but those values are discussed elsewhere and will not be further covered here.

9.5.10 *Activity 4: Wilcoxon Signed Ranks*

Using the *maincafe* dataset in SOFA determine if there is a significant difference in the rating customers awarded for food and service (labeled "svc"). Report the p-value for these two variables.

9.6 DELIVERABLE

Complete the following activities in this lab:

Number	Name	Page
9.5.2	Activity 1: Wizard	98
9.5.6	Activity 2: Kruskal-Wallis H	100
9.5.8	Activity 3: Mann-Whitney U	102
9.5.10	Activity 4: Wilcoxon Signed Ranks	104

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

10

HYPOTHESIS TESTING: PARAMETRIC TESTS

10.1 INTRODUCTION

An important function for statistical analysis is to test a hypothesis to see if it adequately explains some observed phenomenon. The statistical processes for quantitative data are introduced in this lab and the processes for qualitative data are in HYPOTHESIS TESTING: NON-PARAMETRIC TESTS on 93.[1]

10.2 ANOVA

More Than 2 Groups : Ratio/Interval Data

An Analysis of Variance (ANOVA) is used to analyze the difference in three or more groups of observations. For example, imagine three groups of students were in the same class and one group was required to attend tutoring once a week, a second group twice a week, and a third group never. The null hypothesis (H_0) is "The amount of tutoring does not significantly change a student's score on the final exam." The alternate hypothesis (H_a) is "More frequent tutoring significantly changes a student's score on the final test." After the final exam was graded, an ANOVA could be administered and if that showed the test scores for those three groups of students had a significant difference then the null hypothesis would be rejected in favor of the alternate hypothesis.

10.3 T-TEST - INDEPENDENT

2 Groups : Ratio/Interval Data : Not Paired

An independent t-test is one of the most commonly used measures of the difference between two groups. One example of an independent t-test is to compare the spending habits of two similar groups of people. For example, do the residents of Tucson spend more on dining out than the residents of Phoenix? The null hypothesis (H_0) is "People in Phoenix and Tucson spend the same amount of money when dining out." The alternate hypothesis (H_a) is "People in Phoenix and Tucson spend different amounts of money when dining out." Imagine that the dining bills of 100 people from both cities were recorded and it was discovered that the mean bill in Phoenix is $15.13 and in Tucson is $12.47. A t-test would determine if there

1 The definitions of "hypothesis" and the various data types are found in the INTRODUCTION, beginning on page 3.

a significant difference in those two numbers, thus rejecting the null hypothesis, or if this is just a statistical fluke.

10.4 T-TEST - PAIRED

2 Groups : Ratio/Interval Data : Paired

A paired t-test is commonly used to compare the difference between two groups that are paired in some way. Typically, this is used in a test-retest type of experiment. Imagine that a group of students were given a pretest, taught a lesson, and then given a post-test. A paired t-test would be used to compare each student's pretest and post-test scores to see if there is a significant difference in the means for the two tests. Probably the most common use of a paired t-test is in medical drug trials. Typically, some physiological measurement is made of a group of patients (a blood pressure, for example), then the group is split in half and one group receives the drug being tested while the other receives a placebo. At the end of the trial the same measurement is made that was used during the pre-treatment and a paired t-test is used to compare the results of each individual in the two groups to see if the drug brought about a significant improvement.

10.5 PROCEDURE

`SOFA` makes it easy to complete any of the statistical tests listed in this lab exercise. A Wizard provides help in selecting an appropriate test but users can also manually select and configure whatever test they need.

10.5.1 *ANOVA*

1. Start `SOFA` and click the "Statistics" button
2. Select "ANOVA" from the Statistical Test list at the top of the window and then click the "Configure Test" button
3. Data Source Table: email
4. Averaged variable: Line_Breaks
5. Algorithm: Speed
6. Group By variable: Spam
7. From Group: No
8. To: Yes

9. Click "Show Results"

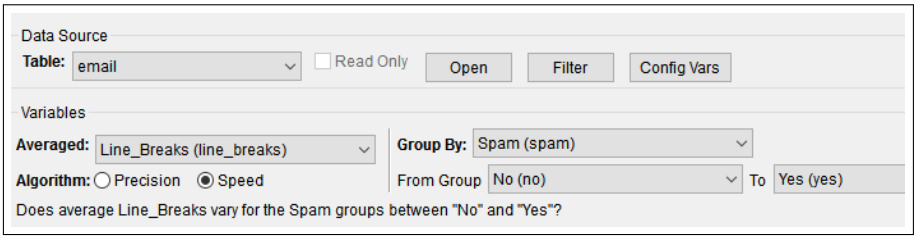


Figure 109: Setting Up the ANOVA Test

10. The ANOVA test generates a lot of information.

    a) **ANOVA Table**. This displays information about two broad partitions of the calculated data. The "Between" Source are values calculated when comparing two (or more) groups and the "Within" Source are values calculated within a single group.

Results of ANOVA test of average Line_Breaks for Spam groups from "No" to "Yes"

Analysis of variance table

Source	Sum of Squares	df	Mean Sum of Squares	F	p[1]
Between	1458157.074	1	1458157.074	155.821	< 0.001 (6.125e-35)
Within	29496188.971	3152	9357.928		

O'Brien's test for homogeneity of variance: 4.731e-14 [2]

Figure 110: ANOVA Table

- **Sum of Squares**. Each value in both groups is subtracted from the mean of that group, then that number is squared (to eliminate negative values), and, finally, all of the squared are summed. For the purposes of this exercise the Sum of Squares is of little value.
- **df**. The degrees of freedom is calculated for the groups being analyzed. The degrees of freedom is one less than the number of groups.[2]
- **Mean Sum of Squares**. This is the mean of the sum of the squares. This is calculated as the sum of squares divided by the degrees of freedom. (Note: the Mean Sum of Squares for the Within Source is frequently referred to as the "residuals").
- **F**. This is the value of the "F-Test" for this ANOVA. An F-Test is the ratio between two mean square values

2 The concept of Degrees of Freedom is explained in Lab 7.3.2, page 77.

and used to indicate if the variance between groups is significantly different from the variance within the groups. In general, if the null hypothesis is true then the F-Test value will be close to one, that is, there is not much difference in the variance between groups when compared to the variance within a group. In the example calculated for this exercise, the F-Test value is 155.821 which is much larger than one and would indicate that there is a significant difference in the variance of the "no" and "yes" groups.

- **p value**. This is computed from the F-value and is used to determine if the null hypothesis can be rejected.[3] *Most research reports include only the P-Value rather than all of the other calculated statistics since it summarizes the result into a single easy-to-understand number.* In most cases a P-Value less than 0.05 (5%) indicates that there is a significant difference in the groups being compared. In the example calculated for this exercise, the P-Value is far less than 0.05 (at 6.125×10^{-35}) so the result is significant and the null hypothesis can be rejected.
- **O'Brien's Test**. This test is commonly used to indicate if the variances within two groups is significantly different. If the value of O'Brien's test is less than 0.05 (5%) then the difference in the two variances is significant. In the example calculated for this exercise, O'Brien's test is well under 0.05 (at 4.731×10^{-14}), so the difference in the variance within the "no" group and "yes" group is significant.

b) **Group Summary Details**. This displays information about the various groups being analyzed.

Group summary details

Group	N	Mean	CI 95%[3]	Standard Deviation[4]	Min	Max	Kurtosis[5]	Skew[6]	p abnormal[7]
No	2809	118.007	114.311 - 121.704	99.955	2.0	452.0	-0.117	0.889	< 0.001 (9.353e-62)
Yes	345	49.119	42.288 - 55.950	64.733	1.0	364.0	4.573	2.198	< 0.001 (2.908e-36)

Figure 111: Group Summary Details

- **Group**. For this exercise there were only two groups: "No" and "Yes."
- **N**. The number of observations in each group.
- **Mean**. The mean for each group.

3 The concept of P-Value was explained in Lab 7.3.1, page 76.

- **CI** 95%. The 95% confidence interval. For example, there is 95% confidence that the actual mean for the "No" group lies between 114.311 and 121.704.
- **Standard Deviation**. The standard deviation for each group.
- **Min**. The minimum value for each group.
- **Max**. The maximum value for each group.
- **Kurtosis**. The Kurtosis[4] for each group.
- **Skew**. This is a measure of the symmetry of the data. While it is problematic to categorically state that some value of skewness is "bad," generally values lower than -1 or higher than $+1$ would be considered non-symmetrical. In the example calculated for this exercise, the "No" skew is $+0.889$ and would be considered reasonably symmetrical while the "Yes" skew is $+2.198$ and would be considered non-symmetrical.
- **p Abnormal**. This is a measure of the normality of the distribution. If the value is less than 0.05 then the data are not normally distributed. In the example calculated for this exercise, both groups have a value well below 0.05 and would be considered non-normally distributed. A histogram of each group's distribution is included in the ANOVA results and shown in the next two figures. It is easy to see that these data are not normally distributed.

4 Kurtosis and skew were described in Lab 1.3.2.1, page 6.

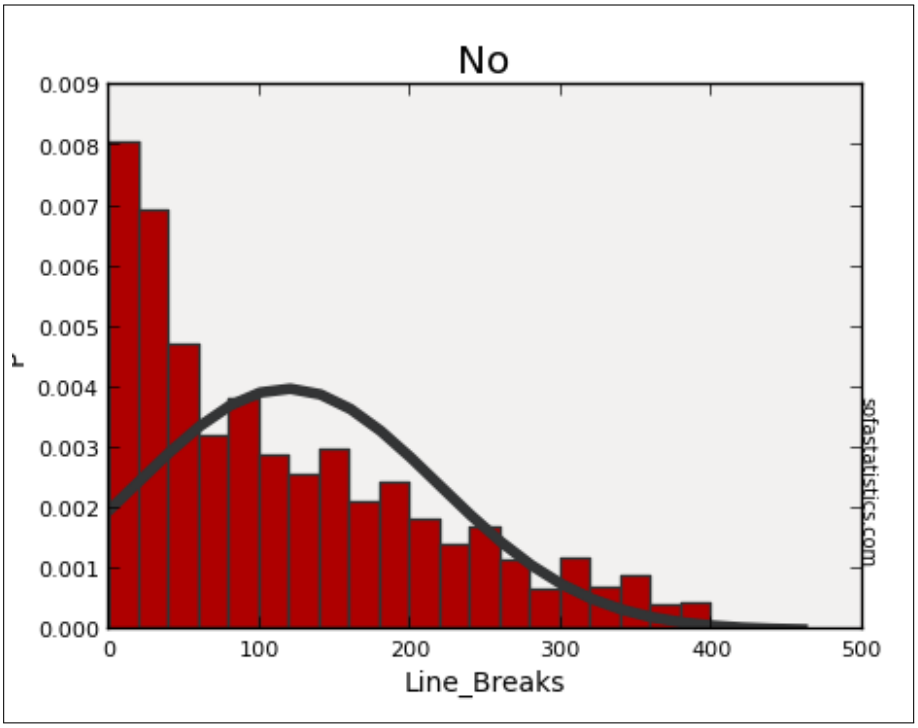


Figure 112: "No" Histogram

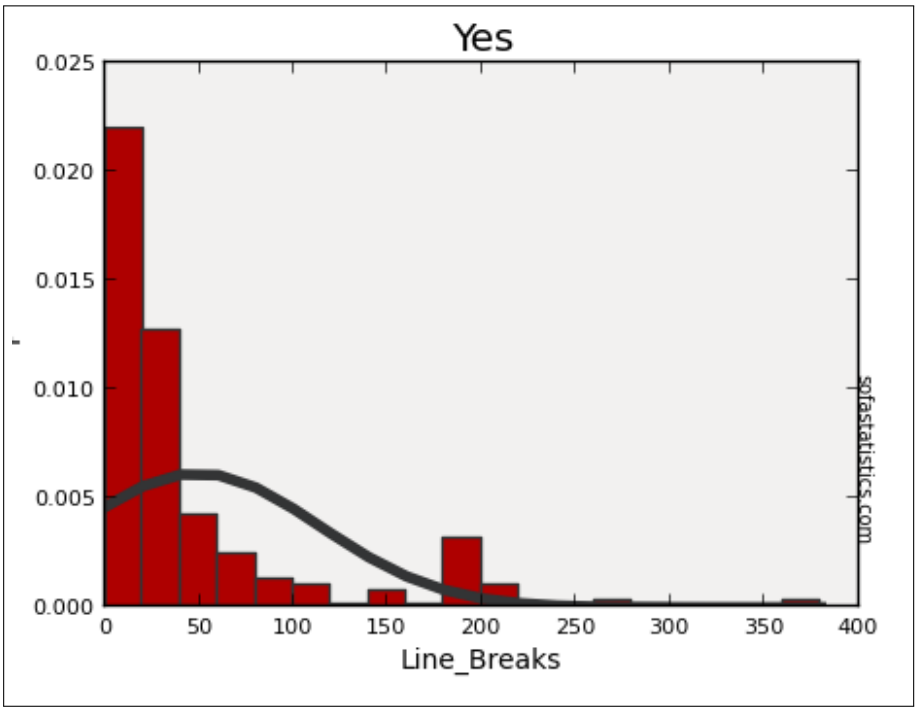


Figure 113: "Yes" Histogram

11. As another example where three groups are compared, consider an ANOVA using the "email" data, "Line_Breaks" Averaged variable, and "Attach" as the "Group By" variable. For the

grouped variable, choose to group from 0.0 to 2.0 only. Here are the results:

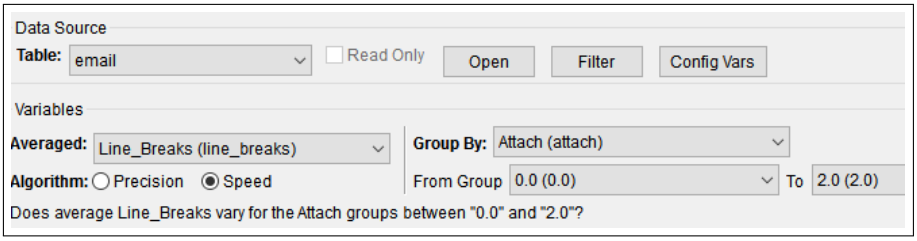


Figure 114: Setting Up ANOVA for Attach

Analysis of variance table

Source	Sum of Squares	df	Mean Sum of Squares	F	p[1]
Between	13726.623	2	6863.312	0.704	0.4948
Within	30511496.783	3129	9751.197		

O'Brien's test for homogeneity of variance: 0.6083 [2]

Figure 115: ANOVA Table

Group summary details

Group	N	Mean	CI 95%[3]	Standard Deviation[4]	Min	Max	Kurtosis[5]	Skew[6]	p abnormal[7]
0.0	2915	110.169	106.593 - 113.745	98.509	1.0	439.0	0.037	0.973	< 0.001 (4.090e-73)
1.0	138	114.674	98.768 - 130.579	95.330	2.0	396.0	-0.130	0.837	< 0.001 (8.971e-4)
2.0	79	98.392	73.560 - 123.225	112.610	3.0	430.0	0.202	1.182	< 0.001 (4.425e-4)

Figure 116: ANOVA Group Summary

10.5.2 *Activity 1: ANOVA*

Using the *maincafe* dataset in `SOFA` conduct an ANOVA and report the Between Source *p-value* for the following variables. Note: in the Word document submitted for this lab, Activity 1 should have a simple listing, something like this. (Notes: these are not the correct answers to the listed tests. To indicate tiny values use scientific notation in a form like $1.6e-5$ since that is easier to type.)

1	0.35
2	$1.27e-8$
3	237.4
4	0.73

Here are the variables to test:

Num	Averaged	Group By	From	To
1	Age	Food	1.0	5.0
2	Age	Svc	1.0	5.0
3	Age	Day	Friday	Wednesday
4	Food	Day	Friday	Wednesday

10.5.3 *Correlation - Pearson's*

The Pearson's R statistic is covered in Lab 7.2.1, Pearson's R, page 73. The SOFA activity related to Pearson's R is on page 80.

10.5.4 *t-test - Independent*

1. Start SOFA and click the "Statistics" button.
2. Select "t-test - independent" from the Statistical Test list at the top of the window and then click the "Configure Test" button.
3. Data Source Table: births
4. Averaged Variable: Weight
5. Group By Variable: Habit
6. Group A: Nonsmoker
7. Group B: Smoker

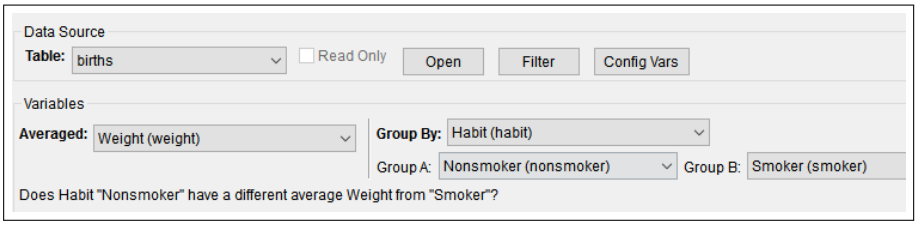


Figure 117: Setting Up Independent t-Test

8. Following is the result of that test.

Results of Independent Samples t-test of average "Weight" for Habit groups "Nonsmoker" vs "Smoker"

p value: 0.02779 [1]

t statistic: 2.203

Degrees of Freedom (df): 997

O'Brien's test for homogeneity of variance: 0.4154 [2]

Group	N	Mean	CI 95%[3]	Standard Deviation[4]	Min	Max	Kurtosis[5]	Skew[6]	p abnormal[7]
Nonsmoker	873	7.144	7.044 - 7.245	1.519	1.0	11.75	2.893	-1.185	< 0.001 (4.135e-44)
Smoker	126	6.829	6.587 - 7.071	1.386	1.69	9.19	1.742	-1.002	< 0.001 (3.349e-6)

Figure 118: Results of an Independent t-Test

9. **p Value**. The most important statistic in the results window is the p-value. As always, when this number is encountered it is desired for it to have a value less than 0.05 (5%). In the example calculated for this exercise, the p-value is 0.02779, which is less than 0.05, so there is a significant relationship between Birth Weight and Smoking Habit in this dataset.

10. **t Statistic**. This number is of little value except, perhaps, to compare the results of one independent t-Test to another. In general, the greater the t-statistic the more likely the null hypothesis can be rejected, but it is challenging to find an appropriate "cutoff" score. Therefore, the most important result of an independent t-Test is the p-value.

11. **Degrees of Freedom (df)**. The degrees of freedom is calculated as the number of different birth weights times one less than the number of groups.[5]

12. **O'Brien's Test**. This test is commonly used to indicate if the variances within two groups is significantly different. If the value of O'Brien's test is less than 0.05 (5%) then the difference in the two variances is significant. In the example calculated for this exercise, O'Brien's test greater than 0.05 (at 0.4154), so the difference in the variance between the "Nonsmoker" and "Smoker" groups is not significant.

13. **Other Statistics**. The table shows a number of statistical values for the two groups and each of the types of values have been described elsewhere.

14. **Histograms**. The results window includes a histogram for each of the groups in the calculation, similar that that found for an ANOVA (see page 112). Those histograms are not reproduced here.

5 The concept of Degrees of Freedom is explained in Lab 7.3.2, page 77.

10.5.5 *Activity 2: t-test - Independent*

Using the *maincafe* dataset in SOFA conduct a t-Test, Independent, and report the *p-value* for the following variables. Note: in the Word document submitted for this lab, Activity 2 should have a simple listing, something like this. (Notes: these are not the correct answers to the listed tests. To indicate tiny values use scientific notation in a form like $1.6e-5$ since that is easier to type.)

1 0.35
2 $1.27e-8$
3 237.4
4 0.73

Here are the variables to test:

Num	Averaged	Group By	From	To
1	Age	Food	1.0	5.0
2	Length	Meal	Breakfast	Other
3	Miles	Sex	Female	Other
4	Bill	Meal	Breakfast	Other

10.5.6 *t-test - Paired*

1. Start SOFA and click the "Statistics" button.
2. Select "t-test - paired" from the Statistical Test list at the top of the window and then click the "Configure Test" button.
3. Data Source Table: tutoring
4. Group A: Pretest
5. Group B: Test_1

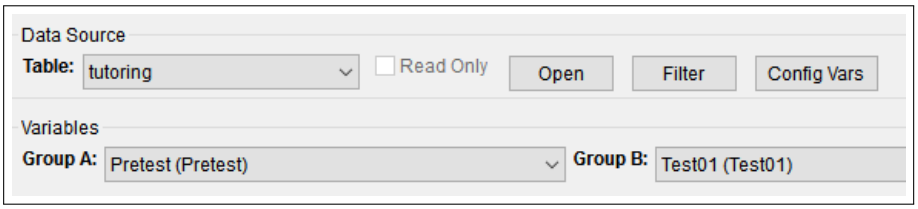


Figure 119: Setting Up a Paired t-Test

6. SOFA calculates the following result.

Results of Paired Samples t-test of "Pretest" vs "Test01"

p value: 0.01870 [1]

t statistic: -2.391

Degrees of Freedom (df): 99

Group	N	Mean	CI 95%[2]	Standard Deviation[3]	Min	Max
Pretest	100	48.08	46.214 - 49.946	9.523	35.0	65.0
Test01	100	48.97	46.933 - 51.007	10.392	30.0	71.0

Figure 120: Results of a Paired t-Test

7. **p-Value**. As in all other statistical tests, the goal is a p-value less than 0.05 (5%). In the example calculated for this exercise, the p-value is 0.01870, which is well below the 0.05 threshold, so the null hypothesis would be rejected.

8. **t Statistic**. This number is of little value except, perhaps, to compare the results of one paired t-Test to another. In general, the greater the t-statistic the more likely the null hypothesis can be rejected, but it is challenging to find an appropriate "cutoff" score. Therefore, the most important result of an paired t-Test is the p-value.

9. **Degrees of Freedom**. This is calculated as the number of observations minus one. Thus, 99 degrees of freedom indicate 100 observations in the dataset.

10. **Other Statistics**. `SOFA` also calculates a number of statistics for each of the two groups, but those values are discussed elsewhere and will not be further covered here.

11. **Graph**. `SOFA` also generates a graph of the results. In the following figure, the differences between pairs is graphed on the X-Axis. Ideally, the result would be a normal distribution, as indicated by the dark line. In this particular example, though, the results are rather flat which would indicate a possible problem with the experiment that the researcher would want to consider.

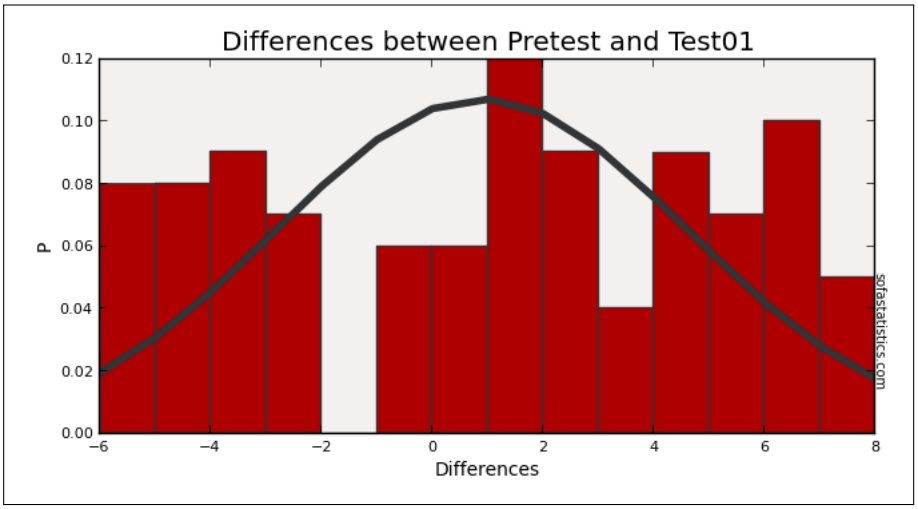


Figure 121: Graphic Results of a Paired t-Test

10.6 DELIVERABLE

Complete the following activities in this lab:

Number	Name	Page
10.5.2	Activity 1: ANOVA	113
10.5.5	Activity 2: t-test - Independent	116

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

11

FINAL

11.1 INTRODUCTION

The final lab requires the review and completion of several SOFA procedures. The specific activities will be provided by the instructor. Following are examples of the types of activities that may be required:

1. Using the *maincafe* dataset, create a frequency table of "Svc."
2. Using the *maincafe* dataset, create a histogram of "Age."
3. Using the *maincafe* dataset, calculate the mean, median, and standard deviation of "Age."

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

Part II

APPENDIX

12 APPENDIX

12.1 APPENDIX A: DATASETS

There are a number of datasets used in the lab exercises and this appendix lists the following information about those datasets:

1. Type of data (Nominal, Ordinal, Interval, Ratio)
2. Whether the data are normally distributed
3. Skew
4. Kurtosis

12.1.1 *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**. (Ratio : Normal : 1.131 : 0.96) The patient's age in years.
- **ank.di**. (Ratio : Normal : 0.07 : −0.372) The patient's ankle diameter in centimeters, measured as sum of two ankles.
- **ank.gi**. (Ratio : Normal : 1.131 : 0.96) The patient's ankle minimum girth in centimeters, measured as average of right and left girths.
- **bia.di**. (Ratio : Normal : 1.156 : −0.83) The patient's biacromial (shoulder width) in centimeters.
- **bic.gi**. (Ratio : Normal : 0.221 : −0.821) The patient's bicep girth in centimeters, measured when flexed as the average of right and left girths.
- **bii.di**. (Ratio : Normal : −0.417 : 1.09) The patient's biiliac (pelvic width) in centimeters.
- **bit.di**. (Ratio : Normal : −0.087 : 0.036) The patient's bitrochanteric (femur width) in centimeters.
- **cal.gi**. (Ratio : Normal : 0.278 : 0.301) The patient's calf maximum girth in centimeters, measured as average of right and left girths.

- **che.de**. (Ratio : Normal : 0.495 : −0.128) The patient's chest depth in centimeters, measured between spine and sternum at nipple level, mid-expiration.
- **che.di**. (Ratio : Normal : 0.259 : −0.727) The patient's chest diameter in centimeters, measured at nipple level, mid-expiration.
- **che.gi**. (Ratio : Normal : 0.239 : −0.879) The patient's chest girth in centimeters, measured at nipple line in males and just above breast tissue in females, mid-expiration.
- **elb.di**. (Ratio : Normal : 0.053 : −0.731) The patient's elbow diameter in centimeters, measured as sum of two elbows.
- **for.gi**. (Ratio : Normal : 0.153 : −0.972) The patient's forearm girth in centimeters, measured when extended, palm up as the average of right and left girths.
- **hgt**. (Ratio : Normal : 0.152 : −0.45) The patient's height in centimeters.
- **hip.gi**. (Ratio : Normal : 0.498 : 0.774) The patient's hip girth in centimeters, measured at at level of bitrochanteric (femur width) diameter.
- **kne.di**. (Ratio : Normal : 0.341 : 0.081) The patient's knee diameter in centimeters, measured as sum of two knees.
- **kne.gi**. (Ratio : Normal : 0.469 : 1.019) The patient's knee diameter in centimeters, measured as sum of two knees.
- **nav.gi**. (Ratio : Normal : 0.45 : 0.14) The patient's navel (abdominal) girth in centimeters, measured at umbilicus and iliac crest using iliac crest as a landmark.
- **sex**. (Nominal : Not Normal : N/A : N/A) The patient's sex.
- **sho.gi**. (Ratio : Normal : 0.14 : −0.896) The patient's shoulder girth in centimeters, measured over deltoid muscles.
- **thi.gi**. (Ratio : Normal : 0.691 : 0.731) The patient's thigh girth in centimeters, measured below gluteal fold as the average of right and left girths.
- **wai.gi**. (Ratio : Normal : 0.541 : −0.218) 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**. (Ratio : Normal : 0.402 : −0.348) The patient's weight in kilograms.
- **wri.di**. (Ratio : Normal : 0.048 : −0.483) The patient's wrist diameter in centimeters, measured as sum of two wrists.

- **wri.gi**. (Ratio : Normal : 0.152 : −0.719) The patient's wrist minimum girth in centimeters, measured as average of right and left girths.

12.1.2 *births*

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

- **Fage**. (Ratio : Normal : 0.292 : −0.246) The father's age.
- **Fageord**. (Ordinal : Not Normal : N/A : N/A) Father's age groups. This is not in the original dataset and was added via recoding.
- **Gained**. (Ratio : Normal : 0.461 : 0.752) The mother's weight gain, in pounds.
- **Gainedord**. (Ordinal : Not Normal : N/A : N/A) Weight gained groups. This is not in the original dataset and was added via recoding.
- **Gender**. (Nominal : Not Normal : N/A : N/A) The gender of the baby.
- **Habit**. (Nominal : Not Normal : N/A : N/A) Whether the mother was a smoker.
- **Lowbirthweight**. (Nominal : Not Normal : N/A : N/A) Whether the baby had a low birth weight.
- **Mage**. (Ratio : Normal : 0.263 : −0.636) The mother's age.
- **Mageord**. (Ordinal : Not Normal : N/A : N/A) Mother's age groups. This is not in the original dataset and was added via recoding.
- **Marital**. (Nominal : Not Normal : N/A : N/A) Whether the mother was married.
- **Mature**. (Nominal : Not Normal : N/A : N/A) The mother's maturity status.
- **Premie**. (Nominal : Not Normal : N/A : N/A) Whether the baby was premature.
- **Visits**. (Interval : Normal : 0.263 : −0.636) The number of hospital visits made by the mother.
- **Weeks**. (Interval : Normal : −2.016 : 7.681) Length of pregnancy, in weeks. This data are normally distributed, but the skew and kurtosis are very poor so the data may be better treated as not normal.

- **Weight**. (Interval : Normal : −1.145 : 2.7) Birth weight of the baby, in pounds.
- **Whitemom**. (Nominal : Not Normal : N/A : N/A) Whether the mother was white.

12.1.3 *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 : Not Normal : N/A : N/A) Vehicle drive train with levels 4WD, front, and rear.
- **mpgCity**. (Ratio : Normal : 1.45 : 1.968) City mileage (miles per gallon).
- **passengers**. (Ordinal : Not Normal : N/A : N/A) The vehicle passenger capacity.
- **price**. (Ratio : Normal : 1.327 : 1.897) Vehicle price in U.S. dollars.
- **type**. (Ordinal : Not Normal : N/A : N/A) The vehicle type with levels large, midsize, and small.
- **weight**. (Ratio : Normal : −0.233 : −1.168) Vehicle weight in pounds. The skew and kurtosis for this data are good, but the histogram is rather flat. This may be better considered a not normal distribution.

12.1.4 *doorsurvey*

This is simulated data. Students recorded observations about the customers entering a store between 9 AM and 11 AM for a one week period.

- **agecat**. (Ordinal : Not Normal : N/A : N/A) An estimate of the age category for the customer where 1 is less than 30 years old, 2 is between 30 and 50 years old, and 3 is more than 50 years old. If a group entered then the age of the first adult in the group was estimated.
- **day**. (Ordinal : Not Normal : N/A : N/A) The date the customers were observed.
- **grpnum**. (Ordinal : Not Normal : N/A : N/A) The number of people in the group.

- **gender**. (Nominal : Not Normal : N/A : N/A) The sex of the customer: *m* or *f*. If a group entered then the sex of the first adult in the group was recorded.

12.1.5 *email*

This dataset contains 3154 observations from the email received by one person over a period of several months in 2012.

- **attach**. (Ratio : Not Normal : 16.278 : 502.528) The number of attached files.
- **cc**. (Ratio : Not Normal : 13.51 : 225.074) How many people were CCed on the message.
- **dollar**. (Ratio : Not Normal : 4.456 : 22.826) The number of times a dollar sign or the word "dollar" appeared in the email.
- **exclaim_mess**. (Ratio : Not Normal : 4.107 : 23.946) The number of exclamation points in the email message.
- **exclaim_subj**. (Nominal : Not Normal : N/A : N/A) *No* if the email subject did not have an exclamation point, otherwise *yes*.
- **format**. (Nominal : Not Normal : N/A : N/A) *Text* if the message was sent in text format and *html* if it used HTML format.
- **image**. (Ratio : Not Normal : 9.621 : 121.313) The number of images attached.
- **inherit**. (Ratio : Not Normal : 18.188 : 478.488) The number of times "inherit" (or an extension, such as "inheritance") appeared in the email.
- **line_breaks**. (Ratio : Normal : 0.974 : 0.046) The number of line breaks in the email (does not count text wrapping).
- **num_char**. (Ratio : Normal : 0.944 : −0.017) The number of characters in the email, in thousands.
- **number** (Ordinal : Not Normal : N/A : N/A) *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**. (Ratio : Not Normal : 16.173 : 333.185) The number of times "password" appeared in the email.
- **re_subj**. (Nominal : Not Normal : N/A : N/A) *No* if the subject did not include any of these: "Re:", "RE:", "re:", or "rE:", otherwise *yes*.

- **sent_email**. (Nominal : Not Normal : N/A : N/A) *No* if no email was sent to the sender in the last 30 days, otherwise *yes*.
- **spam**. (Nominal : Not Normal : N/A : N/A) *Yes* if the message is spam, otherwise *no*.
- **spamnum**. (Ordinal : Not Normal : N/A : N/A) Numeric representation of the "spam" data item. This is not in the original dataset and was added via recoding.
- **to_multiple**. (Nominal : Not Normal : N/A : N/A) *No* for mail that was sent to only one person, otherwise *yes*.
- **urgent_subj**. (Nominal : Not Normal : N/A : N/A) *Yes* if the subject included the word "urgent," otherwise *no*.
- **viagra**. (Ratio : Not Normal : 56.134 : 3149.0) The number of times "viagra" appeared in the email.
- **winner**. (Nominal : Not Normal : N/A : N/A) *Yes* if the word "winner" appeared in the email, otherwise *no*.

12.1.6 *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**. (Ratio : Normal : 0.053 : −0.551) Average number of hours per week the child watched cartoons on TV during the past three months.
- **Count**. (Interval : Normal : −0.197 : −0.653) Age in months when the child first counted to ten successfully.
- **Edutv**. (Ratio : Normal : −0.053 : −0.375) Average number of hours per week the child watched an educational program on TV during the past three months.
- **Fatheriq**. (Interval : Normal : 1.091 : 1.459) Father's IQ.
- **Motheriq**. (Interval : Normal : 0.247 : 0.035) Mother's IQ.
- **Read**. (Ratio : Normal : −0.576 : −0.449) Average number of hours per week the child's mother or father reads to the child.
- **Score**. (Ratio : Normal : −0.016 : −0.528) The score in test of analytical skills.

- **Speak**. (Interval : Normal : -0.627 : -0.153) Age in months when the child first said "mummy" or "daddy."

12.1.7 *maincafe*

This is simulated data. Customers of the Main Street Café completed surveys over a one week period. Note: this dataset is occasionally re-created so the values for skew and kurtosis are only estimates.

- **age**. (Ratio : Normal : ≈ 0.2 : ≈ -0.6) The age in years of the person completing the survey.
- **bill**. (Interval : Normal : ≈ 1.0 : ≈ 0.5) The bill for the meal.
- **day**. (Nominal : Not Normal : N/A : N/A) The day of the week the person visited the cafe. The levels are written names of days, like "Sunday" and "Monday."
- **food**. (Ordinal : Not Normal : N/A : N/A) A rating for the food, from one to five "stars."
- **length**. (Interval : Normal : ≈ -0.5 : ≈ -0.2) The length of the visit in minutes.
- **meal**. (Nominal : Not Normal : N/A : N/A) The meal eaten stored as Breakfast, Lunch, Dinner, and Other.
- **miles**. (Interval : Not Normal : ≈ 8.2 : ≈ 71.0) The number of miles from the visitor's home and the cafe.
- **pref**. (Nominal : Not Normal : N/A : N/A) A binary item for the preference in tables of the diner. The levels are "table" and "booth."
- **ptysize**. (Interval : Normal : ≈ 0.7 : ≈ 0.1) The size of the dining party.
- **recmd**. (Nominal : Not Normal : N/A : N/A) A binary item for whether the customer would recommend the café to other people, stored as "no" and "yes."
- **sex**. (Nominal : Not Normal : N/A : N/A) The gender of the person completing the survey. The levels are: male, female, and other.
- **svc**. (Ordinal : Not Normal : N/A : N/A) A rating for the service, from one to five "stars."
- **tip**. (Ratio : Normal : ≈ 1.4 : ≈ 1.2) The amount of tip left.

12.1.8 *rivers*

The Rivers dataset is a list of the lengths of the longest 141 rivers in the United States. These data are not normally distributed.

> 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

12.1.9 *tutoring*

This is simulated data. A professor designed an experiment where she first administered a pretest to all students. She then required all students to work with an online tutoring service every week. At the end of four weeks she administered a post-test and labeled the scores on that test as "Testo1." She then required students to attend weekly face-to-face tutoring sessions. After four weeks she administered a different, but similar, post-test and labeled the scores on that test as "Testo2." After each of the three tests she asked the students what grade they expected to earn. This dataset includes the test scores and grade predictions for each of the students.

- **ExpGrao1**. (Ordinal : Not Normal : N/A : N/A) This is the expected grade the student reported after Testo1. The levels are A=4, B=3, C=2, and D=1.
- **ExpGrao2**. (Ordinal : Not Normal : N/A : N/A) This is the expected grade the student reported after Testo2. The levels are A=4, B=3, C=2, and D=1.
- **ExpGraPre**. (Ordinal : Not Normal : N/A : N/A) This is the expected grade the student reported during the pretest. The levels are A=4, B=3, C=2, and D=1.
- **Pretest**. (Ratio : Normal : 0.199 : -0.591) This is the student's pretest score.
- **StuID**. (Interval : Not Normal : N/A : N/A) This is just a one-up number assigned to identify each student in the study.

- **Test01**. (Ratio : Normal : 0.063 : −0.205) This is the score for students on the first post-test.
- **Test02**. (Ratio : Normal : −0.05 : −0.109) This is the score for students on the second post-test.

12.2 APPENDIX B: RECODING VARIABLES

12.2.1 *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.

12.2.2 *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).

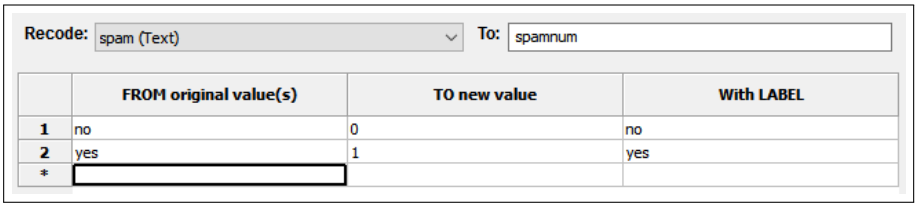


Figure 122: Recoding Spam Field

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

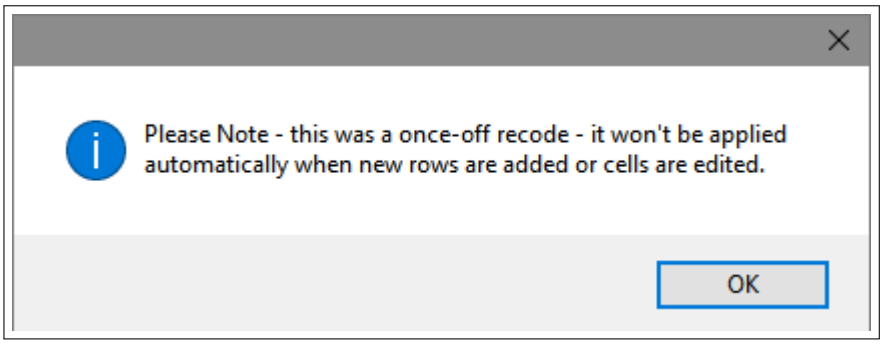


Figure 123: Recode Warning on Save

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."

12.3 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.

12.3.1 *Styles*

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

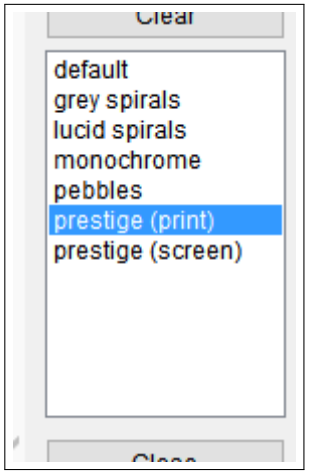


Figure 124: SOFA Styles

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

12.3.2 *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.

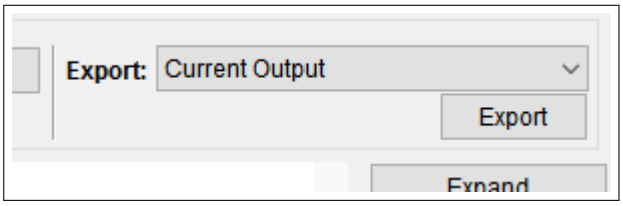


Figure 125: Selecting Export Type

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

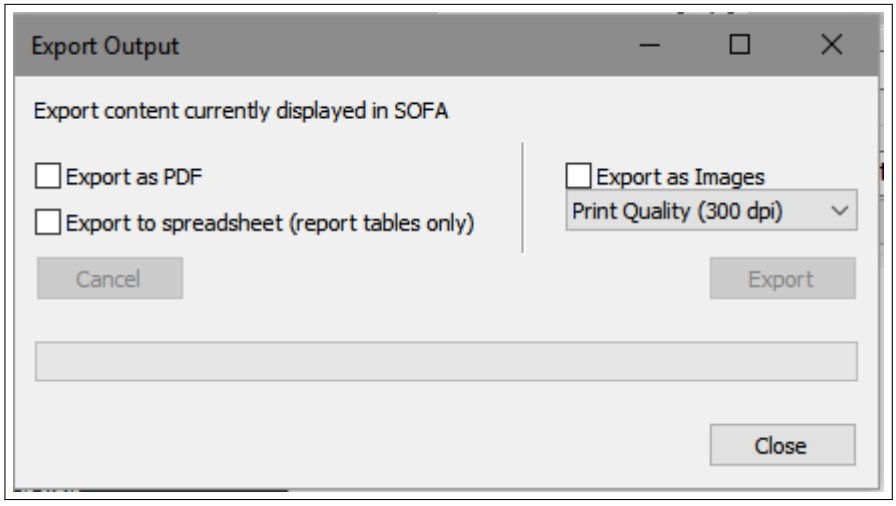


Figure 126: Specifying Desired Export

- 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.

12.3.3 *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.

12.3.4 *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 drop-down 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.