

# Introduction to Data Analysis

*(Second Edition)*

**By Matt Teachout  
College of the Canyons  
Santa Clarita, CA, USA**



This material is from *Introduction to Data Analysis*, 2<sup>nd</sup> edition by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a “CC-BY” [Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by/4.0/) – 3/17/2021

Special thanks to all of the people that made this book possible.

Thanks to all of the *Intermediate Algebra for Statistics* students and teachers at College of the Canyons for pioneering this material and giving great suggestions for improvement.

Thanks to the COC statistics team  
(Joe Gerda, Kathy Kubo, Ambika Silva, Dustin Silva)  
for your leadership and unending work to improve  
statistics education. I will always be grateful  
to be part of the best team ever.

Thank you to the original “honey badgers” Myra Snell and Katie Hern  
at the California Acceleration Project. You inspired so many  
teachers and programs. You made us believe we could  
change the system and taught us how to truly help students.

Thank you to Yousef Alasfoor, Kayla Teachout, Kathy Kubo and Udani Ranasinghe  
for helping me navigate through massive amounts of social justice data.

Thank you to Udani Ranasinghe for your help with Canvas and making answer keys.

Thank you to our “resident statistician” and fearless leader Joe Gerda.  
Your statistics expertise and leadership have been incredible.  
We could not have done this without you.

Special thanks to Kathy Kubo, Ralph (Randy) Ades,  
Udani Ranasinghe, Rupa Sinha, and Joe Gerda  
for your support, encouragement, help and suggestions.

Thank you to James Glapa-Grossklag, Brian Weston  
and the COC OER office staff for your support and help.



This material is from *Introduction to Data Analysis*, 2<sup>nd</sup> edition by Matt Teachout,  
College of the Canyons, Santa Clarita, CA, USA, and is licensed under a “CC-BY”  
[Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by/4.0/) – 3/17/2021

## Introduction

We live in the age of computers and the internet. We are exposed to huge volumes of data every day. How do we make sense of this massive amount of information? How can we tell the difference between helpful and misleading information? How can businesses know what their customers want and need, or hospitals analyze various types of infections and which treatments are working and which are not? All of these questions revolve around the study of data and statistics. A good understanding of statistics is vital to anyone living in the modern world, however very few people understand how to analyze data. The shortage of trained statisticians, data analysts, and data scientists is a huge problem worldwide.

There are many fabulous books on statistics and analyzing data. Unfortunately, they are extremely expensive and most people cannot afford the cost. I wrote this book to help people learn to analyze data. It is free to use the material in this book, update it, add to it, print it or just read it. It is an open educational resource (OER) and so anyone can use it.

Many college students struggle to balance work and family with their education. One of the biggest roadblocks for many students is the cost of textbooks. Students today cannot afford the cost of textbooks and so chose to attend classes without purchasing books and materials needed for the class. It goes without saying, that this is a major impediment to passing their classes, but the students have no choice. They simply cannot afford \$100-\$200 for a textbook. For this reason, I believe strongly in open educational resources (OER). Open source materials like this book are available and are virtually free for students.

### Notes about OER and Creative Commons Licensing

This textbook is licensed through Creative Commons as “Attribution CC-BY”. Creative Commons describes this license as follows: “This license lets others distribute, remix, tweak, and build upon (the author’s) work, even commercially, as long as they (give) credit (to the author) for the original creation.” This is the most accommodating of licenses offered. Recommended for maximum dissemination and use of licensed materials.” If you need to see the license deed or legal code, please go to <https://creativecommons.org/licenses/> and look under the “CC-By” section.

### Pre-Statistics or Intermediate Algebra for Statistics

I tell my beginning statistics students all the time that the study of statistics is a deep well of knowledge, and they are only playing in the puddle. Statisticians, data analysts and data scientists are life-long learners and spend years and years studying this subject.

This is an introduction to some very basic data analysis techniques. It is a book designed for anyone new to statistics. It can be used with a pre-statistics class or an intermediate algebra for statistics class.

Pre-statistics classes focus on helping students understand and analyze categorical and quantitative data sets.

Intermediate algebra for statistics has the same information as a pre-statistics class but often includes some intermediate algebra curve analysis and regression techniques. Many statisticians and statistics educators feel that curve analysis and regression is a topic better addressed in more advanced level statistics classes since this is a topic explored by many graduate level statistics students.

If your college requires intermediate algebra for statistics, I have included that material in chapter 6. If your college is using a pre-statistics class, then chapters 1-5 should suffice.

### Important Note about Technology

We live in the age of computers, internet and a huge volume of data. No practicing statistician or data scientist uses a calculator or tables to analyze data. You cannot even begin to analyze a data set with a million values by hand with a calculator. You need high-powered computer software. There are many statistics software programs on the market, but very few of them are free.



This material is from *Introduction to Data Analysis*, 2<sup>nd</sup> edition by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a “CC-BY” [Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by/4.0/) – 3/17/2021

If you read the history of statistics, you will find brilliant scientists, mathematicians and people in business who had to try to figure out data, but had no access to a computer. (Computers had not been invented yet.) Our pioneers of statistics dreamed of the day that they could compute statistics and graphs and analyze data with the touch of a button. They invented complicated techniques for analyzing data because they had no choice. Today, computers can calculate statistics and graphs directly.

Here is the problem. Most statistics classes taught in high schools, community colleges and even some universities are teaching statistics as if computers have not been invented yet. They are teaching the techniques developed by our pioneers of statistics before the computer age. This is a terrible approach to the subject, especially for the thousands of students that actually want to work in the field. A statistics class should be a study of how to practically collect and analyze data with a computer, not a class on what to do if computers have not been invented yet.

Are formulas important in statistics? Yes. We study formulas to understand what they tell us about the data and the world around us. The pioneers of statistics did an amazing job of addressing the major ideas of statistics with formulas and inventive calculations. However, we should not use a formula and a calculator to calculate a statistic for a data set with 10,000 values or use charts that list critical values and P-values. High-powered computers with statistics software can calculate the statistic and make graphs directly. Then students can focus on the analysis part, and explore and discover the meaning behind the data.

This book will show students how to use statistics software to calculate statistics and graphs. I want students to learn to analyze the data and not spend all their time just trying to calculate something. Remember, no one pays a data analyst to calculate something a computer can already do. They are paid to explore and explain what the data may be telling us.

### **Data Sets**

The national (GAISE) guidelines for teaching statistics recommend that you use real data. Allowing students to learn statistics principles through analysis of real data is key. With that being said, there are many places where raw data can be found and used. The key data sets throughout this book are located at my website [www.matt-teachout.org](http://www.matt-teachout.org). Just click on "Int Alg for Stats" and then "Data Sets".

### **The Computer Dilemma**

#### Face to Face Classes

A statistics or pre-statistics class should be taught in a computer lab. It is important to allow the computers to do the difficult calculations. Students need to focus on interpretation and discovering the meaning behind the data. They cannot do that if they spend all their time trying to calculate with a formula or making graphs by hand.

If your school wants to teach statistics or pre-statistics, but you cannot teach in a computer lab, here are some suggestions for you.

1. Reserve unused computer labs. Some schools may have a couple computer labs that are not always in use. Schedule your statistics and pre-statistics classes in order to use the computer lab. Even if you can only reserve the lab once a week or once every two weeks, it will be a huge help to students.
2. Have groups of students share computers. If you do have a few computers in your classroom, you can divide the class up into groups and share computers. This has many advantages like encouraging explanations to one another and teamwork.
3. Teachers can use their own computer or laptop to project statistics software on a screen and have the class analyze the data with you. Teachers without any computer can make printed copies of the software printouts for your class and for exams. It is a poor substitute for a computer lab, but it is much better than teaching statistics as if computers have not been invented yet.

### **Free Statistics Software**

Teaching statistics with computer software is very important, but many schools and students cannot afford to pay for software. If you are teaching pre-stats, intermediate algebra for stats, or intro stats online or face-to-face, it is vital that students have access to statistics software. I highly recommend a free statistics software that is easy to use. Most free software is not OER licensed, but they are still free for students. My favorite free statistics software



This material is from *Introduction to Data Analysis*, 2<sup>nd</sup> edition by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a "[CC-BY](https://creativecommons.org/licenses/by/4.0/)" [Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by/4.0/) – 3/17/2021

programs are StatKey ([www.lock5stat.com](http://www.lock5stat.com)) and Statcato ([www.statcato.org](http://www.statcato.org)). I use both of these programs throughout the textbook.

### Notes about the 2<sup>nd</sup> Edition

The second edition of Introduction to Data Analysis is similar to the first edition, but there are a few key differences.

1. **Change in software:** The 1st edition only used Statcato. Statcato is a great program and free for students, but is difficult to use with online classes. It works great in the classroom when every computer in the class has Statcato installed. However, some students have a hard time downloading it on their home computers. So for calculations, I moved the 2nd edition to using the free program StatKey ([www.lock5stat.com](http://www.lock5stat.com)). It is free, online hosted and so does not need to be downloaded. It works great on MAC and PC. It is ideal for online classes. There is still some Statcato in the book, but students only have to analyze Statcato printouts provided and not calculate. If they are calculating, they are using StatKey. The book shows students and faculty how to use StatKey.
  2. **Added chapter 1 on Data:** I wanted to expand on the ideas of types of data, random, bias, collecting data and experimental design in the textbook, so I included a new chapter 1 on data. The 2nd edition of the book has 7 chapters while the 1st edition has 6 chapters. Chapters 2-7 in the 2<sup>nd</sup> edition have similar content as chapters 1-6 in the 1<sup>st</sup> edition.
  3. **Added social justice questions:** The 2nd edition of the book has social justice questions dealing with racism and discrimination.
  4. **Updated Projects:** There are optional projects available for chapters 2-6 in the 2<sup>nd</sup> edition of the textbook. The projects have updated directions for both online and face-to-face classes.
- 



## Table of Contents

### Introduction to Data Analysis (2<sup>nd</sup> edition)

#### Chapter 1: Data

- Chapter 1 Introduction
- Section 1A: Two Types of Data (Categorical and Quantitative)
- Section 1B: Collecting Data
- Section 1C: Bias
- Section 1D: Experimental Design

#### Chapter 2: Categorical Data Analysis

- Chapter 2 Introduction
- Section 2A: Proportions and Percentages
- Section 2B: Bar Charts and Pie Charts with Technology
- Section 2C: Comparing Percentages (% Ratio, % of Increase)
- Section 2D: Estimating Amounts with Percentages

#### Chapter 3: Categorical Relationships

- Chapter 3 Introduction
- Section 3A: Contingency Tables with Technology
- Section 3B: Marginal and Joint Percentages
- Section 3C: Conditional Percentages and Categorical Relationships

#### Chapter 4: Normal Quantitative Data Analysis

- Chapter 4 Introduction
- Section 4A: Finding Shape with Dot Plots and Histograms
- Section 4B: Shapes and Centers
- Section 4C: Understanding the Mean Average
- Section 4D: Spread, Standard Deviation and Typical Values for Normal Quantitative Data
- Section 4E: Finding Unusual Values (Outliers) and Summarizing Normal Quantitative Data

#### Chapter 5: Non-normal and Skewed Quantitative Data Analysis

- Chapter 5 Introduction
- Section 5A: Review of Shapes and Centers, Dot Plots and Histograms
- Section 5B: Understanding the Median Average
- Section 5C: Spread and Typical Values for Skewed Quantitative Data, Quartiles, Interquartile Range, and the Five Number Summary
- Section 5D: Box Plots, Finding Unusual Values (Outliers) for Skewed Quantitative Data
- Section 5E: Various Quantitative Statistics (Measures of Center, Spread and Position)

#### Chapter 6: Linear Quantitative Relationships (Correlation and Regression)

- Chapter 6 Introduction
- Section 6A: Rectangular Coordinate System, Scatterplots, Explanatory and Response Variables
- Section 6B: Strength and Direction of Linear Quantitative Relationships, Correlation Coefficient ( $r$ )
- Section 6C: Coefficient of Determination ( $r^2$ ), Confounding Variables, Correlation is not Causation
- Section 6D: Best Fit Regression Line with Technology, Slope and Y-intercept Interpretation
- Section 6E: Residuals, Residual Plots, Histogram of the Residuals, and the Standard Deviation of the Residual Errors ( $S_e$ )
- Section 6F: Predictions, Scope of the X values, Extrapolation and Prediction Error

#### Chapter 7: Non-linear Curved Quantitative Relationships

- Chapter 4 Introduction
- Section 4A: Exponential Quantitative Relationships



This material is from *Introduction to Data Analysis*, 2<sup>nd</sup> edition by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a “CC-BY” [Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by/4.0/) – 3/17/2021

## Chapter 1: Collecting and Analyzing Data

### Vocabulary

Data: Information in all forms.

Population: The collection of all people or objects to be studied.

Census: Collecting data from everyone in a population.

Sample: Collecting data from a small subgroup of the population.

Bias: When data does not represent the population.

**Chapter 1 Introduction:** The goal of collecting and analyzing data is to understand the world around us. How data is collected is very important. The goal of collecting data is to get “unbiased” data that represents the population. Analyzing biased data may result in incorrect conclusions and lead to a misguided view of the world around us. It is also important to have a goal in mind when you collect data. Are we trying to find a population percentage from categorical data or a population average from quantitative data? Are we trying to show that two variables are related or are we trying to show cause and effect? Data needs to be collected differently depending on what goal you have in mind.

---



## Section 1A – Two Types of Data – Categorical and Quantitative

### Vocabulary

Data: Information in all forms.

Analyzing data is an important skill in the modern world. Companies, hospitals, sports teams all need to analyze data in order to make good decisions. But what is data? A good way to think of data is information in all forms. It is often a list of answers to a question or it may be organized in a spread sheet.

One of the most important factors when analyzing data is to determine what type of data you have and how many variables you are analyzing. Let us start with the types of data.

There are two general types of data, categorical and quantitative.

### Categorical Data

Categorical data (or qualitative data) are generally labels that tell us something about the people or objects in the data set. For example, what country do they live in, what is the person's occupation, or what kind of pet they have?

Usually categorical data is made up of words (do you smoke - yes or no), but occasionally a number can be used in place of a word. For example, a zip code can be used instead of the place a person lives. The numbers "1" and "2" may be used instead of yes and no. Or the number of the month instead of the name of the month. Notice though it is a number it really represents a word.

Do you Smoke Cigarettes (Yes or No)    What type of Car do you drive?    What Month were you born in?

Yes	Ford	11
No	Honda	2
No	Dodge	5
Yes	Toyota	7
No	Chevy	9
No	Tesla	10
No	Mercedes	1
No	Chevy	6
No	Toyota	3
No	Ford	2

### Quantitative Data

Quantitative data are numbers that measure or count something. They usually have units and taking an average makes sense. For example: a list of people's heights in inches, or temperature in degrees Celsius, or a list of how many dogs are there in various animal shelters across Los Angeles. Notice in each of these cases the data is numerical and an average seems appropriate in the context. We can find the average height, the average temperature, or the average number of dogs in animal shelters in Los Angeles.

Height (Inches)    Temperature (Celsius)    Number of Dogs at Animal Shelters in LA

60

21.4

122



This material is from *Introduction to Data Analysis*, 2<sup>nd</sup> edition by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a "CC-BY" [Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by/4.0/) – 3/17/2021



65	25	74
68	38.2	68
59	30	39
62.5	29.6	147
73	31.9	26
61.25	36.4	73
70	28	91
64	20.1	31
66	27.5	44

### Numbers used as categories

Remember, not all numeric data is quantitative. Ask yourself if the numbers are measuring or counting something and if an average would make sense. For example, averaging a list of the months people are born in would not really tell us anything. In addition, identity numbers like hospital ID numbers, student ID numbers or social security numbers are not measuring anything and an average would not make sense in the context so they are not quantitative.

### New Vocabulary

Data: Information in all forms.

Categorical Data: Data consisting of words describing people or objects. Numbers may sometimes be used in place of words.

Quantitative Data: Numerical measurement data, often including units, counts or averages.

---



### Practice Problems Section 1A

1. The American black bear is a medium-sized bear from North America. The following data was taken from 54 American black bears. Classify each column of data as categorical or quantitative. If the column is quantitative, what are the units? If the column is categorical, indicate how many different options there are in that category.



AGE (months)	Month Data Taken	Gender	Head Length (In)	Head Width (In)	Neck Circum (in)	Length (in)	Chest (in)	Weight (Lbs)
19	July	male	11	5.5	16	53	26	80
55	July	male	16.5	9	28	67.5	45	344
81	September	male	15.5	8	31	72	54	416
115	July	male	17	10	31.5	72	49	348
104	August	female	15.5	6.5	22	62	35	166
100	April	female	13	7	21	70	41	220
56	July	male	15	7.5	26.5	73.5	41	262
51	April	male	13.5	8	27	68.5	49	360
57	September	female	13.5	7	20	64	38	204
53	May	female	12.5	6	18	58	31	144
68	August	male	16	9	29	73	44	332
8	August	male	9	4.5	13	37	19	34
44	August	female	12.5	4.5	10.5	63	32	140
32	August	male	14	5	21.5	67	37	180
20	August	female	11.5	5	17.5	52	29	105
32	August	male	13	8	21.5	59	33	166
45	September	male	13.5	7	24	64	39	204
9	September	female	9	4.5	12	36	19	26
21	September	male	13	6	19	59	30	120
177	September	male	16	9.5	30	72	48	436
57	September	female	12.5	5	19	57.5	32	125
81	September	female	13	5	20	61	33	132
21	September	male	13	5	17	54	28	90
9	September	male	10	4	13	40	23	40
45	September	male	16	6	24	63	42	220
9	September	male	10	4	13.5	43	23	46
33	September	male	13.5	6	22	66.5	34	154
57	September	female	13	5.5	17.5	60.5	31	116
45	September	female	13	6.5	21	60	34.5	182
21	September	male	14.5	5.5	20	61	34	150
10	October	male	9.5	4.5	16	40	26	65
82	October	female	13.5	6.5	28	64	48	356
70	October	female	14.5	6.5	26	65	48	316
10	October	male	11	5	17	49	29	94
10	October	male	11.5	5	17	47	29.5	86
34	October	male	13	7	21	59	35	150
34	October	male	16.5	6.5	27	72	44.5	270
34	October	male	14	5.5	24	65	39	202
58	October	female	13.5	6.5	21.5	63	40	202
58	October	male	15.5	7	28	70.5	50	365
11	November	male	11.5	6	16.5	48	31	79
23	November	male	12	6.5	19	50	38	148
70	October	male	15.5	7	28	76.5	55	446
11	November	female	9	5	15	46	27	62
83	November	female	14.5	7	23	61.5	44	236
35	November	male	13.5	8.5	23	63.5	44	212
16	April	male	10	4	15.5	48	26	60
16	April	male	10	5	15	41	26	64
17	May	male	11.5	5	17	53	30.5	114
17	May	female	11.5	5	15	52.5	28	76
17	May	female	11	4.5	13	46	23	48
8	August	female	10	4.5	10	43.5	24	29
83	November	male	15.5	8	30.5	75	54	514
18	June	male	12.5	8.5	18	57.3	32.8	140

2. The following data was taken from various cereals. Classify each column of data as categorical or quantitative. If the column is quantitative, what are the units? If the column is categorical, indicate how many different options there are in that category.



This material is from *Introduction to Data Analysis*, 2<sup>nd</sup> edition by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a “CC-BY” [Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by/4.0/) – 3/17/2021

Name	Manufacturer	Target (Adult or Child)	Shelf displayed at store	Calories per serving	Carbs (grams per serving)	Fat (grams per serving)	Fiber (grams per serving)	Potassium (milligrams per serving)	Protein (grams per serving)	Sodium (milligrams per serving)	Sugar (grams per serving)	Vitamin (Percent of daily need per serving)	Consumer Report Magazine Rating	Serving Size (Cups per serving)	Weight (Ounces per serving)
Cap'n Crunch	Quaker	Child	Middle	120	11	2	0	35	1	220	12	25	10	0.75	1
Cocoa Puffs	General	Child	Middle	110	12	1	0	35	1	180	13	25	23	1	1
Tit	General	Child	Middle	110	13	1	0	25	1	140	12	25	20	1	1
Apple Jacks	Wollogg	Child	Middle	110	11	0	1	30	2	125	14	25	33	1	1
Corn Chex	Ralston	Adult	Bottom	110	21	0	0	25	2	280	3	25	41	1	1
Corn Flakes	Wollogg	Adult	Bottom	100	21	0	1	35	2	290	2	25	46	1	1
Nut & Honey	Wollogg	Adult	Middle	120	15	1	0	40	2	190	9	25	30	0.67	1
Cracks	Wollogg	Child	Middle	110	9	1	1	40	2	70	15	25	31	0.75	1
Mult-Gran	General	Adult	Bottom	100	15	1	2	90	2	220	6	25	40	1	1
Cracklin	Wollogg	Adult	Top	110	10	3	4	160	3	140	7	25	40	0.5	1
Grape-Nuts	Post	Adult	Top	110	17	0	3	90	3	170	3	25	53	0.25	1
Honey Nut	General	Child	Bottom	110	11.5	1	1.5	90	3	250	10	25	31	0.75	1
Mult-Gran	Wollogg	Adult	Top	140	21	2	3	130	3	220	7	25	41	0.67	1.33
Product-19	Wollogg	Adult	Top	100	20	0	1	45	3	320	3	100	42	1	1
Total Raisin	General	Adult	Top	140	15	1	4	230	3	190	14	100	29	1	1.5
Wheat Chex	Ralston	Adult	Bottom	100	17	1	3	115	3	230	3	25	50	0.67	1
Ortmeal	General	Adult	Top	130	13.5	2	1.5	120	3	170	10	25	30	0.5	1.25
Life	Quaker	Child	Middle	100	11	2	2	95	4	150	6	25	45	0.67	1
Whego	America	Adult	Middle	100	16	1	0	95	4	0	3	25	55	1	1
Quaker Oats	Quaker	Adult	Top	100	14	1	2	110	4	135	6	25	50	0.5	1
Wheat R	Ralston	Adult	top	150	16	3	3	170	4	150	11	25	34	1	1
Quaker Oatmeal	Quaker	Adult	Bottom	100	14	2	2.7	110	5	120	0	0	51	0.67	1
Chexes	General	Child	Bottom	110	17	2	2	105	6	290	1	25	51	1.25	1
Special K	Wollogg	Adult	Bottom	110	16	0	1	55	6	230	3	25	53	1	1

3. Determine if each of the following variables are quantitative or categorical.

- The number of milligrams of Aspirin given to heart attack patients.
- The various types of cars being sold at a used car lot.
- Determining if a person smokes marijuana or not.
- The number of bicycles sold at various bicycle stores in Seattle, WA.
- The types of birds observed in Florida.
- The number of grams of gold found in various streams across northern California.
- The various types of cardio classes offered at gyms across Los Angeles, CA.
- The number of cardio classes offered at gyms across Los Angeles, CA.
- The city a person lives in.
- The amount of money in peoples' bank accounts.
- The various zip codes from addresses at a post office.
- The drivers' license numbers from various taxi drivers.
- The number of taxis driven in New York City on various days of the week.

## Section 1B – Collecting Data

### Vocabulary



This material is from *Introduction to Data Analysis*, 2<sup>nd</sup> edition by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a “CC-BY” [Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by/4.0/) – 3/17/2021

Data: Information in all forms

Categorical Data: Data consisting of words describing people or objects. Numbers may sometimes be used in place of words.

Quantitative Data: Numerical measurement data, often including units, counts or averages.

Population: The collection of all people or objects you want to study.

Census: Collecting data from everyone in the population.

Sample: Collecting data from a small subgroup of the population.

Bias: When data does not reflect the population.

Random: When everyone in the population has a chance to be included in the sample.

Sample Size: The total number of people, animals or objects you collect data from.

Sampling Bias: Using a bad method to collect data like convenience or voluntary response. Not incorporating randomization in your sample.

One of the most important goals in data science is to learn about the world around us (populations). A population is the collection of all people or objects you want to study. It is very difficult to understand populations sometimes because data may be biased and not reflect the population very well. Bias can occur in many different ways, but certain ways people collect data have more bias than others do. Using a method for collecting data that increases bias is sometimes called “sampling bias”. It is important to be aware of various methods used to collect data, the good and the bad.

#### Method 1: Census

A census is the best way to collect data if it is possible. If our goal is to learn about the population, it makes sense to collect data from everyone in the population. There are ways for a census to be biased, but in terms of the collecting method, a census is the best. Unfortunately, it is almost impossible to collect a census if your population is large. Most statisticians and data scientists are only able to collect a sample, data collected from a small subgroup of the population.

#### Method 2: Simple Random Sample

If a statistician or data scientist cannot collect a census, the preferred method is to collect a random sample. A random sample is one where everyone in the population has a chance to be in the sample, so it tends to represent the population better than other non-random samples. It is nowhere near as good as a census, but as I said, a census is usually not possible.

We should probably start with discussing the word “random”. “Random” in data science and statistics is used very differently than the way people generally use the word. Selecting data randomly means that everyone in the population has a chance to be included in the sample data. You are not collecting data from the millions of people or objects in the population. That would be a census. You are still collecting a sample (small subgroup of the population), but everyone of the millions in the population has a chance to be included. Random samples are often difficult to set up.

Example: A person may walk into a store and say “I chose a person randomly to talk to”. They mean that they talked to whoever they ran into. This is not random in statistics. Not everyone of the millions in your population has a chance to be in that store and bump into the person.

A simple random sample is the most common type of random sample. In a simple random sample, individuals in the population are selected randomly. This can be a difficult process. The usual method is to assign everyone in a population a number and then use a random number generator in a computer program to pick random numbers.



This material is from *Introduction to Data Analysis*, 2<sup>nd</sup> edition by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a “CC-BY” [Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by/4.0/) – 3/17/2021

Computer programs have many built in randomization functions for this purpose. If you have a spreadsheet of the entire population, a computer can also randomly select individuals from the list. The key with a “simple random sample” is that you are selecting people or objects one at a time. Collecting data randomly and one at a time gives greater flexibility to your sample. Almost any grouping is possible with a simple random sample, so it tends to represent populations better than other samples.

There are many examples of a simple random sample. Many statistics companies use a random phone number generator that randomly gives phone numbers. They then call the phone numbers randomly chosen and try to get information from people that answer the phone. The U.S. government may have a computer randomly select social security numbers to select individuals for a sample. A company may have a computer randomly select employee ID numbers to select individuals for a sample.

### Method 3: Convenience Sample

People often find collecting a census or a simple random sample difficult, so they chose to collect data in whatever way seems easiest. A sample collected this way is often called a “convenience sample” and is popular with people not trained in statistics. A convenience sample usually has much more bias than a random sample and may not represent the population very well.

An example of a convenience sample is collecting data from your friends and family. This is fine if your population of interest is your friends and family, but will by no means represent a large population. Another example might be standing outside of a store or post office and collecting data from people that leave the store. Beginning statistics students may walk into a mall and collect data from whomever they bump into. They mistakenly think that these are random samples, but they are not. A random sample means everyone in the population has a chance to be included in the sample. Not everyone in the population has a chance to bump into you at a mall or come out of a store at 2:30 pm on a Tuesday afternoon. These are convenience samples and generally do not reflect the population very well.

### Method 4: Voluntary Response Sample

Some say that all surveys are bad, but that is not the case. A survey is just a form to collect data from people. When a company takes a census of all its employees, it may require all of the employees to fill out a survey. That is a census. As long as no other forms of bias creep into the data, a census will probably be a very good representation of the population. The point is that giving a survey is not the issue. The issue is whom you give the survey to and who is allowed to fill out the survey.

A voluntary response sample puts a survey out into the world and allow anyone to respond. The usual method used today is to put a survey on a website and allow anyone that comes across the survey to answer. The survey can also be mailed to every address in a given population. Again, those that fill it out self-select themselves to be in our data.

On the surface, a voluntary response sample may seem like a good way of collecting data. It usually gives a large amount of data. Does this really allow everyone in the population a chance to answer? It turns out the answer is no. Ask yourself the following question. When you are surfing the web and a survey pops up, do you fill it out? I have been asking my statistics classes that question for years and rarely have anyone that says that they do fill out surveys. The key problem is that only certain types of people will fill out a survey voluntarily. It may be a person who is bored and has nothing better to do. It is certainly not a person with three children, working a full time job and going to college full time. It may also be a person who is upset by or feels very passionate about the topic in the voluntary response survey. They are so upset by the lack of pay for teachers that they are willing to fill out a survey to tell you what they think. The point is that voluntary response surveys tend to over-sample people that are bored or upset and under-sample everyone else. For this reason, voluntary response samples can be very biased and may not represent the population very well.

Note about sample size:



This material is from *Introduction to Data Analysis*, 2<sup>nd</sup> edition by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a “CC-BY” [Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by/4.0/) – 3/17/2021

Students often ask me about the importance of how many people or objects you collect data from. This is called “sample size”. More data is usually better. A simple random sample of 250 people is better than a simple random sample of 50 people. Is a voluntary response sample of five thousand people better than a random sample of fifty people?” I would tell them that though sample size is important, method is important also. The voluntary response sample of five thousand would tend to over-represent people that are bored or upset about the topic. It does not represent typical people in the population. The random sample of fifty people, while a small sample size, at least does not have that bias.

## Summary

So let us summarize the various methods.

- An unbiased census is the best way to collect data to represent a population, because we are collecting data from everyone in the population. An unbiased census is generally better than a random sample.
  - If you cannot do a census, then use an unbiased random sample. A simple random sample is most common. The main thing is that if you are collecting a sample, randomization needs to be involved. Random means that everyone in the population has to have a chance to be included in the sample.
  - Voluntary response samples and convenience samples tend to be very biased and should be avoided if possible.
- 

## Practice Problems Section 1B



This material is from *Introduction to Data Analysis*, 2<sup>nd</sup> edition by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a “CC-BY” [Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by/4.0/) – 3/17/2021

Directions: For each of the following, identify the population of interest. Then identify the method used to collect the data (census, convenience, voluntary response, or simple random). Explain why you chose your answer. Was this a good or bad way to collect the data?

1. The admissions department at a college wants to see how many of their students would be in favor of using a new program to register for classes. They put a link on their website so that any students that want to try out the program can. The students can then take a survey and say how well they like the new system.

- a) What was the population of interest?
- b) What method was used to collect the data? Explain.
- c) Was this a good or bad way to collect the data?

2. Michelle, a teacher at a local high school, wants to see how many students at her high school will be attending community college. She gives the students in her one section of advanced placement U.S. History a questionnaire to fill out that asks where they will be attending college.

- a) What was the population of interest?
- b) What method was used to collect the data? Explain.
- c) Was this a good or bad way to collect the data?

3. Jamie is working at the Republican recruiting committee in her city. She is curious how many people that live in her city will vote for the Republican candidate in the next election. She uses a computer to randomly select phone numbers in her city. She then calls those phone numbers to ask people about their voting preferences.

- a) What was the population of interest?
- b) What method was used to collect the data? Explain.
- c) Was this a good or bad way to collect the data?

4. Micah is the CEO of large software development company. He wants to see if his employees have any ideas about areas of software development that the company should pursue. He has every single employee in his company fill out a questionnaire outlining his or her ideas. He gives the employees a stipend on their paycheck to pay them for their time it took to fill out the questionnaire.

- a) What was the population of interest?
- b) What method was used to collect the data? Explain.
- c) Was this a good or bad way to collect the data?

5. Tara wants to collect data on people living in Portland Oregon. She wants to know how many cups a coffee they drink per day. She went to a few supermarkets close to her house and asked people as they were leaving the store.

- a) What was the population of interest?
- b) What method was used to collect the data? Explain.
- c) Was this a good or bad way to collect the data?

6. Julius works for a company in Toronto, Canada that manufactures eyeglasses. He wants to know what styles of glasses people in Toronto prefer. He randomly selects phone numbers in Toronto and calls them to ask about glasses preference.

- a) What was the population of interest?
- b) What method was used to collect the data? Explain.
- c) Was this a good or bad way to collect the data?

7. Hugo works at a public library and wants to collect data on all of the people that come to the library. He looks up every single person in the library database and notes the number of books that he or she has checked out in the last six months.





- a) What was the population of interest?
- b) What method was used to collect the data? Explain.
- c) Was this a good or bad way to collect the data?

8. A company is designing a new type of smart phone. They want to know how much memory people prefer in their smart phones. They put a question up on several search engines and allow anyone to answer.

- a) What was the population of interest?
- b) What method was used to collect the data? Explain.
- c) Was this a good or bad way to collect the data?

9. A college wants to collect data on their students to see how often they use the various student services offered by the college. They randomly select 50 student ID numbers and collected data from all of the students chosen.

- a) What was the population of interest?
  - b) What method was used to collect the data? Explain.
  - c) Was this a good or bad way to collect the data?
- 

## Section 1C – Bias

### Vocabulary



This material is from *Introduction to Data Analysis*, 2<sup>nd</sup> edition by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a “CC-BY” [Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by/4.0/) – 3/17/2021

Data: Information in all forms

Categorical Data: Data consisting of words or numbers used in place of words.

Quantitative Data: Numerical measurement data.

Population: The collection of all people or objects you want to study.

Census: Collecting data from everyone in the population.

Sample: Collecting data from a small subgroup of the population.

Bias: When data does not reflect the population.

Random: When everyone in the population has a chance to be included in the sample.

Sample Size: The total number of people, animals or objects you collect data from.

Sampling Bias: Using a bad method to collect data like convenience or voluntary response. Not incorporating randomization in your sample.

The purpose of collecting data is to learn about the world around us, to learn about populations. The problem is that many people that collect data may not have had any training in Statistics or Data Science. The result is that many data sets collected do not reflect the population very well. When this happens, we say that the data is biased.

Many people think that if you collect a random sample or a census, it will guarantee that you will have an unbiased data set. This is not true. There are many types of bias and it is possible to have a census or a random sample that does not reflect the population very well. It is critical that we be aware of these other forms of bias and to try our best to make sure they are not incorporated into our data sets.

### Sampling Bias

In the last section, we said that the best way to collect data is a census. This means that we collected data from everyone in the population. If we cannot collect a census then we should try to collect a random sample or at least a sample that represents the population. We said that convenience samples or voluntary response samples are inherently biased and usually do not reflect populations very well. Using a bad data collecting method like convenience or voluntary response gives rise to sampling bias. When sampling bias occurs, it usually means the technique for collecting the data was poor.

### Question Bias

It has been said that there are lies, bad lies, and then there is statistics. There is some truth in this. People with specific agendas may twist data and statistical analysis to suit their purpose. One way to do this is question bias.

A question bias occurs when someone phrases a question in a specific way to force people to answer the way they want.

For example, suppose a politician wants to show that most people in her city agree with her policy on raising taxes to improve health care. She may collect a great simple random sample, but ask the question this way.

“Health care in our city is extremely bad. Hospitals and urgent cares are in bad need of renovation and need better supplies. The elderly need to know that we have not forgotten them. We need to improve the quality of care for our children. Will you support my policy for improving health care across our city?”

Phrasing the question this way, no one would guess that the real issue was whether to raise taxes. People, hearing this question, think about helping the children and elderly, not about taxes. When a large percentage of people answer that they support her plan, she now has data to support her agenda.



When you collect data, you want to ask questions in a neutral way that does not attempt to sway people in one direction or another. It also should not leave out key information like what the real question is. If the politician had simply asked people in the simple random sample if they would be in favor of raising taxes to improve health care, she likely would have gotten a much smaller percentage of people to agree.

Notice that in this example, the data was a simple random sample. This is a good data collection method, as methods go. However, the incorporation of a question bias into the data makes the data very bad. This simple random sample does not reflect the population at all. The data has been manipulated to support an agenda.

### Response Bias

Many topics are very difficult to get data on because people do not feel comfortable answering truthfully. If you ask people if they are addicted to alcohol or drugs, they are likely to deny it even if they do struggle with substance addiction. People may lie about their age, weight, or salary. When a large percentage of people in your data lie, you have a response bias in your data.

Suppose a church wants to collect data on how many hours per week their congregation spends helping the homeless. They decide to have every person in their congregation fill out a survey listing how many hours per week they help the homeless. Remember a census is usually the best way to collect data about a population, but this census has a problem. It is a topic that people are likely to lie about. People may put a higher number of hours on the survey than they really do so that they will not look bad to the church leaders. The average number of hours calculated from this data will likely be larger than the population average number of hours. Even though this is a census, it probably does not reflect the population very well.

When dealing with topics that people are likely to lie about, the data scientist needs to have a plan to deal with the response bias. Instead of asking people their weights, maybe they weigh them on a scale. Instead of asking people about their salary, maybe they look at paycheck stubs. Instead of asking people about substance abuse, they may collect data from agencies that support people with addiction.

### Deliberate Bias

We have stated already that people may misuse statistics and data in order to support their agenda. Deliberate bias is another example of this. Deliberate bias can take on a variety of forms. It could be someone deliberately leaving out groups from the data. The most common is collecting data and then leaving out the data of people that disagreed with you. It can also be deliberately lying about the results of the data report. Maybe the data makes your restaurant or hospital or school look bad, so people just falsify their records and deliberately lie about the results of the study. The data may be census or a random sample but the conclusions have been falsified and the data distorted.

Deliberate bias is a major problem in statistics. It is also a good reason to have an independent statistics company collect the data and do the analysis. Use a statistics company that is not tied to the government, business, hospital, restaurant or politician in question. An independent statistics company is less likely to lie about the results or to falsify the data, though it is naive to think that it never happens.

I tend to be suspicious about internal statistics reports that come out where the company, government or politician refuses to share the data. We are supposed to take their word for it and agree with the findings. There are good reasons why companies do not share data, but I always wonder if they are they afraid that someone analyzing that data would come to a very different conclusion?

There is large worldwide discussion of ethics for people that work in the fields of statistics or data science. Statistical analysis is a powerful tool and is a vital discipline to understand and improve the world around us, but falsifying records or manipulating data should never be an option. It is not only unethical, but also makes people question the integrity of our science.

Sometimes specific groups in the population may not be represented very well in the data. This also falls under the umbrella of deliberate bias. For example, suppose a person may wish to collect data on adults living in a city. However, they only collected data from people living in the wealthier areas of that city. It may not have been done deliberately. It could just be that the person collecting the data did not think about certain groups in the population



This material is from *Introduction to Data Analysis*, 2<sup>nd</sup> edition by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a “CC-BY” [Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by/4.0/) – 3/17/2021

that are not being represented. In large cities, the homeless are often difficult to get data on. A person collecting data has to have a plan for getting data that will represent all the groups in their population, including the homeless.

### Non-response Bias

Non-response bias is becoming a huge problem for all people that collect data. A computer may randomly select people to collect data from, but more often than not, the person does not want to participate. They may fear identity theft or are just too busy to participate. It is a huge problem. We need data. We need to understand the world around us, but it now becoming increasing difficult to get unbiased data. Many people that collect data report that sometimes only one in every five randomly selected people will participate and give data. The problem of non-response bias continues to get worse. This makes us consider what type of person gives data and if that person is truly reflective of all people in the population.

To combat the problem of non-response bias, many people that collect data offer a reward system for people that will participate and give data. This may help a little, but then offering a reward may incorporate its own bias into the data.

### Summary

There are many reasons why data may not reflect a population. It is a mistake to think that a random sample or a census will always be devoid of bias. It is increasingly important to be aware of possible sources of bias and to strive to keep them out of our data as much as possible. The goal of data collecting is to collect unbiased data that reflects the population. Always phrase questions in a neutral way that avoids question bias. Have a plan for collecting data about topics where people are likely to lie. We have to have a good plan on how we will collect data. It should be a census or a random sample, but we should also think about groups that may not be represented. We need to avoid deliberate bias and never falsify reports or distort data to support someone's agenda.

---

### Practice Problems Section 1C

1. Define each of the following and give an example of each.



This material is from *Introduction to Data Analysis*, 2<sup>nd</sup> edition by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a “CC-BY” [Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by/4.0/) – 3/17/2021

- a) Population
- b) Census
- c) Sample
- d) Bias
- e) Question Bias
- f) Response Bias
- g) Sampling Bias
- h) Deliberate Bias
- i) Non-response Bias

*Directions for #2-8: For each of the following scenarios, describe the population of interest and all of the types of bias that the data may have (Question, Response, Sampling, Deliberate or Non-response). There may be more than one type of bias involved. Explain your answers.*

2. We are interested in finding what percent of people in the U.S. agree or disagree with vaccinating children. To figure this out, we randomly selected 350 people in the U.S. and asked them the following question: "In order to save children from devastating diseases, do you agree that all children should be vaccinated?"
  3. We are interested in finding out what percent of Americans use Cocaine. We randomly chose 400 Americans and asked them if they use Cocaine or not.
  4. What is the average age of college students in Canada? Since my cousin lives in Canada, I asked him to drive to two colleges near his house and ask people he bumps into what their age is.
  5. Julie is interested in calculating the yearly income of adults in Palmdale. She drives around Palmdale, stops at certain streets, and then asks people that live on that street what their yearly income is? She skips streets that look "sketchy" as she is worried about her safety.
  6. A college wants to collect data on their students to see how often they use the health office for mental health counseling. They took a simple random sample of college students and asked the following question. "It is very important for all college students to have mental health support. College students report having depression, anxiety and high stress levels. The college offers free mental health counseling at the health office. Have you taken advantage of these mental health services?"
  7. A pharmaceutical company took random samples of their pills to check that the pill has the correct type and amount of medicine. They noticed that several of their pills did not have the correct amount of medicine, but decided to delete this data.
  8. An auto manufacturer wants to collect data on the type and number of mechanical problems in their cars. They decide to keep data only on all cars brought to their dealerships nationwide.
  9. A computer algorithm called COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) was created by Northpointe, Inc. The algorithm assesses whether defendants have a higher or lower risk of repeating crimes. Northpointe, Inc. did the validation study to show that the algorithm works.
- 

## Section 1D – Experimental Design



This material is from *Introduction to Data Analysis*, 2<sup>nd</sup> edition by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a "CC-BY" [Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by/4.0/) – 3/17/2021

## Vocabulary

**Explanatory Variable:** The independent or treatment variable. In an experiment, this is the variable causes the effect.

**Response Variable:** The dependent variable. In an experiment this the variable that measures the effect.

**Confounding Variables** (or lurking variables): Other variables that might influence the response variable other than the explanatory variable being studied.

**Experimental Design:** A scientific method for controlling confounding variables and proving cause and effect.

**Random assignment:** Take a group of people or objects and randomly split them into two or more groups. This creates similar groups and helps control confounding variables.

**Placebo:** A fake medicine or fake treatment used to control the placebo effect.

**Placebo Effect:** The capacity of the human brain to manifest physical responses based on the person believing something is true. A placebo (fake medicine) is often given to control the placebo effect.

In statistics, we often want to determine if there is a relationship or association between two variables. We also may want to measure the strength of the relationship. For example, we may want to know if there is a relationship between blood pressure and heart rate. We may want to see if living in tropical climates is associated with having nut allergies.

In order to show that two variables are related or associated we use an observational study. We would collect data and use statistical methods to analyze and measure the strength of the relationship. However, showing that two variables are related does not prove that one causes the other.

### **Association ≠ Causation!!!!**

Why?

Let us suppose that we have shown that there is a strong relationship between drinking alcohol and getting into a car accident. This tells us that alcohol consumption is an important factor to be considered when studying car accidents. However, this does not prove that drinking alcohol causes car accidents. Many factors go into having a car accident besides how much alcohol they consume. Can you name a few?

Other factors that may influence having a car accident besides alcohol: age of driver, experience of the driver, condition of the car, traffic, road conditions, weather, other drivers, distractions (like texting, eating or changing a radio station), using drugs, ...

These are called “confounding variables”. Confounding variables are factors that might influence your response variable other than the explanatory variable you are studying. In this case, factors that might influence having a car accident other than how much alcohol the driver consumed. Some statistics books call these “confounding variables” or “lurking variables”.

Note: The explanatory variable (alcohol consumption) is not a confounding variable. Alcohol is the explanatory variable we were studying. Confounding variables are factors other than alcohol that might influence the response (car accident).

Here is the point. If many variables were involved in having a car accident, it would be wrong to say that the alcohol was solely responsible for the car accident. Alcohol is just one of many factors involved. We have shown that drinking alcohol is related but we have not proven cause and effect. To prove cause and effect we need to deal with the confounding variables.

## Experimental Design



This material is from *Introduction to Data Analysis*, 2<sup>nd</sup> edition by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a “CC-BY” [Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by/4.0/) – 3/17/2021

So how do we prove cause and effect? It is difficult. You would need to prove that each confounding variables is not involved and so it is only the explanatory variable that is causing the response. The key is controlling the confounding variables. Thankfully, scientists have put a great deal of thought into this process of controlling confounding variables and proving cause and effect. We call this process “experimental design”.

Experimental design is a scientific method for controlling confounding variables and proving cause and effect. A key component to experimental design is the creation of similar groups through random assignment.

To control confounding variables, we will need to create two or more groups of people or objects that are very alike. One way to do this is by “random assignment”. Random assignment is a process where you take a group of people or objects and randomly split them into two or more groups. The randomly assigned groups tend to be very similar. If we do not think the groups are similar enough, we can use techniques like blocking or direct control to make the groups even more alike.

Another way to make alike groups is to use the same group of people twice. Think about it. The two groups would be perfectly alike. They would have the same ages, same amount of stress, same genetics, same blood pressures and the same jobs.

### Example

Let us look at the previous example. How do we prove that drinking alcohol does cause car accidents?

Explanatory (treatment) Variable: Drinking alcohol or not

Response Variable (what we will measure): Did the person get into a car accident or not?

So how do we set up an experiment to prove that drinking alcohol causes car accidents? The first thing is to list out your possible confounding variables.

Possible Confounding Variables: age of driver, experience of the driver, condition of the car, traffic, road conditions, weather, other drivers, distractions (like passengers, texting, eating or changing a radio station), other drugs, gender, race, genetics, reflexes.

To control the confounding variables, we need to create two groups of people. The two groups should be the same (or at least as similar as possible) in all areas that the confounding variables address. Therefore, the groups should have similar ages, similar driving experience, similar cars and car condition, similar road conditions and similar distractions, similar genders, similar race and ethnicity, similar genetics and reflexes.

There are two ways to go about this. Let us suppose we have a group of 80 adult paid volunteers to conduct this experiment. One option would be to randomly put the volunteers into two groups and try to make the groups as similar as possible. A better option in this case would be to use the same people twice.

We had the people in the experiment drive an obstacle course sober. They must have no alcohol or other drugs in their system. They all used the same car on the same track with the same weather. The course was designed with cones and we will monitor how many cones the people hit. They all were not allowed to have any other person in the car. There was no other distractions as radios and phones were not allowed. We will monitor how many car accidents they had by checking how many cones they hit.

Now we will have all the people drink a certain amount of alcohol and then drive the course again. It is important to see that the alcohol (treatment) group was made up of exactly the same people as the sober (control) group. The response variable we measured was the number of cones they hit.

Conclusion



This material is from *Introduction to Data Analysis*, 2<sup>nd</sup> edition by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a “CC-BY” [Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by/4.0/) – 3/17/2021

The results found that the alcohol group hit significantly more cones (significantly more car accidents) than the sober group. We have now proven that drinking alcohol causes car accidents.

Think about it. It cannot be the ages of the drivers or driving experience. The two groups had the exact same ages and the exact same driving experience. It cannot be gender, race, genetics, or reflexes. The two groups had the exact same genders, race, genetics, and reflexes. It cannot be drugs or other distractions like phones or radios. Neither group had drugs or any other distractions. If you notice, every one of the confounding variables is the same in the two groups. The only difference was that one group had alcohol and the other did not. Therefore, the only reason why the alcohol group had significantly more accidents is the alcohol. The experiment has proven that drinking alcohol causes car accidents.

Note: It is easy to confuse the two variables in an experiment with the two groups. They are not the same thing.

In this case, the explanatory variable is having alcohol or not. The response variable is the number of cones (accidents) the drivers had. The two groups are decided by those that have explanatory variable (alcohol) and those that do not. In this case, the two groups are the exact same people measured twice.

We usually call the group that has the explanatory variable the “treatment group” and the group that does not have the explanatory variable the “control group”.

### Example 2

When a pharmaceutical company needs to prove that a medicine works, they must use experimental design. In the United States, pharmaceutical companies have to prove to the Food and Drug Administration (FDA) that their medicine has the effect it is supposed to and is relatively safe with few side effects.

Suppose a company has a new blood pressure medicine on the market and needs to prove to the FDA that taking it does decrease a person’s blood pressure. The company needs to prove cause and effect.

If we have to prove cause and effect, we need an experiment. The first step is to think about the possible confounding variables. What are some reasons why a person’s blood pressure might decrease other than taking this new medicine?

Possible Confounding Variables? Stress, Diet, Exercise, Genetics, Age, Gender, Race, Genetics, taking other medicines ...

To set up the experiment we need to create two groups of people that are similar in these areas. We start with a group of volunteers with high blood pressure that want to try out this new medicine. We randomly assign the people into two groups. Amazingly when scientists randomly assign people into two groups, the groups tend to be a lot alike. The two groups would have similar numbers of people in each race, similar number of males and females, similar numbers of stressed out people, similar numbers of people that exercise a lot or do not exercise. The people running the experiment can also exercise direct control and intentionally assign people to certain groups to make the groups even more alike.

### Human Brain (placebo effect)

There is a problem with our experiment. If a person believes something is true, their brain can tell the body to manifest physical responses. We call this the “placebo effect”. Think of it this way. The group that thinks they are getting blood pressure medicine will not be as stressed out about it and their blood pressure may decrease slightly because of that belief. Similarly, the group that thinks they are not getting blood pressure medicine will be more stressed and worried and their blood pressure may increase because of that belief. In a sense, the human brain is a confounding variable that we need to control.

### Placebo (fake medicine)

To control the placebo effect as a confounding variable, we need the groups to believe the same thing. One group cannot think they are getting medicine, and the other group cannot believe they are not getting medicine. So we introduce a placebo or fake medicine. The treatment group gets the real blood pressure medicine and the control



This material is from *Introduction to Data Analysis*, 2<sup>nd</sup> edition by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a “CC-BY” [Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by/4.0/) – 3/17/2021



group gets a fake medicine (placebo). No one in the experiment knows if he or she will be receiving real medicine or a placebo. Some may ask, “Won’t that make them more stressed and increase their blood pressure?” Yes. The key is that the two groups will be equally stressed and believe the same thing. That way we control the placebo effect.

For this to work, the people in the experiment cannot know if they are getting the medicine or a placebo. This is called “single blind”. When scientists first started using placebos, they were shocked to find that the people in the experiments somehow knew if it was a placebo. This defeated the whole purpose. It turned out they could tell by the body language of the person giving the medicine. The person giving the medicine tended to act differently if they were giving the real medicine versus a placebo. So the standard for an experiment about medicines is to use a “double blind” approach. A double blind experiment means that neither the people in the experiment, nor the people giving the medicine, know if it is a placebo or not. Someone knows though. The scientists keep very careful track of who receives a placebo and who receives the medicine. The person directly giving the medicine or placebo cannot know if it is a placebo or not.

Double blind works well. The people in the experiment no longer know if they are receiving a placebo or the real medicine. The experimental design has controlled the placebo effect.

### Conclusion

Since we have controlled all of the confounding variables, the experiment has the possibility of proving cause and effect. We still need to see the blood pressures of both groups and make a conclusion. If the treatment group had a significantly lower average blood pressure than the control group, this would prove that taking the medicine does cause a person to have lower blood pressure. If the treatment group and control group have relatively the same average blood pressure, then we may conclude that the medicine is not effective in lowering blood pressure. This would be bad news for the pharmaceutical company. Deciding if one group is significantly higher than another can be very difficult. We will study confidence intervals, test statistics and P-value in later chapters to address this.

### Summary

Use an experiment to control confounding variables and prove cause and effect. The groups in the experiment should be the same people either measured multiple times or separated by random assignment. The main idea is that the groups should be very similar in all areas that involve confounding variables. Experiments with medicines should be double blind with a placebo to control the placebo effect.

Use an observational study to see if there is a relationship (association) between two things. Remember observational studies do not control confounding variables, so cannot prove cause and effect.

How can I tell if a study is an experiment or not? Generally, look for random assignment. An experiment usually does not have a random sample of people from the population. The people in the experiment are usually volunteer. The volunteers are then randomly assigned into two or more groups. Random assignment means that they are not trying to apply something to the population, but instead are trying to use experimental design in order to prove cause and effect. If a study takes a random sample from the population, but does not randomly assign, it is probably just an observational study and cannot prove cause and effect.

Note: It should be noted that there are more complex forms of experiments than the types listed in this section. It may not be possible to randomly assign people into two groups. In that case, the scientist need to prove that each confounding variable is not involved. That is a more complex case that you may see in more advanced statistics classes.

---

## Practice Problems Section 1D

(#1-10) Define the following terms.



This material is from *Introduction to Data Analysis*, 2<sup>nd</sup> edition by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a “CC-BY” [Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by/4.0/) – 3/17/2021

1. Observational Study
2. Experiment
3. Explanatory Variable
4. Response Variable
5. Confounding Variables
6. Random Assignment
7. Placebo
8. Placebo Effect
9. Single Blind
10. Double Blind

(#11-12) Directions: Answer the following questions about the experiments described.

11. College students in the United States have long claimed that listening to music while studying causes them to retain information at a higher rate. We want to prove that this is not true. Listening to music while studying does not cause a person to retain information at a higher rate. We took a group of volunteer college students and randomly put them into three groups. The people in each group had to memorize the same information. They were then ranked as high retention or low retention. One group had to listen to their favorite music, another group had to listen to a music they hated, and the third group had no music at all. The volume of music was the same for all of the people.

- a) Was random assignment used in the experiment?
- b) List as many confounding variables as you can for this experiment?
- c) What is the explanatory variable (cause) and the response variable (effect) in this experiment?
- d) Describe the treatment group and the control group. Were they alike in the confounding variables?
- e) Describe how the confounding variables were controlled.
- f) The results of the experiment were that the hated music group and the liked music group did about the same. Both music groups did much worse than the no music group. The no music group had significantly better retention than either of the music groups. Does this prove that listening to music does not cause a person to memorize information better? Why or why not?

12. Dramamine is a common medication used in preventing and treating nausea, vomiting and dizziness caused by motion sickness. This medication has become a staple for thousands of people who travel by boat, car or plane. We need to prove that Dramamine is effective in preventing and treating the symptoms of motion sickness. Volunteers



This material is from *Introduction to Data Analysis*, 2<sup>nd</sup> edition by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a “CC-BY” [Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by/4.0/) – 3/17/2021

were randomly assigned into two groups. One group received Dramamine and the other received a placebo. The amount of motion was the same for all of the people. They were then asked to rank their motion sickness on a scale of 1 to 10.

- a) Was random assignment used in the experiment?
- b) List as many confounding variables as you can for this experiment?
- c) What is the explanatory variable (cause) and the response variable (effect) in this experiment?
- d) Describe the treatment group and the control group. Were they alike in the confounding variables?
- e) Describe how the confounding variables were controlled.
- f) If the Dramamine group has significantly less motion sickness than the placebo group, does this prove that taking Dramamine causes a person to have less motion sickness? Why or why not?

13. An experiment was done on labor market racial discrimination. Statisticians created fictitious resumes to help-wanted ads in Boston and Chicago newspapers. Resumes were randomly assigned to either have a very African American sounding name or a very white sounding name. The results that the percentage of callbacks for resumes with white names was significantly higher than for African American names.

- a) Was random assignment used in the experiment?
  - b) List as many confounding variables as you can for this experiment?
  - c) What is the explanatory variable (cause) and the response variable (effect) in this experiment?
  - d) Describe the treatment group and the control group. Were they alike in the confounding variables?
  - e) Describe how the confounding variables were controlled.
  - f) Does this experiment prove that there is racial discrimination against African Americans when applying for a job in Boston and Chicago? Why or why not?
- 

## Chapter 1 Review Sheet

### Key Vocabulary Terms



This material is from *Introduction to Data Analysis*, 2<sup>nd</sup> edition by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a “CC-BY” [Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by/4.0/) – 3/17/2021

Data: Information in all forms.

Categorical data: Also called qualitative data. Data in the form of labels that tell us something about the people or objects in the data set. For example, the country they live in, occupation, or type of pet.

Quantitative data: Data in the form of numbers that measure or count something. They usually have units and taking an average makes sense. For example, height, weight, salary, or the number of pets a person has.

Population: The collection of all people or objects to be studied.

Census: Collecting data from everyone in a population.

Sample: Collecting data from a small subgroup of the population.

Random: When everyone in the population has a chance to be included in the sample.

Simple Random Sample: Sample data in which individuals are selected randomly. This method tends to minimize sampling bias and is generally considered a good way to collect data.

Convenience Sample: Sample data that is collected in a way that is easy or convenient. This method tends to have a significant amount of sampling bias and is generally considered a bad way to collect data.

Voluntary Response Sample: Sample data that is collected by putting a survey out into the world and allowing anyone to fill it out. This method tends to have a significant amount of sampling bias and is generally considered a bad way to collect data.

Bias: When data does not represent the population.

Sampling Bias: A type of bias that results from collecting data without using a census or random sample. The method of collecting is flawed. For example, using convenience or voluntary response method to collect the data. We can minimize this bias by collecting the data with a census or random sample.

Question Bias: A type of bias that results when someone phrases the question or gives extra information with the goal of tricking the person into answering a certain way. We can minimize this bias by phrasing our questions in a neutral way and not attempt to sway the person giving data.

Response Bias: A type of bias that results when people giving the data do not answer truthfully or accurately. To minimize this bias, we should collect the data anonymously and assure the person giving the data that the data will be used for scientific purposes and will not be released.

Non-response Bias: A type of bias that results when people refuse to participate or give data. To minimize non-response bias, you may give an incentive like a gift card to encourage people to give data.

Deliberate Bias: A type of bias that results when the people collecting the data falsify the reports, delete data, or decide to not collect data from certain groups in the population. To minimize deliberate bias, the people collecting and analyzing the data need to have good ethics. They should not falsify reports, delete data or leave out groups from the population.

Experimental Design: A scientific method for controlling confounding variables and proving cause and effect.

Observational Study: Collecting data without controlling confounding variables. This type of data cannot prove cause and effect.

Explanatory Variable: The independent or treatment variable. In a cause and effect experiment, this is the cause variable.

Response Variable: The dependent variable. In a cause and effect experiment, this the variable that measures the effect.

Treatment Group: The group of people or objects that has the explanatory variable. In an experiment involving medicine, this would be the group that receives the medicine.



Control Group: The group of people or objects that is used to compare and does not have the explanatory variable. In an experiment involving medicine, this would be the group that receives the placebo.

Confounding Variables: Also called lurking variables. Other variables that might influence the response variable other than the explanatory variable being studied.

Random assignment: A process for creating similar groups where you take a group of people or objects and randomly split them into two or more groups.

Placebo Effect: The capacity of the human brain to manifest physical responses based on the person believing something is true.

Placebo: A fake medicine or fake treatment used to control the placebo effect.

---

## Chapter 1 Review Problems

1. Tell if the following data is categorical or quantitative and explain why.

- a) The types of cars in the different parking lots.
- b) The average number of hours spent practicing ping-pong.
- c) Areas in North Dakota that have wild mustangs.
- d) Each person is asked if he or she wear glasses, contacts, neither, or both.
- e) The average speed of racecars at the Indianapolis 500.
- f) Exam scores for various students on a history exam.

2. Jim wants to know how much money the average working COC student makes. Describe how Jim could use each of the following techniques to collect data. For each technique, will there be a significant amount of sampling bias or not too much sampling bias?

- a) Systematic
- b) Voluntary Response
- c) Random Sample
- d) Convenience Sample
- e) Cluster Sample
- f) Stratified Sample
- g) Simple Random Sample
- h) Census

3. Define the following key terms and give an example of each.



This material is from *Introduction to Data Analysis*, 2<sup>nd</sup> edition by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a “CC-BY” [Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by/4.0/) – 3/17/2021

- a) Population
- b) Census
- c) Sample
- d) Random
- e) Bias
- f) Statistic

4. Describe and give an example of each of the following types of bias. Also state how a person collecting and analyzing data, can avoid these biases.

- a) Sampling Bias
- b) Question Bias
- c) Response Bias
- d) Deliberate Bias
- e) Non-Response Bias

5. Rachael needs to do an experiment that will show that wearing nicotine patches cause a person to stop smoking. Set up the experiment for Rachael. What is the explanatory variable? What is the response variable? Write a description of the experiment and include the following. What are some confounding variables that she will need to control? How can Rachael control the confounding variables? Include a description of how Rachael use a double blind placebo to control the placebo effect. Describe the treatment group and the control group in the experiment.

6. Compare and contrast the similarities and differences between an experiment and an observational study. How can we tell if we should use an experiment or an observational study?

---



## Chapter 2: Categorical Data Analysis

### Vocabulary

Data: Information in all forms.

Categorical Data: Data consisting of words describing people or objects. Numbers may sometimes be used in place of words.

Sample Size: The total number of people, animals or objects you collect data from.

**Introduction**: In our last chapter, we learned that data is information in all forms. We also learned that there were two types of data, categorical and quantitative. In this chapter we will look at some of the basic principles for analyzing categorical data. Categorical data consists of words describing people, animals or objects. Sometimes numbers may be used in place of words like using 1 for January, 2 for February and 3 for March. To analyze categorical data we need to look at the amount (frequency) of people or objects that have a certain description, the total number of people or objects in the data (sample size), percentages, and decimal proportions.

### Note about Terminology:

*Percentages are a vital link to understanding categorical data. Most students think of percentages as a calculation of probability, like the probability of drawing an ace from a deck of cards. In statistics, we want to know the proportion of people or objects that have a certain characteristic in a data set. I find that if I ask my class to calculate a probability, they seem to understand the idea, but if I ask what is the proportion of people that want to purchase a particular car, they do not understand. Most students think of solving an equation when they hear the term “proportion”. In statistics, a proportion is an amount (frequency) divided by the total (sample size) or a percentage divided by 100. Do not think of proportions as an equation you need to solve.*

*Though you can think of percentages and proportions as calculating a probability, we will focus on the more common statistics terminology of “proportion”. Also, remember that though decimal proportions and percentages are equivalent, they are not the same thing. If a computer program asks for the sample proportion, it will say “error” if you put the percentage.*

*Decimal proportion = amount / total (or a percentage divided by 100)*

*Percentage = decimal proportion x 100%*

---



## Section 2A – Proportions and Percentages

### Vocabulary

Data: Information in all forms

Categorical Data: Data consisting of words describing people or objects. Numbers may sometimes be used in place of words.

Sample Size: The total number of people, animals or objects you collect data from.

Percentage (%): An amount out of 100.

Proportion: The decimal equivalent of a percentage.

To analyze categorical data, we focus on exploring various types of percentages and compare them. In statistics, the decimal equivalent to a percentage is often called a “proportion”.

### Calculating a decimal proportion from Categorical data

To find a decimal proportion you will need to find the amount divided by the total.

$$\text{Decimal Proportion} = \frac{\text{Amount}}{\text{Total}}$$

Counting how many people share a certain characteristic or even a total number of cars in a data set can take a long time in a big data set, however technology can help. Statistics software can count much quicker and easily than we can. In this section, we will assume we know the amount and the total.

Suppose a health clinic has seen 326 people in the last month and 41 of them had the flu. If we were analyzing their data, the first thing we would like to do is find what proportion of the patients have the flu. It is not a difficult calculation and can be done with a small calculator.

$$\text{Decimal Proportion} = \frac{\text{Amount}}{\text{Total}} = \frac{41}{326} = 0.12576687$$

Should we round the answer? Proportions and Percentages are usually rounded to the three significant figures. Proportions are usually rounded to the thousandths place (3<sup>rd</sup> place to the right of the decimal).

Let us review rounding. We want to round the above answer to the thousandths place, which is the “5”. Always look at the number to the right of the place you are rounding to. If the number to the right is 5-9, round up (add 1 to the place value). If the number is 0-4, round down (leave the place value alone). After rounding cut off the rest of the decimals.

Therefore, in the previous answer we want to round to the thousandths place (5). The number to the right of the 5 is a 7. So should we round up or down? If you said round up, you are correct. Therefore, we will add 1 to the place value and the 5 becomes a 6. Now we cut off the rest of the decimal and our approximate answer is 0.126.

$$\text{Decimal Proportion} = \frac{\text{Amount}}{\text{Total}} = \frac{41}{326} = 0.12576687 \approx 0.126$$

Decimal proportions are vital in the analysis of categorical data, but many people have trouble understanding the implications of a decimal proportion like 0.126. That is why we often convert the proportion into a percentage.





### Convert a decimal proportion into a percentage

To convert a decimal proportion into a percentage, multiply by 100 and put on the “%” symbol. Think of it like taking 100% of the decimal proportion. When you multiply by 100, the decimal moves two places to the right. Some people prefer to move the decimal, but I find students make fewer errors when they just multiply by 100 with their calculator.

$$\text{Percentage} = \text{Decimal Proportion} \times 100\%$$

Look at our previous example of the number of cases of the flu at a health clinic. We used the amount and total to calculate the decimal proportion.

$$\text{Decimal Proportion} = \frac{\text{Amount}}{\text{Total}} = \frac{41}{326} = 0.12576687 \approx 0.126$$

So what percentage of the patients had the flu? All we need to do is multiply the decimal proportion 0.126 by 100% to get the percentage equivalent.

$$\text{Percentage} = \text{Decimal Proportion} \times 100\% = 0.126 \times 100\% = 12.6\%$$

So 12.6% of the patients at the health clinic were seen for the flu. This can be alarming information to the health clinic if that is an unusually high percentage.

Notice that the percentage still has three significant figures, but is rounded to the tenths place (one place to the right of the decimal). Rounding to the tenth of a percent is a common place to round percentages in statistics.

If you want to calculate the percentage directly from the categorical data, here is another formula you may use.

$$\text{Percentage} = \frac{\text{Amount}}{\text{Total}} \times 100\%$$

### Convert a Percentage into a Proportion

The word “percent” mean “per 100” or “out of 100”. So the “%” sign means out of 100 or divide by 100.

To convert a percentage into proportion: Remove the % symbol and divide by 100.  
(Or move the decimal point two places to the left)

#### Example 1

Convert 29.5% into a decimal proportion.

All we need to do is remove the % symbol and divide by 100.

$$29.5\% = 29.5 / 100 = 0.295$$

#### Example 2

Convert 0.97% into a decimal proportion. (This is less than 1%)

All we need to do is remove the % symbol and divide by 100.

$$0.97\% = 0.97 / 100 = 0.0097$$

*Note: Some students prefer to move the decimal point two places to the left. This is fine as well, though I find students make more mistakes with decimal point moving than with dividing a number by 100 with their calculator. Look at this example.*



### Example 3

Convert 5% into a decimal proportion.

Many students do not know where to move the decimal because there is no decimal shown. (They need to remember that 5% is the same as 5.0%)

A better way is to remove the % symbol and divide by 100.

$$5\% = 5 / 100 = 0.05$$

### Important Note: Fraction, Proportion and Percentage

There are three ways to describe categorical data: fraction, decimal, and percentage. Notice for the flu data example above, we have the three ways of describing the data: the fraction  $41/326$ , the decimal proportion 0.126, and the percentage 12.6%. All of them are equivalent. It is important to be comfortable with fractions, decimal proportions and percentages when describing categorical data. They are a foundation for more advanced categorical analysis later on.

Vocabulary to Remember:

Percentage (%): An amount out of 100.

- To calculate a percentage, multiply the proportion by 100 and adding the “%” symbol.

Proportion: The decimal equivalent of a percentage. There are two ways of calculating a proportion.

- To calculate the proportion from categorical data: Amount (frequency) divided by the total (sample size).
  - To calculate the proportion from a percentage: Divide the percentage by 100 and removing the “%” symbol.
- 



## Problem Set Section 2A

*Directions for #1-10: Convert the following proportions into percentages by multiplying the proportion by 100 and putting on the “%” sign. Do NOT round your answers.*

$$\text{Percentage} = \text{Proportion} \times 100\%$$

1. 0.039
2. 0.883
3. 0.0061
4. 0.092
5. 0.217
6. 0.0038
7. 0.651
8. 0.0705
9. 0.00014
10. 0.7005

*Directions for #11-20: Convert the following percentages into proportions by removing the “%” sign and dividing by 100. Do NOT round your answers.*

$$\text{Proportion} = \text{Percentage} \div 100$$

11. 58%
12. 92.6%
13. 8.104%
14. 0.772%
15. 3.19%
16. 8%
17. 62.5%
18. 3.52%
19. 0.044%
20. 3%



*Directions for #21-30: Round the following proportions to the thousandths place. There should be three numbers to the right of the decimal point in your proportion.*

- 21. 0.35419
- 22. 0.02581
- 23. 0.003527
- 24. 0.026114
- 25. 0.19963

*Directions for #26-30: Round the following percentages to the tenths place. There should be one numbers to the right of the decimal point in your percentage.*

- 26. 5.671%
- 27. 12.3499%
- 28. 73.955%
- 29. 2.732%
- 30. 0.287%

*Directions for #31-34: Use the amount and total to calculate a proportion. Round the proportion to the thousandths place (3 numbers to the right of the decimal). Then convert the rounded proportion into a percentage. Your percentage should have one number to the right of the decimal. Your answers should show the fraction, proportion and percentage.*

$$\text{Decimal Proportion} = \frac{\text{Amount}}{\text{Total}}$$

To convert proportion into percentage, multiply by 100 and put on the “%” sign.

- 31. In the 2015 National School Climate Survey by GLSEN, a total of 10,528 LGBTQ students between the ages of 13 and 21 years old were asked a series of questions. Over the past year at school, 6,064 of the LGBTQ students said they feel unsafe at school because of their sexual orientation. Calculate the proportion and percentage of LGBTQ students that feel unsafe.
- 32. In the 2015 National School Climate Survey by GLSEN, a total of 10,528 LGBTQ students between the ages of 13 and 21 years old were asked a series of questions. Over the past year at school, 8,970 of the LGBTQ students said they were verbally harassed (called names or threatened) at school based on a personal characteristic, sexual orientation, or gender expression. Calculate the proportion and percentage of LGBTQ students that experienced verbal harassment.
- 33. In the 2015 National School Climate Survey by GLSEN, a total of 10,528 LGBTQ students between the ages of 13 and 21 years old were asked a series of questions. Over the past year at school, 5,117 of the LGBTQ students said they experienced electronic harassed (cyberbullying) via text messages or postings on social media. Calculate the proportion and percentage of LGBTQ students that experienced cyberbullying.
- 34. In the 2015 National School Climate Survey by GLSEN, a total of 10,528 LGBTQ students between the ages of 13 and 21 years old were asked a series of questions. Over the past year at school, 1,369 of the LGBTQ students were physically assaulted (punched, kicked or injured with a weapon). Calculate the proportion and percentage of LGBTQ students that were physically assaulted.



*Directions for #35-37: Convert the following percentages into a proportion by dividing by 100 and removing the percent symbol %.*

35. In the 2015 National School Climate Survey by GLSEN, LGBTQ students between the ages of 13 and 21 years old were asked a series of questions. 57.6% of LGBTQ students who were harassed or assaulted in school did not report the incident to school staff, most commonly because they doubted that effective intervention would occur or the situation could become worse if reported.

36. In the 2015 National School Climate Survey by GLSEN, LGBTQ students between the ages of 13 and 21 years old were asked a series of questions. 63.5% of the students who did report an incident said that school staff did nothing in response or told the student to ignore it.

37. A computer algorithm called COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) was created by Northpointe, Inc. The algorithm assesses whether defendants have a higher or lower risk of repeating crimes. Judges sometimes use this program when setting bail or jail time. Statisticians analyzed data from 10,000 defendants assessed by the COMPAS program. They determined that 45% of African American defendants were misclassified as high risk.

---



## Section 2B – Bar Charts and Pie Charts with Technology

### Vocabulary

Data: Information in all forms

Categorical Data: Data consisting of words describing people or objects. Numbers may sometimes be used in place of words.

Amount (Frequency): The number of people or objects that have a certain characteristic.

Sample Size: The total number of people, animals or objects you collect data from.

Percentage (%): An amount out of 100.

- To calculate a percentage, multiply the proportion by 100 and adding the “%” symbol.

Proportion: The decimal equivalent of a percentage. There are two ways of calculating a proportion.

- To calculate the proportion from categorical data: Amount (frequency) divided by the total (sample size).
- To calculate the proportion from a percentage: Divide the percentage by 100 and removing the “%” symbol.

A quick way to count how many people or objects have a certain label is to create a Bar Chart or Pie Chart. There are many statistics software that we could use to create these graphs. They are useful to show the characteristics of categorical data. Data scientists are often asked to explore data with thousands or even millions of values. It would take a long time to count the amounts in a categorical data set of this size. That is why we use statistics software to calculate for us. In this class we will primarily be using “StatKey” to calculate.

StatKey: Statistics software located at [www.lock5stat.com](http://www.lock5stat.com). When you get to the website click on the “StatKey” link. StatKey works great on both MAC and PC and never needs to be saved on a computer.

### Creating a Bar Chart with Raw Data and StatKey

StatKey does not create pie charts, but does have a nice bar chart feature. It not only creates the bar chart from the raw data but also calculates the counts (frequencies) from each category as well as the decimal proportions.

To make a bar chart with raw data, go to [www.lock5stat.com](http://www.lock5stat.com) and click on the “StatKey” button. Now click on “one categorical variable” under the descriptive statistics and graphs button. If you have raw categorical data, click the “edit data” tab and paste your raw categorical data into StatKey. Make sure to check “raw data” at the bottom. If your data has a title, also check “data has a header row”. Now click “OK”.

For example, I went to [www.matt-teachout.org](http://www.matt-teachout.org) and opened the data set “Math 075 Survey Data Fall 2015”. Your instructor may have this data set also saved in Canvas. This is a survey of pre-stat students taken in Fall 2015. I copied and pasted the column of data that says “transportation type to campus” into StatKey and created the bar chart. Notice it not only created the graph, but also gave me the counts (frequencies) and the decimal proportions.



M
Transportation type to campus
Drive alone
Drive alone
Drive alone
Drive alone
Drive alone
Public transportation
Drive alone
Drive alone
Dropped off by someone
Carpool
Drive alone
Drive alone
Drive alone
Drive alone
Carpool
Drive alone
Carpool
Drive alone
Drive alone
Drive alone
Carpool
Drive alone
Drive alone

**StatKey** to accompany [Statistics: Un](#)  
by Lo

Descriptive Statistics and Graphs

- One Quantitative Variable
- One Categorical Variable**
- One Quantitative and One Categorical Variable
- Two Categorical Variables
- Two Quantitative Variables

### Descriptive Statistics for One Categorical Variable



Edit data
✕

Transportation type to campus

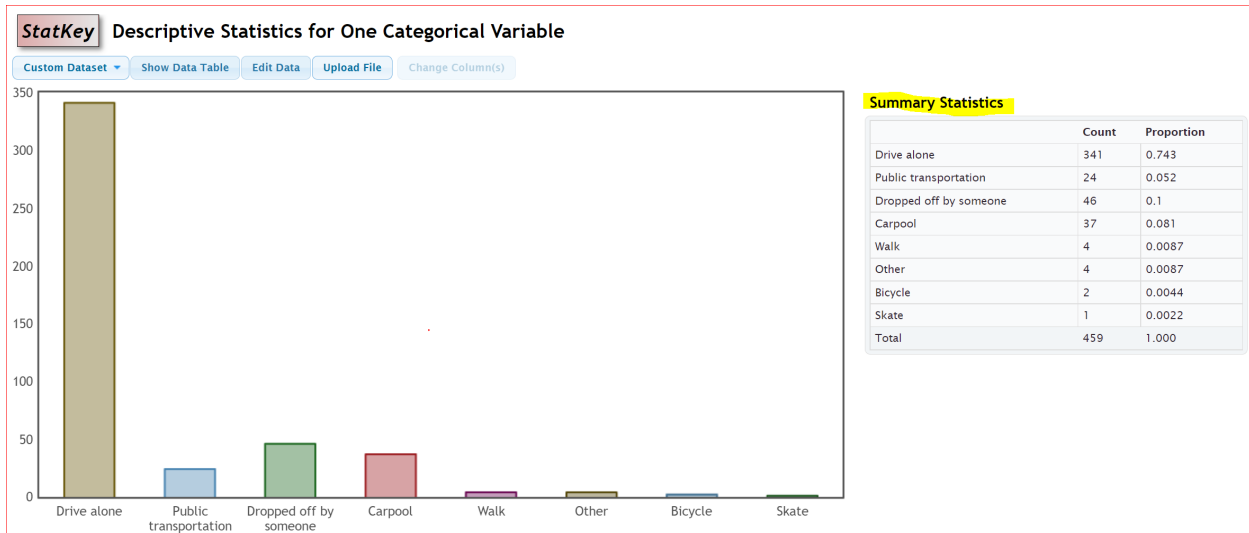
- Drive alone
- Drive alone
- Drive alone
- Drive alone
- Drive alone
- Drive alone
- Public transportation
- Drive alone
- Drive alone
- Dropped off by someone
- Carpool
- Drive alone
- Drive alone
- Drive alone
- Drive alone
- Drive alone
- Carpool
- Drive alone
- Carpool
- Drive alone
- Drive alone

Raw Data

Data has header row

Manually edit the values above or paste a tab or comma separated file into the box and click Ok. For raw data, the file must have only one column. A summary counts table should contain two columns, where the first column contains categories and the second column contains counts.

Ok



Notice we can answer all sorts of questions about this categorical data by using the StatKey printout.

Which type of transportation was most common? Driving alone was most common for math 075 students. (We see in the bar chart that “drive alone” was the highest bar.)

How many math 075 students were dropped off at school? 46 math 075 students were dropped off at school.

What proportion of math 075 students drive alone? 0.743 of the math 075 students drive alone.





What percentage of math 075 students carpool? We know the proportion is 0.081. To convert to a percentage, multiply by 100 and add on the “%” symbol.

$$0.081 \times 100\% = 8.1\% \text{ of math 075 students carpool.}$$

### Creating a Bar Chart with Summary Data and StatKey

Categorical data is often summarized by the counts for each variable. When a data analyst receives categorical data to analyze, it may not be in raw form. Often it is just a list of the categorical variables and the counts (frequencies). In that case, when you go to the “edit data” button in StatKey, you will need to type in the variables and counts as shown below. Uncheck the “raw data” box at the bottom and push “OK”. Note that you need only one space after the comma and do not type in the totals. Notice you will get the exact same graphs, counts and proportions as shown above.

Response, Frequency  
Drive alone, 341  
Public Transportation, 24  
Dropped off by someone, 46  
Carpool, 37  
Walk, 4  
Other, 4  
Bicycle, 2  
Skate, 1

Response, Frequency  
Drive alone, 341  
Public Transportation, 24  
Dropped off by someone, 46  
Carpool, 37  
Walk, 4  
Other, 4  
Bicycle, 2  
Skate, 1

Raw Data  
 Data has header row

Manually edit the values above or paste a tab or comma separated file into the box and click Ok. For raw data, the file must have only one column. A summary counts table should contain two columns, where the first column contains categories and the second column contains counts.

Ok

Statcato

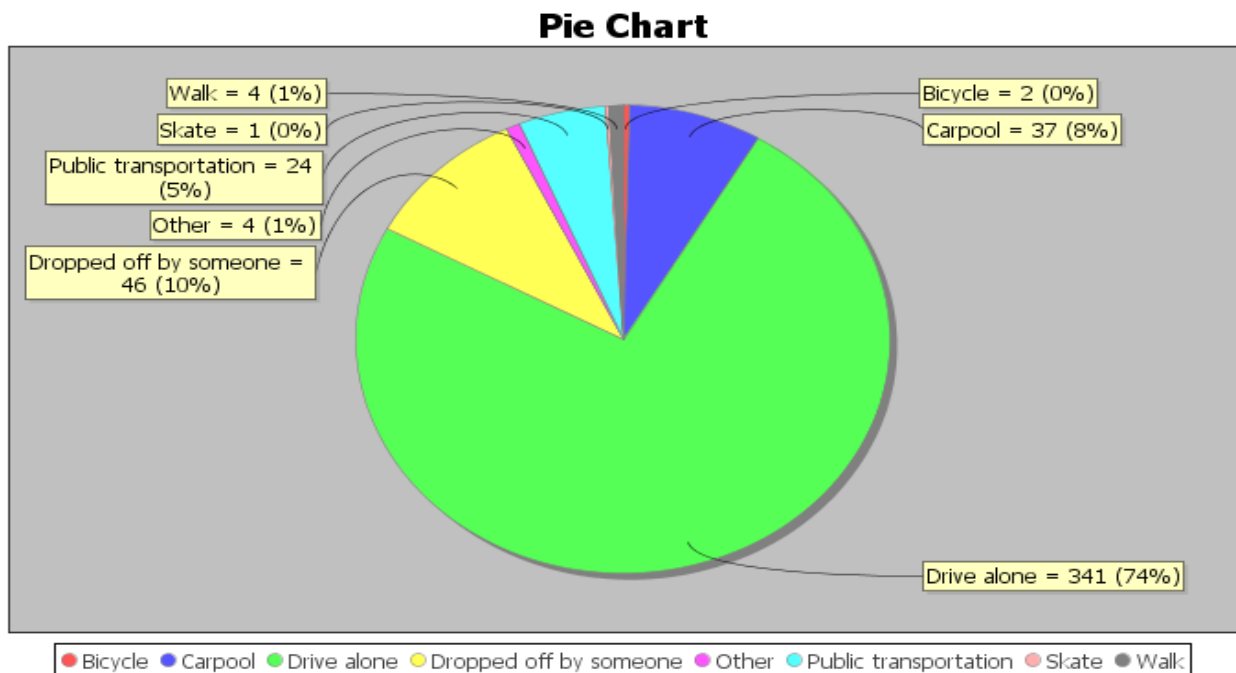


This material is from *Introduction to Data Analysis*, 2<sup>nd</sup> edition by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a “CC-BY” [Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by/4.0/) – 3/17/2021

No computer program does everything. Another free program that is very useful is Statcato ([www.statcato.org](http://www.statcato.org)). Statcato is a great program but is difficult to use. It must be saved on the computer and does not work well on MAC computers. You will often see graphs and printouts from Statcato in the homework. However, you will only need to analyze the Statcato graphs and statistics provided. You will not need to calculate with Statcato. So you do not need to save Statcato on your computer. Calculations will always be done with StatKey. StatKey is much easier to use and does not need to be saved on your computer.

### Pie Charts

Another graph often used when analyzing categorical data is the pie chart. The following Pie Chart was created using the same “Transportation Type to Campus” data from the Math 075 Survey Data Fall 2015. For this graph, we used the statistics software program Statcato. Notice the pie chart from Statcato gives us the same counts as StatKey, but instead of proportions, it gives us the approximate percentages for each variable. Notice the percentages are rounded to the nearest percent and have less accuracy than the proportions in StatKey.



Notice at the touch of a button, the computer can tell us all of the counts (frequencies) and all of the percentages. We can now answer all sorts of questions about how these math 075 students get to the college.

What type of transportation was used the least? Skating (*Notice it had the smallest piece of the pie.*)

How many math 075 students used public transportation? 24 math 075 students used public transportation.

What percentage of math 075 students carpool to campus? Approximately 8% of math 075 students carpool. (*Notice this answer has less accuracy.*)

What proportion of the math 075 students used public transportation? Approximately 0.05 of the math 075 students use public transportation. (*We know from the pie chart that the percentage is approximately 5%. We will convert the 5% into a proportion by removing the “%” sign and dividing by 100.*)

$$5\% = 5 \div 100 = 0.05$$



## Problem Set Section 2B

*Directions: Open the Math 075 Survey Data Fall 2015 on Canvas or at [www.matt-teachout.org](http://www.matt-teachout.org). Go to [www.lock5stat.com](http://www.lock5stat.com) and use StatKey to create a bar chart and summary statistics. Click on the “Edit Data” tab in StatKey and paste the column of data into StatKey. Click on “Raw Data” if it is a column of data. Click on “Data has a header row” if the data has a title. Then push OK. Make a rough sketch of the bar chart and summary statistics on a piece of paper and answer the questions.*

1. Use the column of data that says “Campus” in the Math 075 Survey Data Fall 2015 and StatKey to create a bar chart and find the summary counts and proportions. Make a rough sketch of the bar chart on a piece of paper and answer the following questions.

- a) Were there more math 075 students at the Valencia campus or at the Canyon Country campus?
- b) How many math 075 students went to the Valencia campus?
- c) How many math 075 students went to the Canyon Country campus?
- d) What proportion of the math 075 students went to the Valencia campus?
- e) What proportion of the math 075 students went to the Canyon Country campus?
- f) What percent of the 075 students went to the Valencia campus?
- g) What percent of the 075 students went to the Canyon Country campus?

2. Use the column of data that says “Gender” in the Math 075 Survey Data Fall 2015 and StatKey to create a bar chart and find the summary counts and proportions. Make a rough sketch of the bar chart on a piece of paper and answer the following questions.

- a) Were there more female math 075 students or male math 075 students?
- b) How many math 075 students were female?
- c) How many math 075 students were male?
- d) What proportion of the math 075 students were female?
- e) What proportion of the math 075 students were male?
- f) What percent of the 075 students were female?
- g) What percent of the 075 students were male?

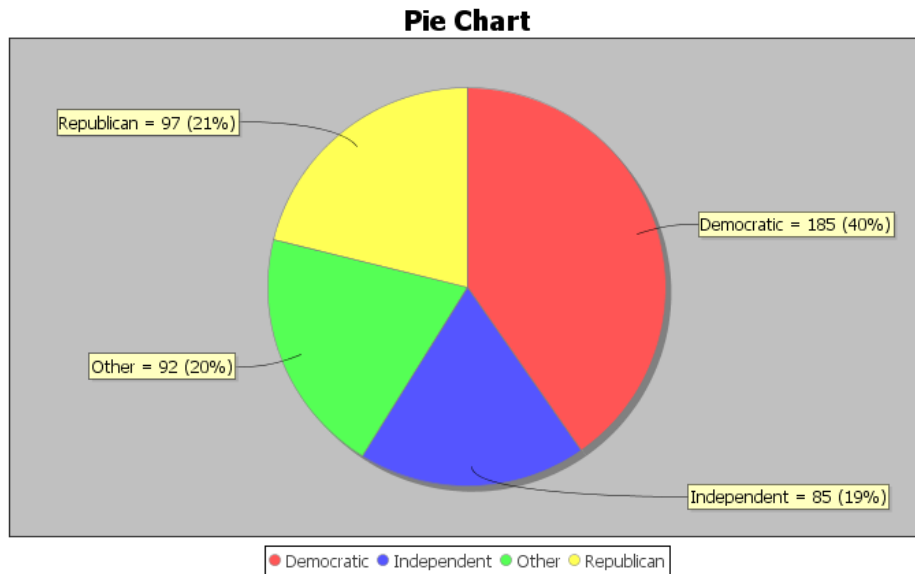
3. Use the column of data that says “Hair Color” in the Math 075 Survey Data Fall 2015 and StatKey to create a bar chart and find the summary counts and proportions. Make a rough sketch of the bar chart on a piece of paper and answer the following questions.

- a) Which hair color had the most students?
- b) Which hair color had the least?
- c) How many of the math 075 students have brown hair?
- d) How many of the math 075 students have blond hair?
- e) What proportion of the math 075 students have red hair?
- f) What proportion of the math 075 students have black hair?
- g) What percentage of math 075 students have red hair?



h) What percentage of the math 075 students have black hair?

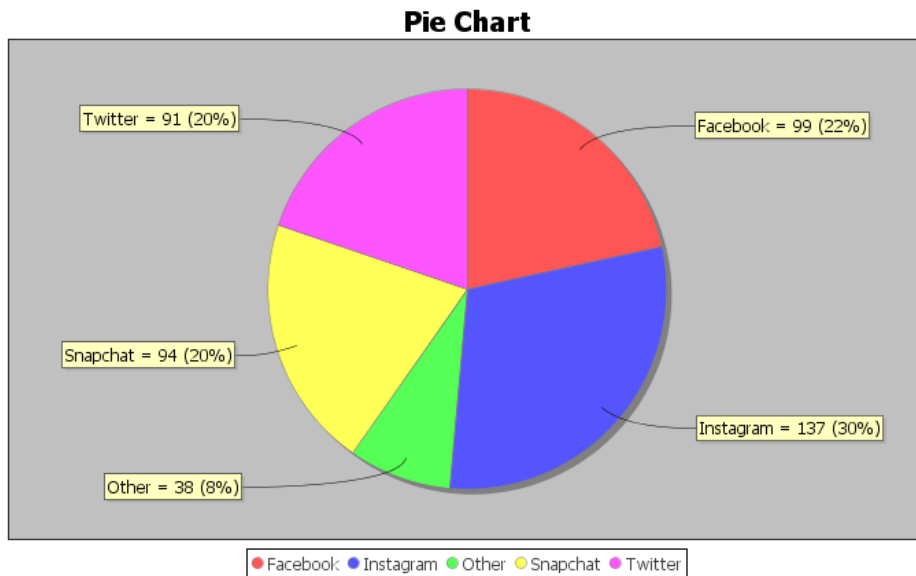
4. The following pie chart was created with Statcato and the column of data that says "Political Party" in the Math 075 Survey Data Fall 2015. Use the pie chart to answer the following questions.



- a) Which political party had the most students?
- b) Which political party had the least students?
- c) How many of the math 075 students were republican?
- d) How many of the math 075 students were democrat?
- e) What percentage of the math 075 students identified as independent political party?
- f) What percentage of the math 075 students identified as "other" political party?
- g) What proportion of math 075 students were democrat?
- h) What proportion of the math 075 students were republican?



5. The following pie chart was created with Statcato and the column of data that says “Social Media Favorite” in the Math 075 Survey Data Fall 2015. Use the pie chart to answer the following questions.



- Which social media was most popular with math 075 students in Fall 2015?
- Which social media was least popular with math 075 students in Fall 2015?
- How many of the math 075 students prefer snapchat?
- How many of the math 075 students prefer instagram?
- What percentage of the math 075 students prefer twitter?
- What percentage of the math 075 students prefer “other” social media?
- What proportion of math 075 students prefer instagram?
- What proportion of the math 075 students were snapchat?

*Directions for #6-7: Enter the given categorical summary data into StatKey to create a bar chart with summary counts and proportions. The summary data must be typed correctly with only one space after commas. Then use the bar chart and summary proportions to answer the questions.*

6. We looked at a sample of 83 retired NFL football players and found that only 18 of them were still doing ok financially, but 65 of them had gone bankrupt. Go to [www.lock5stat.com](http://www.lock5stat.com) and create a bar chart and calculate the summary statistics. Make a rough sketch of the bar chart on a piece of paper and answer the following questions. Go to [www.lock5stat.com](http://www.lock5stat.com) and click on StatKey. Then click on “one categorical variable” and edit data. Type in the following in order to make the bar chart and summary proportions. Do not check the box that says “raw data”. Do click the box that says header row. Then push “ok”. Make a rough sketch of the bar chart on a piece of paper and answer the following questions.

Response, Frequency  
 OK Financially, 18  
 Bankrupt, 65

- What proportion of the retired NFL players in the data had gone bankrupt?
- What percentage of the retired NFL players in the data had gone bankrupt?



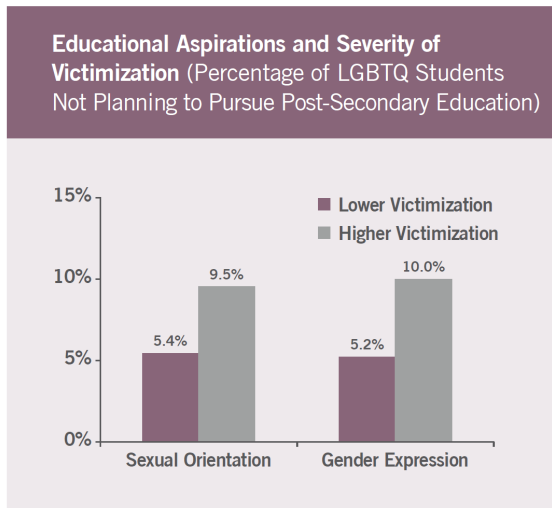
- c) What proportion of the retired NFL players in the data were doing OK financially?
- d) What percentage of the retired NFL players in the data were doing OK financially?

7. When a math 075 student asked COC students what their favorite coffee shop in Santa Clarita was, 41 said they preferred Starbucks, 27 said Coffee Bean, and 19 said Peets Coffee. Go to [www.lock5stat.com](http://www.lock5stat.com) and click on StatKey. Then click on “one categorical variable” and edit data. Type in the following in order to make the bar chart and summary proportions. Make a rough sketch of the bar chart on a piece of paper and answer the following questions. Go to [www.lock5stat.com](http://www.lock5stat.com) and click on StatKey. Then click on “one categorical variable” and edit data. Type in the following in order to make the bar chart and summary proportions. Do not check the box that says “raw data”. Do click the box that says header row. Then push “ok”. Make a rough sketch of the bar chart on a piece of paper and answer the following questions.

Response, Frequency  
 Starbucks, 41  
 Coffee Bean, 27  
 Peets Coffee, 19

- a) What proportion of the COC students in the data preferred Starbucks?
- b) What percentage of the COC students in the data preferred Starbucks?
- c) What proportion of the COC students in the data preferred Coffee Bean?
- d) What percentage of the COC students in the data preferred Coffee Bean?
- e) What proportion of the COC students in the data preferred Peet’s Coffee?
- f) What percentage of the COC students in the data preferred Peet’s Coffee?

8. In the 2015 National School Climate Survey by GLSEN, over ten-thousand LGBTQ students between the ages of 13 and 21 years old from all 50 states in the U.S. were asked a series of questions. The following bar chart was created from this data.

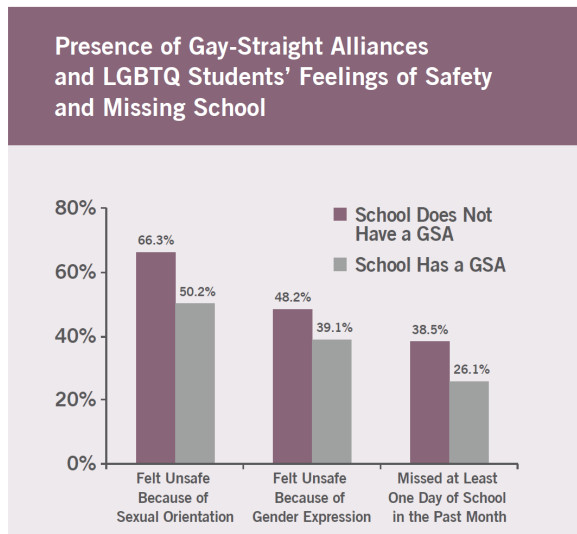


- a) What percentage of the LGBTQ students were not planning to continue their education due to high victimization against their sexual orientation? Convert the percentage into a decimal proportion.
- b) What percentage of the LGBTQ students were not planning to continue their education due to high victimization against their gender expression? Convert the percentage into a decimal proportion.



- c) What percentage of the LGBTQ students were not planning to continue their education due to lower level victimization against their sexual orientation? Convert the percentage into a decimal proportion.
- d) What percentage of the LGBTQ students were not planning to continue their education due to lower level victimization against their gender expression? Convert the percentage into a decimal proportion.

9. In the 2015 National School Climate Survey by GLSEN, over ten-thousand LGBTQ students between the ages of 13 and 21 years old from all 50 states in the U.S. were asked a series of questions. The following bar chart was created from this data.



- a) What percentage of the LGBTQ students attend a school without a Gay-Straight Alliance program and feel unsafe because of sexual orientation? Convert the percentage into a decimal proportion.
- b) What percentage of the LGBTQ students attend a school with a Gay-Straight Alliance program and feel unsafe because of sexual orientation? Convert the percentage into a decimal proportion.
- c) What percentage of the LGBTQ students attend a school without a Gay-Straight Alliance program and feel unsafe because of gender expression? Convert the percentage into a decimal proportion.
- d) What percentage of the LGBTQ students attend a school with a Gay-Straight Alliance program and feel unsafe because of gender expression? Convert the percentage into a decimal proportion.



## Section 2C – Comparing Percentages

### Vocabulary

Data: Information in all forms

Categorical Data: Data consisting of words describing people or objects. Numbers may sometimes be used in place of words.

Amount (Frequency): The number of people or objects that have a certain characteristic.

Sample Size: The total number of people, animals or objects you collect data from.

Percentage (%): An amount out of 100.

- To calculate a percentage, multiply the proportion by 100 and adding the “%” symbol.

Proportion: The decimal equivalent of a percentage. There are two ways of calculating a proportion.

- To calculate the proportion from categorical data: Amount (frequency) divided by the total (sample size).
- To calculate the proportion from a percentage: Divide the percentage by 100 and removing the “%” symbol.

Statistics is based on the idea of answering questions. One of the most common questions that is often asked of a data analyst is to compare a categorical variable from multiple groups. Do men in the data have a higher percentage of Type 2 diabetes than women? Is the percentage of people that own guns lower in large cities than in rural communities? Which high schools in your community give students the best opportunity to get a scholarship to college?

These are all important questions that can be answered with technology and a good understanding of categorical data and percentages.

**Note about populations:** *At this point, we are learning to analyze data. For example, we can look at the percentage of men in the data set with Type 2 Diabetes versus the percentage of women. This gives us an idea about gender and diabetes but we should not apply that to all men or all women. It takes a much greater knowledge of statistical methods to apply data to millions of people. Your data set may not represent all men and all women on planet earth.*

Let us learn to think about questions we can answer from the data. Let us look at an example using the hospital data.

*The data includes gender, blood type (A, B, AB, O), Rhesus factor (Rh + or Rh -) and part of the hospital (Medical/Surgical, Intensive Care Unit, Same Day Surgery, Emergency Room).*

Gender	Blood Type	Rh Factor	Floor
M	A	-	SDS
M	O	+	ER
F	AB	+	Med/Surg
M	O	-	ICU
F	O	+	SDS
F	O	+	Med/Surg
M	A	+	SDS
F	O	+	Med/Surg
F	O	+	ER
M	B	+	SDS
F	A	-	Med/Surg
M	O	+	ICU
M	A	+	Med/Surg
F	O	-	SDS



This material is from *Introduction to Data Analysis*, 2<sup>nd</sup> edition by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a “CC-BY” [Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by/4.0/) – 3/17/2021



F	B	+	ICU
M	O	+	ER
F	AB	-	ER
M	O	+	SDS
M	O	+	Med/Surg
M	A	+	ER

**Example 1**

Do male patients have a higher chance of getting admitted to the emergency room than female patients?

Remember how to find a percentage.

$$\text{Decimal Proportion} = \frac{\text{Amount}}{\text{Total}}$$

To convert proportion into percentage, multiply by 100%.

Let us start by finding the total number of male patients and then see how many of them were admitted to the emergency room.

*Note: This is often called a “Conditional Proportion” since we are only looking at only the male patients and not everyone in the data set.*

Gender	Blood Type	Rh Factor	Floor
M	A	-	SDS
M	O	+	ER
F	AB	+	Med/Surg
M	O	-	ICU
F	O	+	SDS
F	O	+	Med/Surg
M	A	+	SDS
F	O	+	Med/Surg
F	O	+	ER
M	B	+	SDS
F	A	-	Med/Surg
M	O	+	ICU
M	A	+	Med/Surg
F	O	-	SDS
F	B	+	ICU
M	O	+	ER
F	AB	-	ER
M	O	+	SDS
M	O	+	Med/Surg
M	A	+	ER

Total number of male patients: 11

How many of those 11 male patients were admitted to the emergency room? 3

Decimal Proportion =  $3/11 = 0.2727272... \approx 0.273$  (Remember to round proportion to three decimal places.)

Percentage of male patients admitted to ER?  $0.273 \times 100\% \approx 27.3\%$  (Convert to percentage by multiplying the proportion by 100 and adding the “%” sign.)

Now let us compare this percentage to female patients admitted to the emergency room.



Gender	Blood Type	Rh Factor	Floor
M	A	-	SDS
M	O	+	ER
F	AB	+	Med/Surg
M	O	-	ICU
F	O	+	SDS
F	O	+	Med/Surg
M	A	+	SDS
F	O	+	Med/Surg
F	O	+	ER
M	B	+	SDS
F	A	-	Med/Surg
M	O	+	ICU
M	A	+	Med/Surg
F	O	-	SDS
F	B	+	ICU
M	O	+	ER
F	AB	-	ER
M	O	+	SDS
M	O	+	Med/Surg
M	A	+	ER

Total number of female patients: 9

How many of those 9 female patients were admitted to the emergency room? 2

Decimal Proportion =  $2 \div 9 = 0.2222222... \approx 0.222$  (Remember to round proportion to three decimal places.)

Percentage of female patients admitted to ER?  $0.222 \times 100\% \approx 22.2\%$  (Convert to percentage by multiplying the proportion by 100 and adding the “%” sign.)

So what does this tell us?

First, remember these percentages do not apply to all patients in every hospital, but we can see what this data suggests about the patients in this data set from this single hospital.

The percentage of male patients admitted to ER (27.3%) is higher than the percentage of female patients admitted to the ER (22.2%). This is important information for this hospital and in particular the emergency room to know. However we are not sure if the percentage of males admitted to ER is significantly higher than the percentage of females.

[How can we tell if there is a significant difference between groups?](#)

This is a very difficult question to answer. Many statisticians studied and worked on methods to determine significance. They invented hypothesis tests (significance tests), Confidence Intervals, P-values, and many other ways to check significance. We are not at that level yet, but I find that taking a ratio of the percentages is a good way to compare. We can also calculate a percentage of increase.

#### Note on Practical Significance and Sample Size

Sometimes there may be a large difference between percentages, but it may not be significant. It is always important to consider the total number of people or objects in your data (sample size). Suppose we have a ratio of 2. The



This material is from *Introduction to Data Analysis*, 2<sup>nd</sup> edition by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a “CC-BY” [Creative Commons Attribution 4.0 International License](#) – 3/17/2021

higher percentage is twice as large as the lower. At first glance, we might think this is significant. However, if our sample size is only 10 people, then it may not be significant.

Some people refer to this principle as “statistical significance” versus “practical significance”.

Sometimes when there is a statistically significant difference, it does not necessarily mean it is of practical use. Suppose a company makes backpacks and collects data from a small sample of 12 customers. 8 of the customers liked the new line of backpacks and 4 did not. The percentage of customers that liked the backpacks is 66.7% and the percentage that did not like the backpacks is 33.3%. This would be a 100% increase and looks significant. However, should the company really make a decision based on 12 people where 4 more liked the backpack than not? This data does not have practical significance.

### Ratio of Two Percentages

A ratio of two percentages tells us how many times larger the higher % is than the lower %. For example, if the ratio comes out to be 3.5, then the higher % is 3.5 times greater than the lower %. If the ratio comes out to be around 1.5 or higher, that may indicate a significant difference. If the ratio comes out around 1, that is usually not very significant. Remember, this is not the most accurate way to determine significance, but it can give us an idea.

Here is the formula for calculating the percentage ratio.

$$\text{Ratio of Two Percentages} = \frac{\text{Higher \%}}{\text{Lower \%}} \text{ or } \frac{\text{Higher Proportion}}{\text{Lower Proportion}}$$

A ratio of two percentages can be difficult to interpret. Sample size is important to consider.

Ratio close to 1 = Usually NOT significant

Ratio of 1.5 or higher = Usually significant if the sample size was large enough.

Ratio between 1.2 and 1.4 = Might be significant if the sample size was large enough.

Let us look at the ratio for the previous example.

$$\text{Ratio} = 27.3\% / 22.2\% \approx 1.23$$

It is important to consider sample size. The sample size was only 20 patients. This is a relatively small sample size. The ratio of the percentages was 1.23. This tells us that male patients that go to ER are only 1.23 times more than the female patients that go to ER. This ratio 1.23 might be significant if we had thousands of people in the data set. However, this was a small data set. I would not tell the hospital to make changes to their care based on this data. It does not have practical significance.

*Note: We can also calculate the ratio from the decimal proportions. Be careful to either compare the percentages or compare the decimal proportions. Do not compare a percentage to a decimal proportion.*

$$\text{Ratio using decimal proportions} = 0.273 / 0.222 \approx 1.23 \text{ (same correct answer)}$$

*However, 27.3% / 0.222 does not equal 1.23!!!*

### Percentage of Increase

A “percentage of increase” is another common calculation that is sometimes used when compare categorical variables and see if one variable has a significantly higher proportion or percentage than another. To compare



proportion or percentages, many people often calculate the “percentage of increase”. As with the ratio, the percent of increase can be calculated from the percentages or from the proportions. These formulas give the same answer.

$$\text{Percent of Increase} = \frac{(\text{Higher Proportion} - \text{Lower Proportion})}{\text{Lower Proportion}} \times 100\%$$

$$\text{Percent of Increase} = \frac{(\text{Higher \%} - \text{Lower \%})}{\text{Lower \%}} \times 100\%$$

Percent of Increase can be difficult to interpret. Sample size is important to consider.

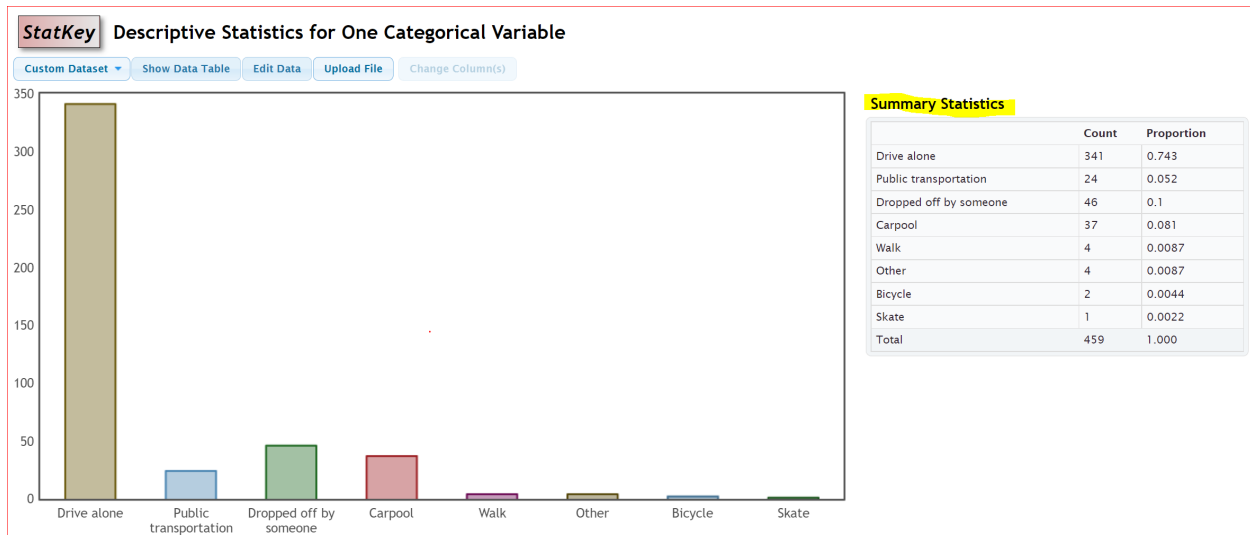
Percent of Increase less than 10% = Usually NOT significant

Percent of Increase greater than 50% = Usually significant if the sample size was large enough.

Percent of increase between 10% and 50% = Might be significant if the sample size was large enough.

### Example

In the last section, we used StatKey to calculate summary statistics for the “transportation type to campus” data from the Math 075 Survey Data Fall 2015. Suppose we want to compare the percentage of math 075 students that carpool versus the percentage that were dropped off. We can calculate the percent of increase from the proportions or percentages. It is important to recognize which is the lower proportion and which is the higher proportion. In this case, the proportion of students that were dropped off (0.1 or 10%) was higher than the proportion of students that carpooled (0.081 or 8.1%). The key question is was it significantly higher.



We can calculate the percent of increase from either the proportions or the percentages.

$$\begin{aligned} \text{Percent of Increase} &= \frac{(\text{Higher Proportion} - \text{Lower Proportion})}{\text{Lower Proportion}} \times 100\% = \frac{(0.1 - 0.081)}{0.081} \times 100\% = \\ &= \frac{(0.019)}{0.081} \times 100\% \approx 23.5\% \text{ increase} \end{aligned}$$

$$\text{Percent of Increase} = \frac{(\text{Higher \%} - \text{Lower \%})}{\text{Lower \%}} \times 100\% = \frac{(10\% - 8.1\%)}{8.1\%} \times 100\% =$$



$$= \frac{(1.9\%)}{8.1\%} \times 100\% \approx 23.5\% \text{ increase}$$

In this problem we had a total of 459 students and a 23.5% increase. The sample size seems large enough. This looks significant, both statistically significant and practically significant.

*Note: We are at the Pre-Stat level. In higher levels of statistics, you will learn how to use confidence intervals, test statistics, and P-values to determine significant differences. These are generally more accurate and easier to read than the ratio or the percent of increase.*

---



## Practice Problems Section 2C

(#1-4) Directions: Use the following formulas to calculate the ratio of the two percentages and the percent of increase. Then answer the questions.

$$\text{Ratio of the Percentages} = \frac{\text{Higher Proportion}}{\text{Lower Proportion}} \text{ or } \frac{\text{Higher Percentage}}{\text{Lower Percentage}}$$

$$\text{Percent of Increase} = \frac{(\text{Higher \%} - \text{Lower \%})}{\text{Lower \%}} \times 100\% \text{ or } \frac{(\text{Higher Proportion} - \text{Lower Proportion})}{\text{Lower Proportion}} \times 100\%$$

1. An article at [www.seattletimes.com](http://www.seattletimes.com) was addressing the issue of whether women in the U.S. prefer traditional jeans or athletic wear like yoga pants, sweat pants or leggings. Assume that a random sample of 213 total women were asked if they prefer traditional jeans or athletic wear. Assume 139 of the women (0.653 or 65.3%) said they prefer athletic wear. 74 of the women (0.347 or 34.7%) said they prefer traditional jeans.

- What is the ratio of the percentages? Write a sentence to explain the ratio.
- What is the percent of increase? Does the percent of increase look high or low?
- The sample size was large enough in this case. Does the ratio and percent of increase indicate that the percentages were significantly different?
- How would you advise a women's clothing company to act on this data?

2. A hospital is trying to decide how to allocate resources to various departments. In particular, they are comparing the medical/surgical ward to the telemetry (heart monitor) ward since these wards have similar costs per patient. Assume we looked at a random sample of patients admitted to the hospital. Of the 350 total patients, 57 of the patients (0.163 or 16.3%) were admitted to the medical/surgical ward. 49 of the patients (0.14 or 14%) were admitted to telemetry.

- What is the ratio of the percentages? Write a sentence to explain the ratio.
- What is the percent of increase? Does the percent of increase look high or low?
- The sample size was large enough in this case. Does the ratio and percent of increase indicate that the percentages were significantly different?
- How would you advise the hospital to act on this data?

3. A company found that of their 348 total employees, 96 employees (0.276 or 27.6%) have health insurance and 252 employees (0.724 or 72.4%) do not have health insurance.

- What is the ratio of the percentages? Write a sentence to explain the ratio.
- What is the percent of increase? Does the percent of increase look high or low?
- The sample size was large enough in this case. Does the ratio and percent of increase indicate that the percentages were significantly different?
- How would you advise the company to act on this data?

4. An experiment was done to test the effectiveness of a new medicine to treat depression. They found that of the 57 people that received the medicine, 13 of them (0.228 or 22.8%) indicated significant improvement in their depression symptoms. Of the 61 people in the placebo group, 11 of them (0.180 or 18.0%) indicated significant improvement in their depression symptoms.

- What is the ratio of the percentages? Write a sentence to explain the ratio.
- What is the percent of increase? Does the percent of increase look high or low?
- The amount of people that showed improvement was rather small. Does the ratio and percent of increase indicate that the percentages were significantly different?
- Should the pharmaceutical company release this medicine for public use? Why or why not?



5. A computer algorithm called COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) was created by Northpointe, Inc. The algorithm assesses whether defendants have a higher or lower risk of repeating crimes. Judges sometimes use this program when setting bail or jail time. Statisticians analyzed data from 10,000 defendants assessed by the COMPAS program. They determined that 45% of African American defendants were misclassified as high risk, while 23% of white defendants were misclassified as high risk.

- What is the ratio of the percentages? Write a sentence to explain the ratio.
- What is the percent of increase? Does the percent of increase look high or low?
- The amount of defendants assessed was large enough. Does the ratio and percent of increase indicate that the percentage of African American defendants was significantly higher than for white defendants?
- What does this data tell us about the use of this program in assessing whether or not a defendant will repeat their crime?

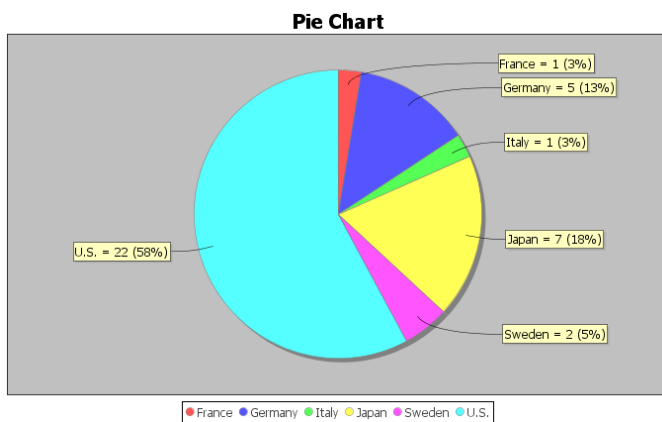
(#6-10) Use the following formulas, pie charts and bar charts to calculate the ratio of the percentages and the percent of increase. Then answer the questions.

$$\text{Ratio of the Percentages} = \frac{\text{Higher Proportion}}{\text{Lower Proportion}} \text{ or } \frac{\text{Higher Percentage}}{\text{Lower Percentage}}$$

$$\text{Percent of Increase} = \frac{(\text{Higher \%} - \text{Lower \%})}{\text{Lower \%}} \times 100\% \text{ or } \frac{(\text{Higher Proportion} - \text{Lower Proportion})}{\text{Lower Proportion}} \times 100\%$$

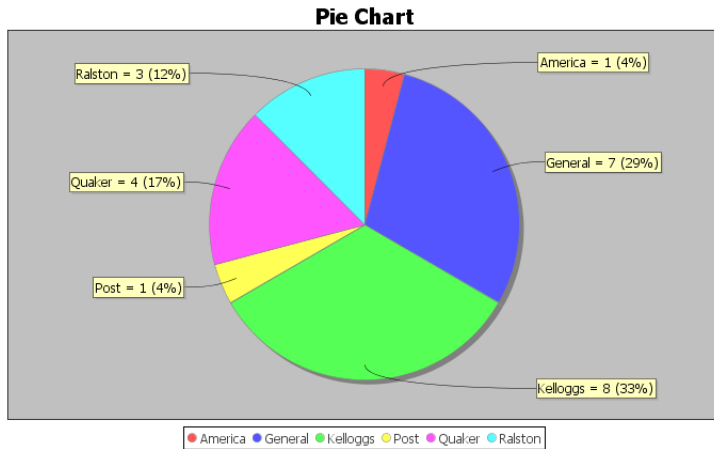
6. The following pie chart was created from the “car data”. It is data taken from a random sample of various types of cars around the world. Use the pie chart to answer the following questions.

- What is the ratio of the percentages for Japan and Germany? Write a sentence to explain the ratio.
- What is the percent of increase for Japan and Germany? Does the percent of increase look high or low?
- The sample size was rather small. Does the ratio and percent of increase indicate that the percentages were significantly different?



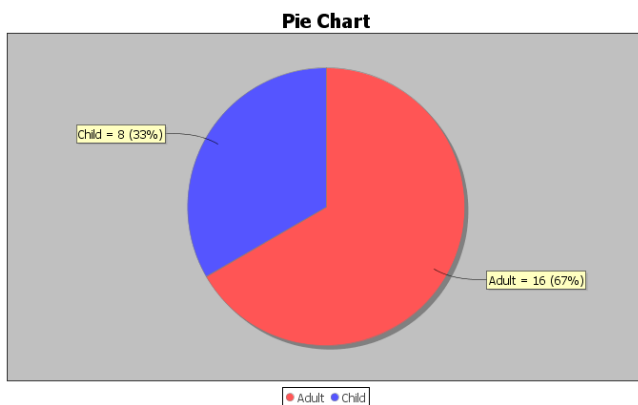
7. The following pie chart was created from the “cereal data” using Statcato. The data was taken from a random sample of various cereals. Use the pie chart to answer the following questions.

- What is the ratio of the percentages for cereals made by Kelloggs and General? Write a sentence to explain the ratio.
- What is the percent of increase for cereals made by Kelloggs and General? Does the percent of increase look high or low?
- The sample size was rather small. Does the ratio and percent of increase indicate that the percentages were significantly different?



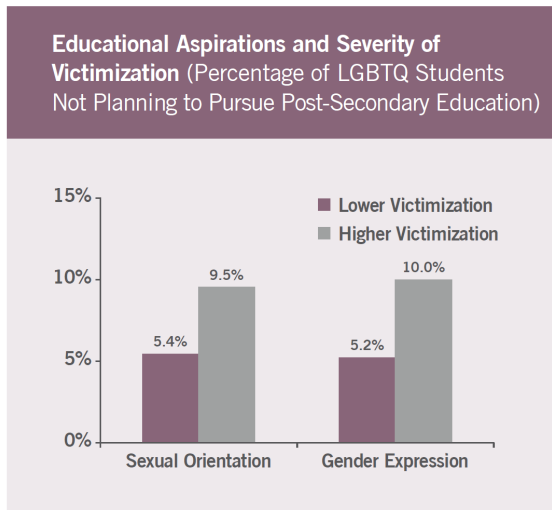
8. The following pie chart was created from the “cereal data”. The data was taken from a random sample of various cereals. Use the pie chart to answer the following questions.

- What is the ratio of the percentages for cereals made for adults verses children? Write a sentence to explain the ratio.
- What is the percent of increase for cereals made for adults verses children? Does the percent of increase look high or low?
- The sample size was rather small. Does the ratio and percent of increase indicate that the percentages were significantly different?





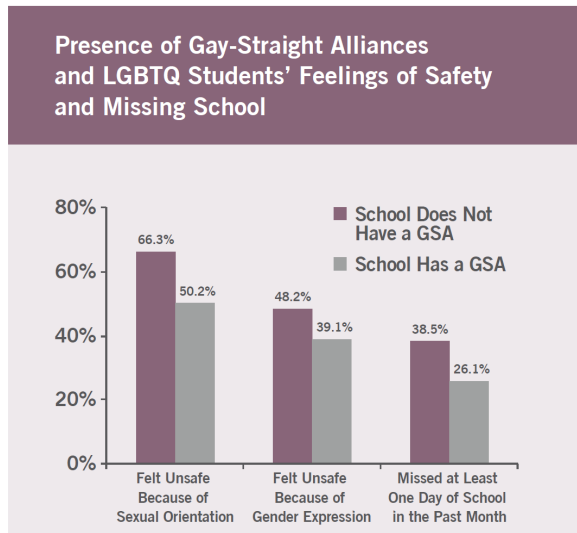
9. In the 2015 National School Climate Survey by GLSEN, over ten-thousand LGBTQ students between the ages of 13 and 21 years old from all 50 states in the U.S. were asked a series of questions. The following bar chart was created from this data.



- Focus on LGBTQ students that are not planning to continue school due to victimization regarding their sexual orientation. What is the ratio of the percentages for higher victimization versus lower level victimization? Write a sentence to explain the ratio.
- Focus on LGBTQ students that are not planning to continue school due to victimization regarding their sexual orientation. What is the percent of increase for higher victimization versus lower level victimization? Does the percent of increase look high or low?
- Focus on LGBTQ students that are not planning to continue school due to victimization regarding their sexual orientation. The sample size was large enough. Does the ratio and percent of increase indicate that the percentages were significantly different?
- Focus on LGBTQ students that are not planning to continue school due to victimization regarding their gender expression. What is the ratio of the percentages for higher victimization versus lower level victimization? Write a sentence to explain the ratio.
- Focus on LGBTQ students that are not planning to continue school due to victimization regarding their gender expression. What is the percent of increase for higher victimization versus lower level victimization? Does the percent of increase look high or low?
- Focus on LGBTQ students that are not planning to continue school due to victimization regarding their gender expression. The sample size was large enough. Does the ratio and percent of increase indicate that the percentages were significantly different?



10. In the 2015 National School Climate Survey by GLSEN, over ten-thousand LGBTQ students between the ages of 13 and 21 years old from all 50 states in the U.S. were asked a series of questions. The following bar chart was created from this data.



- Focus on LGBTQ students that feel unsafe due to their sexual orientation. What is the ratio of the percentage for schools that do not have a Gay-Straight Alliance program to the percentage of schools that do have a Gay-Straight Alliance program? Write a sentence to explain the ratio.
- Focus on LGBTQ students that feel unsafe due to their sexual orientation. What is the percent of increase for schools that do not have a Gay-Straight Alliance program to the percentage of schools that do have a Gay-Straight Alliance program? Does the percent of increase look high or low?
- Focus on LGBTQ students that feel unsafe due to their sexual orientation. The sample size was large enough. Does the ratio and percent of increase indicate that the percentages were significantly different?
- Focus on LGBTQ students that feel unsafe due to their gender expression. What is the ratio of the percentage for schools that do not have a Gay-Straight Alliance program to the percentage of schools that do have a Gay-Straight Alliance program? Write a sentence to explain the ratio.
- Focus on LGBTQ students that feel unsafe due to their gender expression. What is the percent of increase for schools that do not have a Gay-Straight Alliance program to the percentage of schools that do have a Gay-Straight Alliance program? Does the percent of increase look high or low?
- Focus on LGBTQ students that feel unsafe due to their gender expression. The sample size was large enough. Does the ratio and percent of increase indicate that the percentages were significantly different?



11. An experiment was done on labor market discrimination. The created fictitious resumes to help-wanted adds in Boston and Chicago newspapers. Each resume was assigned either a very African American sounding name or a very white sounding name. The following table summarizes the results.

**Table 1**  
**Mean Call-Back Rates By Racial Soundingness of Names**

	<i>Call-Back Rate for White Names</i>	<i>Call-Back Rate for African American Names</i>	<i>Ratio</i>
Sample:			
All sent resumes	10.06%	6.70%	1.50

- The ratio of the percentages for white versus African American names was already calculated as 1.50. Write a sentence to explain the ratio.
  - What is the percent of increase for white versus African American names? Does the percent of increase look high or low?
  - The sample size was large enough. Does the ratio and percent of increase indicate that the percentage of callbacks for white applicants was significantly higher than for African American applicants?
  - Since the experiment controlled confounding variables like age, experience, education, etc., does this data indicate that there is racial discrimination in the labor market in Boston and Chicago?
- 



## Section 2D – Estimating Amounts with Percentage Data

If you pick up any newspaper or magazine or click on any news or sports link online, you are likely to see information summarized with percentages.

How can we use these percentages to give us a better understanding of the categorical data?

One of the most common uses of percentages is to estimate amounts from a total. Before we can do this, we need to remember how to convert the percentage back into a proportion.

**Convert a Percentage into a Proportion: Remove “%” sign and divide by 100.**

Example: Convert 13.7% into a proportion.

$$13.7\% = 13.7 \div 100 = 0.137$$

### Estimating an amount from percentage information

Recall the following formula.

$$\text{Proportion} = \text{Amount (Frequency)} \div \text{Total (Sample Size)}$$

If you do a little algebra and multiply both sides of that formula by the Total, you get the following formula.

$$\text{Amount (Frequency)} = \text{Proportion} \times \text{Total}$$

In other words, to find an amount, convert the percentage into a proportion and then multiply by the total. This is a common use of percentage information and a great way to bring meaning to articles that you read.

### Example

According to the Center for Disease Control (CDC), about 32% of Americans have hypertension (high blood pressure). According to suburbanstats.org, Tulsa Oklahoma has approximately 603,403 people living in it. If the CDC is correct and 32% of Americans have hypertension, then how many people do we expect to have hypertension in Tulsa?

Step 1: Convert 32% into a decimal proportion.

$$32\% = 32 / 100 = 0.32$$

Step 2: Multiply the decimal proportion by the total.

$$\text{Amount of people with hypertension} = 0.32 \times 603403 = 193088.96$$

***Rounding Rule:** When dealing with estimated amounts, we should remember that this is the number of people or cars or objects. It sounds weird to say we estimate that 193088.96 people have high blood pressure. If an estimated amount has numbers to the right of the decimal, we prefer to round to the ones place. The ones place is the first number to the left of the decimal. So our answer should have no numbers to the right of the decimal point.*

$$\text{Amount of people with hypertension} = 0.32 \times 603403 = 193088.96 \approx 193,089$$

So approximately 193,089 people in Tulsa have high blood pressure. This is vital information for hospitals, urgent cares and doctors in the Tulsa, Oklahoma area.



## Practice Problems Section 2D

*Directions #1-7: Round the following estimated amounts to the ones place.  
(There should be no numbers to the right of the decimal.)*

1. 658.31 cars
2. 1471.83 people
3. 259.64 dogs
4. 77.42 cats
5. 314.73 bears
6. 20,246.15 car accidents
7. 10,799.622 cases of flu

*Directions #8-17: Convert the given percentages into proportions. Then use the estimated amount formula to find the estimated amounts. If your amount calculation comes out as a decimal, round your estimated amount to the ones place (no numbers to the right of the decimal).*

$$\text{Proportion} = \text{Percentage} \div 100$$

$$\text{Estimated Amount} = \text{Proportion} \times \text{Total}$$

8. According to an article by CBS news, approximately 15% of Americans still do not have health insurance. If approximately 78,300 people live in Chino Hills CA, then how many people in Chino Hills would we expect to not have health insurance? Round your answer to the ones place.
9. According to an article online, about 30% of Americans own at least one gun. About 305,700 people live in Stockton CA. If the article was accurate, then approximately how many people in Stockton do we expect to own at least one gun? Round your answer to the ones place.
10. An article by the American Diabetes Association estimates that as of 2012, about 9.3% of Americans have diabetes. College of the Canyons has approximately 18,400 students. If the percentage were correct, how many COC students would we expect to have diabetes? Round your answer to the ones place.
11. According to a news report by [www.nielsen.com](http://www.nielsen.com), about 15.9% of Americans struggle with hunger. Lancaster CA has approximately 161,000 people living in it. If the percentage from the Nielsen report is accurate, then how many people in Lancaster CA may be struggling with hunger? Round your answer to the ones place.
12. According to an article by the Autism Society, about 1.47% of people in the U.S. have autism. The article also stated that the percentage is increasing every year and that Autism is one of the fastest growing disorders in the U.S. Van Nuys, CA has approximately 136,400 people living in it. If the percentage by the Autism Society is correct, how many do we expect to have autism? Round your answer to the ones place.
13. According to a recent article, about 0.51% of airbags in the U.S. are defective. According to vehicle registration data, there are approximately 1,769,000 cars in San Francisco, CA. How many of them do we expect to have defective airbags? Round your answer to the ones place.
14. According to a recent U.S. census, about 14.8% of people in the U.S. live below the poverty line. About 305,700 people live in Stockton CA. If the census was accurate, then approximately how many people in Stockton are living in poverty? Round your answer to the ones place.
15. According to an article by the American Medical Association, approximately 33% of medical doctors in the U.S. have been sued by patients for malpractice. Suppose a hospital has currently 147 doctors on staff. How many of them do we expect to have been sued for malpractice? Round your answer to the ones place.



16. Sports Illustrated estimates that 78% of retired NFL football players are either bankrupt or under financial stress within two years of retirement. Pro-football-reference.com indicates that there are 26,682 NFL football players all time. How many of them will we expect to be bankrupt or under financial stress? Round your answer to the ones place.

17. Sports Illustrated estimates that 60% of retired NBA basketball players are broke within five years of leaving the sport. An article online claims that there are a total of 4,374 NBA players all time. How many of them do we expect to have gone broke? Round your answer to the ones place.

18. In the 2015 National School Climate Survey by GLSEN, LGBTQ students from all states in the U.S. were asked a series of questions about their experiences over one year of school. 57.6% of the LGBTQ students said they feel unsafe at school because of their sexual orientation. According to the Williams Institute at UCLA, there is an estimated to be 244,000 LGBTQ students in California between the ages of 13 and 17 years old. If the climate survey is correct, how many of them feel unsafe at school?

19. In the 2015 National School Climate Survey by GLSEN, LGBTQ students from all states in the U.S. were asked a series of questions about their experiences over one year of school. 85.2% of the LGBTQ students said they were verbally harassed (called names or threatened) at school based on a personal characteristic, sexual orientation, or gender expression. According to the Williams Institute at UCLA, there is an estimated to be 1,994,000 LGBTQ students in the U.S. between the ages of 13 and 17 years old. If the climate survey is correct, how many of them have been verbally harassed at school?

20. In the 2015 National School Climate Survey by GLSEN, LGBTQ students from all states in the U.S. were asked a series of questions about their experiences over one year of school. 48.6% of the LGBTQ students said they experienced electronic harassment (cyberbullying) via text messages or postings on social media. According to the Williams Institute at UCLA, there is an estimated to be 114,000 LGBTQ students in Florida between the ages of 13 and 17 years old. If the climate survey is correct, how many of them experienced cyberbullying?

21. In the 2015 National School Climate Survey by GLSEN, LGBTQ students from all states in the U.S. were asked a series of questions about their experiences over one year of school. 13.0% of the LGBTQ students said they were physically assaulted (punched, kicked or injured with a weapon). According to the Williams Institute at UCLA, there is an estimated to be 113,000 LGBTQ students in New York between the ages of 13 and 17 years old. If the climate survey is correct, how many of them were physically assaulted?

---



## Chapter 2 Review Sheet

### Chapter 2 Categorical Data Analysis Key Terms

- Data: Information in all forms.
- Categorical Data: Data usually made up of words that describe people or objects and compare groups.
- Quantitative Data: Numerical measurement data with units that can be used to find averages.
- Frequencies: Also called the number of successes or number of events. The number of people or objects in a categorical data set with a specific characteristic.
- Sample Size: Also called the total number of trials. The total number of people or objects in a data set.
- Percentage: An important statistic for categorical data analysis that gives the amount out of one-hundred.
- Proportion: The decimal equivalent of a percentage.

Here is a list of important ideas in this chapter.

- You should be able to distinguish between categorical data and quantitative data.
- You should be comfortable with the following terms: Frequency (count), Amount, Total, Decimal Proportion, Percentage
- You should be able to calculate a decimal proportion for a category from the frequency (amount) and the total.

$$\text{Decimal Proportion} = \frac{\text{Amount}}{\text{Total}}$$

- You should be comfortable converting a decimal proportion into a percentage and a percentage into a decimal proportion.  
Percentage => Decimal Proportion: *Remove % symbol and divide by 100.*  
Decimal Proportion => Percentage: *Multiply by 100 and put on the % symbol.*
- You should be able create bar charts for categorical data with StatKey.
- You should be more comfortable reading and using pie charts and bar charts.
- You should be able to estimate an amount from a percentage and a total.  
*Convert the percentage into a decimal proportion by dividing by 100.*  
*Amount = Decimal Proportion x Total*
- You should be more comfortable reading and understanding articles online or in print that involve percentages.
- You should be able to compute the percentage ratio and the percent of increase in order to judge if two percentages are significantly different.



### Problem Set Chapter 1 Review

Directions: Show your work and circle your answers. You will need a scientific calculator. Formulas are given below.

(#1-4) Classify the following variables as Categorical or Quantitative.

1. The amount of money spent by customers in restaurants across the San Fernando Valley.
2. Whether or not a person uses Marijuana.
3. The types of frogs in Florida.
4. The number of cattle on various cattle ranches in Nebraska.

(#5-7) To convert a percentage into a decimal proportion: Divide by 100 and take off the % symbol

5. Convert 3.85% into the equivalent decimal proportion.
6. Convert 92.6% into the equivalent decimal proportion.
7. Convert 0.51% into the equivalent decimal proportion.

(#8-10) To convert a decimal proportion into a percentage: Multiply by 100 and put on the % symbol

8. Convert the decimal proportion 0.558 into a percentage.
9. Convert the decimal proportion 0.0032 into a percentage.
10. Convert the decimal proportion 0.093 into a percentage.





(#11-12) Missy works for a shoe store and is wondering what percent of her customers prefer Adidas shoes. She asked 47 customers what their favorite shoe was and 17 said Adidas.

$$\text{Proportion} = \frac{\text{Amount}}{\text{Total}}$$

$$\text{Percentage} = \frac{\text{Amount}}{\text{Total}} \times 100\%$$

11. What is the decimal proportion of the customers that prefer Adidas? Round your answer to the thousandths place (3<sup>rd</sup> decimal to the right of the decimal point)

12. What percent of the customers prefer Adidas? Round your percentage answer to the tenths place (1<sup>st</sup> decimal to the right of the decimal point)

(#13-14) According to an article by [www.who.int](http://www.who.int), people with HIV are highly susceptible to Tuberculosis. In fact, they say that approximately 33.3% of HIV deaths are from Tuberculosis.

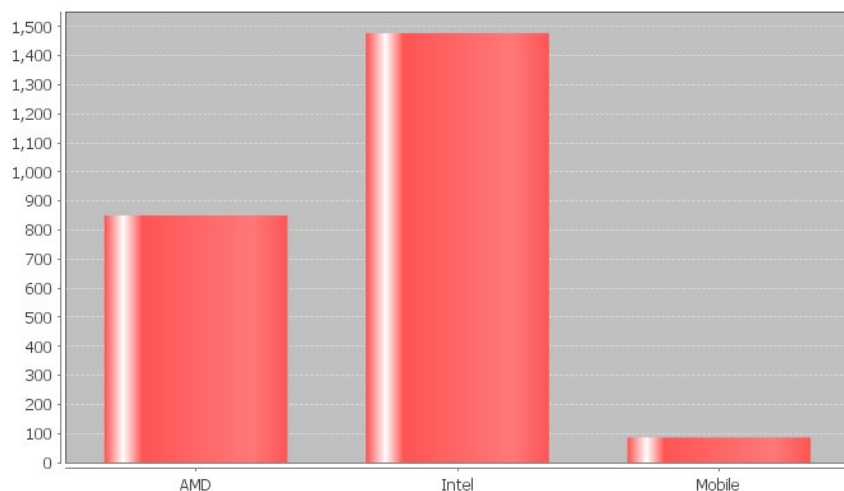
13. Convert 33.3% into a decimal proportion. (Divide by 100 and take off the % symbol.)

14. If a hospital has 58 HIV deaths, how many do we expect to be from Tuberculosis? (Round your answer to the ones place)

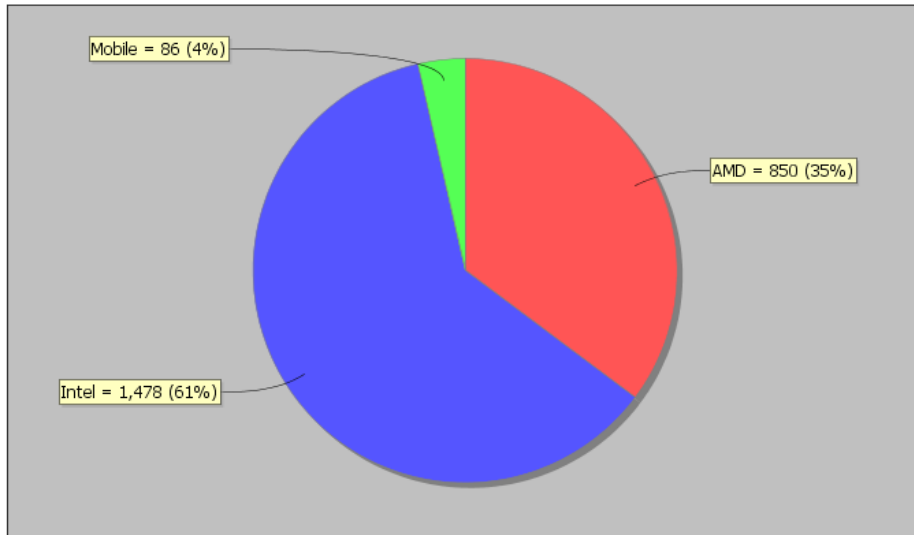
$$\text{Amount} = \text{Decimal Proportion} \times \text{Total}$$

(#15-18). Three of the largest producers of computer CPU's (central processing units) worldwide are AMD, Intel, and Mobile. Use the bar plot and pie chart to answer the following questions.

**Bar Chart**



**Pie Chart**



15. How many different processors are made by Intel?
  16. How many different processors are made by AMD?
  17. What percentage of the CPU's are made by Mobile?
  18. What percentage of the CPU's are made by AMD?
  19. What percentage of the CPU's are made by Intel?
  20. Which of the three companies makes the most CPU's?
  21. Which of the three companies makes the least CPU's?
  22. Calculate the percentage ratio for Intel and AMD. Is there a significant difference between the percentages of processors made by the two companies? Explain why.
- 



## Project Chapter 2 - Categorical Data Analysis

**Directions for Online Classes:** *This will be an individual project. Each student will analyze one categorical data set from the math 075-survey data fall 2015 create a poster summarizing their findings. Students can chose from the following columns of data: Tattoo, Texting While Driving, Favorite Social Media, Transportation to School, Car Accident, Cigarettes, Eat Breakfast, Glasses/Contacts, High School in Santa Clarita, Living with parents*

*After submitting the project to their instructor, students will then go to the “Chapter 2 Project Class Discussion” in Canvas and discuss their findings with other students in the class.*

### The Individual Poster Should Have

- The poster does not have to be extremely large.
- Your first and last name on the poster
- What column of data did you chose?
- A few sentences explaining why this data is important or interesting to you?
- Go to StatKey at [www.lock5stat.com](http://www.lock5stat.com) and enter the column of data under “One Categorical Variable” in the “Descriptive Statistics” menu. Then enter the column of data under “edit data”. Remember to check the box that says “raw data” before pushing “OK”. Also check “header row” if data has a title.
- Copy the “Summary Statistics” table from StatKey onto your poster in large letters.
- Draw the bar chart onto your poster.
- Convert all of the proportions listed on the Summary Statistics table into percentages.
- Use the percentages to draw a Pie chart on your poster. Make sure to label each piece of the pie with the name, count and percentage.
- Does you think that the percentages are significantly different?
- Do any of the percentages seem unusual or surprising?
- Can you think of any reasons why the percentages are different?
- Decorate Poster

Now take a picture of your poster project and submit the picture to your instructor in Canvas.

After submitting the picture of the poster, go to the discussion menu in Canvas and complete the “Chapter 3 Project Discussion”. You will be discussing your findings with other students in the class.

---



**Directions for Face to Face Classes:** *The class will be separated into groups. Each group is required to pick a “team name” for their group and analyze one column of categorical data from the math 075-survey data fall 2015, create a poster summarizing their findings, and present the poster to other students in the class.*

*Each group will have a different topic and will pick one of the following data sets to present it to their classmates: Tattoo, Texting While Driving, Favorite Social Media, Transportation to School, Car Accident, Cigarettes, Eat Breakfast, Glasses/Contacts, High School in Santa Clarita, Living with parents*

#### The Group Poster Should Have

- **The Team Name**
- **First and Last Name of each team members on the poster**
- **What column of data did your group chose?**
- **A few sentences explaining why this data is important or interesting to your group?**
- **Go to StatKey at [www.lock5stat.com](http://www.lock5stat.com) and enter the column of data under “One Categorical Variable” in the “Descriptive Statistics” menu. Then enter the column of data under “edit data”. Remember to check the box that says “raw data” before pushing “OK”. Also check “header row” if data has a title.**
- **Copy the “Summary Statistics” table from StatKey onto your poster in large letters.**
- **Draw the bar chart onto your poster.**
- **Convert all of the proportions listed on the Summary Statistics table into percentages.**
- **Use the percentages to draw a Pie chart on your poster. Make sure to label each piece of the pie with the name, count and percentage.**
- **Does your group think that the percentages are significantly different?**
- **Do any of the percentages seem unusual or surprising?**
- **Can your group think of any reasons why the percentages are different?**
- **Decorate Poster**

#### Presentation

*Make sure each person on the team understands the poster and can present your findings. Bring your poster to a designated presentation area in the classroom and hang or tape your poster to a wall. One person at a time will present the poster. We will then rotate so that each member of the team gets to present. Everyone else will listen to presentations and give feedback with sticky notes. (Be Nice!)*

---



## Chapter 3 – Relationships between Categorical Variables

**Introduction:** An important field of exploration when analyzing data is the study of relationships between variables. A lot of thought has been put into determining which variables have relationships and the scope of that relationship. Is a person's diet related to having high blood pressure? Is the city a person lives in related to whether or not they have tuberculosis? Is being in a car accident related to texting while driving? These are all important questions that statisticians, data analysts and data scientists explore.

We can study relationships between two categorical variables like texting while driving and having a car accident. We can also study relationships between two quantitative variables like the weight of a person and their blood pressure. A third relationship we can study is the relationship between a categorical variable and a quantitative variable. For example, we can study the relationship between the type of job you have and your annual income. In this chapter, we will begin to explore the relationships between two categorical variables.

Remember, statistics is a deep well of mathematics and knowledge learned by years of study. There are much more advanced techniques for studying relationships, but we will be focusing on a basic introduction to the topic. You will find that a good understanding of this chapter will help tremendously when you go on to the more advanced techniques later on. For example, I find my advanced statistics students do not understand the Chi-Square distribution because they lack the foundational understanding of contingency tables and analyzing the differences between categories.

**Note on Terminology:** *When studying relationships between variables you will hear different words used to describe the relationship. The most common are "relationship", "association", or "correlation". "Correlation" is often used for describe a relationship between two quantitative variables, while "relationship" and "association" are used for two categorical variables or for a categorical - quantitative relationship study.*

*In this chapter, we will be using the terms "relationship" and "association".*

**Note on Causation:** *One of the most famous statements in statistics is that "correlation is not causation". Proving that one thing causes another is a much more complex kind of study and involves controlling confounding variables and experimental design. The main thing to remember is that just because there is a relationship, that does not prove causation. There may be many other factors involved.*

---



## Section 3A – Contingency Tables with Technology

When studying relationships between categorical variables, we start by creating a contingency table. Some people refer to this table as a “two-way” table, but contingency table is more common. A contingency table is a summary of counts or frequencies for two categorical data sets. Let us look at the hospital data again from the last chapter.

### Example 1

Patient ID#	Age	Gender	Blood Type	Rh Factor	Floor
1	23	M	A	-	SDS
2	68	M	O	+	ER
3	51	F	AB	+	Med/Surg
4	74	M	O	-	ICU
5	49	F	O	+	SDS
6	62	F	O	+	Med/Surg
7	35	M	A	+	SDS
8	46	F	O	+	Med/Surg
9	72	F	O	+	ER
10	61	M	B	+	SDS
11	43	F	A	-	Med/Surg
12	81	M	O	+	ICU
13	65	M	A	+	Med/Surg
14	59	F	O	-	SDS
15	44	F	B	+	ICU
16	26	M	O	+	ER
17	58	F	AB	-	ER
18	45	M	O	+	SDS
19	55	M	O	+	Med/Surg
20	71	M	A	+	ER

Suppose we want to analyze the relationship and proportions for a patient’s gender and their blood type. Notice gender is one categorical variable with two options (male and female). Blood type is another categorical variable with four options (A, B, AB, and O). To make a contingency table, pick one of the variables to be the row and the column. I am going to pick gender to be my rows and blood type to be my columns. Since there are two options for the rows and four options for the columns, we will have a “2 by 4” table (2 rows and 4 columns, not counting totals).

	Type A	Type B	Type AB	Type O
Female				
Male				

Now we just need to count and fill out the table. It should be noted that no data analyst or statistician does this by hand. All use either excel or a statistics software. Remember we live in the age of “big data”. No one wants to count variables in a data set with twenty thousand values, and that is not even “big”.

Since we are introducing the topic, see if you can count the amount for each box. You can use tally marks if you wish. Where the “Female” row meets the “Type A” column we should put how many female patients had type A blood. (There was only one.) Where the “Male” row meets the “Type O” column we should put how many male patients had type O blood. (There was six.)

	Type A	Type B	Type AB	Type O
Female	1			
Male				6



See if you can find the rest of the counts (frequencies) for the table.

You should get the following table. There were twenty patients so the numbers in the two-way table should add up to twenty. This is called the “grand total”. Also, notice there were no males with type AB blood, so we needed to put a zero in that cell.

	Type A	Type B	Type AB	Type O
Female	1	1	2	5
Male	4	1	0	6


Before we can analyze the relationship and proportions, we need to calculate all the row and column totals. This is automatically done with excel or statistics software programs. Notice the “grand total” is always in the bottom right corner of the table. Keep in mind that this is still considered a two-by-four table. Totals are not included in the size of a table.

	Type A	Type B	Type AB	Type O	Total
Female	1	1	2	5	9
Male	4	1	0	6	11
Total	5	2	2	11	Grand Total = 20

Notice a few things about this table. The row totals (9 and 11) add up to the grand total (20). Also the column totals (5, 2, 2, and 11) add up to the grand total. Be careful. The row totals plus the column totals does not add up to the grand total.

[Creating a contingency table with raw data and StatKey](#)

Let us look at an example. Go to [www.matt-teachout.org](http://www.matt-teachout.org) or Canvas and click on the “math 075 survey data fall 2015”. We want to explore the relationship between the campus a person goes to and their political party.

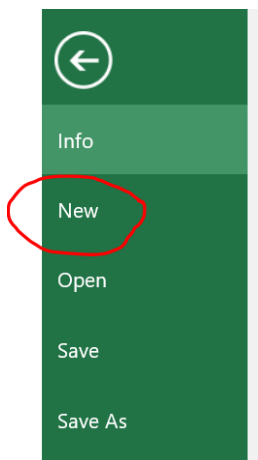
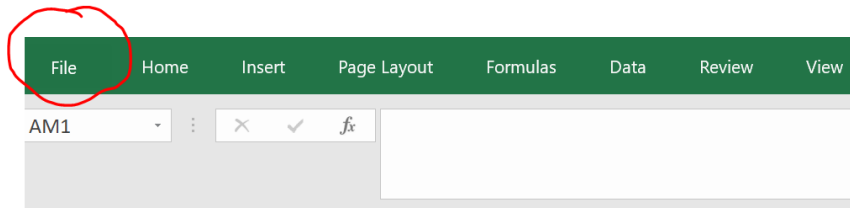


A	B	C	D	E
Campus	Gender	Age (in years)	Month of birthday	Weight (in pounds)
Canyon Country Campus	Female	20	4	144
Canyon Country Campus	Female	19	3	120
Canyon Country Campus	Female	50	10	135
Canyon Country Campus	Male	22	1	155
Canyon Country Campus	Female	25	6	125
Valencia Campus	Male	18	10	180
Valencia Campus	Male	20	5	155
Valencia Campus	Male	19	6	172
Valencia Campus	Female	19	5	135
Valencia Campus	Female	17	10	149
Valencia Campus	Female	18	11	106
Valencia Campus	Male	19	4	165
Valencia Campus	Male	18	12	250

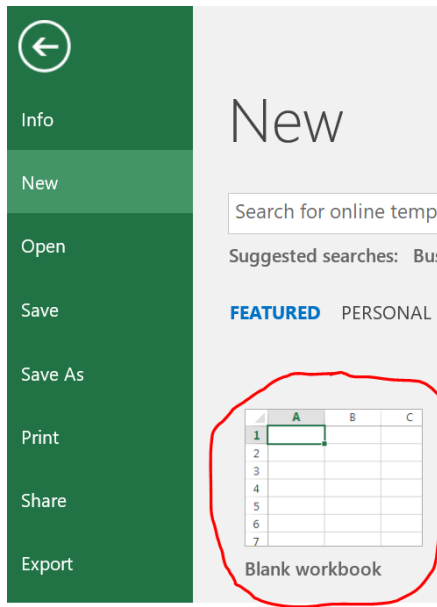


Z	AA	AB
Number alcoholic beverages (per week on average)	Political Party	Math Intimidation? (on scale of 1-10)
0	Democratic	10
0	Democratic	7
1	Democratic	8
0	Other	4
6	Other	6
0	Other	3
0	Republican	3
0	Republican	7
0	Democratic	7
0	Republican	5
0	Democratic	7
0	Democratic	1
0	Democratic	2
0	Republican	7
0	Other	6
0	Independent	7

First, we will need to check the data. When exploring relationships between two data sets, the data needs to be ordered pair. This usually means the data came from the same people. In this data values in the same row came from the same math 075 pre-stat student. We also need to be careful of blanks. This means a person did not answer one or both of the questions. Start by copy and pasting the campus data and political party data into a new spreadsheet. In excel, you would go to the “file” menu on the top left corner and then press “new” and click on the “blank workbook”.







A good rule of thumb is never mess up an original data set. Always copy and paste into a new excel file if you want to change things. The two columns of categorical data need to be in next to each other in the new spreadsheet. Otherwise, StatKey will not accept it. The larger the data set, the more difficult it is to copy and paste columns of data. In Excel, if you hold your cursor at the title at the top of the column you will see it turn into a downward arrow ↓. When you see the downward arrow, left click and it will highlight the entire column. You can also click and drag, but the larger the data set, the longer it will take to drag to the very bottom. I prefer the downward arrow method. Once the data set is highlighted, hold the control key down and your keyboard and push “C”. Control C is an easy way to copy. You can also right click and push copy. Your new spreadsheet should now look like this. Be careful. Did you copy and paste all of the data? Your columns should go all the way to row 460!

	A	B
1	<b>Campus</b>	<b>Political Party</b>
2	Canyon Country Campus	Democratic
3	Canyon Country Campus	Democratic
4	Canyon Country Campus	Democratic
5	Canyon Country Campus	Other
6	Canyon Country Campus	Other
7	Valencia Campus	Other
8	Valencia Campus	Republican
9	Valencia Campus	Republican
10	Valencia Campus	Democratic
454	Canyon Country Campus	Democratic
455	Canyon Country Campus	Democratic
456	Canyon Country Campus	Republican
457	Canyon Country Campus	Independent
458	Canyon Country Campus	Democratic
459	Canyon Country Campus	Democratic
460	Canyon Country Campus	Independent
461		

Go through the data and make sure there are no blanks. If there is a blank, delete that entire row. If you remember from chapter 1, this is called non-response bias. This process of deleting out missing cells is sometimes called “cleaning the data”. This data did not seem to have any blanks that needed deleting.

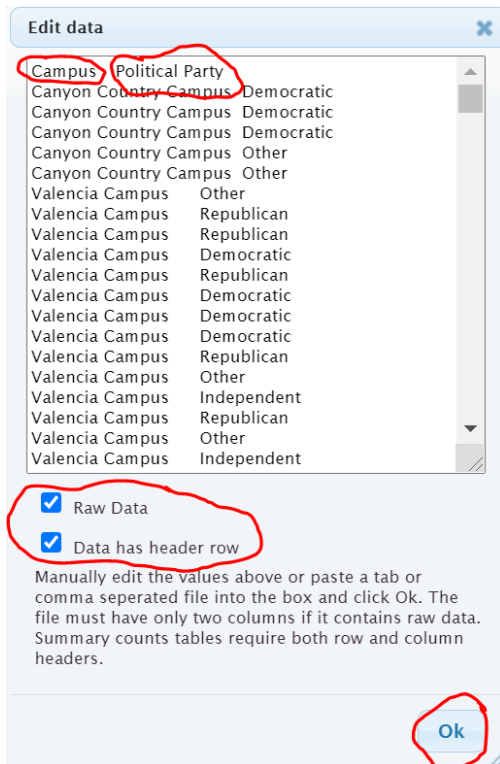


To make a contingency table with StatKey, go to [www.lock5stat.com](http://www.lock5stat.com) and click the “StatKey” button. Now click on “Two Categorical Variables” under the “Descriptive Statistics and Graphs” menu. Click on the “edit data” button. If there is data already there, push “Control A” on your keyboard and then delete. Make sure your cursor is at the very top of the edit data field.



Now we will copy and paste our data into StatKey. Remember to hold your cursor right above the column you want to copy until you see the downward arrow and then left click. Hold the control key down on your keyboard and do the same thing for the second column. Now push “Control C”. Both columns are now copied together.

Then go back to the “edit data” field in the “Two Categorical Variables” menu in StatKey and paste the columns into StatKey.



It is important to know what type of data you have put in. The data we pasted is not a list of the summary counts. This data is the actual column of words. We call that “raw categorical data”. So we will need to check the box that says “raw data”. It is also to note whether the titles are included in the data we pasted. We see the titles “campus” and “political party” at the top of our data. StatKey refers to titles as “header rows”. Since our titles are there, we will need to check the box that says “data has header row”. Now push “OK”.

Note: If the data did not have the titles, we should uncheck the box that says “data has a header row”. If this was summary counts of our categorical data and not the actual column of words, we would also uncheck the box that says “raw data”.



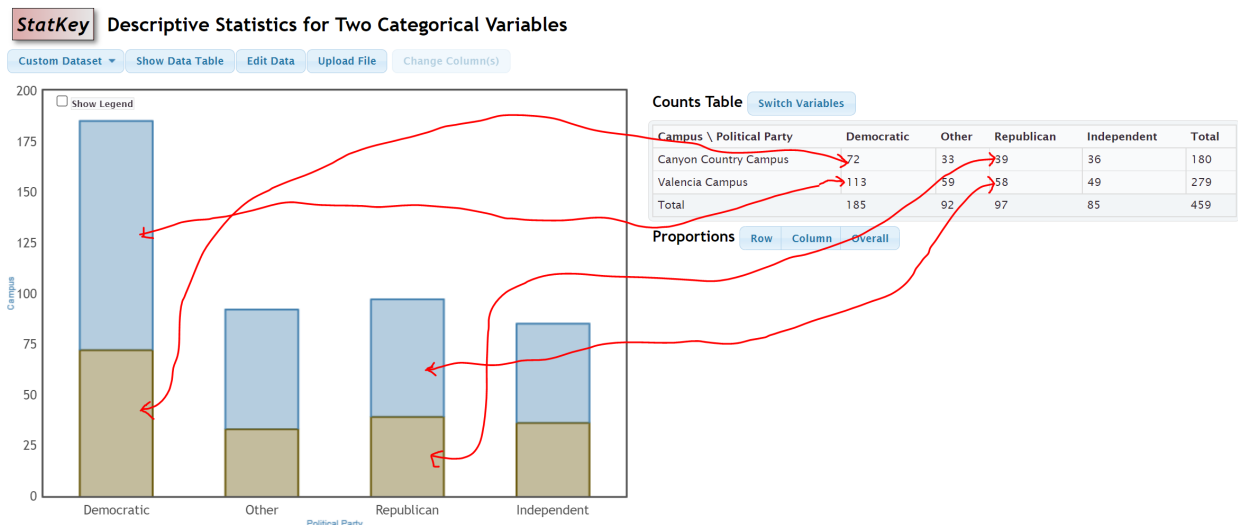
You should now see the contingency table. StatKey refers to the table as a “Counts Table”. Notice StatKey has counted all of the categorical variables for us. We know there were 72 democrats that went to the Canyon Country campus. We know there was 58 republican students that went to the Valencia campus. We know there was a total of 92 students that identified as supporting a political party other than democrat, republican or independent. We also see the grand total of 459 students. (Remember your columns of data had 460 rows. That is because the first row was the title.)

## Counts Table [Switch Variables](#)

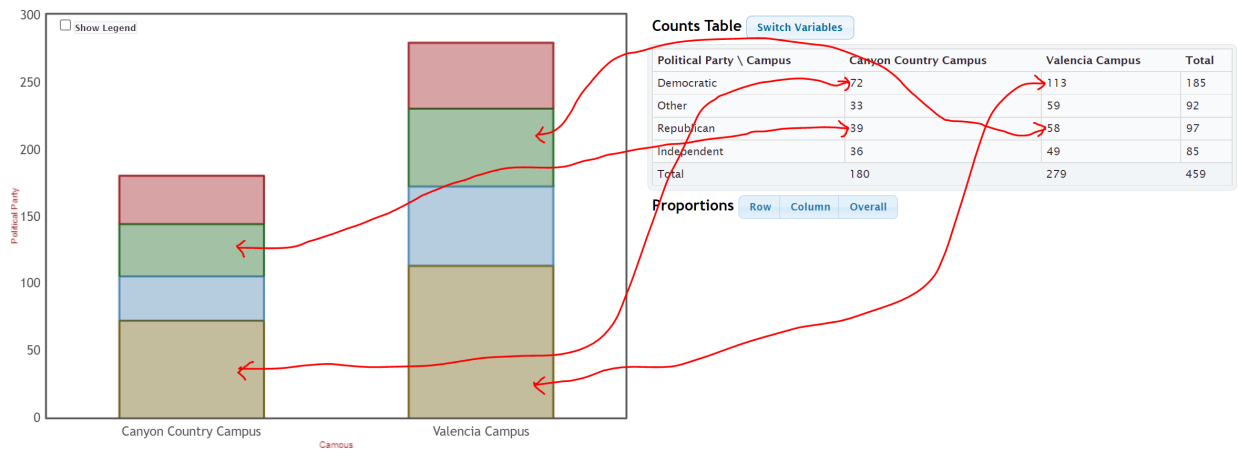
Campus \ Political Party	Democratic	Other	Republican	Independent	Total
Canyon Country Campus	72	33	39	36	180
Valencia Campus	113	59	58	49	279
Total	185	92	97	85	459

The size of a contingency table is the number of rows by the number of columns. Totals are not included. This table has two rows (CCC and Valencia) and four columns (Independent, Other, Republican, and Democratic), so this is a “2 by 4” or “2×4” contingency table.

StatKey has several cool features with the contingency table. Notice it has created a stacked bar chart. This stacked bar chart gives a visual representation of a contingency table. Notice if you place your cursor on any section of the graph the corresponding count lights up in the contingency table.



Another feature is the “switch variables” button in StatKey. Clicking on this button will switch the rows and columns. So the rows will now be political party and the columns will be campus. The stacked bar chart will also adjust to the new switched contingency table. Notice all of the counts are the same.



Most statistics software programs can make contingency tables. Here is what the contingency tables would look like in Statcato. Notice the counts are identical to the tables with StatKey.

	Democratic	Independent	Other	Republican	All
Canyon Country Campus	72	36	33	39	180
Valencia Campus	113	49	59	58	279
All	185	85	92	97	459

	Canyon Country Campus	Valencia Campus	All
Democratic	72	113	185
Independent	36	49	85
Other	33	59	92
Republican	39	58	97
All	180	279	459

### Putting an existing contingency table into StatKey

Suppose you have an existing contingency table that you want to put into StatKey in order to create the stacked bar chart. You can go to “Two categorical Variables” under the “Descriptive Statistics and Graphs” menu. Click on edit data and type in the contingency table. This is no longer “raw data”. It is the counts of the categories. So we will NOT check the box that says “raw data”. Every program has a different way of typing in data. StatKey uses commas.

#### Example

Earlier in this section we created a contingency table by counting patients in terms of their gender and blood type. We can type this contingency table into StatKey using commas. It must be typed in a certain way though.

	Type A	Type B	Type AB	Type O
Female	1	1	2	5
Male	4	1	0	6



Here is what we would type. We need [blank] in the top left corner corresponding to the empty cell in the top left corner. Comma means we are going to the next cell in that row. There should only be one space after the comma. You also want to avoid typing other punctuation marks. Never type in the totals. Those are calculated automatically. This is not raw categorical data. That would be columns of words. This is a contingency table with summary counts. So we should NOT check the box that says "Raw Data". It does have titles at the top, so we will check the box that says "Data has header row".

[blank], Type A, Type B, Type AB, Type O

Female, 1, 1, 2, 5

Male , 4, 1, 0, 6

**Edit data**

[blank], Type A, Type B, Type AB, Type O  
Female, 1, 1, 2, 5  
Male , 4, 1, 0, 6

Raw Data **NO!!**

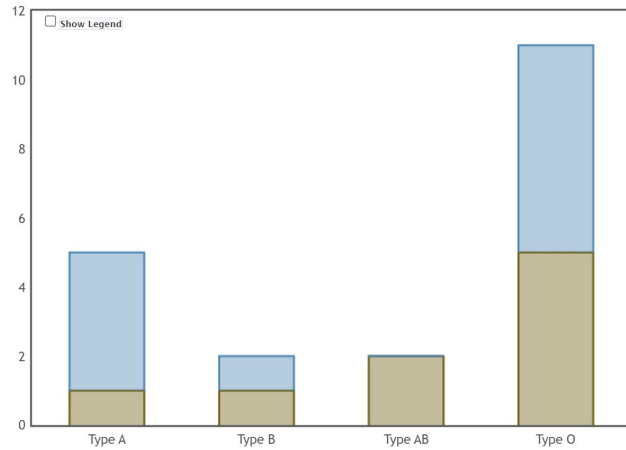
Data has header row

Manually edit the values above or paste a tab or comma separated file into the box and click Ok. The file must have only two columns if it contains raw data. Summary counts tables require both row and column headers.

Ok

Once we press OK, we have the contingency table in StatKey and the stacked bar chart is automatically created. Notice that even though we did not type them, the totals are now there.



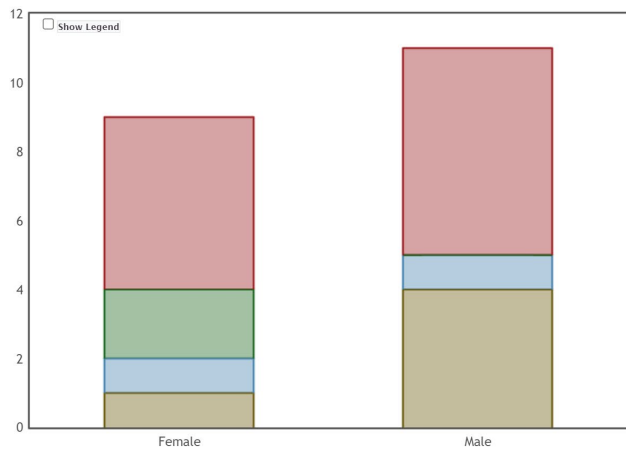


Counts Table [Switch Variables](#)

undefined \ undefined	Type A	Type B	Type AB	Type O	Total
Female	1	1	2	5	9
Male	4	1	0	6	11
Total	5	2	2	11	20

Proportions [Row](#) [Column](#) [Overall](#)

We can also press the “Switch Variables” button to switch the blood type to the rows and gender to the columns if we prefer.



Counts Table [Switch Variables](#)

undefined \ undefined	Female	Male	Total
Type A	1	4	5
Type B	1	1	2
Type AB	2	0	2
Type O	5	6	11
Total	9	11	20

Proportions [Row](#) [Column](#) [Overall](#)



## Practice Problems Section 3A

Directions #1-4: Here is some data taken from the medical records department at a local hospital. The data includes gender, blood type (A, B, AB, O), Rhesus factor (Rh + or Rh -) and part of the hospital the patient was in (Medical/Surgical, Intensive Care Unit , Same Day Surgery, Emergency Room).

Gender	Blood Type	Rh Factor	Floor
M	A	-	SDS
M	O	+	ER
F	AB	+	Med/Surg
M	O	-	ICU
F	O	+	SDS
F	O	+	Med/Surg
M	A	+	SDS
F	O	+	Med/Surg
F	O	+	ER
M	B	+	SDS
F	A	-	Med/Surg
M	O	+	ICU
M	A	+	Med/Surg
F	O	-	SDS
F	B	+	ICU
M	O	+	ER
F	AB	-	ER
M	O	+	SDS
M	O	+	Med/Surg
M	A	+	ER

1. Create a contingency table that we could use to compare Rh factor (Rh+ or Rh-) to blood type (A,B,AB or O). Make the rows represent the Rh factor and the columns represent the blood type. Label the rows and columns with titles and include the grand total and all of the row and column totals in your table. What is the size of the table (# rows by # columns) not counting totals?
2. Create a contingency table that we could use to compare gender to the part of the hospital the patient went to. Make the rows represent gender and the columns represent the part of the hospital. Label the rows and columns with titles and include the grand total and all of the row and column totals in your table. What is the size of the table (# rows by # columns) not counting totals?
3. Create a contingency table that we could use to compare the Rh factor (Rh+ or Rh-) to the part of the hospital the patient went to (SDS, ER, MedSurg, ICU). Make the rows represent the Rh factor and the columns represent the part of the hospital. Label the rows and columns with titles and include the grand total and all of the row and column totals in your table. What is the size of the table (# rows by # columns) not counting totals?
4. Create a contingency table that we could use to compare the part of the hospital (SDS, ER, MedSurg, ICU) to the blood type (A,B,AB or O). Make the rows represent the blood type and the columns represent the part of the hospital. Label the rows and columns with titles and include the grand total and all of the row and column totals in your table. What is the size of the table (# rows by # columns) not counting totals?



Directions #5-8: Open the “Math 075 Survey Data Fall 2015” in Canvas or at [www.matt-teachout.org](http://www.matt-teachout.org). Use StatKey to create a contingency table and stacked bar chart for the following variables. Make a rough sketch of the stacked bar chart and table on your paper. Then use the table and graph to answer the questions.

Directions for creating contingency table with StatKey:

- Open the “Math 075 Survey Data Fall 2015”. Copy and paste the two columns next to each other in a new spreadsheet. Then copy both columns together.
- Go to [www.lock5stat.com](http://www.lock5stat.com) and click on “StatKey”. Under the “Descriptive Statistics and Graphs” menu, click on “Two Categorical Variables”. Click the “Edit Data” button. Push “Control A” and “Delete” on your keyboard to delete out any existing data. Then paste in your two columns of data. Check the box that says “Raw Data”. If your data has a title, check the box that says “Data has a header row”. Then push “OK”. If your rows and columns are backward, push the “Switch Variables” button.

5. Use StatKey and the “Math 075 Survey Data Fall 2015” to create a contingency table and stacked bar chart for campus (Valencia or Canyon Country) and at least one tattoo (yes or no). Let the rows represent tattoo status and let the columns represent the campus.

- Draw a sketch of the contingency table including titles and totals.
- Draw a sketch of the stacked bar chart.
- What was the grand total?
- How many total students went to the Valencia campus?
- How many total students have at least one tattoo?
- How many students both did not have a tattoo and went to the Canyon Country campus?

6. Use StatKey and the “Math 075 Survey Data Fall 2015” to create a contingency table and stacked bar chart for contact lenses or glasses (yes or no) and hair color (brown, black, blond(e), red, other). Let the rows represent glasses/contacts status and the columns represent hair color.

- Draw a sketch of the contingency table including titles and totals.
- Draw a sketch of the stacked bar chart.
- What was the grand total?
- How many total students need contacts or glasses?
- How many total students have brown hair?
- How many students both did not need glasses and have black hair?

7. Use StatKey and the “Math 075 Survey Data Fall 2015” to create a contingency table and stacked bar chart for texting while driving (yes or no) and being in a car accident (yes or no). Let the car accident status represent the rows and texting while driving represent the columns.

- Draw a sketch of the contingency table including titles and totals.
- Draw a sketch of the stacked bar chart.
- What was the grand total?
- How many total students said that do not text and drive? Do you believe them?
- How many total students have not been in a car accident?
- How many students both text and drive and have been in a car accident?





8. Use StatKey and the “Math 075 Survey Data Fall 2015” to create a contingency table and stacked bar chart for live with parents (yes or no) and political party (democrat, republican, independent, other). Let the political party represent the rows and living with parents status represent the columns.

- a) Draw a sketch of the contingency table including titles and totals.
  - b) Draw a sketch of the stacked bar chart.
  - c) What was the grand total?
  - d) How many total students do not live with their parents?
  - e) How many total students identify with independent political party?
  - f) How many students are both democrat and live with their parents?
- 



## Section 3B – Marginal and Joint Percentages from Contingency Tables

Analyzing two categorical data sets involves not only creating contingency tables, bar charts and pie charts, but also being able to find and analyze proportions and percentages.

Remember that a proportion is found by taking the amount (frequency) and dividing by the total (sample size).

$$\text{Proportion} = \frac{\text{Amount (Frequency)}}{\text{Total}}$$

To convert that proportion into a percentage, simply multiply the proportion by 100%.

### Marginal Percentages

Let us start with looking at basic marginal proportions. These are proportions where the amount involves only a single variable and the total is everyone in the data (grand total).

Look at the following contingency table created with StatKey from the Fall 2015 Math 075 Survey data. This table describes the relationship between smoking and political party for Math 075 pre-stat students.

#### Counts Table [Switch Variables](#)

Smoke cigarettes? \ Political Party	Democratic	Other	Republican	Independent	Total
No	176	82	90	78	426
Yes	9	10	7	7	33
Total	185	92	97	85	459

Remember, analyzing data involves asking questions and finding the answers to those questions.

For example. Here are a few questions that came to mind when I looked at this table.

#### [Example 1](#)

What percentage of the pre-stat students smoke cigarettes?

Notice we are looking at all of the students (not just democrats), so we should use the grand total as our total. Where do we find the amount of pre-stat students that smoke cigarettes? Smoking cigarettes (yes) is a row, so we should look in the margin at the total for that row.

#### Counts Table [Switch Variables](#)

Smoke cigarettes? \ Political Party	Democratic	Other	Republican	Independent	Total
No	176	82	90	78	426
Yes	9	10	7	7	33
Total	185	92	97	85	459

Notice the amount and the grand total are found in the margins where the totals are. This is why this is often called a “marginal proportion” or a “marginal percentage”. Notice the marginal percentage only involves one variable (smoking) and does not include political party.



Proportion of students that smoke = Amount of Smokers  $\div$  Grand Total =  $33 \div 459 = 0.07189524 \approx 0.072$

Percentage of students that smoke  $\approx 0.072 \times 100\% = 7.2\%$

### Example 2

What percentage of the pre-stat students identified as other political party?

Notice we are looking at all of the pre-stat students, so we should use the grand total again as our total.

Where will we find the amount of pre-stat students that support “other” political party? Other political party is a column so we will have to look at the total for that column.

### Counts Table

[Switch Variables](#)

Smoke cigarettes? \ Political Party	Democratic	Other	Republican	Independent	Total
No	176	82	90	78	426
Yes	9	10	7	7	33
Total	185	92	97	85	459

Proportion of students that support other political party = Amount of other political party  $\div$  Grand Total =  $92 \div 459 = 0.200435729 \approx 0.200$

Percentage of students that are other political party  $\approx 0.200 \times 100\% = 20.0\%$

Notice we only looked at one variable (other political party), and the amount of students that identified as other political party and the grand total were both found in the margins. So this is again a “marginal percentage”.

Note: Some students may ask why we did not write the answer as 0.2 or 20%. These are equivalent to 0.200 and 20.0%, but these answers tell us that the answer was rounded to three significant figures.

### Formula

**Single Variable Marginal Proportion = Total for Row or Column  $\div$  Grand Total**

### Joint Percentages

Sometimes we want to find a proportion or percentages where the amount (frequency) involves more than one variable. These are often called “joint proportions” or “joint percentages”.

There are two types of joint proportions.

AND: This is when we want to know the proportion or percentage involving two things being true about a person or object.

OR: This is when we want to know the proportion of percentage involving either one variable or another variable being true about the person or object.

Let us look at the political party and cigarette data again.



### Example 3 ("AND" Joint %)

What percentage of all the pre-stat students both smoked cigarettes and were Republican?

Notice there are two variables involved, republican and smoking. The key though is that we want the proportion for both things being true about the person. We cannot look at only smokers and we cannot look at only republicans. We need the amount of smoking republicans. This is a classic "AND" proportion since both things need to be true about the student.

Notice also we are picking from all pre-stat students, so our total should be the grand total again.

#### Counts Table Switch Variables

Smoke cigarettes? \ Political Party	Democratic	Other	Republican	Independent	Total
No	176	82	90	78	426
Yes	9	10	7	7	33
Total	185	92	97	85	459

Notice that to find the smoking republicans, we need to look where the republican column meets the yes smoking row. This is why "AND" proportions are often referred to as an intersection.

Proportion of pre-stat students that both smoke cigarettes and are republican =

$$\text{amount of smoking republicans} \div \text{grand total} = 7 \div 459 \approx 0.015250544 \approx 0.015$$

Percentage of pre-stat students that both smoke and are republican  $\approx 0.015 \times 100\% \approx 1.5\%$

#### Formula

**"AND" Intersecting Proportion = Amount where row and column intersect  $\div$  Grand Total**

### Example 4 ("OR" Joint %)

Suppose we only wanted to know the percentage of students that either smoke or are republican. (Not both)

This would be a classic "OR" joint proportion. The key is that we will now need to include everyone that smokes, as well as everyone that is republican. This is why an OR joint proportions are often referred to as a union. When calculating an "OR" joint proportion, you will need to do some adding to find the amount.

#### Counts Table Switch Variables

Smoke cigarettes? \ Political Party	Democratic	Other	Republican	Independent	Total
No	176	82	90	78	426
Yes	9	10	7	7	33
Total	185	92	97	85	459



Proportion of students that either smoke or are republican =  
amount of students that either smoke or are republican / grand total

$$= (90 + 9 + 10 + 7 + 7) \div 459 = 123 \div 459 \approx 0.267973856 \approx 0.268$$

Percentage of students that either smoke or are republican  $\approx 0.268 \times 100\% \approx 26.8\%$

**Important Note:** Notice that we did not use the row and column totals when calculating an “OR” joint proportion. If we added the total for smokers (33) plus the total for republicans (97), we would have gotten 130 as our amount. This would be wrong. The correct amount was 123. Adding the row and column totals gives you the wrong answer because we would have added the 7 smoking republicans twice.

Here are some other formulas that may be used to calculate an OR (union) proportion.

### Formulas

“OR” Union Proportion = Add up all of the values in the row or column without using totals  $\div$  Grand Total

“OR” Union Proportion = (Row Total + Column Total – Intersection amount)  $\div$  Grand Total

“OR” Union Proportion = 1<sup>st</sup> Variable Proportion + 2<sup>nd</sup> Variable Proportion – Intersecting “AND” Proportion

In the previous example here is how we could have used the other formulas to get the same answer.

What proportion of the pre-stat students either smoke cigarettes or are republican?

### Counts Table Switch Variables

Smoke cigarettes? \ Political Party	Democratic	Other	Republican	Independent	Total
No	176	82	90	78	426
Yes	9	10	7	7	33
Total	185	92	97	85	459

“OR” Union Proportion = (Row Total + Column Total – Intersection amount)  $\div$  Grand Total

$$= (97 + 33 - 7) \div 459 = 123 \div 459 = 123 \div 459 \approx 0.267973856 \approx 0.268$$

Percentage of students that either smoke or are republican  $\approx 0.268 \times 100\% \approx 26.8\%$

Notice we got the same answer as before.

“OR” Union Proportion = 1<sup>st</sup> Variable Proportion + 2<sup>nd</sup> Variable Proportion – Intersecting “AND” Proportion

= Proportion Smoke + Proportion Republican – Proportion that smoke and are Republican

$$= \frac{33}{459} + \frac{97}{459} - \frac{7}{459} \approx 0.072 + 0.211 - 0.015 = 0.268 = 26.8\%$$

Notice we got the same answer as before. This formula is particularly useful, especially when a statistics program calculates the marginal and intersecting proportions for you.



## Calculating Marginal and Joint Proportions with StatKey

StatKey can calculate the marginal and intersecting proportions for you. Under the “Counts table” (Contingency Table) you will see a “Proportions” menu. Click the button that says “Overall”. We put the smoking and political party columns from the Math 075 Summary Data Fall 2015 into StatKey.

### Counts Table Switch Variables

Smoke cigarettes? \ Political Party	Democratic	Other	Republican	Independent	Total
No	176	82	90	78	426
Yes	9	10	7	7	33
Total	185	92	97	85	459

### Proportions Row Column Overall

Smoke cigarettes? \ Political Party	Democratic	Other	Republican	Independent	Total
No	0.383	0.179	0.196	0.17	0.928
Yes	0.02	0.022	0.015	0.015	0.072
Total	0.403	0.2	0.211	0.185	1

When you click the Overall button, StatKey calculates the marginal proportions and the “AND” intersecting proportions. However, it does not calculate the “OR” union proportions.

Our first example in this section asked what percentage of the pre-stat students smoke cigarettes. Yes (smoking) is a row in this table so we just need to look at the margin (end of the row) to find the answer. Notice it says ) 0.072 or 7.2%. This is the same answer we calculated earlier in the section.

### Counts Table Switch Variables

Smoke cigarettes? \ Political Party	Democratic	Other	Republican	Independent	Total
No	176	82	90	78	426
Yes	9	10	7	7	33
Total	185	92	97	85	459

### Proportions Row Column Overall

Smoke cigarettes? \ Political Party	Democratic	Other	Republican	Independent	Total
No	0.383	0.179	0.196	0.17	0.928
Yes	0.02	0.022	0.015	0.015	0.072
Total	0.403	0.2	0.211	0.185	1

Earlier in this section we asked what percentage of pre-stat students both smoke cigarettes and are republican. To find this answer we just need to go to where Yes (smoking) and Republican intersect. Notice the answer is given as 0.015 or 1.5%. This is again the same answer we calculated earlier in the section.



## Counts Table Switch Variables

Smoke cigarettes? \ Political Party	Democratic	Other	Republican	Independent	Total
No	176	82	90	78	426
Yes	9	10	7	7	33
Total	185	92	97	85	459

## Proportions Row Column Overall

Smoke cigarettes? \ Political Party	Democratic	Other	Republican	Independent	Total
No	0.383	0.179	0.196	0.17	0.928
Yes	0.02	0.022	0.015	0.015	0.072
Total	0.403	0.2	0.211	0.185	1

Earlier in the section we wanted to find out what percentage of pre-stat students either smoke cigarettes or are republican. StatKey does not calculate “OR” (union) proportions, but we can use the proportions calculated and the following formula.

**“OR” Union Proportion = 1<sup>st</sup> Variable Proportion + 2<sup>nd</sup> Variable Proportion – Intersecting “AND” Proportion**

= Proportion Smoke + Proportion Republican – Proportion that smoke and are Republican

## Counts Table Switch Variables

Smoke cigarettes? \ Political Party	Democratic	Other	Republican	Independent	Total
No	176	82	90	78	426
Yes	9	10	7	7	33
Total	185	92	97	85	459

## Proportions Row Column Overall

Smoke cigarettes? \ Political Party	Democratic	Other	Republican	Independent	Total
No	0.383	0.179	0.196	0.17	0.928
Yes	0.02	0.022	0.015	0.015	0.072
Total	0.403	0.2	0.211	0.185	1

The proportion that smoke will be at the end of the Yes (smoke) row. The proportion of republicans will be at the bottom of the “Republican” column. The AND proportion will be where the smoking row and republican column intersect. Notice we got the same answer as before.

$$= 0.072 + 0.211 - 0.015 = 0.268 = 26.8\%$$

**Note:** Categorical data is often given to a data scientist as a contingency table with summary counts. Most data scientists do not calculate things by hand. Recall that in section 3A, we learned we can type in an existing contingency table into StatKey using commas. Typing the table into StatKey allows us to not only have access to the stacked bar chart, but also the proportion button that can calculate proportions automatically for us.



## Formulas

Single Variable Marginal Proportion = Total for Row or Column  $\div$  Grand Total

“AND” Intersecting Proportion = Amount where row and column intersect  $\div$  Grand Total

“OR” Union Proportion = Add up all of the values in the row or column without using totals  $\div$  Grand Total

“OR” Union Proportion = (Row Total + Column Total – Intersection amount)  $\div$  Grand Total

“OR” Union Proportion = 1<sup>st</sup> Variable Proportion + 2<sup>nd</sup> Variable Proportion – Intersecting “AND” Proportion

---



This material is from *Introduction to Data Analysis*, 2<sup>nd</sup> edition by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a “CC-BY” [Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by/4.0/) – 3/17/2021



## Practice Problems Section 3B

### Formulas

Single Variable Marginal Proportion = Total for Row or Column  $\div$  Grand Total

“AND” Intersecting Proportion = Amount where row and column intersect  $\div$  Grand Total

“OR” Union Proportion = Add up all of the values in the row or column without using totals  $\div$  Grand Total

“OR” Union Proportion = (Row Total + Column Total – Intersection amount)  $\div$  Grand Total

“OR” Union Proportion = 1<sup>st</sup> Variable Proportion + 2<sup>nd</sup> Variable Proportion – Intersecting “AND” Proportion

To convert a proportion into a percentage, multiply by 100%.

Directions #1-12: The following contingency table was created from the Math 075 Survey Data Fall 2015 and describes the student’s favorite social media and whether or not they have a tattoo. Use the table to find the given proportions and percentages. Show your work.

Tattoo (at least one)? \ Social Media favorite?	Twitter	Instagram	Facebook	Snapchat	Other	Total
Yes	13	38	33	15	9	108
No	78	99	66	79	29	351
Total	91	137	99	94	38	459

### Basic Marginal Proportions

- How many students have at least one tattoo?
  - What proportion of the students have a tattoo?  
(Show how you calculated the answer. Round your proportion to the thousandths place.)
  - What percentage of the students have a tattoo.  
(Show how you calculated the answer. Round your percent to the tenths place.)
- How many students prefer Facebook?
  - What proportion of the students prefer Facebook?  
(Show how you calculated the answer. Round your proportion to the thousandths place.)
  - What percentage of the students prefer Facebook?  
(Show how you calculated the answer. Round your percent to the tenths place.)
- How many students do not have a tattoo?
  - What proportion of the students do not have a tattoo?  
(Show how you calculated the answer. Round your proportion to the thousandths place.)
  - What percentage of the students do not have a tattoo?  
(Show how you calculated the answer. Round your percent to the tenths place.)



4.
  - a) How many students prefer Instagram?
  - b) What proportion of the students prefer Instagram?  
(Show how you calculated the answer. Round your proportion to the thousandths place.)
  - c) What percentage of the students prefer Instagram?  
(Show how you calculated the answer. Round your percent to the tenths place.)

Joint Proportions “AND”

Tattoo (at least one)? \ Social Media favorite?	Twitter	Instagram	Facebook	Snapchat	Other	Total
Yes	13	38	33	15	9	108
No	78	99	66	79	29	351
Total	91	137	99	94	38	459

5.
  - a) How many students both have a tattoo and prefer Facebook?
  - b) What proportion of the students both have a tattoo and prefer Facebook?  
(Show how you calculated the answer. Round your proportion to the thousandths place.)
  - c) What percentage of the students both have a tattoo and prefer Facebook?  
(Show how you calculated the answer. Round your percent to the tenths place.)
  
6.
  - a) How many students both do not have a tattoo and prefer Instagram?
  - b) What proportion of the students both do not have a tattoo and prefer Instagram?  
(Show how you calculated the answer. Round your proportion to the thousandths place.)
  - c) What percentage of the students both do not have a tattoo and prefer Instagram?  
(Show how you calculated the answer. Round your percent to the tenths place.)
  
7.
  - a) How many students both do not have a tattoo and prefer Snapchat?
  - b) What proportion of the students both do not have a tattoo and prefer Snapchat?  
(Show how you calculated the answer. Round your proportion to the thousandths place.)
  - c) What percentage of the students both do not have a tattoo and prefer Snapchat?  
(Show how you calculated the answer. Round your percent to the tenths place.)
  
8.
  - a) How many students both have a tattoo and prefer “Other” social media?
  - b) What proportion of the students both have a tattoo and prefer “Other” social media?  
(Show how you calculated the answer. Round your proportion to the thousandths place.)
  - c) What percentage of the students both have a tattoo and prefer “Other” social media?  
(Show how you calculated the answer. Round your percent to the tenths place.)



Joint Proportions “OR”

Tattoo (at least one)? \ Social Media favorite?	Twitter	Instagram	Facebook	Snapchat	Other	Total
Yes	13	38	33	15	9	108
No	78	99	66	79	29	351
Total	91	137	99	94	38	459

9. a) How many total students either have a tattoo or prefer Facebook?  
(Show how you calculated the answer.)
- b) What proportion of the students either have a tattoo or prefer Facebook?  
(Show how you calculated the answer. Round your proportion to the thousandths place.)
- c) What percentage of the students either have a tattoo or prefer Facebook?  
(Show how you calculated the answer. Round your percent to the tenths place.)
10. a) How many students either do not have a tattoo or prefer Instagram?  
(Show how you calculated the answer.)
- b) What proportion of the students either do not have a tattoo or prefer Instagram?  
(Show how you calculated the answer. Round your proportion to the thousandths place.)
- c) What percentage of the students either do not have a tattoo or prefer Instagram?  
(Show how you calculated the answer. Round your percent to the tenths place.)
11. a) How many students prefer either Twitter or Snapchat?  
(Show how you calculated the answer.)
- b) What proportion of the students prefer either Twitter or Snapchat?  
(Show how you calculated the answer. Round your proportion to the thousandths place.)
- c) What percentage of the students prefer either Twitter or Snapchat?  
(Show how you calculated the answer. Round your percent to the tenths place.)
12. a) How many students either have a tattoo or prefer “Other” social media?  
(Show how you calculated the answer.)
- b) What proportion of the students either have a tattoo or prefer “Other” social media?  
(Show how you calculated the answer. Round your proportion to the thousandths place.)
- c) What percentage of the students either have a tattoo or prefer “Other” social media?  
(Show how you calculated the answer. Round your percent to the tenths place.)



13. Copy and paste the gender and month data taken columns from the “Bear” data into StatKey. Use StatKey to calculate the following. Let gender represent the rows and the month data taken represent the columns.

Directions for creating contingency table with StatKey from Raw Data:

- Open the “Math 075 Survey Data Fall 2015”. Copy and paste the two columns next to each other in a new spreadsheet. Then copy both columns together.
- Go to [www.lock5stat.com](http://www.lock5stat.com) and click on “StatKey”. Under the “Descriptive Statistics and Graphs” menu, click on “Two Categorical Variables”. Click the “Edit Data” button. Push “Control A” and “Delete” on your keyboard to delete out any existing data. Then paste in your two columns of data. Check the box that says “Raw Data”. If your data has a title, check the box that says “Data has a header row”. Then push “OK”.
- Click on the “Overall” proportions button and use the proportions provided to answer the questions.

- a) What proportion of the bears had data taken in September? Convert the proportion into a percentage.
- b) What proportion of the bears were female? Convert the proportion into a percentage.
- c) What proportion of the bears were both female and had data taken in September? Convert the proportion into a percentage.
- d) What proportion of the bears were either female or had data taken in September? Use the following formula and your answers from parts (a), (b) and (c). Convert the proportion into a percentage.

**“OR” Union Proportion = 1<sup>st</sup> Variable Proportion + 2<sup>nd</sup> Variable Proportion – Intersecting “AND” Proportion**

14. Type in the following contingency table into StatKey and use the “Overall Proportions” button in StatKey to calculate the following proportions.

Directions for putting a contingency table into StatKey:

- Go to [www.lock5stat.com](http://www.lock5stat.com) and click on “StatKey”. Under the “Descriptive Statistics and Graphs” menu, click on “Two Categorical Variables”.
- Click the “Edit Data” button. Push “Control A” and “Delete” on your keyboard to delete out any existing data. Then type in the contingency table with commas as seen below. Do NOT check the box that says “Raw Data”. Check the box that says “Data has a header row”. Then push “OK”.
- Click on the “Overall” proportions button and use the proportions provided to answer the questions.

Contingency Table (Credit Card by Server)

[Blank], Cash, Credit Card

Server A, 39, 21

Server B, 50, 15

Server C, 17, 15

- a) What proportion of the bills were paid with cash? Convert the proportion into a percentage.
- b) What proportion of the bills had server B as the server? Convert the proportion into a percentage.
- c) What proportion of the bills were both served by server B and paid in cash? Convert the proportion into a percentage.
- d) What proportion of the bills were either served by server B or paid in cash? Use the following formula and your answers from parts (a), (b) and (c). Convert the proportion into a percentage.

**“OR” Union Proportion = 1<sup>st</sup> Variable Proportion + 2<sup>nd</sup> Variable Proportion – Intersecting “AND” Proportion**



## Section 3C – Conditional Percentages from Contingency Tables and Categorical Relationships

### Conditional Proportions and Percentages

Conditional proportions and percentages are the key to understanding categorical relationships. A condition is thought of as prior knowledge about the person or situation that may change the percentage. Let us say that the Los Angeles Lakers have a 75% chance of beating the Phoenix Suns. If the Lakers best player LeBron James does not play, will that change the percentage? Of course. Knowing that LeBron James will not play is called a condition.

In contingency tables, a condition involves restricting to one particular group before you calculate the percentage.

#### Example:

What percentage of the Canyon Country campus Math 075 pre-stat students prefer Twitter as their favorite social media?

First notice that this is not a joint proportion. It does NOT ask for the percentage of all students both prefer Twitter and go to the Canyon Country campus.

The key is to identify which group we are restricting ourselves to. In other words, what is the condition? Look for words that say “if” or “given this is true” or “out of”. This designates the condition. In this example, notice that the problem said “of the Canyon Country students”. That means that we are supposed to only look at the Canyon Country students when we find our amount (frequency) and total. A commonly used method for calculating conditional percentages from a contingency table is to circle the row or column that has your condition (Canyon Country). Then only use numbers in that row or column.

#### Counts Table Switch Variables

Campus \ Social Media favorite?	Twitter	Instagram	Facebook	Snapchat	Other	Total
Canyon Country Campus	35	52	43	36	14	180
Valencia Campus	56	85	56	58	24	279
Total	91	137	99	94	38	459

Notice that the Canyon Country Campus counts are in the first row. So we should highlight or circle the first row and only use numbers in the first row when we calculate. We should not use the grand total anymore. We need the total number of students that attend the Canyon Country campus. In other words, the total from our condition. The amount will be the number of students that prefer Twitter in the Canyon Country row. In other words the intersection cell frequency.

#### Counts Table Switch Variables

Campus \ Social Media favorite?	Twitter	Instagram	Facebook	Snapchat	Other	Total
Canyon Country Campus	35	52	43	36	14	180
Valencia Campus	56	85	56	58	24	279
Total	91	137	99	94	38	459

$$\text{Conditional Proportion} = \frac{\text{Amount in Intersection Cell (Canyon Country meets Twitter)}}{\text{Row or Column Total (Row total Canyon Country)}} = \frac{35}{180} \approx 0.19444444 \approx 0.194$$



Conditional Percentage = Conditional Proportion x 100% = 0.194 x 100% = 19.4%

So 19.4% of the Canyon Country pre-stat students prefer Twitter as their favorite social media.

We can also have StatKey calculate conditional proportions for us by using the “row” and “column” proportion buttons. We need to ask ourselves if the condition is a row or a column? In the last question we were restricting ourselves to only Canyon Country students. This is a row. Since the condition is a row, we should click the “row” proportion button. If the condition had been a column, we would have clicked on the “Column” proportion button.

### Counts Table Switch Variables

Campus \ Social Media favorite?	Twitter	Instagram	Facebook	Snapchat	Other	Total
Canyon Country Campus	35	52	43	36	14	180
Valencia Campus	56	85	56	58	24	279
Total	91	137	99	94	38	459

### Proportions Row Column Overall

Campus \ Social Media favorite?	Twitter	Instagram	Facebook	Snapchat	Other	Total
Canyon Country Campus	0.194	0.289	0.239	0.2	0.078	1
Valencia Campus	0.201	0.305	0.201	0.208	0.086	1
Total	0.198	0.298	0.216	0.205	0.083	1

Notice that all of the rows add up to 1 (100%). This confirms that the computer is calculating the conditional proportions for the rows. We are looking for the proportion of Canyon Country pre-stat students that prefer Twitter. Notice the answer we are looking for is given in the intersecting cell. If we restrict ourselves to considering only the Canyon Country students, 0.194 or 19.4% of them prefer twitter. This is the same answer we got earlier in the section.

Example:

What proportion of the Snapchat math 075 pre-stat students attend the Valencia campus?

To answer this we need to recognize that we are no longer considering all the students. We are restricting our proportion to considering only the Snapchat students (“out of”). Notice that the student that prefer Snapchat are in a column. Since the condition is preferring Snapchat, we should only use numbers in the Snapchat column to calculate the proportion. Notice we highlighted the numbers in Snapchat column. The total will now be the total number of Snapchat students and the amount will be the amount of Snapchat students that attend the Valencia campus.

### Counts Table Switch Variables

Campus \ Social Media favorite?	Twitter	Instagram	Facebook	Snapchat	Other	Total
Canyon Country Campus	35	52	43	36	14	180
Valencia Campus	56	85	56	58	24	279
Total	91	137	99	94	38	459

$$\text{Conditional Proportion} = \frac{\text{Amount in Intersection Cell (Snapchat meets Valencia)}}{\text{Row or Column Total (column total Snapchat)}} = \frac{58}{94} \approx 0.617021276 \approx 0.617$$



Conditional Percentage = Conditional Proportion x 100% = 0.617 x 100% = 61.7%

We can also use StatKey to find the proportion of the Snapchat math 075 pre-stat students attend the Valencia campus. Notice our condition is now Snapchat (“out of”). This is a column so I will click the “column” proportion button in StatKey.

**Counts Table** Switch Variables

Campus \ Social Media favorite?	Twitter	Instagram	Facebook	Snapchat	Other	Total
Canyon Country Campus	35	52	43	36	14	180
Valencia Campus	56	85	56	58	24	279
Total	91	137	99	94	38	459

**Proportions** Row Column Overall

Campus \ Social Media favorite?	Twitter	Instagram	Facebook	Snapchat	Other	Total
Canyon Country Campus	0.385	0.38	0.434	0.383	0.368	0.392
Valencia Campus	0.615	0.62	0.566	0.617	0.632	0.608
Total	1	1	1	1	1	1

Notice that when we click the “Column” proportion button, all of the columns add up to 1 (100%). This lets us know that StatKey has calculated all of the conditional proportions for the columns.

The conditional proportion we are looking for is where Snapchat and Valencia intersect. 0.617 or 61.7%. Notice this is the same answer as our earlier calculation.

Note: Categorical data is often given to a data scientist as a contingency table with summary counts. Most data scientists do not calculate things by hand. Recall that in section 3A, we learned we can type in an existing contingency table into StatKey using commas. Typing the table into StatKey allows us to not only have access to the stacked bar chart, but also the proportion button that can calculate proportions automatically for us.

Relationship Principle

Let us go back to the LeBron James example. The key to understanding categorical relationships is to judge how close or far apart conditional percentages are.

- Chances of Los Angeles Lakers beating the Phoenix Suns if LeBron James plays ≈ 75%
- Chances of Los Angeles Lakers beating the Phoenix Suns if LeBron James does not play ≈ 25%

These percentages are significantly different, so it tells us that the condition of LeBron James playing in the game is related to the Lakers winning.

Note: Does this mean that LeBron playing is the only factor that CAUSES the Lakers to win? No. Remember related (associated) does NOT prove cause and effect. There are many confounding variables that go into the Lakers winning or losing. (Health of LeBron, Health of other players, the team the Lakers are playing, home game or away game, Number of games played, etc...) We can say that LeBron James playing is related to the Lakers winning, but the data does not prove that LeBron James playing is the only factor that causes the Lakers to win.



Let us look at another example using the Lakers chances of beating the Phoenix Suns.

Chances of Lakers winning if it snows in Nebraska  $\approx 75\%$

Chances of Lakers winning if it does not snow in Nebraska  $\approx 75\%$

These percentages are not significantly different, so it tells us that the condition of snowing in Nebraska is not related to the Lakers winning. The condition does not matter.

**Relationship Principle:**

**Close Conditional Percentages from same variable = Condition is NOT related to the categorical variable**

**Significantly Different Conditional Percentages from same variable = Condition IS related to the categorical variable**

*Note: You cannot compare any conditional percentages you want. They must be the same variable for the percentage and from different groups (different condition). You cannot compare the percentage of Snapchat students from the Canyon Country campus to the percentage of Twitter from the Valencia campus. They are not the same thing and will likely have very different percentages regardless of the relationship. Compare the percentage of Snapchat students from the Canyon Country campus to the percentage of Snapchat students from the Valencia campus. That will give us information about the relationship. Conditional percentage analysis is the basis behind the Chi-Square test statistics in more advanced statistics classes.*

Example:

Look at the following conditional proportions StatKey calculated based on the rows (Canyon Country campus and Valencia campus).

**Counts Table** Switch Variables

Campus \ Social Media favorite?	Twitter	Instagram	Facebook	Snapchat	Other	Total
Canyon Country Campus	35	52	43	36	14	180
Valencia Campus	56	85	56	58	24	279
Total	91	137	99	94	38	459

**Proportions** Row Column Overall

Campus \ Social Media favorite?	Twitter	Instagram	Facebook	Snapchat	Other	Total
Canyon Country Campus	0.194	0.289	0.239	0.2	0.078	1
Valencia Campus	0.201	0.305	0.201	0.208	0.086	1
Total	0.198	0.298	0.216	0.205	0.083	1

We can only compare Twitter to Twitter, Instagram to Instagram, Facebook to Facebook, Snapchat to Snapchat, Other to Other. Notice the proportions look very close. This gives us the idea that a pre-stat students favorite social media may not be related to the campus they go to. If the proportions were significantly different, that may indicate a relationship.





## Practice Problems Section 3C

### Formulas

Conditional Proportion = Intersection of the row and column  $\div$  Row or column total for Condition

Circle the row or column that has the condition. Use only numbers in that row or column when calculating the conditional proportion.

To convert a proportion into a percentage, multiply by 100%.

Directions #1-4: The following contingency table was created from the Math 075 Survey Data Fall 2015 and describes the student's favorite social media and whether or not they have a tattoo. Use the table to find the given proportions and percentages. Show your work.

Tattoo (at least one)? \ Social Media favorite?	Twitter	Instagram	Facebook	Snapchat	Other	Total
Yes	13	38	33	15	9	108
No	78	99	66	79	29	351
Total	91	137	99	94	38	459

- How many total students have at least one tattoo?
  - How many students both have a tattoo and prefer Instagram?
  - What proportion of the tattoo students prefer Instagram? (Show how you calculated the answer.)
  - What percentage of the tattoo students prefer Instagram? (Show how you calculated the answer.)
- How many total students prefer Twitter?
  - How many students both do not have a tattoo and prefer Twitter?
  - What proportion of the Twitter students do not have a tattoo? (Show how you calculated the answer.)
  - What percentage of the Twitter students do not have a tattoo? (Show how you calculated the answer.)
- How many total students do not have a tattoo?
  - How many students both do not have a tattoo and prefer Facebook?
  - What proportion of the no tattoo students prefer Facebook? (Show how you calculated the answer.)
  - What percentage of the no tattoo students prefer Facebook? (Show how you calculated the answer.)
- How many total students prefer Snapchat?
  - How many students both have a tattoo and prefer Snapchat?
  - What proportion of the Snapchat students have a tattoo? (Show how you calculated the answer.)
  - What percentage of the Snapchat students have a tattoo? (Show how you calculated the answer.)



Directions #5-8: The following contingency table was created from the Math 075 Survey Data Fall 2015 and describes the campus the student attended and the type of transportation they took to get to school. Use the table to find the given proportions and percentages. Show your work.

Campus \ Transportation type to campus	Drive alone	Public transportation	Dropped off by someone	Carpool	Walk	Other	Bicycle	Skate	Total
Canyon Country Campus	138	7	14	15	1	4	1	0	180
Valencia Campus	203	17	32	22	3	0	1	1	279
Total	341	24	46	37	4	4	2	1	459

5.
  - a) How many total students went to the Canyon Country campus?
  - b) How many students both drive alone and went to the Canyon Country campus?
  - c) What proportion of the Canyon Country campus students drove alone to school?  
(Show how you calculated the answer.)
  - d) What percentage of the Canyon Country campus students drove alone to school?  
(Show how you calculated the answer.)
  
6.
  - a) How many total students were dropped off by someone?
  - b) How many students were both dropped off and went to the Canyon Country campus?
  - c) What proportion of the dropped off students went to the Canyon Country campus?  
(Show how you calculated the answer.)
  - d) What percentage of the dropped off students went to the Canyon Country campus?  
(Show how you calculated the answer.)
  
7.
  - a) How many total students went to the Valencia campus?
  - b) How many students both carpool and went to the Valencia campus?
  - c) What proportion of the Valencia campus students carpool to school?  
(Show how you calculated the answer.)
  - d) What percentage of the Valencia campus students carpool to school?  
(Show how you calculated the answer.)
  
8.
  - a) How many total students used public transportation to school?
  - b) How many students both used public transportation and went to the Valencia campus?
  - c) What proportion of the public transportation students went to the Valencia campus?  
(Show how you calculated the answer.)
  - d) What percentage of the public transportation students went to the Valencia campus?  
(Show how you calculated the answer.)



9. Copy and paste the gender and month data taken columns from the “Bear” data into StatKey. Use StatKey to calculate the following. Let bear gender represent the rows and month data was taken represent the columns.

Directions for creating contingency table with StatKey from Raw Data:

- Open the “Math 075 Survey Data Fall 2015”. Copy and paste the two columns next to each other in a new spreadsheet. Then copy both columns together.
- Go to [www.lock5stat.com](http://www.lock5stat.com) and click on “StatKey”. Under the “Descriptive Statistics and Graphs” menu, click on “Two Categorical Variables”. Click the “Edit Data” button. Push “Control A” and “Delete” on your keyboard to delete out any existing data. Then paste in your two columns of data. Check the box that says “Raw Data”. If your data has a title, check the box that says “Data has a header row”. Then push “OK”. The rows should be gender and the columns should be month data taken. If it is not, simply click the “Switch Variables” button.
- Click on the “Row” proportions button and use the conditional row proportions to answer the questions.

a) What proportion of the female bears were measured in August?

b) What proportion of the male bears were measured in August?

c) Compare your answer in letter (a) to your answer in letter (b). Do the proportions look close or significantly different?

d) What proportion of the female bears were measured in October?

e) What proportion of the male bears were measured in October?

f) Compare your answer in letter (d) to your answer in letter (e). Do the proportions look close or significantly different?

g) Do your answers in letters (c) and (f) indicate that the bear gender may be related to what month the bears are measured in? Explain your answer.

h) If data indicated that bear gender was related to the month the bears were measured in, would that prove that the gender of the bear causes the bear to be measured in a certain month? Explain your answer.



10. Type in the following contingency table into StatKey and use the “Overall Proportions” button in StatKey to calculate the following proportions.

Directions for putting a contingency table into StatKey:

- Go to [www.lock5stat.com](http://www.lock5stat.com) and click on “StatKey”. Under the “Descriptive Statistics and Graphs” menu, click on “Two Categorical Variables”.
- Click the “Edit Data” button. Push “Control A” and “Delete” on your keyboard to delete out any existing data. Then type in the contingency table with commas as seen below. Do NOT check the box that says “Raw Data”. Check the box that says “Data has a header row”. Then push “OK”. The rows should be the servers and the columns should be the type of payment. If they are not, simply push the “switch variables” button.
- Click on the “Column” proportions button and use the conditional column proportions provided to answer the questions.

Contingency Table (Credit Card by Server)

[Blank], Cash, Credit Card

Server A, 39, 21

Server B, 50, 15

Server C, 17, 15

- a) What proportion of the credit card customers were served by server A?
  - b) What proportion of the credit card customers were served by server B?
  - c) What proportion of the credit card customers were served by server C?
  - d) Do your answers in parts (a), (b) and (c) seem close or significantly different?
  - e) Does your answer in part (d) indicate that paying with a credit card may be related to who the server is? Explain your answer.
- 



### Chapter 3 Review Sheet

Here is a list of important ideas in this chapter.

- Be comfortable creating and analyzing contingency tables with technology from two categorical data sets
- Be able to create and analyze bar charts and pie charts to summarize two way table information
- Be able to find basic marginal proportions, joint proportions (AND / OR), and conditional proportions and be able to convert the proportions into percentages.
- Be able to look at relationships between categorical variables by looking at conditional proportions.
- Relationship Principle  
 Values Significantly different => related  
 Values Close => not related

### Problem Set Chapter 3 Review

1. The following categorical data gives the gender (male or female) of people's pets and who takes care of the pet (caretaker). Create a two-way table from this data. Give the counts and the totals.

Pet Gender	Caretaker
F	Everyone
M	Everyone
F	Parents
F	Parents
M	Everyone
M	Parents
M	Everyone
M	Parents
M	Kids
M	Parents
M	Parents
M	Everyone
F	Everyone

	Kids	Parents	Everyone	Totals
Female Pet				
Male Pet				
Totals				Grand Total =



A total of 280 high school students were asked about their political affiliation. The following two-way table was created from the data. Use the table to answer the following question.

	Democrat	Republican	Other	Total
Freshmen	7	7	28	42
Sophomore	28	21	56	105
Junior	35	28	21	84
Senior	21	14	14	49
Total	91	70	119	280

$$\text{Proportion} = \frac{\text{Amount}}{\text{Total}}$$

$$\text{Percentage} = \frac{\text{Amount}}{\text{Total}} \times 100\%$$

2. What proportion of the students identified with the “Other” political party? (Give your answer as a fraction, decimal proportion and as a percent.)
3. What percent of the students were in their senior year? (Give your answer as a fraction, decimal proportion and as a percent.)
4. What proportion of the students were both democrat and in their junior year? (Both must be true about person) (Give your answer as a fraction, decimal proportion and as a percent.)
5. What percent of the students were both republican and in their sophomore year? (Both must be true about person) (Give your answer as a fraction, decimal proportion and as a percent.)
6. What proportion of the students were either in their freshman year or in their senior year? (Either one can be true about person) (Give your answer as a fraction, decimal proportion and as a percent.)
7. What percent of the students were either democrat or in their senior year? (Either one can be true about person) (Give your answer as a fraction, decimal proportion and as a percent.)

A total of 280 High School Students were asked about their political affiliation. The following two-way table was created from the data. Use the table to answer the following question.

	Democrat	Republican	Other	Total
Freshmen	7	7	28	42
Sophomore	28	21	56	105
Junior	35	28	21	84
Senior	21	14	14	49
Total	91	70	119	280

$$\text{Proportion} = \frac{\text{Amount}}{\text{Total}}$$

$$\text{Percentage} = \frac{\text{Amount}}{\text{Total}} \times 100\%$$

8. If we only look at the sophomores, what percent of them are democrat? (Give your answer as a fraction, decimal proportion and as a percent.)



9. If we only look at the seniors, what percent of them are democrat? (Give your answer as a fraction, decimal proportion and as a percent.)

10. Where the percentages in #8 and #9 close or significantly different?

11. Does the data suggest that grade level is related to being a democrat, or not related?

---



## Project Chapter 3 – Categorical Relationships

**Directions for Online Classes:** *This will be an individual project. Each student will choose two columns of categorical data to analyze from the “Math 075 Survey data Fall 2015” and create a poster summarizing their findings. Students can choose two from the following columns of data: Tattoo, Texting While Driving, Favorite Social Media, Transportation to School, Car Accident, Cigarettes, Eat Breakfast, Glasses/Contacts, High School in Santa Clarita, Living with parents*

*After submitting the project to their instructor, students will then go to the “Chapter 3 Project Class Discussion” in Canvas and discuss their findings with other students in the class.*

### The Individual Poster Should Have

- The poster does not have to be extremely large.
- Your first and last name on the poster
- What two columns of data did you pick?
- Explain why this data is important or interesting to you?
- Copy and paste the two columns of data next to each other in a new spreadsheet. Then copy both columns together. Go to StatKey at [www.lock5stat.com](http://www.lock5stat.com) and click on “Two Categorical Variables” under the “Descriptive Statistics and Graphs” menu. Then paste the two columns of data together under “edit data”. Remember to check the box that says “raw data” before pushing “OK”. Also check “header row” if data has a title.
- Copy the “Counts Table” table from StatKey onto your poster in large letters. (Label this table as your “Contingency Table”.) Pick out a few counts on this table and explain them.
- Draw the stacked bar chart onto your poster.
- Click the “Overall” button where it says Proportions in StatKey. Copy the “Overall Proportions Table” table from StatKey onto your poster in large letters. (Label this table as your “Overall Proportions Table”.) Pick out a proportion under totals. Explain what variable the computer is finding the marginal proportion of and explain how the computer calculated it. Pick out one proportion in the middle of the table. Explain what two variables the computer is finding the “AND” joint proportion for and explain how the computer calculated it.
- Click the “Row” button where it says Proportions in StatKey. Copy the “Row Proportions Table” table from StatKey onto your poster in large letters. (Label this table as your “Conditional Row Proportions Table”.) In the “Row proportion table”, compare proportions that are in the same column. Do they look close or significantly different? What does this indicate about whether or not the two columns of data you chose are related or not?
- Decorate Poster

Now take a picture of your poster project and submit the picture to your instructor in Canvas.

After submitting the picture of the poster, go to the discussion menu in Canvas and complete the “Chapter 3 Project Discussion”. You will be discussing your findings with other students in the class.

---





**Directions for Face to Face Classes:** The class will be broken up into groups of three or four. Each group will pick a team name and one of the following pairs of categorical variables from the Math 075 Survey Data Fall 2015 to study. Each group should have a different pair of variables to study.

Group#	Team Name	Categorical Variable A	Categorical Variable B
1		Political Party	Hair Color
2		Smoking	Political Party
3		Texting/Driving	Car Accidents
4		Smoking	Transportation
5		Gender	Political Party
6		Breakfast	Fixed Intelligence
7		Hair Color	Gender
8		Fixed Intelligence	Political Party
9		Tattoo	Gender
10		Political Party	Tattoo
11		Tattoo	Hair Color
12		Smoking	Tattoo

#### The Group Poster Should Have

- The poster does not have to be extremely large.
- Your first and last name of everyone in your group should be on the poster.
- Which two columns of categorical data did you chose?
- Explain why this data is important or interesting to your group?
- Copy and paste the two columns of data next to each other in a new spreadsheet. Then copy both columns together. Go to StatKey at [www.lock5stat.com](http://www.lock5stat.com) and click on “Two Categorical Variables” under the “Descriptive Statistics and Graphs” menu. Then paste the two columns of data together under “edit data”. Remember to check the box that says “raw data” before pushing “OK”. Also check “header row” if data has a title.
- Copy the “Counts Table” table from StatKey onto your poster in large letters. (Label this table as your “Contingency Table”.) Pick out a few counts on this table and explain them.
- Draw the stacked bar chart onto your poster.
- Click the “Overall” button where it says Proportions in StatKey. Copy the “Overall Proportions Table” table from StatKey onto your poster in large letters. (Label this table as your “Overall Proportions Table”.) Pick out a proportion under totals. Explain what variable the computer is finding the marginal proportion of and explain how the computer calculated it. Pick out one proportion in the middle of the table. Explain what two variables the computer is finding the “AND” joint proportion for and explain how the computer calculated it.
- Click the “Row” button where it says Proportions in StatKey. Copy the “Row Proportions Table” table from StatKey onto your poster in large letters. (Label this table as your “Conditional Row Proportions Table”.) In the “Row proportion table”, compare proportions that are in the same column. Do they look close or significantly different? What does this indicate about whether or not the two columns of data you chose are related or not?
- Decorate Poster



### Presentation Directions

Each group will put their poster up around the room. Chose one person from the group to present first. Everyone else in the class who is not presenting will find a poster that is not their own. Then rotate and have another person from the group present. Keep rotating till each person in every group has presented. Each presentation should take a few minutes. Make sure audience rotates to new posters as well.

---

## Chapter 4 – Analyzing Normal Quantitative Data

**Introduction:** In chapters 2 and 3, we focused on analyzing categorical data and exploring relationships between categorical data sets. We will now be doing the same for quantitative data. Let us start by reviewing the difference between quantitative and categorical data sets.

### Categorical Data

Categorical data are generally labels that tell us something about the people or objects in the data set. For example, what country do they live in, what is the person's occupation, or what kind of pet they have. Usually categorical data is made up of words (do you smoke - yes or no), but occasionally a number can be used as a category. For example, a zip code can be used instead of the place a person lives. The numbers "1" and "2" can be used instead of female and male. Analyzing categorical data involved finding and comparing proportions and percentages.

### Quantitative Data

Quantitative data are numbers that measure or count something. They usually have units and taking an average makes sense. For example: a list of people's heights in inches, or their weights in kilograms, or a list of how many dogs are there in various animal shelters across Los Angeles. Notice in each of these cases the data is numerical and an average seems appropriate in the context. We can find the average height, the average weight, or the average number of dogs in animal shelters in Los Angeles.

We are now moving into quantitative data analysis. Analyzing quantitative data is complex and involves shape, measures of center, averages, measures of spread, measures of position, finding typical values and finding unusual values. It is a very different approach than we took for categorical data.

---



## Section 4A – Finding the Shape of Quantitative Data Sets with Dot Plots and Histograms

When analyzing numerical quantitative data, always start with finding the shape of the data set. Categorical data can be graphed, but does not have a shape. Categorical bar charts can be organized in a variety of ways depending on the order of the categories. Quantitative data is numerical measurement data and does have a shape.

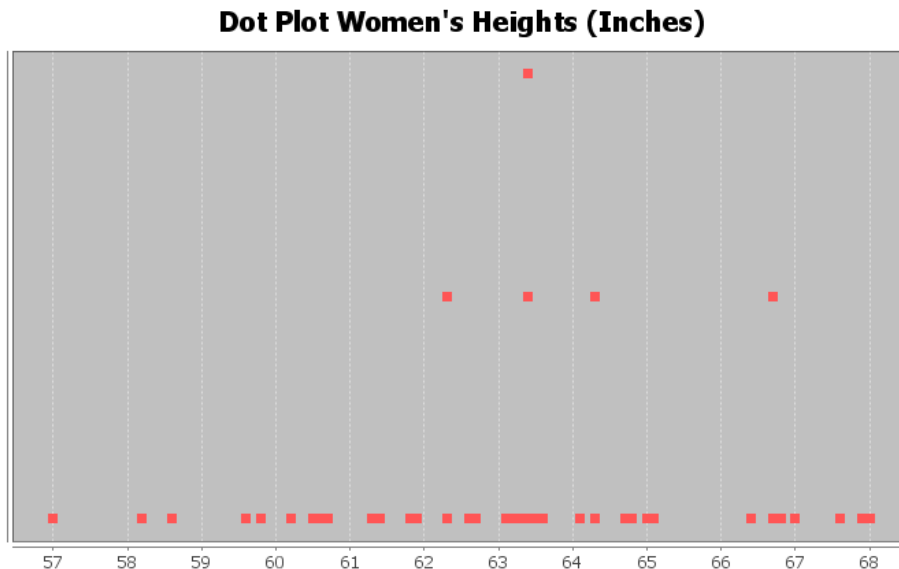
Why should we find the shape?

The goal in analyzing quantitative data is to find the average, spread, typical values and unusual values. In statistics, there are many types of averages, many types of spreads and different ways to find typical and unusual values. Shape helps us determine which averages, spreads and calculations are most accurate for the data.

### Dot plots

The most basic kind of graph for quantitative data is the dot plot. The computer draws the numerical scale usually horizontally. It then draws a dot for every single number in the quantitative data set.

Here is the dot plot for the 40 women's heights created with Statcato.

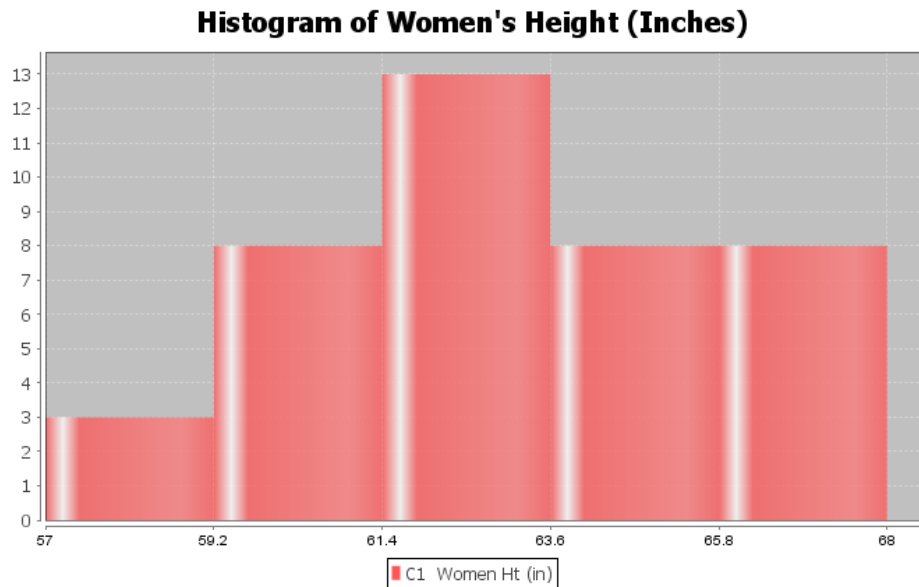


Dot plots are very useful, especially when identifying unusual values in the data set. Most students find them a little difficult to determine shape from though.

When determining shape, it is better to make a histogram. Think of a histogram as braking the scale up into sections and counting how many dots are in each section. Then drawing a bar that represents the number of dots in that section (frequency).



Here is a histogram made with Statcato for the same women's heights data as we used in the dot plot above.



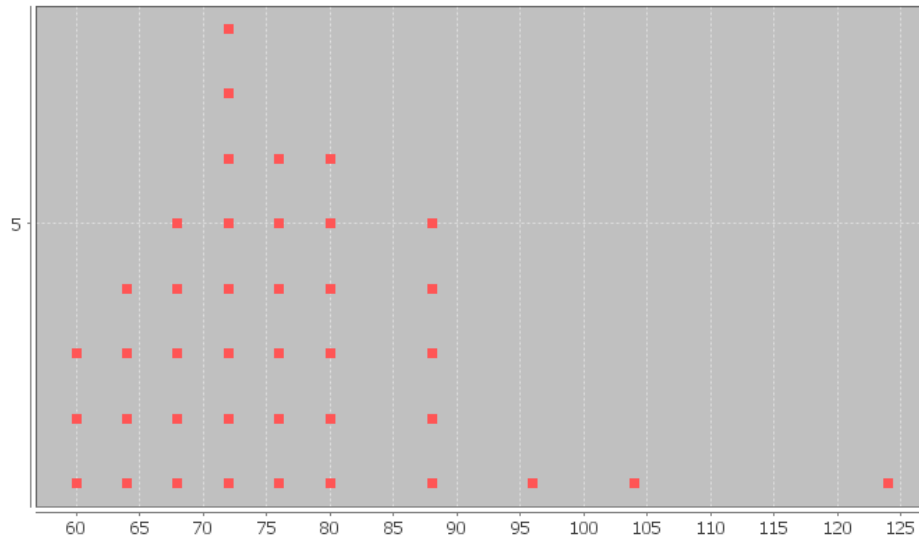
This is a very important shape in statistics. Notice the highest bar is close to the middle and the bars get smaller as we move away from the middle. This is often called “Bell Shaped” or “Normal Data”. Some like to describe this shape as unimodal (1 hill) and symmetric (left and right side look about the same). Most people in statistics call this shape “normal” or “normally distributed”.

Note: Histograms and Bar Charts are different graphs. A bar chart is a graph for categorical data where each bar gives the count or frequency for each categorical variable. A histogram is a graph used to see the shape of quantitative (numerical measurement) data. Do not confuse the two graphs.

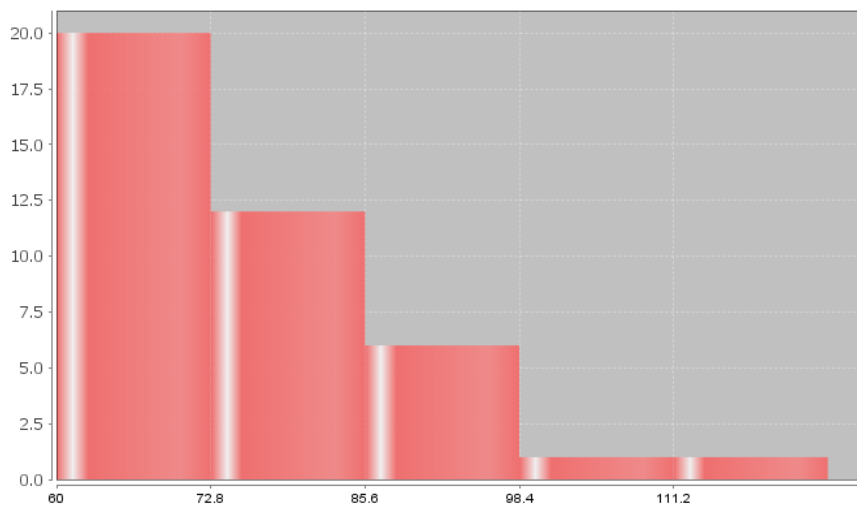
Let us look at another example from the health data. This time we will look at women's pulse rates in beats per minute (BPM). Here is a dot plot and histogram for the data.



**Dot Plot of Women's Pulse Rates (Beats Per Minute)**



**Histogram of Women's Pulse Rates (Beats Per Minute)**



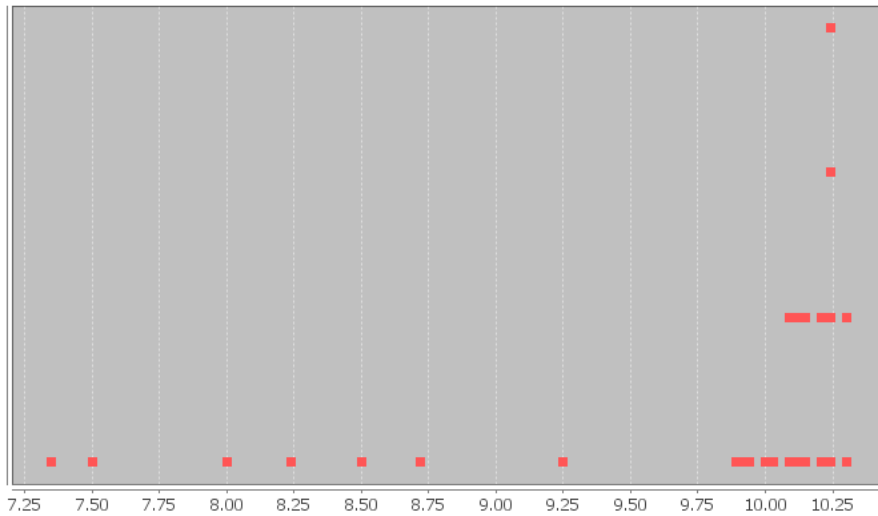
Notice this has a very different shape. There are more dots in the dot plot congregated on the far left. The highest bar in the histogram is on the far left and there are more bars to the right of the highest bar. There is a long tail to the right of the highest bar. This is called “Skewed Right”. Some people also call this “Positively Skewed”. Remember the skew is referring to the long tail. Look for the highest bar. If there is a significantly longer tail to the right, then it is skewed right. If there is a significantly longer tail to the left, then it is skewed left. If the highest hill is in the middle and the tails are approximately the same length, then it is closer to normal.

Let us look at another example.

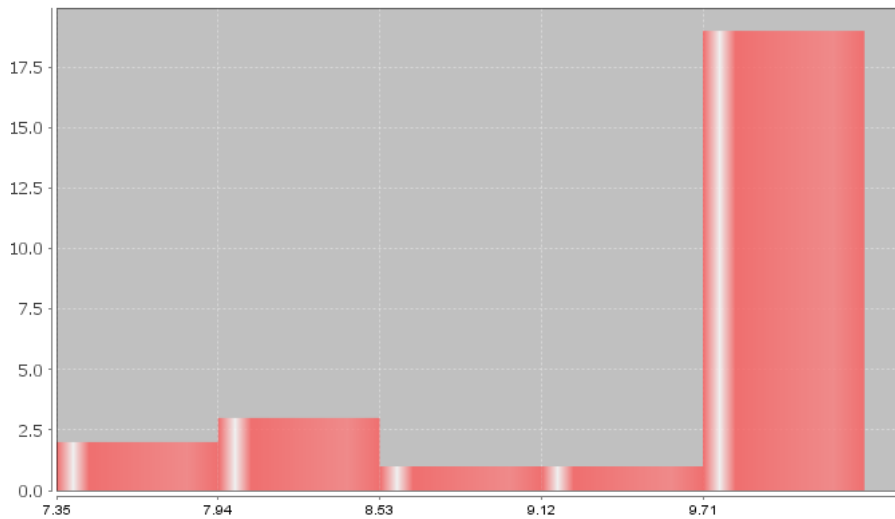
Here is some salary data from a small company with 26 employees. The salaries are given in dollars per hour. We created a dot plot and histogram for this data.



**Dot Plot of Sallary in Dollars per Hour**



**Histogram of Salary in \$ per hour**



What is the shape of these two graphs?

Notice the highest bar and most dots are on the far right, while there is a long tail to the left. Therefore, this is called skewed left.

**Note on Shape:** *Real data rarely has a perfect shape. Most data has a shape somewhere in between bell shaped and skewed, and you will need to make a decision. Look for a significant difference in the length of the tail to classify something as skewed. If my highest hill is toward the middle and I had 2 bars to the right and 3 bars to the left of the highest bar, I would still classify that bell shaped or normal. Some say that is “nearly normal”.*

*If the highest hill is on the far right and I have 2 bars to the right of the highest hill and 7 bars to the left of the highest hill, I would classify that as skewed left. Some call this “negatively skewed” since negative numbers are to the left on the number line.*

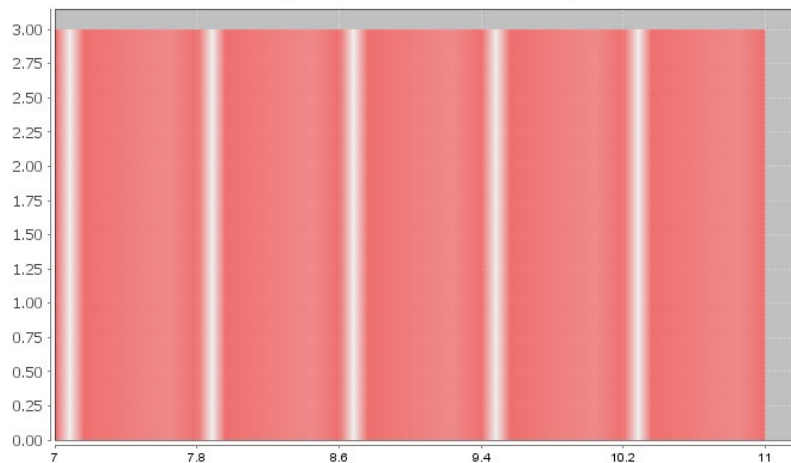
Here are a couple unusual shapes that sometimes appear.

A graph that looks like a rectangle is called “uniform”. A graph with two distinct high bars is called “bimodal”.

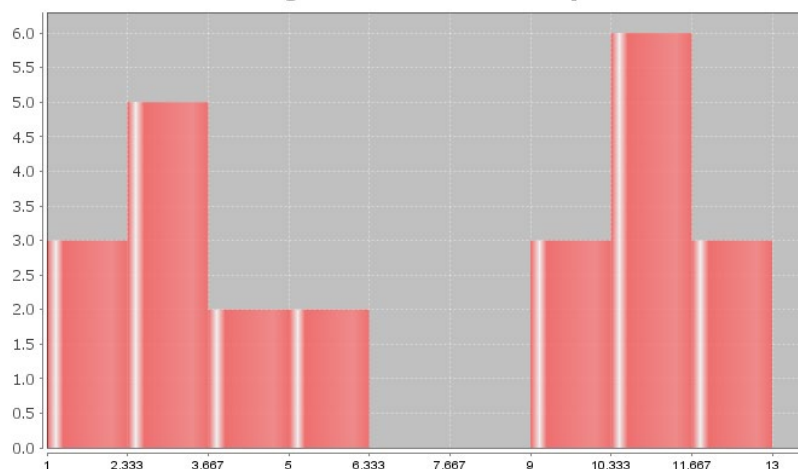


This material is from *Introduction to Data Analysis*, 2<sup>nd</sup> edition by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a “CC-BY” [Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by/4.0/) – 3/17/2021

**Histogram with Uniform Shape**



**Histogram with Bimodal Shape**



**Note:** In this chapter, we will be focusing on the how to analyze normal (bell shaped) quantitative data sets. We will discuss how to analyze skewed quantitative data sets in the next chapter.

### [Finding Quantitative Statistics and Creating Graphs with StatKey](#)

The most basic kind of graph for quantitative data is the dot plot. The computer draws the numerical scale usually horizontally. It then draws a dot for every single number in the data set. Another type of graph is a histogram. This graph counts the number of data values in certain sections and makes a bar telling us how many numbers are in that section. The number of bars is also called “bins” or “buckets”.

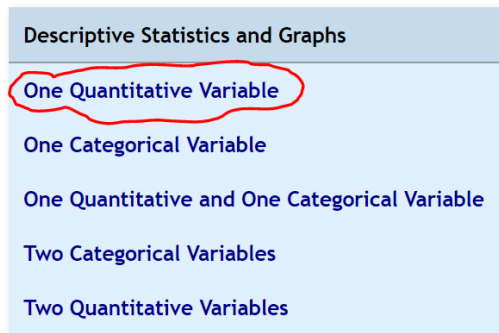
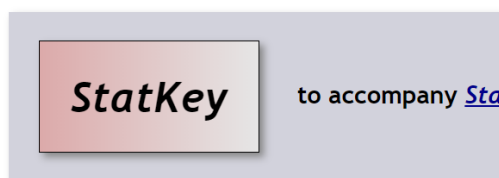
All of these graphs and statistics can be made with StatKey. The heights of women used earlier in this section, may be found in the “Health Data” on Canvas or at [www.matt-teachout.org](http://www.matt-teachout.org). Open the data set and copy the column of data that says women’s heights data. Notice the data is quantitative. The data is made up of numbers that measure the height in inches of the women. It also seems reasonable to look for an average height for these women.



P	Q	R	S	T
Women Age (years)	Women Ht (in)	Women Wt (Lbs)	Women Waist (cm)	Women Pulse (Beats per min)
17	64.3	114.8	67.2	76
32	66.4	149.3	82.5	72
25	62.3	107.8	66.7	88
55	62.3	160.1	93	60
27	59.6	127.1	82.6	72
29	63.6	123.1	75.4	68
25	59.8	111.7	73.6	80
12	63.3	156.3	81.4	64
41	67.9	218.8	99.4	68
32	61.4	110.2	67.7	68
31	66.7	188.3	100.7	80
19	64.8	105.4	72.9	76

To copy the column of data, hold your cursor over the top of the column until it turns into a downward arrow "↓". Left click your mouse and the whole column will be highlighted. Then push "Control C" on your keyboard to copy.

Go to [www.lock5stat.com](http://www.lock5stat.com) and click on the "StatKey" button. Under the "Descriptive Statistics and Graphs" menu, click on "One Quantitative Variable".



### StatKey Descriptive Statistics for One Quantitative Variable



Click on the "Edit Data" button. Push "Control A" on your keyboard and then "delete" in order to get rid of any old data. Make sure your cursor is at the top of the "edit data" field. Copy and paste the women's height data into StatKey. Do NOT check the box that says, "First column is an identifier". An identifier is a word next to every number. This data set does not have that. If your data has a title, then check the box that says, "Data has a header row". Now push "OK". Notice StatKey calculates many sample statistics and creates a dot plot and a histogram.





Edit data
✕

Women Ht (in)

64.3

66.4

62.3

62.3

59.6

63.6

59.8

63.3

67.9

61.4

66.7

64.8

63.1

66.7

66.8

64.7

65.1

61.9

64.3

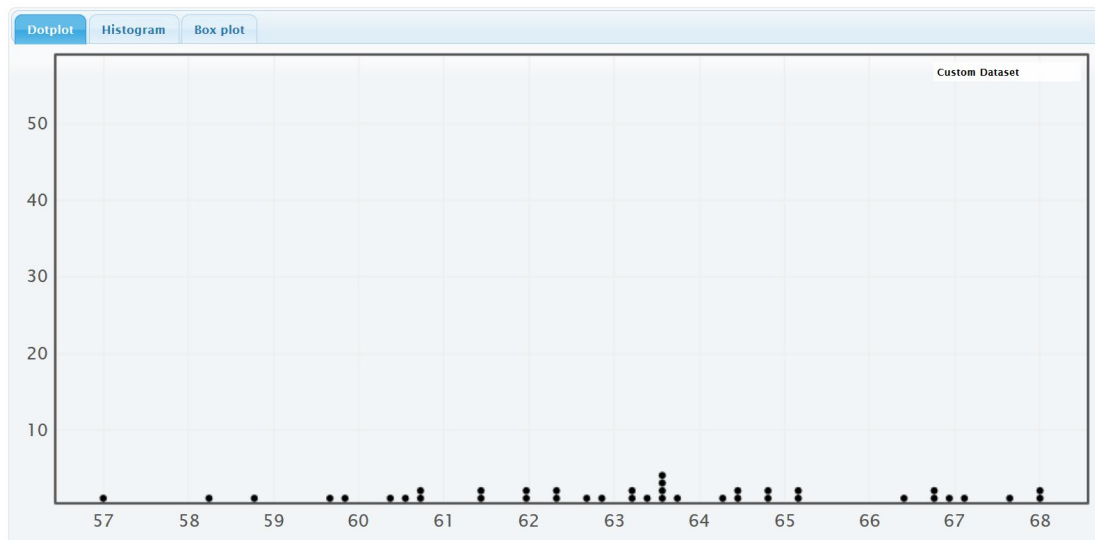
First column is identifier ← NO

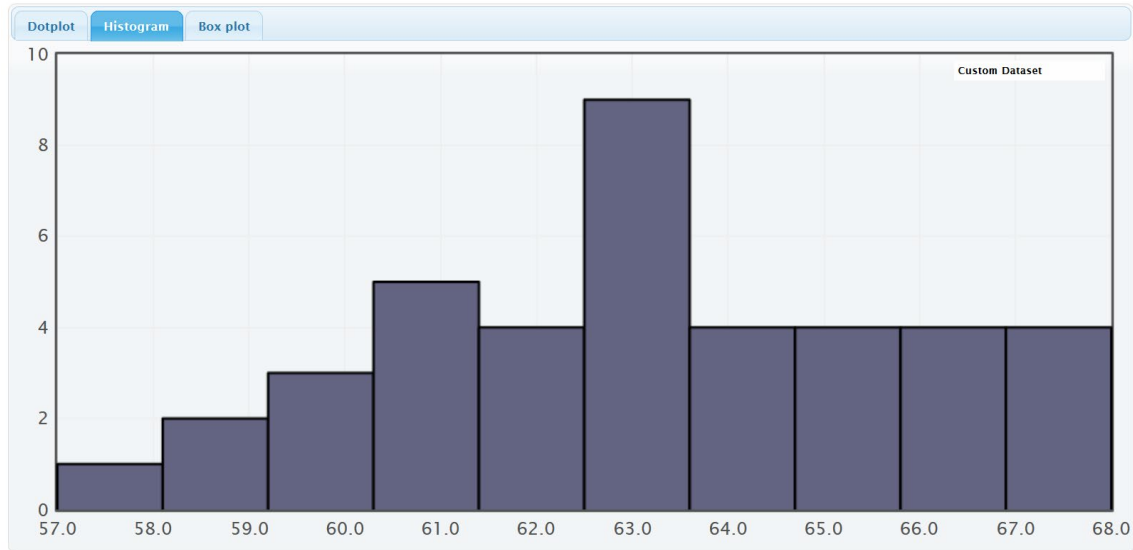
Data has header row

Manually edit the values above or paste a tab or comma separated file into the box and click Ok. The file must have only one column (or two if there is an identifier).

Ok

Here are the graphs that StatKey created. Notice there are buttons at the top to pick which graph you want to see. You can see a Dotplot, a Histogram or a Boxplot. We will focus on the dotplot and histogram. We will discuss boxplots in our next chapter.

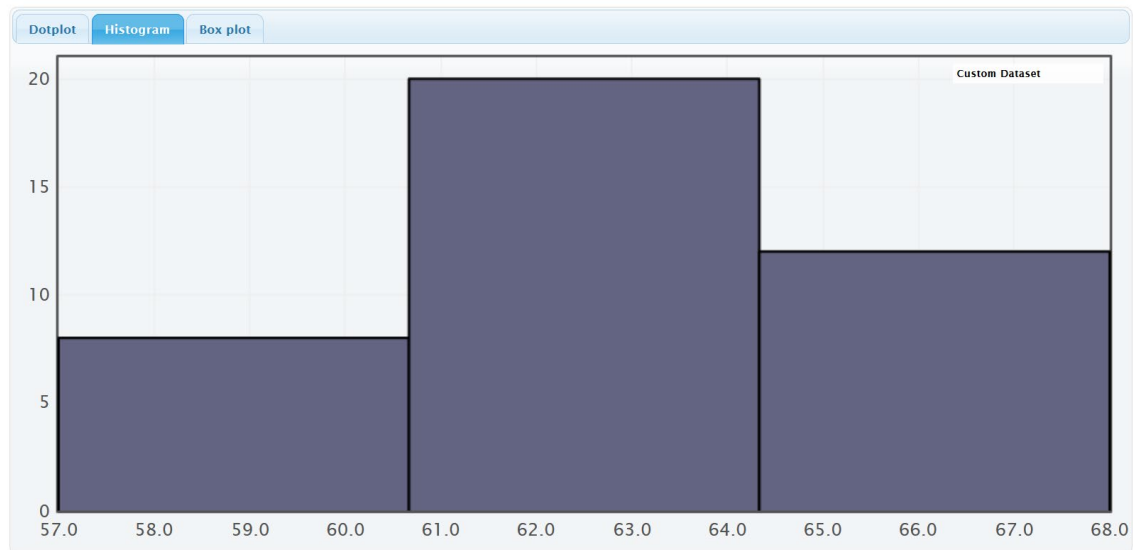




On the right of this histogram, you will see a slider that can adjust the number of bars or “buckets” in your histogram. The smaller the data set the less bins you should have. Also the less bars you have the easier it is to see the shape. This data set only has 40 numbers, so we want only a few bars. If we slide it to 3 buckets (3 bars), we get the following Histogram.

**Histogram Controls** Set Limits

Number of buckets: 3



We see that the highest bar is in the middle and the right and left tails are roughly symmetric. So this is “normal” data.

StatKey has also calculated many summary statistics. We will be discussing these statistics in future sections.



## Summary Statistics

Statistic	Value
Sample Size	40
Mean	63.195
Standard Deviation	2.741
Minimum	57
Q <sub>1</sub>	61.350
Median	63.350
Q <sub>3</sub>	64.900
Maximum	68

---

### Practice Problems Section 4A

Directions: Open the “Health” data in Canvas or at [www.matt-teachout.org](http://www.matt-teachout.org). You will be using the women’s data (columns P – AB) and men’s data (columns AD – AP). Go to [www.lock5stat.com](http://www.lock5stat.com), click on StatKey and then “One Quantitative Variable”. Paste the column of data under “edit data” in StatKey. Click on dotplot and histogram. Draw rough sketch of the dotplot and histogram on a sheet of paper. What is the shape of the data set?

1. Use a StatKey to create a dot plot and histogram of women’s ages in years. Adjust the histogram to have three bars (3 buckets).
  - a) Draw a rough sketch of the dotplot on a sheet of paper.
  - b) Draw a rough sketch of the histogram (with 3 bars) on a sheet of paper.
  - c) What is the shape of the data?
2. Use a StatKey to create a dot plot and histogram of women’s height in inches. Adjust the histogram to have three bars (3 buckets).
  - a) Draw a rough sketch of the dotplot on a sheet of paper.
  - b) Draw a rough sketch of the histogram (with 3 bars) on a sheet of paper.
  - c) What is the shape of the data?
3. Use a StatKey to create a dot plot and histogram of women’s weight in pounds. Adjust the histogram to have three bars (3 buckets).



- a) Draw a rough sketch of the dotplot on a sheet of paper.
  - b) Draw a rough sketch of the histogram (with 3 bars) on a sheet of paper.
  - c) What is the shape of the data?
4. Use a StatKey to create a dot plot and histogram of women's waist size in centimeters. Adjust the histogram to have three bars (3 buckets).
- a) Draw a rough sketch of the dotplot on a sheet of paper.
  - b) Draw a rough sketch of the histogram (with 3 bars) on a sheet of paper.
  - c) What is the shape of the data?
5. Use a StatKey to create a dot plot and histogram of women's pulse rate in beats per minute. Adjust the histogram to have three bars (3 buckets).
- a) Draw a rough sketch of the dotplot on a sheet of paper.
  - b) Draw a rough sketch of the histogram (with 3 bars) on a sheet of paper.
  - c) What is the shape of the data?
6. Use a StatKey to create a dot plot and histogram of women's systolic blood pressure in millimeters of mercury. Adjust the histogram to have three bars (3 buckets).
- a) Draw a rough sketch of the dotplot on a sheet of paper.
  - b) Draw a rough sketch of the histogram (with 3 bars) on a sheet of paper.
  - c) What is the shape of the data?
7. Use a StatKey to create a dot plot and histogram of women's diastolic blood pressure in millimeters of mercury. Adjust the histogram to have three bars (3 buckets).
- a) Draw a rough sketch of the dotplot on a sheet of paper.
  - b) Draw a rough sketch of the histogram (with 3 bars) on a sheet of paper.
  - c) What is the shape of the data?
8. Use a StatKey to create a dot plot and histogram of women's cholesterol in milligrams per deciliter. Adjust the histogram to have three bars (3 buckets).
- a) Draw a rough sketch of the dotplot on a sheet of paper.
  - b) Draw a rough sketch of the histogram (with 3 bars) on a sheet of paper.
  - c) What is the shape of the data?
9. Use a StatKey to create a dot plot and histogram of women's body mass index (BMI) in kilograms per meters squared. Adjust the histogram to have three bars (3 buckets).
- a) Draw a rough sketch of the dotplot on a sheet of paper.
  - b) Draw a rough sketch of the histogram (with 3 bars) on a sheet of paper.
  - c) What is the shape of the data?
10. Use a StatKey to create a dot plot and histogram of women's wrist circumference in inches. Adjust the histogram to have three bars (3 buckets).
- a) Draw a rough sketch of the dotplot on a sheet of paper.
  - b) Draw a rough sketch of the histogram (with 3 bars) on a sheet of paper.
  - c) What is the shape of the data?
11. Use a StatKey to create a dot plot and histogram of men's ages in years. Adjust the histogram to have three bars (3 buckets).



- a) Draw a rough sketch of the dotplot on a sheet of paper.
  - b) Draw a rough sketch of the histogram (with 3 bars) on a sheet of paper.
  - c) What is the shape of the data?
12. Use a StatKey to create a dot plot and histogram of men's height in inches. Adjust the histogram to have three bars (3 buckets).
- a) Draw a rough sketch of the dotplot on a sheet of paper.
  - b) Draw a rough sketch of the histogram (with 3 bars) on a sheet of paper.
  - c) What is the shape of the data?
13. Use a StatKey to create a dot plot and histogram of men's weight in pounds. Adjust the histogram to have three bars (3 buckets).
- a) Draw a rough sketch of the dotplot on a sheet of paper.
  - b) Draw a rough sketch of the histogram (with 3 bars) on a sheet of paper.
  - c) What is the shape of the data?
14. Use a StatKey to create a dot plot and histogram of men's waist size in centimeters. Adjust the histogram to have three bars (3 buckets).
- a) Draw a rough sketch of the dotplot on a sheet of paper.
  - b) Draw a rough sketch of the histogram (with 3 bars) on a sheet of paper.
  - c) What is the shape of the data?
15. Use a StatKey to create a dot plot and histogram of men's pulse rate in beats per minute. Adjust the histogram to have three bars (3 buckets).
- a) Draw a rough sketch of the dotplot on a sheet of paper.
  - b) Draw a rough sketch of the histogram (with 3 bars) on a sheet of paper.
  - c) What is the shape of the data?
16. Use a StatKey to create a dot plot and histogram of men's systolic blood pressure in millimeters of mercury. Adjust the histogram to have three bars (3 buckets).
- a) Draw a rough sketch of the dotplot on a sheet of paper.
  - b) Draw a rough sketch of the histogram (with 3 bars) on a sheet of paper.
  - c) What is the shape of the data?
17. Use a StatKey to create a dot plot and histogram of men's diastolic blood pressure in millimeters of mercury. Adjust the histogram to have three bars (3 buckets).
- a) Draw a rough sketch of the dotplot on a sheet of paper.
  - b) Draw a rough sketch of the histogram (with 3 bars) on a sheet of paper.
  - c) What is the shape of the data?
18. Use a StatKey to create a dot plot and histogram of men's cholesterol in milligrams per deciliter. Adjust the histogram to have three bars (3 buckets).
- a) Draw a rough sketch of the dotplot on a sheet of paper.
  - b) Draw a rough sketch of the histogram (with 3 bars) on a sheet of paper.
  - c) What is the shape of the data?
19. Use a StatKey to create a dot plot and histogram of men's body mass index (BMI) in kilograms per meters squared. Adjust the histogram to have three bars (3 buckets).



- a) Draw a rough sketch of the dotplot on a sheet of paper.
- b) Draw a rough sketch of the histogram (with 3 bars) on a sheet of paper.
- c) What is the shape of the data?

20. Use a StatKey to create a dot plot and histogram of men's wrist circumference in inches. Adjust the histogram to have three bars (3 buckets).

- a) Draw a rough sketch of the dotplot on a sheet of paper.
  - b) Draw a rough sketch of the histogram (with 3 bars) on a sheet of paper.
  - c) What is the shape of the data?
- 

## Section 4B – Shapes and Centers

When analyzing quantitative (numerical measurement) data, we want to find the average. In statistics, we often refer to an average as a "Center". When a person asks about the center, they are really asking about the average.

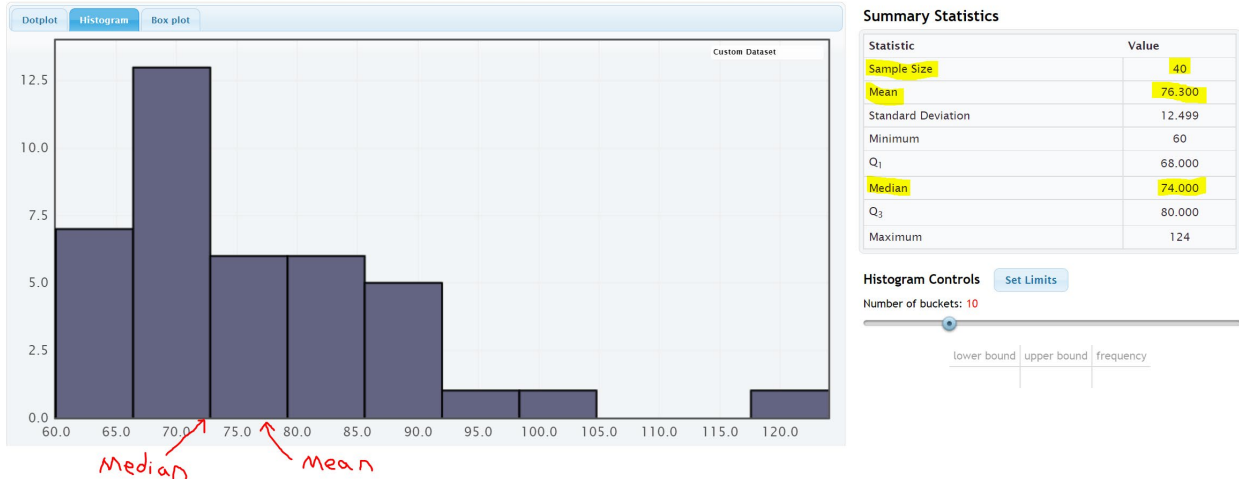
**Definition of Statistics:** The word "statistics" refers to numbers that are calculated to describe sample data sets. For example, a mean average is one of many types of statistics. Therefore, the study of "statistics" is the study of numbers calculated from data sets that help describe the characteristics of that data and hopefully what that data tells us about the world around us. We are not there yet though.

In statistics, there are many types of centers or averages. The two most commonly used centers (averages) are the mean average and the median average. The key is to determine which center (average) is most accurate for the data. An accurate center should be close to your highest bar in your histogram.

### Example 1

The following histogram and statistics were calculated with StatKey and are describing the pulse rates of 40 women in beats per minute. This data was found in the "health data". We see that this quantitative data is skewed right since the histogram has a long right tail.





We see that there was 40 women in the data since the “sample size” was 40. We also see that the mean average was 76.3 bpm and the median average was 74 bpm.

The center of a data set is where the most people or objects are located. The highest bar or bars represent the center of the data. An accurate center or average should be close to the highest bar in the data set and therefore be representative of the data values. An average that is not close to the highest bar is not a very good average.

Let us compare these values to the histogram. Notice a few things. The mean is not very accurate measures of center since they are not close to the highest bar. So the mean is not a very good average for this data. The median seems to be more accurate, since it is closer to the highest bar. So the median is closer to the center of the data.

Here are a couple of things to keep in mind when finding an accurate average for a data set. The women’s pulse data is skewed right. Mean averages get pulled in the direction of the skew (long tail) and tend to not be very accurate for skewed data sets. The median average does not get pulled in the direction of the skew and remains close to the highest bar. All this leads to an important principle. When a data set is skewed, statisticians use the median average as best measure of center and the average of the data set.

### Center Principle for Skewed Data

**If a data set has a skewed shape, the median average is usually the most accurate average (measure of center) and we should use the median as the average for the data set.**

#### Example 2

Let us look at another data set from the health data. Here is the StatKey histogram and sample statistics from the women’s height data. The data set gives the heights in inches of 40 women.





Let us compare these values to the histogram. Notice a few things. First, look at the shape. This data set is bell shaped (normal) data. The highest bar is in the middle and the right and left tails are about the same distance from the center bar.

Notice that both the mean average and the median average are close to the highest bar. It seems like either of these statistics are pretty accurate averages (centers) since they are both close to the highest bar in the histogram. Either of them would be a decently accurate average for this data.

So which one should we use?

If a data set is bell shaped, statisticians prefer to use the mean. There are several reasons for this. One being that people are most familiar with the mean. It is after all the most common type of average. That is not the real reason why we should use the mean for bell shaped data though. The real reason has to do with the spread of the data set. Bell shaped data has a very specific spread that is measured most accurately with standard deviation. Standard deviation is the most accurate spread for normal (bell shaped) data. It measures the typical distance from the mean. Therefore, in a bell shaped data set we need to use the mean as our center or average so that we can use the standard deviation to accurately measure the spread.

### Center Principle for Normal (Bell Shaped) Data

**If a data set is bell shape (normal), then the mean average is usually accurate and we should use the mean as the average and center for the data set.**

**Key: Do not use the mean average unless the data is bell shaped (normal). If the data is not normal then the mean is not accurate.**

### Calculating Centers with Technology

Remember that “statistics” are numbers that describe characteristics of data sets. The calculations though are very difficult by hand or by calculator, especially with large data sets. Always use a statistics software to calculate statistics.

### Finding Quantitative Statistics with StatKey

The heights of women used earlier in this section, may be found in the “Health Data” on Canvas or at [www.matt-teachout.org](http://www.matt-teachout.org). Open the data set and copy the column of data that says women’s heights data. Notice the



This material is from *Introduction to Data Analysis*, 2<sup>nd</sup> edition by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a “CC-BY” [Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by/4.0/) – 3/17/2021

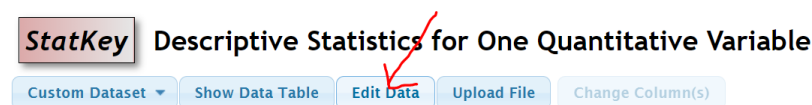
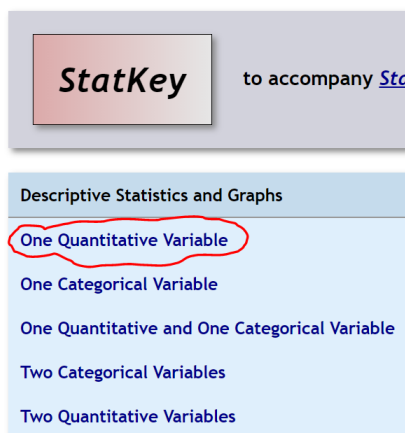


data is quantitative. The data is made up of numbers that measure the height in inches of the women. It also seems reasonable to look for an average height for these women.

P	Q	R	S	T
Women Age (years)	Women Ht (in)	Women Wt (Lbs)	Women Waist (cm)	Women Pulse (Beats per min)
17	64.3	114.8	67.2	76
32	66.4	149.3	82.5	72
25	62.3	107.8	66.7	88
55	62.3	160.1	93	60
27	59.6	127.1	82.6	72
29	63.6	123.1	75.4	68
25	59.8	111.7	73.6	80
12	63.3	156.3	81.4	64
41	67.9	218.8	99.4	68
32	61.4	110.2	67.7	68
31	66.7	188.3	100.7	80
19	64.8	105.4	72.9	76

To copy the column of data, hold your cursor over the top of the column until it turns into a downward arrow "↓". Left click your mouse and the whole column will be highlighted. Then push "Control C" on your keyboard to copy.

Go to [www.lock5stat.com](http://www.lock5stat.com) and click on the "StatKey" button. Under the "Descriptive Statistics and Graphs" menu, click on "One Quantitative Variable".



Click on the "Edit Data" button. Push "Control A" on your keyboard and then "delete" in order to get rid of any old data. Make sure your cursor is at the top of the "edit data" field. Copy and paste the women's height data into StatKey. Do NOT check the box that says, "First column is an identifier". An identifier is a word next to every number. This data set does not have that. If your data has a title, then check the box that says, "Data has a header row". Now push "OK". Notice StatKey calculates many sample statistics and creates a dot plot and a histogram.



Edit data
✕

Women Ht (in)

64.3

66.4

62.3

62.3

59.6

63.6

59.8

63.3

67.9

61.4

66.7

64.8

63.1

66.7

66.8

64.7

65.1

61.9

64.3

First column is identifier ← NO  
 Data has header row

Manually edit the values above or paste a tab or comma separated file into the box and click Ok. The file must have only one column (or two if there is an identifier).

Once we push "OK" we see the quantitative statistics calculated on the right of the page.

### Summary Statistics

Statistic	Value
Sample Size	40
Mean	63.195
Standard Deviation	2.741
Minimum	57
Q <sub>1</sub>	61.350
Median	63.350
Q <sub>3</sub>	64.900
Maximum	68

We will be discussing these statistics in greater detail later on. Notice the summary statistics for one quantitative data set include the following:

Mean Average: Average (or center) used for normal quantitative data.

Median Average: Average (or center) used for skewed or non-normal quantitative data.

Sample Size: Total number of people or objects that we collected data from. (For quantitative data, it also tells us how many numbers were in our quantitative data set.)

Minimum: Smallest number in the quantitative data set.



This material is from *Introduction to Data Analysis*, 2<sup>nd</sup> edition by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a "CC-BY" [Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by/4.0/) – 3/17/2021

Maximum: Largest number in the quantitative data set.

---

#### Practice Problems Section 4B

Directions: Open the Health data in Canvas or at [www.matt-teachout.org](http://www.matt-teachout.org). You will be using the men and women's combined data (columns B - N). This data has columns of 80 values from 40 men and 40 women. Go to [www.lock5stat.com](http://www.lock5stat.com), click on StatKey and then "One Quantitative Variable". Paste the column of data under "edit data" in StatKey. Click on "histogram" and change the histogram to have 3 bars (3 buckets). What was the shape of the data set? Under "summary statistics" in StatKey, make a note of the minimum, maximum, sample size (n), mean average, and median average. Based on the shape, of the data, should we use the mean or the median as our most accurate average?



This material is from *Introduction to Data Analysis*, 2<sup>nd</sup> edition by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a "CC-BY" [Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by/4.0/) – 3/17/2021

1. Use a StatKey to create a histogram and summary statistics for the ages in years. Adjust the histogram to have three bars (3 buckets). Do not copy or draw the histogram. Simply use the histogram to verify the shape.

- a) What is the shape of the data?
- b) What is the sample size ( $n$ )? (This tells us how many total people were measured.)
- c) What is the minimum (smallest value) for the data?
- d) What is the maximum (largest value) for the data?
- d) What is the mean average for this data?
- e) What is the median average for this data?
- f) Based on the shape of this data set, should we use the mean or median as our most accurate average (center)?

2. Use a StatKey to create a histogram and summary statistics for the heights in inches. Adjust the histogram to have three bars (3 buckets). Do not copy or draw the histogram. Simply use the histogram to verify the shape.

- a) What is the shape of the data?
- b) What is the sample size ( $n$ )? (This tells us how many total people were measured.)
- c) What is the minimum (smallest value) for the data?
- d) What is the maximum (largest value) for the data?
- d) What is the mean average for this data?
- e) What is the median average for this data?
- f) Based on the shape of this data set, should we use the mean or median as our most accurate average (center)?

3. Use a StatKey to create a histogram and summary statistics for the weights in pounds. Adjust the histogram to have three bars (3 buckets). Do not copy or draw the histogram. Simply use the histogram to verify the shape.

- a) What is the shape of the data?
- b) What is the sample size ( $n$ )? (This tells us how many total people were measured.)
- c) What is the minimum (smallest value) for the data?
- d) What is the maximum (largest value) for the data?
- d) What is the mean average for this data?
- e) What is the median average for this data?
- f) Based on the shape of this data set, should we use the mean or median as our most accurate average (center)?

4. Use a StatKey to create a histogram and summary statistics for the waist sizes in centimeters. Adjust the histogram to have three bars (3 buckets). Do not copy or draw the histogram. Simply use the histogram to verify the shape.

- a) What is the shape of the data?
- b) What is the sample size ( $n$ )? (This tells us how many total people were measured.)
- c) What is the minimum (smallest value) for the data?
- d) What is the maximum (largest value) for the data?
- d) What is the mean average for this data?
- e) What is the median average for this data?



- f) Based on the shape of this data set, should we use the mean or median as our most accurate average (center)?

5. Use a StatKey to create a histogram and summary statistics for the pulse rates in beats per minute. Adjust the histogram to have three bars (3 buckets). Do not copy or draw the histogram. Simply use the histogram to verify the shape.

- a) What is the shape of the data?
- b) What is the sample size (n)? (This tells us how many total people were measured.)
- c) What is the minimum (smallest value) for the data?
- d) What is the maximum (largest value) for the data?
- d) What is the mean average for this data?
- e) What is the median average for this data?
- f) Based on the shape of this data set, should we use the mean or median as our most accurate average (center)?

6. Use a StatKey to create a histogram and summary statistics for the systolic blood pressure in millimeters of mercury. Adjust the histogram to have three bars (3 buckets). Do not copy or draw the histogram. Simply use the histogram to verify the shape.

- a) What is the shape of the data?
- b) What is the sample size (n)? (This tells us how many total people were measured.)
- c) What is the minimum (smallest value) for the data?
- d) What is the maximum (largest value) for the data?
- d) What is the mean average for this data?
- e) What is the median average for this data?
- f) Based on the shape of this data set, should we use the mean or median as our most accurate average (center)?

7. Use a StatKey to create a histogram and summary statistics for the diastolic blood pressure in millimeters of mercury. Adjust the histogram to have three bars (3 buckets). Do not copy or draw the histogram. Simply use the histogram to verify the shape.

- a) What is the shape of the data?
- b) What is the sample size (n)? (This tells us how many total people were measured.)
- c) What is the minimum (smallest value) for the data?
- d) What is the maximum (largest value) for the data?
- d) What is the mean average for this data?
- e) What is the median average for this data?
- f) Based on the shape of this data set, should we use the mean or median as our most accurate average (center)?

8. Use a StatKey to create a histogram and summary statistics for cholesterol in milligrams per deciliter. Adjust the histogram to have three bars (3 buckets). Do not copy or draw the histogram. Simply use the histogram to verify the shape.

- a) What is the shape of the data?
- b) What is the sample size (n)? (This tells us how many total people were measured.)
- c) What is the minimum (smallest value) for the data?
- d) What is the maximum (largest value) for the data?
- d) What is the mean average for this data?
- e) What is the median average for this data?



- f) Based on the shape of this data set, should we use the mean or median as our most accurate average (center)?

9. Use a StatKey to create a histogram and summary statistics for body mass index (BMI) in kilograms per square meters. Adjust the histogram to have three bars (3 buckets). Do not copy or draw the histogram. Simply use the histogram to verify the shape.

- a) What is the shape of the data?
- b) What is the sample size (n)? (This tells us how many total people were measured.)
- c) What is the minimum (smallest value) for the data?
- d) What is the maximum (largest value) for the data?
- d) What is the mean average for this data?
- e) What is the median average for this data?
- f) Based on the shape of this data set, should we use the mean or median as our most accurate average (center)?

10. Use a StatKey to create a histogram and summary statistics for leg length in inches. Adjust the histogram to have three bars (3 buckets). Do not copy or draw the histogram. Simply use the histogram to verify the shape.

- a) What is the shape of the data?
- b) What is the sample size (n)? (This tells us how many total people were measured.)
- c) What is the minimum (smallest value) for the data?
- d) What is the maximum (largest value) for the data?
- d) What is the mean average for this data?
- e) What is the median average for this data?
- f) Based on the shape of this data set, should we use the mean or median as our most accurate average (center)?

11. Use a StatKey to create a histogram and summary statistics for elbow circumference in inches. Adjust the histogram to have three bars (3 buckets). Do not copy or draw the histogram. Simply use the histogram to verify the shape.

- a) What is the shape of the data?
- b) What is the sample size (n)? (This tells us how many total people were measured.)
- c) What is the minimum (smallest value) for the data?
- d) What is the maximum (largest value) for the data?
- d) What is the mean average for this data?
- e) What is the median average for this data?
- f) Based on the shape of this data set, should we use the mean or median as our most accurate average (center)?

12. Use a StatKey to create a histogram and summary statistics for wrist circumference in inches. Adjust the histogram to have three bars (3 buckets). Do not copy or draw the histogram. Simply use the histogram to verify the shape.

- a) What is the shape of the data?
- b) What is the sample size (n)? (This tells us how many total people were measured.)
- c) What is the minimum (smallest value) for the data?
- d) What is the maximum (largest value) for the data?
- d) What is the mean average for this data?
- e) What is the median average for this data?



- f) Based on the shape of this data set, should we use the mean or median as our most accurate average (center)?

13. Use a StatKey to create a histogram and summary statistics for arm length in inches. Adjust the histogram to have three bars (3 buckets). Do not copy or draw the histogram. Simply use the histogram to verify the shape.

- a) What is the shape of the data?  
b) What is the sample size ( $n$ )? (This tells us how many total people were measured.)  
c) What is the minimum (smallest value) for the data?  
d) What is the maximum (largest value) for the data?  
e) What is the mean average for this data?  
f) What is the median average for this data?  
f) Based on the shape of this data set, should we use the mean or median as our most accurate average (center)?
- 

## Section 4C – Understanding the Mean Average

If you walked up to someone and asked them how to calculate an average, most would tell you to add up the numbers and divide by how many numbers are in the data set. In other words, most people equate the word “average” with the mean average. It is by far the most common average used.

We learned in the last section that in statistics there are many types of averages and the mean average is only accurate when the data is bell shaped (normal). While many people have an idea of how the mean is calculated, very few understand the complexities behind the mean average.



This material is from *Introduction to Data Analysis*, 2<sup>nd</sup> edition by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a “CC-BY” [Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by/4.0/) – 3/17/2021

Since we are in the chapter on analyzing normal (bell shaped) data and data analysts prefer to use the mean average when data is normal, we will focus on understanding the mean average in this section.

**Definition of the Mean Average:** The mean average is the center or average that balances the distances between all of the numbers in a quantitative data set. The mean is only accurate if the data set has a normal (bell) shape.

#### Note on Calculating Statistics

Many people focus on how statistics are calculated instead of the true meaning of the statistic and how to use and explain it properly. Remember, calculations in statistics are extremely time consuming, which is why we prefer to have a computer program do the calculations. What a computer cannot do is tell you what the meaning behind the statistic and when and how it should be used. In statistics, always focus on understanding and being able to explain ideas. That is the real job of a statistician, data scientist, or data analyst.

#### Calculating the Mean Average

Formulas for calculating statistics are very difficult. Focus on understanding the ideas behind the formula, not on using the formula to calculate. Remember, the formulas are already programmed into statistics software programs. The software should be the one doing the calculation. You should be focused on explaining the statistic and what it tells us about the data.

Here are some variables (letters) you often see in statistics formulas for the mean.

n: total frequency or sample size (the number of values in your data set)

x: each individual number in the data set

$\Sigma$  : summation symbol (tells us to add)

$\Sigma x$  : add up all the numbers in your data set

$\bar{x}$  : "x-bar". This symbol is used for the mean average of a data set (sample mean average)

#### **Formula for calculating the mean average**

$$\bar{x} = \frac{\sum x}{n}$$

(Add up all the number in your data set and divide by how many numbers are in your data set.)





### Example 1

As we have said, no statistician calculates the mean with a formula and calculator. The data sets are usually way too large. Since we are just learning about how mean averages work, it would be nice to calculate a couple. If anything, so you have an idea of what the computer is doing.

The following data describes the weights (in kilograms) of various bricks at a building site. Calculate the mean average for the following data:

4.7 , 6.2 , 3.3 , 5.1 , 2.9 , 7.4 , 4.5

How many numbers are in the data set? (This is the total frequency or sample size.)

Seven ( $n = 7$ )

Mean Average =  $(4.7 + 6.2 + 3.3 + 5.1 + 2.9 + 7.4 + 4.5) / 7 = 34.1 / 7 = 4.871428571$

Be sure to add the numbers first and then divide by the frequency.

Where should we round the answer?

**Rounding Rule for Quantitative Data:** *Round statistics calculated from quantitative data to one more decimal place to the right than is present in the original data.*

Notice the numbers in the data set ended in the tenths place (one place to the right of the decimal). This means that we should round our statistic to one more place value to the right. Therefore, we would round to two places to the right of the decimal (hundredths place).

Mean Average Weight of the Bricks =  $4.871428571 \approx 4.87$  kilograms

Remember; focus on interpreting the meaning of this statistic.

What does a mean average of 4.87 kilograms tell us about the data?

A mean average of 4.87 kg tells us that the balancing point for the distances for all the numbers in the data set is 4.87 kg. What does this tell us?

Look at the numbers in the data set above the mean: 6.2, 5.1, and 7.4

Let us look at how far are each of these numbers from the mean? Remember we rounded the mean, so these are just approximate distances.

$$6.2 - 4.87 \approx 1.33$$

$$5.1 - 4.87 \approx 0.23$$



$$7.4 - 4.87 \approx 2.53$$

Therefore, for numbers in the data set above the mean, we have a total approximate distance from the mean of  $1.33 + 0.23 + 2.53 \approx 4.09$

Now look at the numbers in the data set below the mean: 2.9, 3.3, 4.5, and 4.7

Approximately how far are these numbers from the mean? If we subtract in the same order with the value minus the mean we will get negative differences. This issue of negative number differences is a reoccurring problem in statistics that is usually addressed by squaring the values

$$2.9 - 4.87 \approx -1.97$$

$$3.3 - 4.87 \approx -1.57$$

$$4.5 - 4.87 \approx -0.37$$

$$4.7 - 4.87 \approx -0.17$$

Therefore, the total of the differences for numbers below the mean is

$$-1.97 + -1.57 + -0.37 + -0.17 \approx -4.08$$

Technically distances are not negative so the total distance is approximately +4.08

Notice that the total distance for numbers above the mean is almost the same as the total distance for numbers below the mean. This is why the mean is called the “balancing point”. Why is it not perfectly equal? It would be if we used the unrounded version of the mean.

### Understanding the Balancing Point

If you understand that mean is the balancing point, you will not only have a much better understanding of the mean, but you will also be able to estimate the mean in situations and be able to create data sets with a specific mean.

### Example 2

Suppose I want to create a data set five values that has a mean average of 20.

I can pick any numbers I want as long as I balance the distances.

Suppose I use 14, 16, 18, and 19 for my first four numbers. Look at the distance from 20.

14 (six units from 20)

16 (four units from 20)

18 (two units from 20)

19 (one unit from 20)

All these numbers were below 20, so the total distance below so far is  $6 + 4 + 2 + 1 = 13$

If I want a total of five numbers in the data set, I will have to choose one number above 20 that has the same total distance. In this case 13 above 20 or 33.

Therefore, my created data set with five numbers and a mean of 20 is



This material is from *Introduction to Data Analysis*, 2<sup>nd</sup> edition by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a “CC-BY” [Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by/4.0/) – 3/17/2021

14, 16, 18, 19, 33

Let us check it:

$$\text{Mean} = (14 + 16 + 18 + 19 + 33) / 5 = 100 / 5 = 20$$

### More Examples

You can create tons of different data sets, if you understand this principle of the balancing point. Symmetric data sets are probably the easiest to create.

Suppose I want to create a data set with twelve numbers with a mean of 20.

An easy way to do this is to take six numbers above the mean (20) and six numbers below the mean (20). I will pick them so they have the same distances.

Below mean of 20: 14, 15, 16, 17, 18, 19

Above the mean of 20: 21, 22, 23, 24, 25, 26

Notice that 19 and 21 are both one from twenty, 18 and 22 are both two from twenty, and so on. The distances are balanced, so the mean of all of these numbers will be twenty.

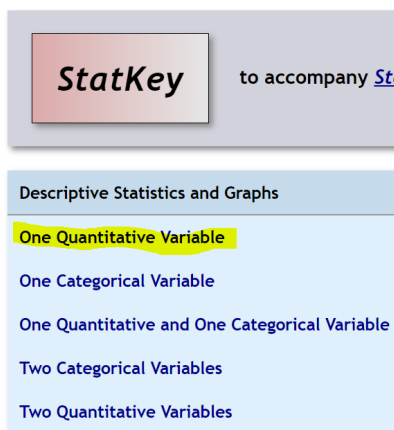
Data set with twelve numbers and a mean of twenty:

14, 15, 16, 17, 18, 19, 21, 22, 23, 24, 25, 26

Let's check and see if this data set has a mean of 20.

Mean Average =  $(14 + 15 + 16 + 17 + 18 + 19 + 21 + 22 + 23 + 24 + 25 + 26) \div 12 = (240) \div 12 = 20$ . The mean is 20.

An easier way would be to go to "One Quantitative Variable" in StatKey at [www.lock5stat.com](http://www.lock5stat.com). If we click the edit data button and type in the numbers. Do not check the box that says identity. Do not check the box that says "data has a header row". Just click "OK".



## StatKey Descriptive Statistics for One Quantitative Variable

Mammal Longevity ▾ Show Data Table Edit Data Upload File Change Column(s)

Edit data ✕

14  
15  
16  
17  
18  
19  
21  
22  
23  
24  
25  
26

First column is identifier

Data has header row

Manually edit the values above or paste a tab or comma separated file into the box and click Ok. The file must have only one column (or two if there is an identifier).

Ok

### Summary Statistics

Statistic	Value
Sample Size	12
Mean	20.000

### Practice Problems Section 4C

Directions for #1-7: Find the mean for the following data sets. You may use a calculator. When rounding is appropriate, round answers to one more decimal place than the numbers in the data set. Then write the definition of the mean average in context to explain the mean for each problem.

$$\text{Mean } (\bar{x}) = \frac{\sum x}{n} = \frac{\text{Sum of the numbers in the data set}}{\text{Sample size (how many numbers are in the data set)}}$$

- Number of dogs at dog hotels: 2, 7, 7, 9, 8, 8, 4, 5, 1, 0, 3, 2, 11, 3, 1, 7, 2, 4
- Number of cars parked in various parking lots: 17, 21, 23, 24, 25, 27, 28, 29, 31, 32, 33, 36
- Temperature in degrees Celsius: 9.4, 3.5, 1.1, 7.8, 3.2, 16.4, 6.6



This material is from *Introduction to Data Analysis*, 2<sup>nd</sup> edition by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a “CC-BY” [Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by/4.0/) – 3/17/2021

4. Grams of medicine: 1.6 , 5.2 , 3.3 , 9.4 , 1.7 , 1.9 , 2.8 , 12.5 , 8.6 , 1.8 , 2.6 , 2.4
5. Dollars spent for a hot dog: 2.54 , 3.14 , 2.49 , 1.98 , 1.46 , 2.27 , 1.83 , 2.63 , 2.87 , 3.25 , 8.75
6. Weight of building stones in kilograms: 1.362 , 5.714 , 3.199 , 2.285 , 4.477 , 9.251
7. Weight of a group of men in pounds: 146 , 157 , 181 , 193 , 226 , 158 , 176 , 187 , 216
  
8. Find a data set with six numbers that has a mean of 13 and without any repeating numbers. Check your answer by calculating the mean to make sure the data set works.
  
9. Add two numbers to your data set in #8, so that the mean remains 13. (You should now have eight numbers in your data set.) There should not be any repeating numbers. Check your answer by calculating the mean to make sure the data set works.
  
10. Find a data set with nine numbers that has a mean of 21.5 and without any repeating numbers. Check your answer by calculating the mean to make sure the data set works.
  
11. Add two numbers to your data set in #10, so that the mean remains 21.5. (You should now have eleven numbers in your data set.) There should not be any repeating numbers. Check your answer by calculating the mean to make sure the data set works.
  
12. Explain how the mean is the balancing point of the data in terms of distances. Look at the following data set. Use the distances to explain how the mean is really 11 without adding the numbers and without calculating the mean directly.

5, 6, 7, 8, 9, 13, 14, 15, 16, 17

---

## Section 4D – Spread, Standard Deviation, and Typical Values for Normal Quantitative Data Sets

When analyzing a quantitative data set, we have seen so far that we want to look at the shape of the data set and we want to find the most accurate center (in which we get the average). There is another description of the data that is important to explore, and that is the “Spread” or “Variability” of a data set.

A measure of spread or variability in a data set tells us how spread out the data is. Why is this important? Let’s look at an example.

Being a teacher, I like to look at quiz scores for my classes.

Class A: 90 , 92 , 99 , 100 , 97 , 96 , 98 , 94 , 91 , 90 , 89 , 100 , 93 , 93 , 88



This material is from *Introduction to Data Analysis*, 2<sup>nd</sup> edition by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a “CC-BY” [Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by/4.0/) – 3/17/2021

This class has a very small spread. Virtually everyone in the class got an A or a high B. These kinds of scores make me very happy as a teacher. A data set with a small spread or small variability means it is more consistent and easier for us to predict future values. I predict quiz scores to be high for this class.

Class B: 26 , 97 , 35 , 84 , 55 , 72 , 61 , 44 , 88 , 69 , 77 , 38 , 51 , 99 , 86

This class has a very large spread with a lot of variability. The quiz scores are all over the place. This class is worrying me. Not only was there many low grades, but the class was very inconsistent. It will be very difficult to predict what quiz grades to expect from these students. I definitely need to review the material more with this class.

### Notes on Spread (Variability)

- **Small Spread** (Small amount of Variability): Tells us the data values are close, more consistent and easier to predict.
- **Large Spread** (Large amount of Variability): Tells us that the data values are very spread out, less consistent, and more difficult to predict.

### Measures of Spread

There are several statistics that measure spread or variability. The most common ones are the range, the interquartile range (IQR), the standard deviation ( $s$ ), and the variance ( $s^2$ ). Which are most accurate? Again it depends on the shape of the data set.

For normal (bell shaped) quantitative data sets, the standard deviation is the most accurate measure of spread.

### **Spread Principle for Normal Quantitative Data**

**When quantitative data is normal, use the standard deviation “s” as our most accurate measure of typical spread.**

**Note: The standard deviation is only accurate if the quantitative data is normal (bell shaped).**

**Definition of Standard Deviation:** The standard deviation is how far typical values are from the mean in a normal (bell shaped) quantitative data set. The standard deviation can be thought of as an average distance from the mean or the typical distance from the mean, but is only accurate if the quantitative data is normal (bell shaped).

Calculations of spread are often even more difficult than measures of center, so it is even more important to use a statistics software program to calculate. For example, calculating standard deviation with a formula and calculator can take a long time, even for a small data set.

Remember how to calculate statistics for one quantitative variable with StatKey?

Go to [www.lock5stat.com](http://www.lock5stat.com) and click on StatKey. Click on “one quantitative variable” under the “descriptive statistics and graphs” menu. Under “edit data”, copy and paste the column of quantitative data into StatKey. Do NOT check the box that says “identifier”. If your data has a title, click on the box that says “data has a header row”. If the data does not have a title, then do NOT check the box that says “data has a header row”.

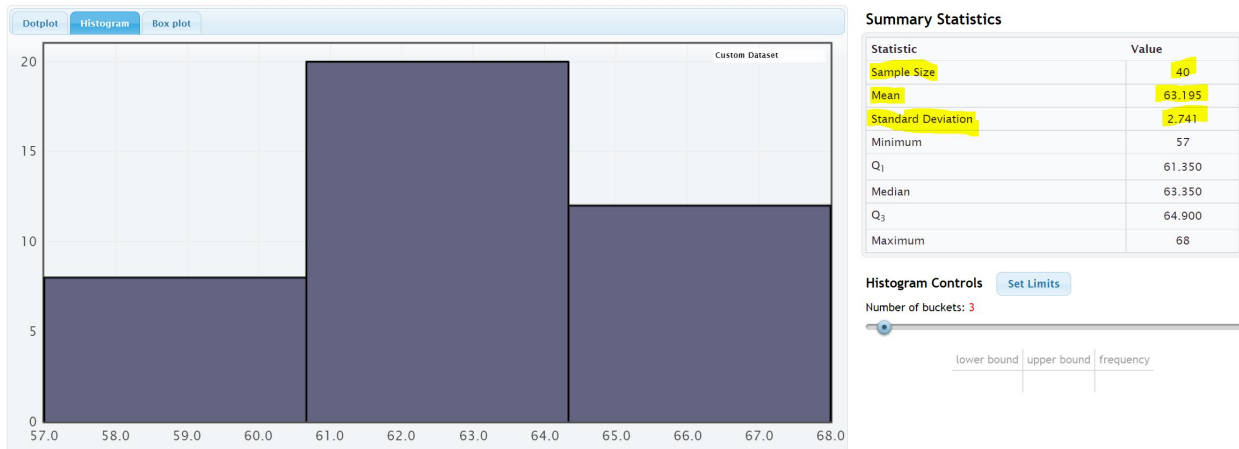
### Example 1

We used StatKey to calculate the mean average and the standard deviation for the women’s heights data. This



This material is from *Introduction to Data Analysis*, 2<sup>nd</sup> edition by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a “CC-BY” [Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by/4.0/) – 3/17/2021

column of data is located in the Health Data Set. Notice first that the data is normal (bell shaped). That tells us that we should use the Standard Deviation as the best measure of typical spread.



Remember to focus on interpretation, not on calculation: In the women's height data, the standard deviation is 2.741 inches. So typical heights for the women were 2.741 inches from the mean on average.

What does this tell us? The mean average for the women's height data was 63.195 inches. So typical women in the data set were within 2.741 inches from 63.195 inches. This gives us a "typical range" (two values that typical numbers in the data are in between).

$$63.195 - 2.741 \leq \text{typical heights for these women} \leq 63.195 + 2.741$$

$$60.454 \leq \text{typical heights for these women} \leq 65.936$$

Typical women in this data set had a height between 60.454 inches (little over 5 feet) and 65.936 inches (little under 5 ½ feet).

**To calculate Typical Values for Normal Quantitative Data Sets: Add and subtract the mean and standard deviation. (Be careful to subtract in the correct order.)**

$$\text{Mean} - \text{Standard Deviation} \leq \text{typical values} \leq \text{Mean} + \text{Standard Deviation}$$

#### Empirical Rule for Normal (Bell Shaped) Quantitative Data Sets

After looking at a lot of bell shaped data sets over the years, statisticians found that usually about 68% of the data values fall within one standard deviation of the mean. This means that in a bell shaped data set, approximately the middle 68% of the values are considered typical. Since this seemed to be the case for most bell shaped data sets, it is often referred to as the "Empirical Rule". The more bell shaped the data set is the more accurate the 68% is. The Empirical Rule does not apply to skewed data sets.

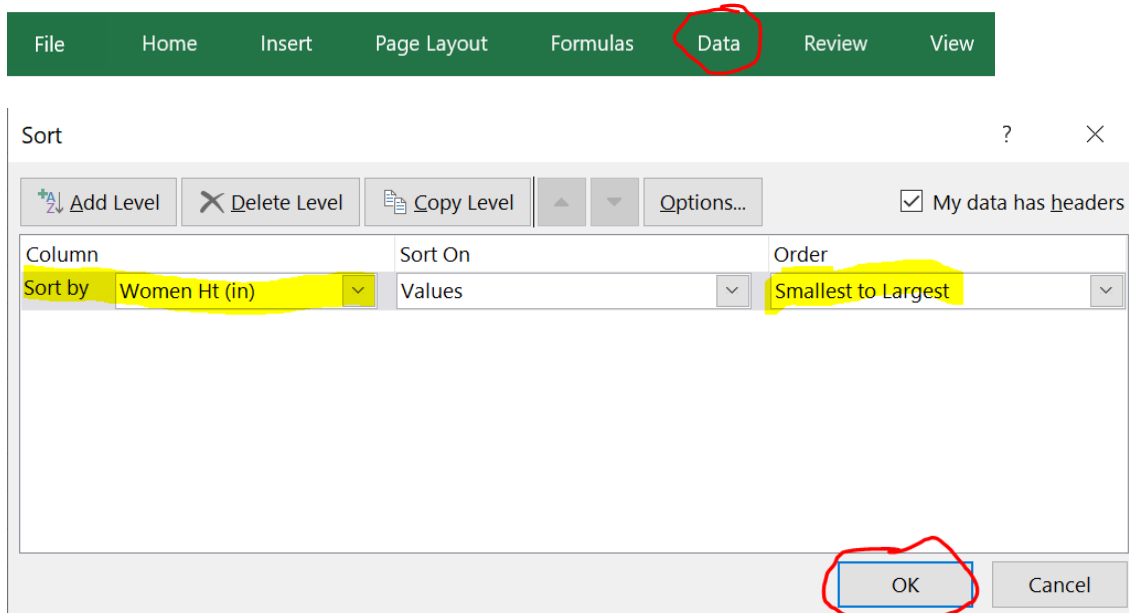


- **Typical Values in normal (bell shaped) data sets make up the middle 68% of the data values and are within ONE standard deviation from the mean.**

### Example

In our last example, we saw that typical women in the health data had a height between 60.454 inches and 65.936 inches. If we look put the column of data in order from smallest to largest, we can identify which heights were considered “typical”.

To put a data set in order in Excel, first highlight the entire column. Then click on the “data” tab, then click on “sort”. You should see the data set you want to sort under “sort by” and under “order”, you should see “Smallest to Largest”. Now just push “OK”.



Now that we have the data in order, we can identify the typical values in the data. The typical heights will be between 60.454 inches and 65.936 inches.





A	B
Women Ht (in)	
57	
58.2	
58.6	
59.6	
59.8	
60.2	
60.5	Typical
60.6	Typical
60.7	Typical
61.3	Typical
61.4	Typical
61.8	Typical
61.9	Typical
62.3	Typical
62.3	Typical
62.6	Typical
62.7	Typical
63.1	Typical
63.2	Typical
63.3	Typical
63.4	Typical
63.4	Typical
63.4	Typical
63.4	Typical
63.5	Typical
63.6	Typical
64.1	Typical
64.3	Typical
64.3	Typical
64.7	Typical
64.8	Typical
65	Typical
65.1	Typical
66.4	
66.7	
66.7	
66.8	
67	
67.6	
67.9	
68	

Notice that the middle 26 heights (65%) out of 40 total women were typical. This data was not perfectly normal, so we are not surprised it is not exactly 68%. Still it is close to what the empirical rule predicts.

### Calculating Standard Deviation

As I said earlier, no one calculates standard deviation by hand. Always use a computer. I will show you the formula and calculation so that you can get a sense of what the computer is doing.

Let us look at the brick weight data from the previous section.

4.7 , 6.2 , 3.3 , 5.1 , 2.9 , 7.4 , 4.5

The standard deviation is the typical distance from the mean, so when calculating the standard deviation you need to know how many numbers are in the data set (seven) and you need to know the mean average.

$$\text{Mean Average} = (4.7 + 6.2 + 3.3 + 5.1 + 2.9 + 7.4 + 4.5) / 7 = 34.1 / 7 = 4.871428571 \approx 4.87$$

I will be using the rounded value of the mean. Computers are always much more accurate since they carry many decimal places of accuracy.

Let us look at how far are each of these numbers from the mean? We will subtract the mean from each number in the data set  $(x - \bar{x})$ . Remember we rounded the mean, so these are just approximate distances.



$$6.2 - 4.87 \approx 1.33$$

$$5.1 - 4.87 \approx 0.23$$

$$7.4 - 4.87 \approx 2.53$$

$$2.9 - 4.87 \approx -1.97$$

$$3.3 - 4.87 \approx -1.57$$

$$4.5 - 4.87 \approx -0.37$$

$$4.7 - 4.87 \approx -0.17$$

Notice that some of the differences are negative and some are positive. In fact, if we were to add the distances now, they would add up to approximately zero. (Remember the mean is the balancing point.)

The negative numbers are a problem. To average the distances we need to get rid of the negatives. There are two ways to deal with negative numbers in mathematics, absolute value or squaring the numbers. Absolute value can have issues with calculus applications, so early statisticians preferred to square all the numbers and then eventually take a square root.

Squares of the distances

$$(1.33)^2 \approx 1.7689$$

$$(0.23)^2 \approx 0.0529$$

$$(2.53)^2 \approx 6.4009$$

$$(-1.97)^2 \approx 3.8809$$

$$(-1.57)^2 \approx 2.4649$$

$$(-0.37)^2 \approx 0.1369$$

$$(-0.17)^2 \approx 0.0289$$

Now we will add up all the squared distances and calculate the “Sum of Squares”  $\sum (x - \bar{x})^2$ . This is a very important technique in statistics and occurs in many different applications.

$$\text{Sum of Squares} \approx 1.7689 + 0.0529 + 6.4009 + 3.8809 + 2.4649 + 0.1369 + 0.0289 \approx 14.6814$$

We now want to take an average of the sum of squares. When dealing with spread, we will divide by one less than the sample size. This is often called “degrees of freedom” in statistics. Therefore, we will divide by  $n-1$  instead of the frequency  $n$ . There are seven numbers in the data set, so we will divide by  $7 - 1$  or 6. Then we will take the square root of the answer.

Standard Deviation Formula



This material is from *Introduction to Data Analysis*, 2<sup>nd</sup> edition by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a “CC-BY” [Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by/4.0/) – 3/17/2021

$$\sqrt{\frac{\sum (x - \bar{x})^2}{(n-1)}} = \sqrt{\frac{14.6814}{(7-1)}} = \sqrt{\frac{14.6814}{6}} = \sqrt{2.4469} \approx 1.564 \text{ kilograms}$$

We calculated the standard deviation with StatKey and got approximately 1.567 kg. StatKey is more accurate since it has less rounding error.

### Degrees of Freedom

Why do we divide by n-1 when calculating the standard deviation? That is a good question. Think of it this way. Suppose you take a history class and your grade is based on six exams. The first five exams can have some variability. Maybe you got an 88 on the first exam, a 93 on the second exam, and so on. You want to get a 90 overall mean average to get an A in the history class. Therefore, once you know your first five exam scores, you can calculate what you need to get on the last exam to get an A in class. In other words, the last exam score is fixed in the sense that we can calculate it.

That is how degrees of freedom works. If we have a given mean average, n-1 of the numbers have variability from that mean, but the last number can be calculated. Therefore, if we have the heights of forty women and we know the mean average, then we should only measure the variability of 39 of those numbers.

What is important?

If a quantitative data set is normal (bell shaped), use the mean as the center or average. Use the standard deviation as the best measure of spread. Do not calculate these with formula and calculator. Use a computer program like StatKey.

Remember focus on interpretation not calculation. You should be able to explain the mean and standard deviation for a data set to someone. You should also be able to use the mean and standard deviation to calculate the typical values by adding and subtracting the mean and standard deviation.

Key: The mean and standard deviation should only be used if the data set is normal. They are not accurate if the data set is skewed or non-normal.

---



## Practice Problems Section 4D

1. What is the definition of standard deviation?
2. For what shape is the standard deviation an accurate measure of typical spread?
3. For what shapes is the standard deviation not an accurate measure of typical spread?
4. How can we use the mean and standard deviation to identify typical values in a normally distributed (bell shaped) data set?
5. How many standard deviations from the mean is considered typical for normally distributed (bell-shaped) data?
6. What percentage of the data values are considered typical for normally distributed (bell shaped) data?

Directions #7-9: Fill out the following tables in order to calculate the Standard Deviation ( $s$ ) for each of the following data sets. The mean average has already been calculated for you. To calculate standard deviation ( $s$ ), subtract each number from the mean and square the differences. Then add up the squared differences (sum of squares). Then divide by the degrees of freedom ( $n-1$ ). Last, take the square root of the answer.

$$\text{(Mean) } \bar{x} = \frac{\sum x}{n} \qquad \text{(Standard Deviation) } s = \sqrt{\frac{\sum (x - \bar{x})^2}{n - 1}}$$

7.      1, 2, 3, 11, 12, 13      Mean ( $\bar{x}$ ) = 7

Values in data set ( $x$ )    Differences: Each Value – mean ( $x - \bar{x}$ )    Squares of Differences ( $(x - \bar{x})^2$ )

1  
2  
3  
11  
12  
13

Sum of squares  $\sum (x - \bar{x})^2 =$

Sample Size (How many numbers in the data set) ( $n$ ) =

Degrees of Freedom ( $n - 1$ ) =

Standard Deviation ( $s$ ) =



8. 2, 5, 6, 7, 17, 18, 19, 22 Mean ( $\bar{x}$ ) = 12

Values in data set (x) Differences: Each Value – mean ( $x - \bar{x}$ ) Squares of Differences ( $(x - \bar{x})^2$ )

2

5

6

7

17

18

19

22

Sum of squares  $\sum (x - \bar{x})^2 =$

Frequency (How many numbers in the data set) (n) =

Degrees of Freedom (n – 1) =

Standard Deviation (s) =

9. 5, 8, 14, 21, 30, 35, 41 Mean ( $\bar{x}$ ) = 22

Values in data set (x) Differences: Each Value – mean ( $x - \bar{x}$ ) Squares of Differences ( $(x - \bar{x})^2$ )

5

8

14

21

30

35

41



Sum of squares  $\sum (x - \bar{x})^2 =$

Sample Size (How many numbers in the data set) (n) =

Degrees of Freedom (n - 1) =

Standard Deviation (s) =

(Directions #10-11): The following histogram and statistics were calculated using the "Bear" data and Statcato. Use the histogram and statistics provided to answer the following questions.

10. Bear neck circumference (inches)

- a) What is the data measuring?
- b) What are the units?
- c) How many numbers are in the data set?
- d) Is the data set normally distributed (bell shaped)? (Yes or No)
- e) Is the mean an accurate average in this data? Why or why not?
- f) Is the standard deviation an accurate measure of typical spread for this data? Why or why not?
- g) Write a sentence to explain the standard deviation in this context.
- h) Use the mean, standard deviation and the following formula to calculate two numbers that typical values in this data fall in between.

Mean - Standard Deviation  $\leq$  Typical Values  $\leq$  Mean + Standard Deviation

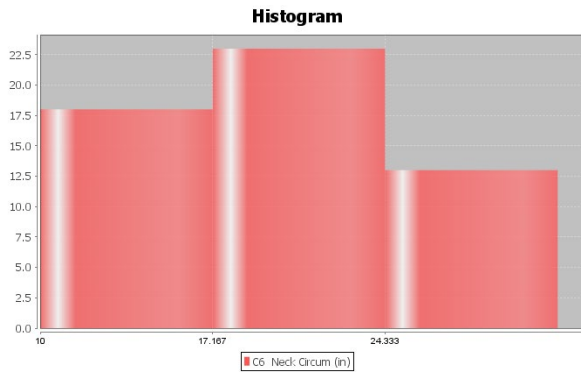
**Descriptive Statistics**

Variable	Mean	Standard Deviation
C6 Neck Circum (in)	20.556	5.641

Variable	Min	Max
C6 Neck Circum (in)	10.0	31.5

Variable	N total
C6 Neck Circum (in)	54





11. Bear Chest Size (inches)

- a) What is the data measuring?
- b) What are the units?
- c) How many numbers are in the data set?
- d) Is the data set normally distributed (bell shaped)? (Yes or No)
- e) Is the mean an accurate average in this data? Why or why not?
- f) Is the standard deviation an accurate measure of typical spread for this data? Why or why not?
- g) Write a sentence to explain the standard deviation in this context.
- h) Use the mean, standard deviation and the following formula to calculate two numbers that typical values in this data fall in between.

$$\text{Mean} - \text{Standard Deviation} \leq \text{Typical Values} \leq \text{Mean} + \text{Standard Deviation}$$

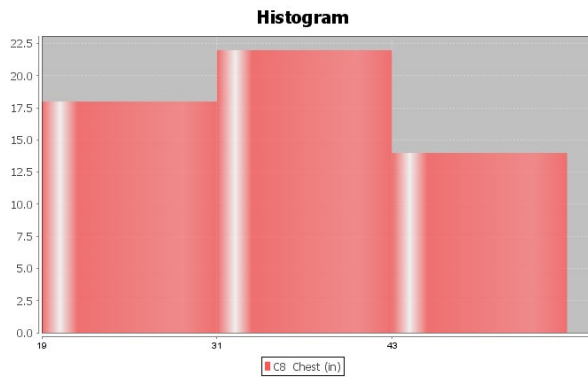
**Descriptive Statistics**

Variable	Mean	Standard Deviation
C8 Chest (in)	35.663	9.352

Variable	Min	Max
C8 Chest (in)	19.0	55.0

Variable	N total
C8 Chest (in)	54





(Directions #12-15): Open “Health” data in Canvas or at [www.matt-teachout.org](http://www.matt-teachout.org) . Go to [www.lock5stat.com](http://www.lock5stat.com) and click on StatKey. Then click on “One Quantitative Variable” and “Edit Data”. Copy and paste the indicated column of data into StatKey and push OK. Create a histogram with only 3 bars (3 buckets) and verify that the data looks normal. Look at the summary statistics to answer the following questions..

12. Women’s Diastolic Blood Pressure (Millimeters of Mercury (mm of Hg))

- What is the data measuring?
- What are the units?
- How many numbers are in the data set?
- Is the data set normally distributed (bell shaped)? (Yes or No)
- Is the mean an accurate average in this data? Why or why not?
- Is the standard deviation an accurate measure of typical spread for this data? Why or why not?
- Write a sentence to explain the standard deviation in this context.
- Use the mean, standard deviation and the following formula to calculate two numbers that typical values in this data fall in between.

$$\text{Mean} - \text{Standard Deviation} \leq \text{Typical Values} \leq \text{Mean} + \text{Standard Deviation}$$

13. Women’s Wrist Circumference (Inches)

- What is the data measuring?
- What are the units?
- How many numbers are in the data set?
- Is the data set normally distributed (bell shaped)? (Yes or No)
- Is the mean an accurate average in this data? Why or why not?
- Is the standard deviation an accurate measure of typical spread for this data? Why or why not?
- Write a sentence to explain the standard deviation in this context.
- Use the mean, standard deviation and the following formula to calculate two numbers that typical values in this data fall in between.

$$\text{Mean} - \text{Standard Deviation} \leq \text{Typical Values} \leq \text{Mean} + \text{Standard Deviation}$$





14. Men's Height (Inches)

- a) What is the data measuring?
- b) What are the units?
- c) How many numbers are in the data set?
- d) Is the data set normally distributed (bell shaped)? (Yes or No)
- e) Is the mean an accurate average in this data? Why or why not?
- f) Is the standard deviation an accurate measure of typical spread for this data? Why or why not?
- g) Write a sentence to explain the standard deviation in this context.
- h) Use the mean, standard deviation and the following formula to calculate two numbers that typical values in this data fall in between.

$$\text{Mean} - \text{Standard Deviation} \leq \text{Typical Values} \leq \text{Mean} + \text{Standard Deviation}$$

15. Men's Weight (Pounds)

- a) What is the data measuring?
- b) What are the units?
- c) How many numbers are in the data set?
- d) Is the data set normally distributed (bell shaped)? (Yes or No)
- e) Is the mean an accurate average in this data? Why or why not?
- f) Is the standard deviation an accurate measure of typical spread for this data? Why or why not?
- g) Write a sentence to explain the standard deviation in this context.
- h) Use the mean, standard deviation and the following formula to calculate two numbers that typical values in this data fall in between.

$$\text{Mean} - \text{Standard Deviation} \leq \text{Typical Values} \leq \text{Mean} + \text{Standard Deviation}$$

---



## Section 4E – Unusual Values in Normal Data, Using the Dot Plot, and Summarizing Quantitative Data

In this section, we will try to summarize how to analyze a normal (bell shaped) quantitative data set. When analyzing a quantitative data set there are a few key things to address.

### Quantitative Data Analysis Summary

- What is the data measuring? What are the units?
- How many numbers are in the data set? (Sample Size “n”)
- What is the shape of the data? Is the data bell shaped (normal), skewed right (long right tail), or skewed left (long left tail). (In this section, all of the data sets are normal.)
- What is the best measure of center? This will be the average. (If the data is normal (bell shaped), these should both be the mean average. Write a sentence to explain the mean average.
- What is the best measure of spread? (If the data set is normal (bell shaped), this should be the standard deviation. Write a sentence to explain the standard deviation.)
- Find two numbers that typical values fall in between. If the data is normal (bell shaped), then we should add and subtract the mean and standard deviation to calculate the two numbers.  
Mean – Standard Deviation  $\leq$  Typical Values in Data  $\leq$  Mean + Standard Deviation
- Find any unusual values in the data set. (Some call these unusual values “outliers”.)
- I usually like to give the smallest and largest numbers in the data set, even if they are not unusual.

### Finding Outliers (Unusual Values) in a Bell Shaped (Normal) Data Set

So how do you find unusual values in a normal (bell shaped) data set? Unusual values are often called “outliers” in statistics. It has long been considered that one standard deviation from the mean is considered typical. Any values that are two standard deviations or more from the mean is considered unusual. So any value in the data that is two standard deviations or more above the mean is considered “unusually high” or a “high outlier”. Any value in the data that is two standard deviations or less below the mean is considered “unusually low” or a “low outlier”.

Remember these rules only apply when data is normal (bell shaped). If the data is skewed right or skewed left, these rules do not apply.



This material is from *Introduction to Data Analysis*, 2<sup>nd</sup> edition by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a “CC-BY” [Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by/4.0/) – 3/17/2021

**High Outlier Cutoff** (for normal quantitative data):  $mean + (2 \times Standard\ Deviation)$

**Low Outlier Cutoff** (for normal quantitative data):  $mean - (2 \times Standard\ Deviation)$

When calculating the cut off, be sure to multiply the standard deviation by 2 before doing the adding or subtracting. Also, be sure to add and subtract in the correct order.

The cutoff's themselves are not necessarily numbers in the data set. Think of them as fences. If a value in the data set is greater than or equal to the unusual high cutoff, then it is considered unusually high (high outlier). If a value in the data set is less than or equal to the unusual low cutoff, then it is considered unusually low (low outlier).

**High Outliers** (for normal quantitative data)  $\geq mean + (2 \times Standard\ Deviation)$

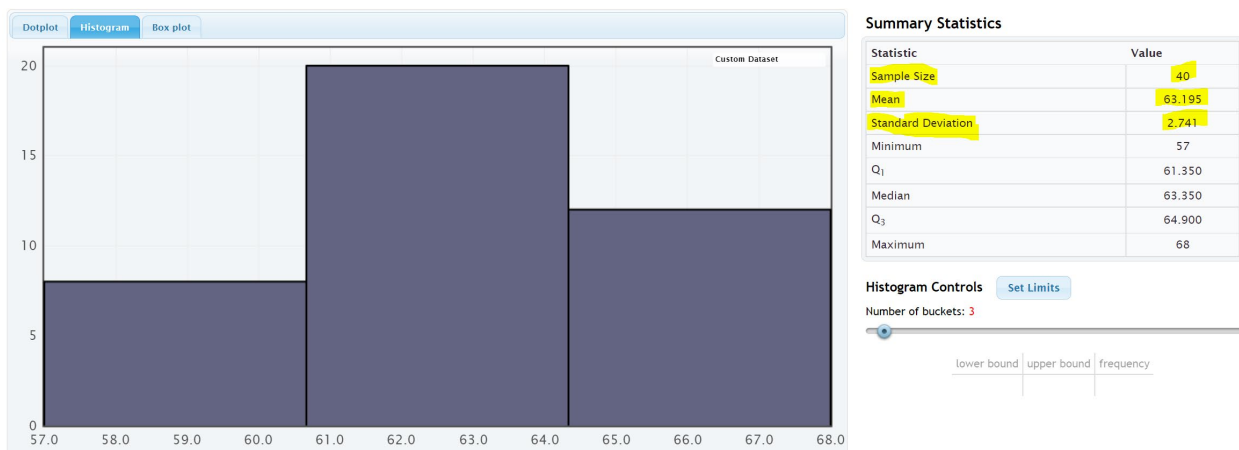
**Low Outliers** (for normal quantitative data)  $\leq mean - (2 \times Standard\ Deviation)$

### Use the column of data

Once calculating the outlier cutoffs, I find it very easy to simply put the column of data in order from lowest to highest and then identify all of the values that are greater than or equal to the high cutoff and all the values that are less than or equal to the low cutoff.

### Example

We used StatKey to calculate the mean average and the standard deviation for the women's heights data. This column of data is located in the "Health" data set. Notice first that the data is normal (bell shaped). That tells us that we should use the Mean as the center (average) and the Standard Deviation as the best measure of typical spread.



Let's see if there are any outliers (unusual values) in this data. Not all data sets have outliers.

$$\text{High Outlier Cutoff} = \text{mean} + (2 \times \text{standard deviation}) = 63.195 + (2 \times 2.741) = 63.195 + (5.482) = 68.677$$

$$\text{Low Outlier Cutoff} = \text{mean} - (2 \times \text{standard deviation}) = 63.195 - (2 \times 2.741) = 63.195 - (5.482) = 57.713$$

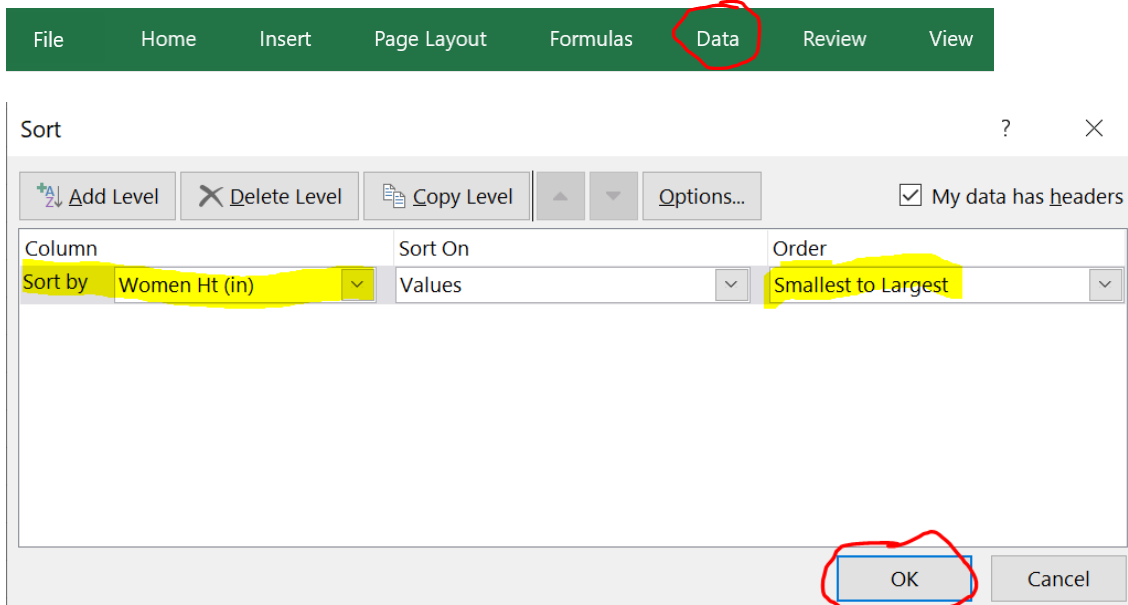
So unusually tall women in this data will have a height of 68.677 inches or higher.

So unusually short women in this data will have a height of 57.713 inches or less.



Note: Remember, these cutoffs only apply to women in this data set and do not apply to all women.

To put a data set in order in Excel, first highlight the entire column. Then click on the “data” tab, then click on “sort”. You should see the data set you want to sort under “sort by” and under “order”, you should see “Smallest to Largest”. Now just push “OK”.



In the last section we saw that typical values were within one standard deviation from the mean. For the women's height data, all values between 60.454 inches and 65.936 inches were considered typical. We can now identify the unusual values (outliers) as well.



Women HT (in)	
57	LOW OUTLIER!!
	Low Outlier Cutoff (57.713)
58.2	
58.6	
59.6	
59.8	
60.2	
	Typical Values Cutoff (60.454)
60.5	Typical
60.6	Typical
60.7	Typical
61.3	Typical
61.4	Typical
61.8	Typical
61.9	Typical
62.3	Typical
62.3	Typical
62.6	Typical
62.7	Typical
63.1	Typical
63.2	Typical
63.3	Typical
63.4	Typical
63.4	Typical
63.4	Typical
63.5	Typical
63.6	Typical
64.1	Typical
64.3	Typical
64.3	Typical
64.7	Typical
64.8	Typical
65	Typical
65.1	Typical
	Typical Values Cutoff (65.936)
66.4	
66.7	
66.7	
66.8	
67	
67.6	
67.9	
68	
	High Outlier Cutoff (68.677)
	NO HIGH OUTLIERS!!

Notice that there is only one number (57) in the data set that is less than or equal to the low outlier cutoff of 57.713. So height is 57 inches is considered unusually low (low outlier) compared to the rest of the data.

Notice also that there are no numbers in the data set that are greater than or equal to the high outlier cutoff of 68.677. So there are no unusually high values in the data set (no high outliers). Notice the tallest woman in the data set was 68 inches tall, but her height is not considered unusual compared to the rest of the women.

Notice also that not all data values are either typical or unusual. There are values in between. These values are not typical and they are not unusual.

Using a dot plot: Once you find the unusual cutoffs, you can also use a dot plot to identify those values in the data set that are unusually high or unusually low. This works ok for smaller data sets, but for larger data sets, I prefer to use the actual column of data. In a data set of 10,000 values for example, you may have 500 outliers. It is hard to pick out 500 dots from a dot plot.





A common questions students ask me is if less than 1 standard deviation or less is considered “typical” and 2 standard deviation or more is considered “unusual”, then what about all the values that are in between 1 and 2 standard deviations away from the mean? They are not typical and they are not unusual.

The Empirical Rule discussed in the last section can shed some light on this issue.

#### Empirical Rule for Normal (Bell Shaped) Data Sets

After looking at a lot of bell shaped data sets over the years, statisticians found that usually about 68% of the data values fall within one standard deviation of the mean. This means that in a bell shaped data set, approximately the middle 68% of the values are considered typical.

Unusually high values (high outliers) in a normal (bell shaped) data set are in the top 2.5% of the data and usually corresponds to about two standard deviations above the mean or higher. Unusually low values (low outliers) in a normal (bell shaped) data set are in the bottom 2.5% of the data and usually corresponds to about two standard deviations below the mean or less. The middle 95% of a bell shaped data set is not considered unusual.

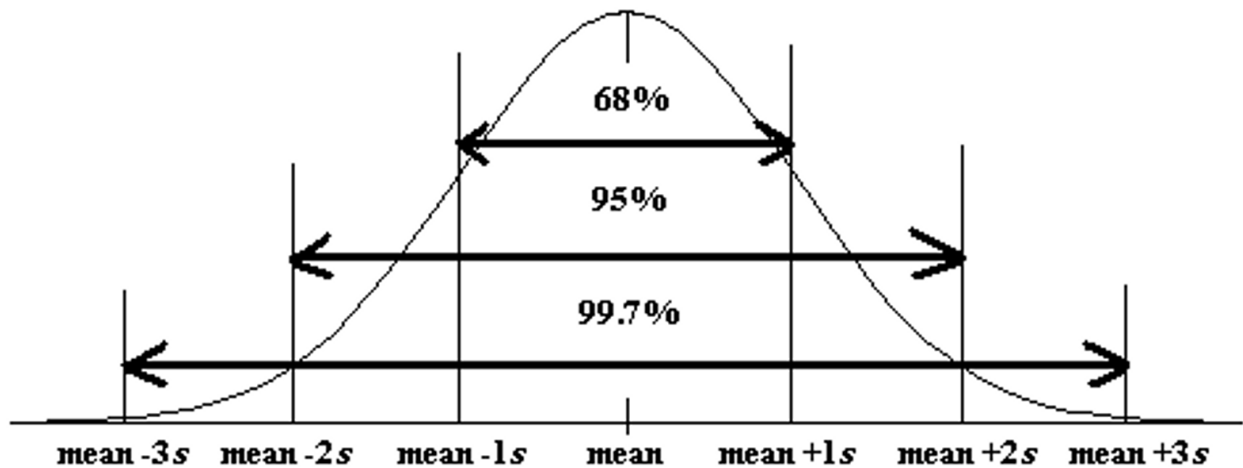
It turns out that almost all of a bell shaped data set (99.7%) is within three standard deviations of the mean. Remember, the empirical rule percentages are rarely perfectly accurate. The more normal distributed the data set, the more accurate the Empirical Rule percentages. The Empirical Rule does not apply to skewed data sets or non-normal data sets. The following diagram describes the Empirical Rule for normal data. Notice the curve looks like a bell.



<= unusually low ==|

|==== typical values ====|

|== unusually high ==>



Now let us see if we can summarize this chapter. This information is often summarized in a “Data Analysis Report” or a “Data Analysis Paragraph”.

- **Normal (normally distributed) quantitative data is unimodal and symmetric and generally has a bell shape. The highest bar in the histogram is in the middle and the right and left tail are about the same length.**
- **For normally distributed data, we will use the mean as our average or center.**
- **For normally distributed data, we will use the standard deviation as our best measure of spread.**
- **For normally distributed data, typical values for normal data will be within one standard deviation from the mean and will make up the middle 68% of the data.**
- **For normally distributed data, unusually high values in the data (high outliers) will be two or more standard deviations above the mean and make up the top 2.5% of the data.**
- **For normally distributed data, unusually low values in the data (low outliers) will be two or more standard deviations below the mean and make up the bottom 2.5% of the data.**
- **We should also include general information like what the data is measuring, the units, how many people or objects were measured (sample size), and the smallest and largest values in the data set.**

#### Quantitative Data Analysis Summary Report

- What is the data measuring? What are the units?
- How many numbers are in the data set? (Frequency “N” or Sample Size)
- What is the shape of the data? (This will be bell shaped in this section.)
- What is the best measure of center? This is the average. (If the data is normal, the center or average should be the mean. Write a sentence to explain the mean average.)
- What is the best measure of spread? (If the data set is normal, this should be the standard deviation. Write a sentence to explain the standard deviation.)
- Find two numbers that typical values fall in between. If the data is normal then we should add and subtract the mean and standard deviation and look for data values in between.  
Mean – Standard Deviation ≤ Typical Data Values ≤ Mean + Standard Deviation
- Find any unusual values (high outliers) and unusually low values (low outliers) in the data set. Calculate the high outlier cutoff and the low outlier cutoff. Put the column of data in order from smallest to largest. Look for values in the column that are greater than or equal to the high cutoff or less than or equal to the low cutoff.



This material is from *Introduction to Data Analysis*, 2<sup>nd</sup> edition by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a “CC-BY” [Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by/4.0/) – 3/17/2021

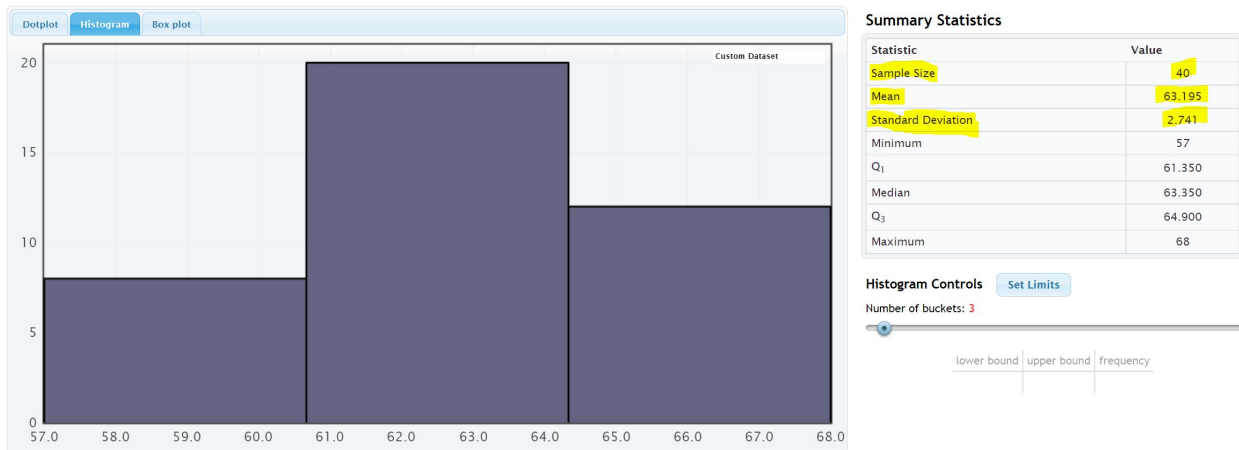
Unusual High Cutoff:  $mean + (2 \times Standard\ Deviation)$

Unusual Low Cutoff:  $mean - (2 \times Standard\ Deviation)$

## Example

Analyze the women's heights data located in the "Health" data.

First we need to put the data into StatKey and create a histogram and find the summary statistics.



Notice that the women's height data has a relatively normal shape. The highest bar is in the middle and the left and right tails are about the same length.

We see that the sample size is 40. So the data describes the heights of 40 women.

The shortest woman was 57 inches and the tallest woman was 68 inches.

Since the data is normally distributed, the best center (average) is the mean of 63.195 inches. The average and the balancing point for this data is 63.195 inches.

Since the data is normally distributed, the best measure of spread is the standard deviation of 2.741 inches. So typical values in the data are within 2.741 inches from the mean.

Typical values in this data are between  $(63.195 - 2.741)$  and  $(63.195 + 2.741)$ . So typical heights are between 60.454 inches and 65.936 inches.

Unusually high values (high outliers)  $\geq 63.195 + (2 \times 2.741)$

Unusually high values (high outliers)  $\geq 68.677$  inches

Looking at the column of data with data values in order, we see that there are no high outliers in this data set.

Unusually low values (low outliers)  $\leq 63.195 - (2 \times 2.741)$

Unusually low values (low outliers)  $\leq 57.713$  inches

Looking at the column of data with data values in order, we see that there is only one low outlier. The height of 57 inches is considered unusually low when compared to the other data values.

## Writing a Summary Paragraph (Report)

Data analysts often summarize their findings in a paragraph. Think of this as a small report that explains the key features of the quantitative data set. You just need to write a sentence for each part of the summary.



This material is from *Introduction to Data Analysis*, 2<sup>nd</sup> edition by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a "CC-BY" [Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by/4.0/) – 3/17/2021



### Example Summary Report Paragraph

Women's Height Summary Report Paragraph: *This data describes the heights in inches of 40 women. The histogram showed a bell shape, so this data is considered normal or normally distributed. The mean average height of the women was 63.195 inches (5 ft. 3 in.). So the center or balancing point for this data was 63.195 inches. The typical spread for this data was 2.741 inches (standard deviation). So typical values in the data were within 2.741 inches from the mean. This means that typical heights for these women were in between 60.454 inches (a little over 5 ft.) and 65.936 inches (about 5 ft. 6 in.). There were no unusually high values (no high outliers). The tallest woman in the data set was 68 inches (5 ft. 8 in), but this was not unusual. The shortest woman in the data set was 57 inches (4 ft. 9 in.). This height was the only outlier in the data. The height of 57 inches was considered to be unusually short (low outlier) compared to the other women.*

---

### Practice Problems Section 4E

1. What is the definition of the mean average?
2. What is the definition of standard deviation?
3. For what shape is the mean an accurate average (accurate measure of center)?
4. For what shapes is the mean NOT an accurate average (not accurate measure of center)?



This material is from *Introduction to Data Analysis*, 2<sup>nd</sup> edition by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a “CC-BY” [Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by/4.0/) – 3/17/2021

5. For what shape is the standard deviation an accurate measure of spread?
6. For what shapes is the standard deviation NOT an accurate measure of spread?
7. High outliers (unusually high values) are how many standard deviations above the mean?
8. What percentage of values in a large normally distributed data set are usually considered high outliers?
9. How can we use the mean and standard deviation to identify unusually high values (high outliers) in a normally distributed data set?
10. Low outliers (unusually high values) are how many standard deviations below the mean?
11. What percentage of values in a large normally distributed data set are usually considered low outliers?
12. How can we use the mean and standard deviation to identify unusually low values (low outliers) in a normally distributed data set?

(Directions #13-14): Open “Bear” data in Canvas or at [www.matt-teachout.org](http://www.matt-teachout.org) . Go to [www.lock5stat.com](http://www.lock5stat.com) and click on StatKey. Then click on “One Quantitative Variable” and “Edit Data”. Copy and paste the indicated column of data into StatKey and push OK. Create a histogram with only 3 bars (3 buckets) and verify that the data looks normal. Use the histogram and the mean and standard deviation from the “summary statistics” printout to answer the following questions.

13. Bear neck circumference (inches)

- a) Is the data set normally distributed (bell shaped)? (Yes or No)
- b) Is the mean an accurate average in this data? Why or why not?
- c) Is the standard deviation an accurate measure of typical spread for this data? Why or why not?
- d) Use the mean, standard deviation and the following formula to calculate the high outlier cutoff for this data.  
High Outlier Cutoff = Mean + (2 x Standard Deviation)
- e) Use the mean, standard deviation and the following formula to calculate the low outlier cutoff for this data.  
High Outlier Cutoff = Mean – (2 x Standard Deviation)
- e) Copy the column of data into a new spreadsheet. Put the data values in order from smallest to largest. List all the values in the data set that are greater than or equal to the high outlier cutoff. These are the high outliers. If there are none, simply say “no high outliers”.
- f) Use the column of data that is in order from smallest to largest. List all the values in the data set that are less than or equal to the low outlier cutoff. These are the low outliers. If there are none, simply say “no low outliers”.

14. Bear Chest Size (inches)

- a) Is the data set normally distributed (bell shaped)? (Yes or No)
- b) Is the mean an accurate average in this data? Why or why not?
- c) Is the standard deviation an accurate measure of typical spread for this data? Why or why not?
- d) Use the mean, standard deviation and the following formula to calculate the high outlier cutoff for this data.  
High Outlier Cutoff = Mean + (2 x Standard Deviation)
- e) Use the mean, standard deviation and the following formula to calculate the low outlier cutoff



for this data.

High Outlier Cutoff = Mean - (2 x Standard Deviation)

- e) Copy the column of data into a new spreadsheet. Put the data values in order from smallest to largest. List all the values in the data set that are greater than or equal to the high outlier cutoff. These are the high outliers. If there are none, simply say "no high outliers".
- f) Use the column of data that is in order from smallest to largest. List all the values in the data set that are less than or equal to the low outlier cutoff. These are the low outliers. If there are none, simply say "no low outliers".

(Directions #15-18): Open "Health" data in Canvas or at [www.matt-teachout.org](http://www.matt-teachout.org) . Go to [www.lock5stat.com](http://www.lock5stat.com) and click on StatKey. Then click on "One Quantitative Variable" and "Edit Data". Copy and paste the indicated column of data into StatKey and push OK. Create a histogram with only 3 bars (3 buckets) and verify that the data looks normal. Use the histogram and the mean and standard deviation from the "summary statistics" printout to answer the following questions.

15. Women's Diastolic Blood Pressure (Millimeters of Mercury (mm of Hg))

- a) Is the data set normally distributed (bell shaped)? (Yes or No)
- b) Is the mean an accurate average in this data? Why or why not?
- c) Is the standard deviation an accurate measure of typical spread for this data? Why or why not?
- d) Use the mean, standard deviation and the following formula to calculate the high outlier cutoff for this data.

High Outlier Cutoff = Mean + (2 x Standard Deviation)

- e) Use the mean, standard deviation and the following formula to calculate the low outlier cutoff for this data.
- f) Use the column of data that is in order from smallest to largest. List all the values in the data set that are less than or equal to the low outlier cutoff. These are the low outliers. If there are none, simply say "no low outliers".

High Outlier Cutoff = Mean - (2 x Standard Deviation)

- e) Copy the column of data into a new spreadsheet. Put the data values in order from smallest to largest. List all the values in the data set that are greater than or equal to the high outlier cutoff. These are the high outliers. If there are none, simply say "no high outliers".
- f) Use the column of data that is in order from smallest to largest. List all the values in the data set that are less than or equal to the low outlier cutoff. These are the low outliers. If there are none, simply say "no low outliers".

16. Women's Wrist Circumference (Inches)

- a) Is the data set normally distributed (bell shaped)? (Yes or No)
- b) Is the mean an accurate average in this data? Why or why not?
- c) Is the standard deviation an accurate measure of typical spread for this data? Why or why not?
- d) Use the mean, standard deviation and the following formula to calculate the high outlier cutoff for this data.

High Outlier Cutoff = Mean + (2 x Standard Deviation)



- e) Use the mean, standard deviation and the following formula to calculate the low outlier cutoff for this data.  

$$\text{High Outlier Cutoff} = \text{Mean} - (2 \times \text{Standard Deviation})$$
- e) Copy the column of data into a new spreadsheet. Put the data values in order from smallest to largest. List all the values in the data set that are greater than or equal to the high outlier cutoff. These are the high outliers. If there are none, simply say “no high outliers”.
- f) Use the column of data that is in order from smallest to largest. List all the values in the data set that are less than or equal to the low outlier cutoff. These are the low outliers. If there are none, simply say “no low outliers”.

#### 17. Men's Height (Inches)

- a) Is the data set normally distributed (bell shaped)? (Yes or No)
- b) Is the mean an accurate average in this data? Why or why not?
- c) Is the standard deviation an accurate measure of typical spread for this data? Why or why not?
- d) Use the mean, standard deviation and the following formula to calculate the high outlier cutoff for this data.  

$$\text{High Outlier Cutoff} = \text{Mean} + (2 \times \text{Standard Deviation})$$
- e) Use the mean, standard deviation and the following formula to calculate the low outlier cutoff for this data.  

$$\text{High Outlier Cutoff} = \text{Mean} - (2 \times \text{Standard Deviation})$$
- e) Copy the column of data into a new spreadsheet. Put the data values in order from smallest to largest. List all the values in the data set that are greater than or equal to the high outlier cutoff. These are the high outliers. If there are none, simply say “no high outliers”.
- f) Use the column of data that is in order from smallest to largest. List all the values in the data set that are less than or equal to the low outlier cutoff. These are the low outliers. If there are none, simply say “no low outliers”.

#### 18. Men's Weight (Pounds)

- a) Is the data set normally distributed (bell shaped)? (Yes or No)
- b) Is the mean an accurate average in this data? Why or why not?
- c) Is the standard deviation an accurate measure of typical spread for this data? Why or why not?
- d) Use the mean, standard deviation and the following formula to calculate the high outlier cutoff for this data.  

$$\text{High Outlier Cutoff} = \text{Mean} + (2 \times \text{Standard Deviation})$$
- e) Use the mean, standard deviation and the following formula to calculate the low outlier cutoff for this data.  

$$\text{High Outlier Cutoff} = \text{Mean} - (2 \times \text{Standard Deviation})$$
- e) Copy the column of data into a new spreadsheet. Put the data values in order from smallest to largest. List all the values in the data set that are greater than or equal to the high outlier cutoff. These are the high outliers. If there are none, simply say “no high outliers”.
- f) Use the column of data that is in order from smallest to largest. List all the values in the data set that are less than or equal to the low outlier cutoff. These are the low outliers. If there are none, simply say “no low outliers”.

Directions #19-20: Use the following information to write a quantitative data analysis paragraph. There should be a sentence written for each of the following.

- What is the quantitative data measuring?
- What are the units?
- How many numbers are in the quantitative data set?
- What is the shape of the quantitative data set?
- What is the average?



- What is the spread?
- What are the two values that typical numbers fall in between?
- What are the unusually high values (high outliers) in the quantitative data set?
- What are the unusually low values (low outliers) in the quantitative data set?

19. Quantitative Data: Restaurant Bill amounts in dollars.

Shape: Bell shaped (normal)

Sample Size (n) = 74

Mean = \$84.31

Standard Deviation = \$13.74

Calculate two numbers that typical values are in between: Mean  $\pm$  Standard Deviation

Calculate the low outlier cutoff: Mean + (2 x standard deviation)

Calculate the high outlier cutoff: Mean – (2 x standard deviation)

High Outliers: \$119.54 , \$136.82

Low Outliers: \$49.67 , \$41.88

20. Quantitative Data: Weights of male lions in pounds

Shape: Bell shaped (normal)

Sample Size (n) = 40

Mean = 452.6 pounds

Standard Deviation = 59.1 pounds

Calculate two numbers that typical values are in between: Mean  $\pm$  Standard Deviation

Calculate the low outlier cutoff: Mean + (2 x standard deviation)

Calculate the high outlier cutoff: Mean – (2 x standard deviation)

High Outliers: 596.5 pounds

Low Outliers: 324.7 pounds

## Chapter 4 Review Sheet

Here is a list of important ideas in this chapter.

- Be able to distinguish between categorical data and quantitative (numerical measurement) data.
- Be able to create histograms and dot plots with technology and find the shape of a quantitative data set.
- Be able to find the mean, standard deviation, minimum, maximum, frequency (N) with technology.



This material is from *Introduction to Data Analysis*, 2<sup>nd</sup> edition by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a “CC-BY” [Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by/4.0/) – 3/17/2021

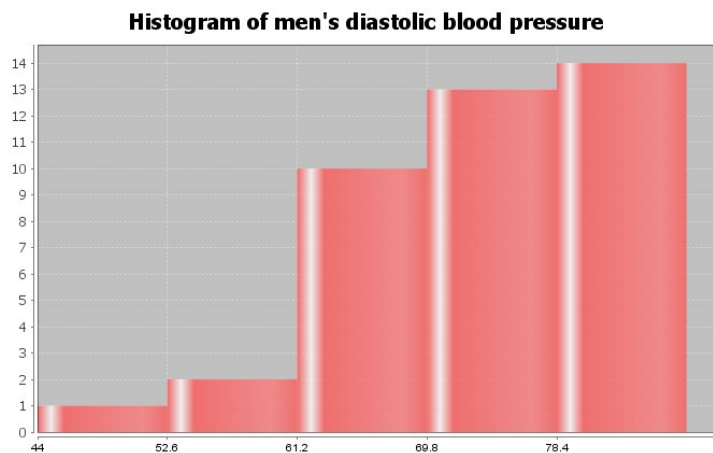
- A center gives an average value for the data set is usually close to the highest bar or bars in the histogram.
- Statistics that measure center: Mean, Median, Mode and Midrange
- If a data set is bell shaped, we should use the mean average as our measure of center and our average for the data set. If a data set is not bell shaped, we should not use the mean.
- Mean Average definition: A statistic that measures the center or average of a bell shaped data set by balancing the distances.
- A measure of spread or variability tells us how spread out the data set is. The more spread out the data is, the less consistent the data is and the harder it is to predict. A small amount of spread tells us that the data is more consistent and easier to predict.
- Statistics that measure spread (variability): Standard Deviation, Variance, Range, Interquartile Range (IQR)
- If a data set is bell shaped, we should use the standard deviation as our measure of spread for the data set. If a data set is not bell shaped, then we should not use the standard deviation.
- Standard Deviation definition: A measure of spread that tells us how far typical values are from the mean in a bell shaped data set.
- Mean – Standard Deviation ≤ Typical Values ≤ Mean + Standard Deviation
- Unusually High Cutoff: Mean + (2 x Standard Deviation)
- Unusually Low Cutoff: Mean – (2 x Standard Deviation)
- Be able to use the unusual cutoffs and a dot plot to identify unusual values in the data set.
- Be able to write a summary report paragraph summarizing the key characteristics of a bell shaped quantitative data set.

#### Problems Chapter 4 Review Sheet

Give the shape of each of the following graphs from the men's health data. Then decide if the mean or the median is the most appropriate average for the data set.

##### 1. Men's Diastolic Blood Pressure

Shape = \_\_\_\_\_ Mean or Median? \_\_\_\_\_

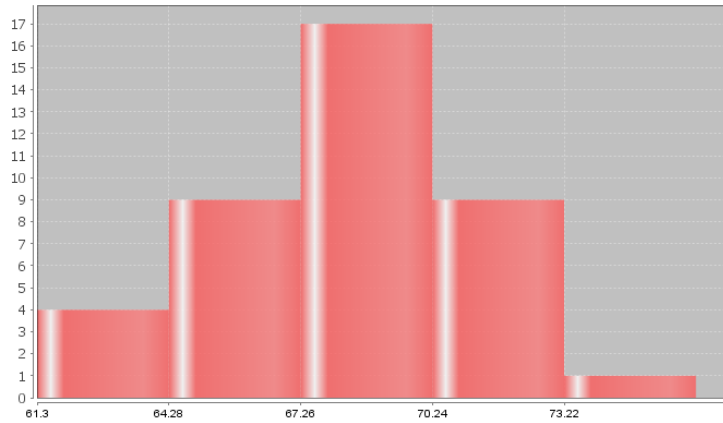


##### 2. Men's Heights (inches)

Shape = \_\_\_\_\_ Mean or Median? \_\_\_\_\_



**Histogram of men's heights in inches**

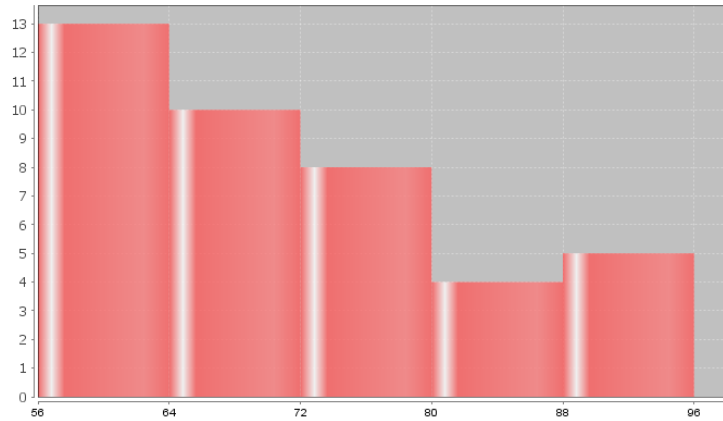


3. Men's Pulse Rates (Beats per Minute)

Shape = \_\_\_\_\_

Mean or Median? \_\_\_\_\_

**Histogram of men's pulse rates in beats per minute**



4. Calculate the Mean Average for the following data. Round your answer to the hundredths place (two numbers to right of decimal).

$$\bar{x} = \frac{\sum x}{n}$$

12.6                      21.8                      20.1                      16.6

16.7                      20.8                      11.2                      9.0

21.2                      12.3                      12.9                      15.2

25.7

Mean Average = \_\_\_\_\_

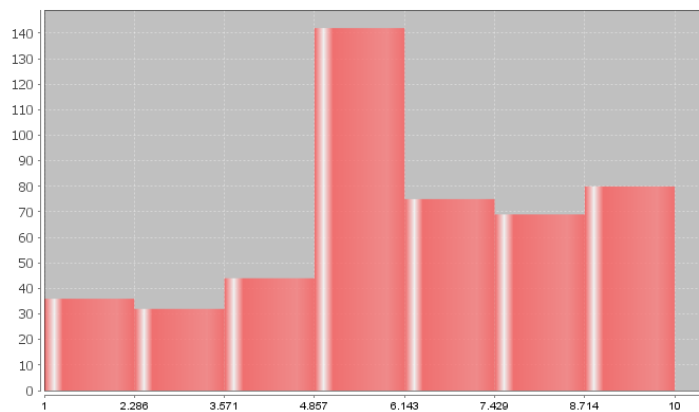
5. Standard Deviation is an important measure of spread or variability in statistics. Give the basic definition of Standard Deviation.



6. How can we tell if the mean and standard deviation are accurate?
7. What percentage of the values in a bell shaped data set are considered typical?
8. What percentage of the values in a bell shaped data set are considered unusually high?
9. What percentage of the values in a bell shaped data set are considered unusually low?

Math 075 Students in the Fall 2015 semester were asked on a scale of one to ten, how intimidated are you about math classes. Here is a histogram, dot plot, mean, standard deviation, frequency, minimum and maximum from Statcato.

**Histogram of Math 075 students Math Intimidation Scale**



**Descriptive Statistics**

Variable	Mean	Standard Deviation
C15 math intimidation	6.159	2.418

Variable	Min	Max
C15 math intimidation	1.0	10.0

Variable	N total
C15 math intimidation	478

10. What is the shape of the data set? \_\_\_\_\_

11. How many numbers are in the data set? \_\_\_\_\_





12. Are the mean and standard deviation accurate for this data? (Yes or No) \_\_\_\_\_

13. What is the average math intimidation score for the students? (Give a number.)

Average math intimidation score = \_\_\_\_\_

14. How far are typical values in the data set from the mean on average? (Give a number.)

Average distance from the mean = \_\_\_\_\_

15. Calculate two numbers that typical values fall in between and put your answer below.

Mean – Standard Deviation  $\leq$  typical math intimidation scores  $\leq$  Mean + Standard Deviation

\_\_\_\_\_  $\leq$  typical math intimidation scores  $\leq$  \_\_\_\_\_

16. What is the cutoff for an unusually high math intimidation score?

Unusual High Cutoff = Mean + (2 x Standard Deviation)

Unusual High Cutoff = \_\_\_\_\_

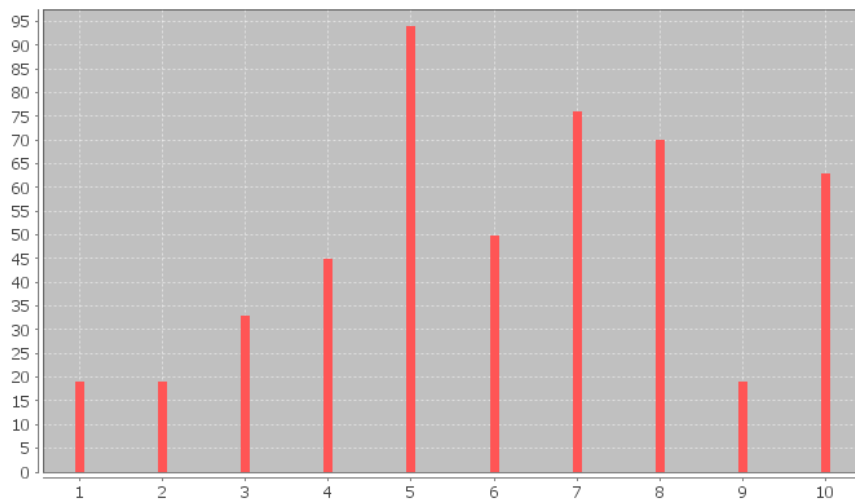
17. What is the cutoff for an unusually low math intimidation score?

Unusual Low Cutoff = Mean – (2 x Standard Deviation)

Unusual Low Cutoff = \_\_\_\_\_

Look at the following Dot Plot for the data and your answers to #16 and #17 to answer the following questions.

**Dot Plot for Math Intimidation Score Data**



18. Are there any unusually high math intimidation scores in the data (yes or no)?

19. If you answered yes to #18, what are the unusually high scores? \_\_\_\_\_

20. Are there any unusually low math intimidation scores in the data (yes or no)?

21. If you answered yes to #20, what are the unusually low scores? \_\_\_\_\_

---



## Chapter 4 Project Normal Quantitative Data Analysis

**Directions for Online Classes:** *This will be an individual project. Each student will analyze one quantitative data set from the “Math 075 Chapter 4 Project Normal Data” and create a poster summarizing their findings, After submitting a picture of the poster to their instructor, students will then go to the “Chapter 4 Project Class Discussion” in Canvas and discuss their findings with other students in the class.*

*Each student will pick one of the following data sets from the Math 075 Chapter 4 Project Normal Data to analyze: Male Body Temp Degrees Fahrenheit, Female Body Temp Degrees Fahrenheit, North Territory Australia Weekly Salary Dollars, Tasmania Australia Weekly Salary Dollars, Chicks Weight Gain (in grams) after 20 days on Normal Corn, January minimum temperature in degrees Fahrenheit of various U.S. Cities, Percent of Female Students at Universities around the world, Salamander Total Length (cm), Fat (grams ) Fast Food Breakfast Items, Soil Surface temperature (degrees Celsius) in Comanche, Texas, NBA All-Star Player Heights.*

### The Individual Poster Should Have

- **First and Last Name of student**
- **Why is this data important or interesting to you?**
- **Go to [www.lock5stat.com](http://www.lock5stat.com) and open StatKey. Click on “one quantitative variable” under the “descriptive statistics and graphs” menu. Click on “edit data” and copy and paste your one column of quantitative data into Statkey. If you data has a title, click on “data has a header row”. Do NOT click the box the says data has identifier. Press OK.**
- **Click on dot plot in StatKey and sketch the dot plot onto your poster.**
- **Click on histogram in StatKey, and pull the slider to “3 buckets”. Your histogram should have 3 bars. Sketch the histogram onto your poster.**
- **Write down the Mean, Standard Deviation, Min, Max and Sample Size onto your poster. You will see them under the “Sample Statistics” menu in StatKey.**
- **What is the data measuring?**
- **What are the units?**
- **How many numbers are in the data set : sample size (n)**
- **What is the Shape? Look at your histogram. Should be normal (bell shaped).**
- **Write a sentence to explain the mean.**
- **What is the average? (Use the mean if normal data.)**
- **What is your spread for the data? (Use the standard deviation if normal data.)**
- **Write a sentence to explain the standard deviation.**
- **Find two numbers that typical values fall in between (Mean – Stand Dev , Mean + Stand Dev)**
- **Calculate Unusually high cutoff (Mean + 2 x Stand Dev)**
- **List all unusually high values (high outliers) in the data set. (Find these on the dot plot or excel spreadsheet.) If there are none, say “No high outliers”.**
- **Calculate Unusually low cutoff (Mean – 2 x Stand Dev)**
- **List all unusually low values (low outliers) in the data set. (Find these on the dot plot or excel spreadsheet.) If there are none, say “No low outliers”.**
- **Decorate Poster**

Now take a picture of your poster project and submit the picture to your instructor in Canvas.

After submitting the picture of the poster, go to the discussion menu in Canvas and complete the “Chapter 4 Project Discussion”. You will be discussing your findings with other students in the class.



This material is from *Introduction to Data Analysis*, 2<sup>nd</sup> edition by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a “CC-BY” [Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by/4.0/) – 3/17/2021

**Directions for Face to Face Classes:** *The class will be separated into groups. Each group is required to pick a “team name” for their group and analyze one quantitative data set from the “Math 075 Chapter 4 Project Normal Data”, create a poster summarizing their findings, and present the poster to other students in the class.*

*Each group will have a different topic and will pick one of the following data sets from the Math 075 Chapter 4 Project Normal Data to present it to their classmates: Male Body Temp Degrees Fahrenheit, Female Body Temp Degrees Fahrenheit, North Territory Australia Weekly Salary Dollars, Tasmania Australia Weekly Salary Dollars, Chicks Weight Gain (in grams) after 20 days on Normal Corn, January minimum temperature in degrees Fahrenheit of various U.S. Cities, Percent of Female Students at Universities around the world, Salamander Total Length (cm), Fat (grams) Fast Food Breakfast Items, Soil Surface temperature (degrees Celsius) in Comanche, Texas, NBA All-Star Player Heights.*

#### The Poster Should Have

- **Group/Team Name**
- **First and Last Name of each team members on the poster**
- **Why is this data important or interesting to your group?**
- **Go to [www.lock5stat.com](http://www.lock5stat.com) and open StatKey. Click on “one quantitative variable” under the “descriptive statistics and graphs” menu. Click on “edit data” and copy and paste your one column of quantitative data into Statkey. If you data has a title, click on “data has a header row”. Do NOT click the box the says data has identifier. Press OK.**
- **Click on dot plot in StatKey and sketch the dot plot onto your poster.**
- **Click on histogram in StatKey, and pull the slider to “3 buckets”. Your histogram should have 3 bars. Sketch the histogram onto your poster.**
- **Write down the Mean, Standard Deviation, Min, Max and Sample Size onto your poster. You will see them under the “Sample Statistics” menu in StatKey.**
- **What is the data measuring?**
- **What are the units?**
- **How many numbers are in the data set : sample size (n)**
- **What is the Shape? Look at your histogram. Should be normal (bell shaped).**
- **Write a sentence to explain the mean.**
- **What is the average? (Use the mean if normal data.)**
- **What is your spread for the data? (Use the standard deviation if normal data.)**
- **Write a sentence to explain the standard deviation.**
- **Find two numbers that typical values fall in between (Mean – Stand Dev , Mean + Stand Dev)**
- **Calculate Unusually high cutoff (Mean + 2 x Stand Dev)**
- **List all unusually high values (high outliers) in the data set. (Find these on the dot plot or excel spreadsheet.) If there are none, say “No high outliers”.**
- **Calculate Unusually low cutoff (Mean – 2 x Stand Dev)**
- **List all unusually low values (low outliers) in the data set. (Find these on the dot plot or excel spreadsheet.). If there are none, say “No low outliers”.**
- **Decorate Poster**

#### Presentation

*Make sure each person on the team understands the poster and can present your findings. Bring your poster to a designated presentation area in the classroom and hang or tape your poster to a wall. One person at a time will present the poster. We will then rotate so that each member of the team gets to present. Everyone else will listen to presentations and give feedback.*

---



*This chapter is from Introduction to Data Analysis, first edition by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/) - 10/1/2017*

## Chapter 5: Analyzing Skewed Quantitative Data

**Introduction:** In our last chapter, we focused on analyzing bell shaped (normal) data, but many data sets are not bell shaped. How do we analyze quantitative data when it is not normal?

The main issue is that the mean and standard deviations are not accurate and should not be used in the analysis. Then what statistics should we use?

We will be introducing a new kind of graph that is specially designed for analyzing skewed data. It is called the “box and whisker plot” or “box plot” for short.

When data sets are not bell shaped, we will focus on the median, quartiles, interquartile range and boxplots to measure center and spread. Quartiles are more accurate because they are based on the order of the numbers instead of distances and so are not as effected by the skewed shape and extremely unusual values.

---



## Section 5A – Review of Shapes and Centers with Histograms and Dot Plots

Let us start by reviewing shapes and centers.

Here are the directions for making dot plots and histograms in StatKey.

**Making a dot plot in StatKey:** Go to [www.lock5stat.com](http://www.lock5stat.com). Click on “One Quantitative Variable” under the “Descriptive Statistics and Graphs” menu. Click on “Edit Data”. Copy and paste in the column of quantitative data you want to analyze. If the data has a title, check the box that says “Data has a header row”. If the data does NOT have a title, do NOT check the box that says “Data has a header row”. Do NOT check the box that says “First column is an identifier”. (You would only check the “identifier” box if there is a word next to every number in the data.) Now push “OK”. At the top left of the graph, click on the “dot plot” tab.

**Making a histogram in StatKey:** Go to [www.lock5stat.com](http://www.lock5stat.com). Click on “One Quantitative Variable” under the “Descriptive Statistics and Graphs” menu. Click on “Edit Data”. Copy and paste in the column of quantitative data you want to analyze. If the data has a title, check the box that says “Data has a header row”. If the data does NOT have a title, do NOT check the box that says “Data has a header row”. Do NOT check the box that says “First column is an identifier”. (You would only check the “identifier” box if there is a word next to every number in the data.) Now push “OK”. At the top left of the graph, click on the “histogram” tab. Use the slider button on the right of the screen to adjust the number of bars (buckets) in your histogram. Three bars is usually the best for seeing the shape.

### Center Principle for Quantitative Data

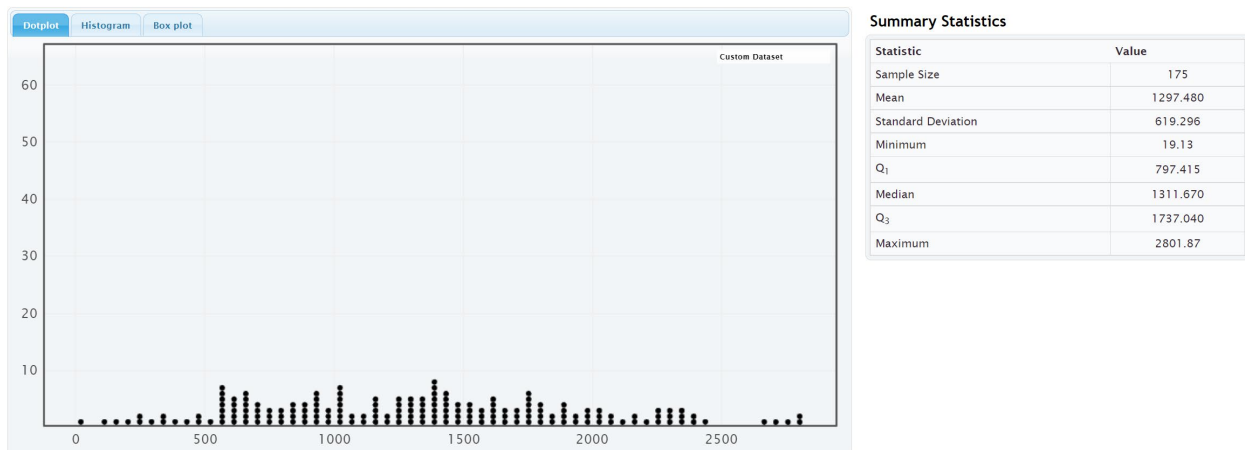
**If a data set is normally distributed (bell shaped), the mean average is usually an accurate measure of center and we should use the mean as the average for the data set.**

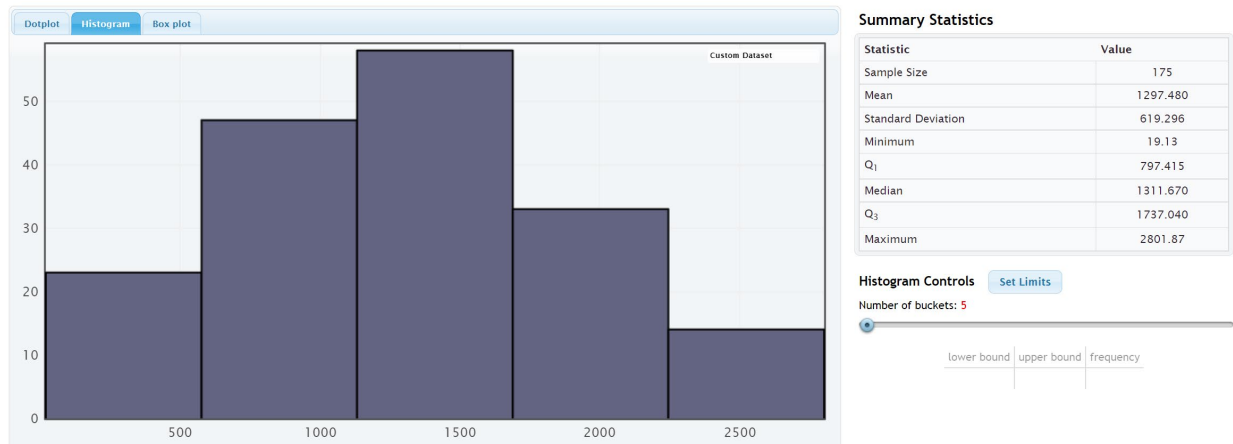
**If a data set has a skewed or irregular shape, the median average is usually the most accurate measure of center and we should use the median as the average for the data set.**

*Note: If a data set is not skewed, but just has an unusual shape like uniform, use the median also. Do not use the mean unless it is bell shaped. The mode is sometimes used as the center for bimodal or multimodal shaped data, since it can have multiple values and represent each hill in the data. That is why it is called bi-modal or multi-modal.*

### Example 1

We copied and pasted a random sample of 175 salaries from Australia into StatKey. Here is the dot plot and histogram created.





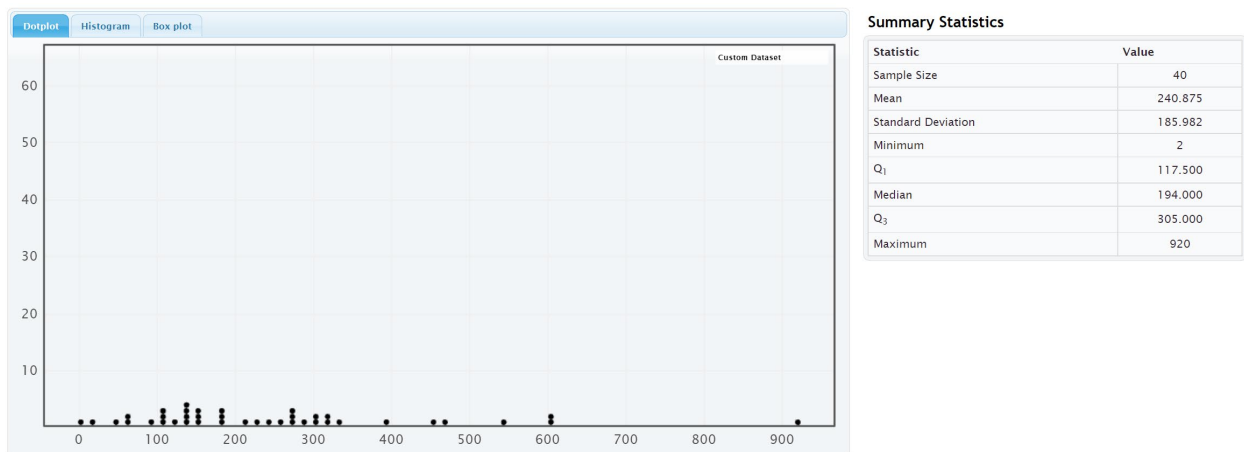
Notice that the salary data from Australia is normally distributed (bell shaped). The highest bar and the largest concentration of dots are in the center. The left and right tails are roughly the same length.

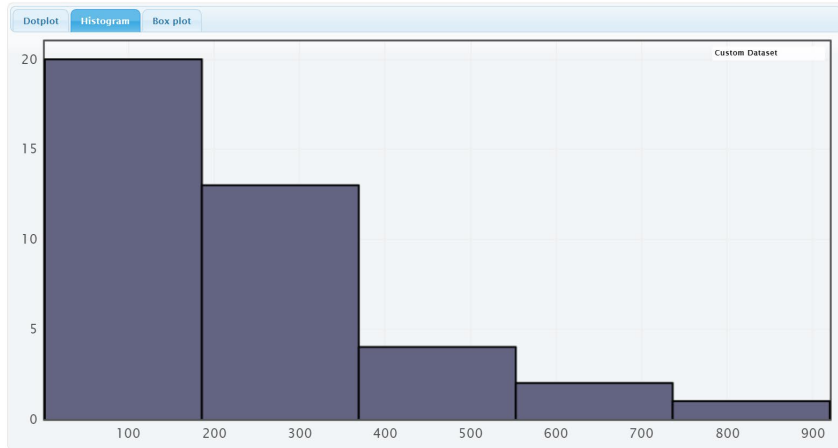
Notice that the mean and median are relatively close when data is normally distributed. However, we should use the mean as our center and average for this data.

So the average salary for this data is \$1297.48 (mean).

### Example 2

We copied and pasted a random sample of 40 women's cholesterol in milligrams per deciliter into StatKey. Here is the dot plot and histogram created.





### Summary Statistics

Statistic	Value
Sample Size	40
Mean	240.875
Standard Deviation	185.982
Minimum	2
Q <sub>1</sub>	117.500
Median	194.000
Q <sub>3</sub>	305.000
Maximum	920

### Histogram Controls

Number of buckets: 5

lower bound    upper bound    frequency

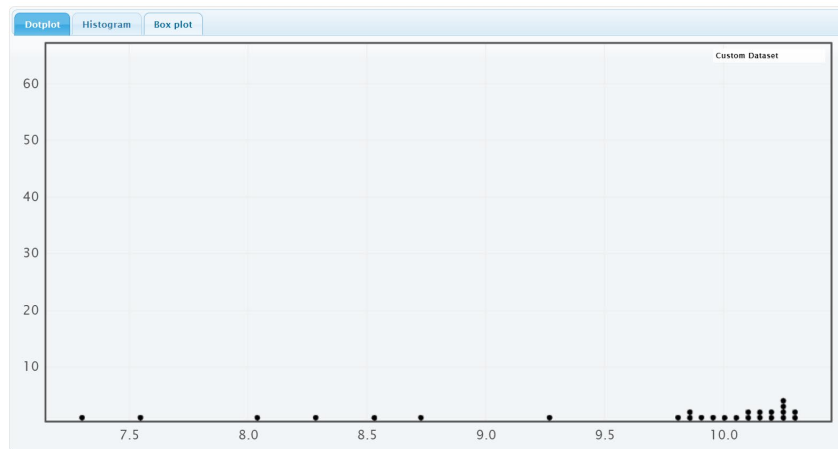
Notice that the shape of this data is not normal. The highest bars and the largest cluster of dots are on the far left and it has a long right tail. We call the shape of this data skewed right or positively skewed.

Notice that the mean average (240.875 mg/dL) is much larger than the median average (194 mg/dL). The mean has been pulled up in the direction of the long tail and is no longer accurate. The median average however is still close to the highest bar is a much more accurate average. Remember the rule for centers, when data is not normal, use the median as your average (center).

The average cholesterol for these 40 women is 194 mg/dL (median).

### Example 3

We copied and pasted a sample of salaries in dollars per hour from a company into StatKey. Here is the dot plot and histogram created.

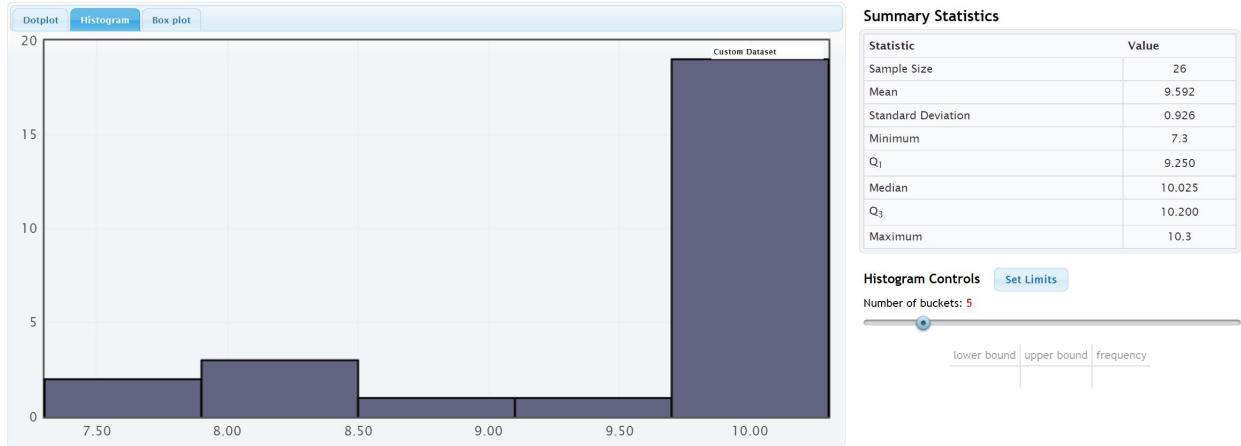


### Summary Statistics

Statistic	Value
Sample Size	26
Mean	9.592
Standard Deviation	0.926
Minimum	7.3
Q <sub>1</sub>	9.250
Median	10.025
Q <sub>3</sub>	10.200
Maximum	10.3







Notice that the shape of this data is not normal. The highest bars and the largest cluster of dots are on the far right and it has a long left tail. We call the shape of this data skewed left or negatively skewed.

Notice that the mean average (9.592 \$/hour) is much smaller than the median average (10.025 \$/hour). The mean has been pulled down in the direction of the long tail and is no longer accurate. The median average however is still close to the highest bar and is a much more accurate average. Remember the rule for centers, when data is not normal, use the median as your average (center).

The average salary for the employees in this company is \$10.025 (median).

---



## Problem Set Section 5A

Directions: Open the “Bear Data” and “Health Data” (columns AD-AP) in Canvas or at [www.matt-teachout.org](http://www.matt-teachout.org). Go to [www.lock5stat.com](http://www.lock5stat.com). Click on “One Quantitative Variable” under the “Descriptive Statistics and Graphs” menu. Click on “Edit Data”. Copy and paste in the column of quantitative data indicated in the problem. If the data has a title, check the box that says “Data has a header row”. If the data does NOT has a title, do NOT check the box that says “Data has a header row”. Do NOT check the box that says “First column is an identifier”. (You would only check the “identifier” box if there is a word next to every number in the data.) Now push “OK”. At the top left of the graph, click on the “dot plot” tab and the “histogram tab”. When you click the histogram tab, pull the slider on the right to “3 buckets” so your histogram has 3 bars.

### 1. Bear Ages in Months

- Draw a rough sketch of the dot plot on a piece of paper or save the graph on a word document.
- Draw a rough sketch of the 3-bar histogram on a piece of paper or save the graph on a word document.
- What is the shape of the data set?
- Should we use the mean or median as the average (center) for the data?

### 2. Bear Head Length in Inches

- Draw a rough sketch of the dot plot on a piece of paper or save the graph on a word document.
- Draw a rough sketch of the 3-bar histogram on a piece of paper or save the graph on a word document.
- What is the shape of the data set?
- Should we use the mean or median as the average (center) for the data?

### 3. Bear Head Width in Inches

- Draw a rough sketch of the dot plot on a piece of paper or save the graph on a word document.
- Draw a rough sketch of the 3-bar histogram on a piece of paper or save the graph on a word document.
- What is the shape of the data set?
- Should we use the mean or median as the average (center) for the data?

### 4. Bear Neck Circumference in Inches

- Draw a rough sketch of the dot plot on a piece of paper or save the graph on a word document.
- Draw a rough sketch of the 3-bar histogram on a piece of paper or save the graph on a word document.
- What is the shape of the data set?
- Should we use the mean or median as the average (center) for the data?

### 5. Bear Length in Inches

- Draw a rough sketch of the dot plot on a piece of paper or save the graph on a word document.
- Draw a rough sketch of the 3-bar histogram on a piece of paper or save the graph on a word document.
- What is the shape of the data set?
- Should we use the mean or median as the average (center) for the data?



6. Bear Chest Size in Inches

- a) Draw a rough sketch of the dot plot on a piece of paper or save the graph on a word document.
- b) Draw a rough sketch of the 3-bar histogram on a piece of paper or save the graph on a word document.
- c) What is the shape of the data set?
- d) Should we use the mean or median as the average (center) for the data?

7. Bear Weight in Pounds

- a) Draw a rough sketch of the dot plot on a piece of paper or save the graph on a word document.
- b) Draw a rough sketch of the 3-bar histogram on a piece of paper or save the graph on a word document.
- c) What is the shape of the data set?
- d) Should we use the mean or median as the average (center) for the data?

8. Men's Ages in Years

- a) Draw a rough sketch of the dot plot on a piece of paper or save the graph on a word document.
- b) Draw a rough sketch of the 3-bar histogram on a piece of paper or save the graph on a word document.
- c) What is the shape of the data set?
- d) Should we use the mean or median as the average (center) for the data?

9. Men's Weight in Pounds

- a) Draw a rough sketch of the dot plot on a piece of paper or save the graph on a word document.
- b) Draw a rough sketch of the 3-bar histogram on a piece of paper or save the graph on a word document.
- c) What is the shape of the data set?
- d) Should we use the mean or median as the average (center) for the data?

10. Men's Height in Inches

- a) Draw a rough sketch of the dot plot on a piece of paper or save the graph on a word document.
- b) Draw a rough sketch of the 3-bar histogram on a piece of paper or save the graph on a word document.
- c) What is the shape of the data set?
- d) Should we use the mean or median as the average (center) for the data?

11. Men's Pulse Rate in Beats per Minute (BPM)

- a) Draw a rough sketch of the dot plot on a piece of paper or save the graph on a word document.
- b) Draw a rough sketch of the 3-bar histogram on a piece of paper or save the graph on a word document.
- c) What is the shape of the data set?
- d) Should we use the mean or median as the average (center) for the data?



12. Men's Systolic Blood Pressure (mm of Hg)

- a) Draw a rough sketch of the dot plot on a piece of paper or save the graph on a word document.
- b) Draw a rough sketch of the 3-bar histogram on a piece of paper or save the graph on a word document.
- c) What is the shape of the data set?
- d) Should we use the mean or median as the average (center) for the data?

13. Men's Diastolic Blood Pressure (mm of Hg)

- a) Draw a rough sketch of the dot plot on a piece of paper or save the graph on a word document.
- b) Draw a rough sketch of the 3-bar histogram on a piece of paper or save the graph on a word document.
- c) What is the shape of the data set?
- d) Should we use the mean or median as the average (center) for the data?

14. Men's Cholesterol (mg per dL)

- a) Draw a rough sketch of the dot plot on a piece of paper or save the graph on a word document.
- b) Draw a rough sketch of the 3-bar histogram on a piece of paper or save the graph on a word document.
- c) What is the shape of the data set?
- d) Should we use the mean or median as the average (center) for the data?

15. Men's Body Mass Index (kg per  $m^2$ )

- a) Draw a rough sketch of the dot plot on a piece of paper or save the graph on a word document.
  - b) Draw a rough sketch of the 3-bar histogram on a piece of paper or save the graph on a word document.
  - c) What is the shape of the data set?
  - d) Should we use the mean or median as the average (center) for the data?
- 



## Section 5B – Understanding the Median Average

In the last section, we said that we should use the median as our center and average, when data is skewed or not normally distributed, but what is the median? Why is it more accurate than the mean for skewed data?

Let us see if we can get a better understanding of the median average.

**Definition of the Median average:** The median average is the center of the data when the values are put in order from smallest to largest. The median is also called the “50<sup>th</sup> percentile” since approximately 50% of the numbers in the data set will be greater than the median and 50% of the numbers in the data set will be less than the median. Think of the median as a marker that divides the data in half. That is why it is often called the true center of the data.

Skewed data tends to have extremely unusual values. These unusual values (outliers) are very far from the mean. That is why the mean and standard deviation (typical distance from the mean) are not accurate for skewed data. The median is based on how many numbers are in the data set (frequency) and the order of the numbers. If the highest value were 40 or 4000, it would not change the median.

**Names for the Median Average:** “Median”, The 50<sup>th</sup> percentile ( $P_{50}$ ), or the Second Quartile ( $Q_2$ ).

### How to Calculate the Median Average

As with all statistics, rely on technology to calculate. No statistician calculates the median by hand, especially for large data sets. All of them use statistics software or computer software programs. To get a better understanding of the median, we will look at a couple examples where we calculate the median with small data sets.

To calculate the median, put the data in order from smallest to largest. Computer programs like excel can sort the data for you if you do not want to put it in order. Once the data is in order, you will look for the center of the data.

**Odd Number of Values:** If you have an odd number of values in the data set, then your median will be the number in the exact middle of the data when it is in order. Suppose we have 17 numbers in order from smallest to largest in the data set. Then our median would be the ninth number in the data set. That would give us eight numbers below the median and eight numbers above the median. Remember the median separates the data into two equal groups.

**Even Number of Values:** If you have an even number of values in the data set, then your median will not be a value in the data set. The median will be half way in between the two numbers in the middle. Suppose you have 26 numbers in order from smallest to largest in the data set. Then the median will be half way between the 13<sup>th</sup> and 14<sup>th</sup> numbers in the data set. That way thirteen numbers will be below the median and thirteen numbers will be above the median. If you cannot think of what half way in between would be, you could use the following formula. Remember this formula only works if the data values are in order.

Median (even # of data values) = (first number in middle + second number in middle) / 2

### Example 1

Find the median for the brick weight (in kilograms) data from last chapter.

4.7 , 6.2 , 3.3 , 5.1 , 2.9 , 7.4 , 4.5

The first thing to notice is that the data is not in order. It needs to be put in order before we can find the median.

Data in order:

2.9, 3.3, 4.5, 4.7, 5.1, 6.2, 7.4

Since there are seven numbers in the data set. The fourth number (4.7) will be the median.

Median Average = 4.7 kilograms



Notice there are three numbers in the data set greater than the median (5.1, 6.2 and 7.4) and there are three numbers in the data set less than the median (2.9, 3.3 and 4.5).

### Example 2

Let us look at a second example.

Here are the yearly salaries in thousands of dollars for employees from a small company.

36.5 , 51.2 , 47.9 , 44.1 , 37.2 , 39.6 , 41.8 , 45.4 , 43.2 , 253.5

*(This last salary of 253.5 thousand dollars was the CEO of the company.)*

Remember to put the numbers in order first.

Yearly Salary Data in order:

36.5 , 37.2 , 39.6 , 41.8 , 43.2 , 44.1 , 45.4 , 47.9 , 51.2 , 253.5

Since there are ten numbers (even), the median will not be a number in the data set. It will be half way between the two middle numbers that can divide the data in half. The two numbers in the middle are 43.2 and 44.1 thousand dollars.

Median Average =  $(43.2 + 44.1) / 2 = (87.3) / 2 = 43.65$  thousand dollars

Notice again that there are five numbers above the median (44.1 , 45.4 , 47.9 , 51.2 , 253.5) and five numbers below the median (36.5 , 37.2 , 39.6 , 41.8 , 43.2). The data has been split in half.

This is a good example to explain why the median is a better average than the mean. The CEO is a large unusual value in the data set, making the data very skewed right. Let us compare the mean and median averages.

Mean Average =

$(36.5 + 37.2 + 39.6 + 41.8 + 43.2 + 44.1 + 45.4 + 47.9 + 51.2 + 253.5) / 10$

=  $640.4 / 10 = 64.04$  thousand dollars.

Median Average =

$(43.2 + 44.1) / 2 = (87.3) / 2 = 43.65$  thousand dollars

Notice no one in the company makes 64 thousand dollars. The mean is not a good average for this data. The median however is very accurate. Many people in the company make around 43 or 44 thousand dollars. Recently, companies have been using the median average as their “average salary” on their websites for this very reason.

### Calculating the median average with technology

All statistics software programs can calculate the median. This is a much better way to find the median, especially if you have larger data sets.

Here are the steps to calculating graphs and quantitative statistics with StatKey.

Go to [www.lock5stat.com](http://www.lock5stat.com). Click on “One Quantitative Variable” under the “Descriptive Statistics and Graphs” menu. Click on “Edit Data”. Copy and paste in the column of quantitative data you want to analyze. If the data has a title, check the box that says “Data has a header row”. If the data does NOT has a title, do NOT check the box that says “Data has a header row”. Do NOT check the box that says “First column is an identifier”. (You would only check the “identifier” box if there is a word next to every number in the data set.) Push “OK”. You should see the median in the list of “Summary Statistics” on the top right of the page.



## Example

In the last section, we looked at a random sample of 40 women's cholesterol in milligrams per deciliter. Let's put this data into StatKey, verify the shape and calculate the median. First we need to find the column of data in the "Health" data set.

W
Women Cholesterol (mg per deciliter)
264
181
267
384
98
62
126
89
531
130
175
44
8
112
462
62
98
447

Now we will go to [www.lock5stat.com](http://www.lock5stat.com) and copy and paste the data into StatKey.

StatKey to accompany [Stat](#)

Descriptive Statistics and Graphs

- One Quantitative Variable
- One Categorical Variable
- One Quantitative and One Categorical Variable
- Two Categorical Variables
- Two Quantitative Variables

Edit data

Women Cholesterol (mg per deciliter)

264  
181  
267  
384  
98  
62  
126  
89  
531  
130  
175  
44  
8  
112  
462  
62  
98  
447  
125

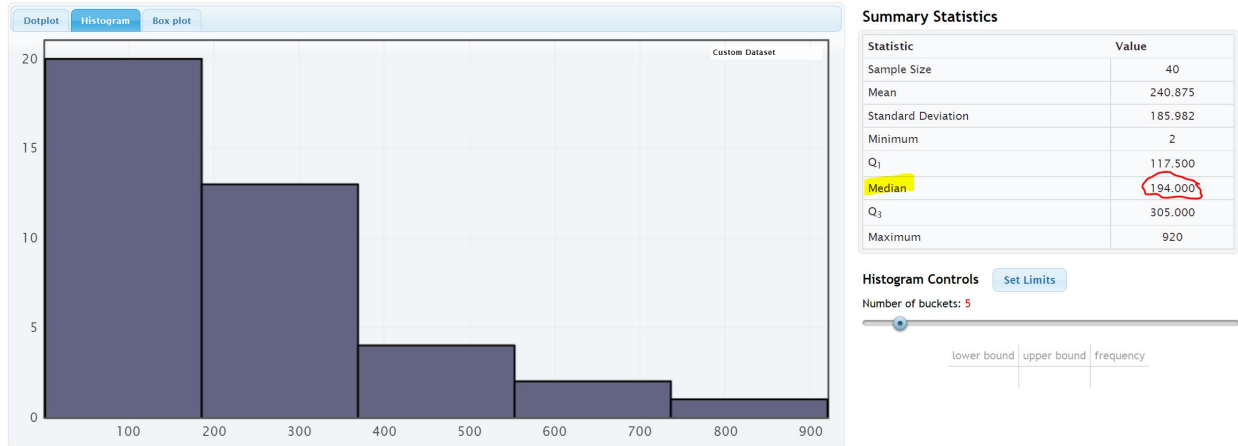
First column is identifier

Data has header row

Manually edit the values above or paste a tab or comma separated file into the box and click Ok. The file must have only one column (or two if there is an identifier).

Ok





Notice that the shape of this data is skewed right.

Notice that the mean average (240.875 mg/dL) is much larger than the median average (194 mg/dL). The mean has been pulled up in the direction of the long tail and is no longer accurate. The median average however is still close to the highest bar is a much more accurate average. Remember the rule for centers, when data is not normal, use the median as your average (center).

The average cholesterol for these 40 women is 194 mg/dL (median).

*Note: In a skewed left data set, the mean will also be pulled in the direction of the skew. This will make the mean average too small. You can often get a good idea of the shape of a data set by just looking at the mean and median.*

*Normal Data: The mean and median are very close.*

*Both the mean and median are accurate, but we use the mean.*

*Skewed Right Data: The mean is significantly larger than the median. Only the median is accurate.*

*Skewed Left Data: The mean is significantly smaller than the median. Only the median is accurate.*





## Problem Set Section 5B

1. Put each of the following data sets in order from smallest to largest. Then calculate the median average (50<sup>th</sup> percentile).

- a) 5 , 7 , 8 , 8 , 9 , 11 , 14 , 16 , 17 , 19 , 21 , 25 , 26 , 29 , 31 , 33 , 36
- b) 2.1 , 3.8 , 5.1 , 6.9 , 7.2 , 10.4 , 11.3 , 14.7 , 15.1 , 16.0
- c) 31 , 34 , 41 , 52 , 68 , 71 , 79 , 83 , 88 , 90 , 103
- d) 150 , 152 , 154 , 155 , 157 , 159 , 163 , 164 , 165
- e) 7.5 , 2.3 , 4.6 , 1.9 , 2.8 , 9.4 , 8.3 , 6.1
- f) 21 , 29 , 23 , 26 , 25 , 19 , 28 , 31 , 32 , 20 , 18

2. Use StatKey and the “Bear Data” to calculate median average for each of the following data sets.

Directions: Open the “Bear Data” in Canvas or at [www.matt-teachout.org](http://www.matt-teachout.org). Go to [www.lock5stat.com](http://www.lock5stat.com). Click on “One Quantitative Variable” under the “Descriptive Statistics and Graphs” menu. Click on “Edit Data”. Copy and paste in the column of quantitative data indicated in the problem. If the data has a title, check the box that says “Data has a header row”. If the data does NOT have a title, do NOT check the box that says “Data has a header row”. Do NOT check the box that says “First column is an identifier”. (You would only check the “identifier” box if there is a word next to every number in the data.) Now push “OK”. The median will be listed under “summary statistics” on the top right of StatKey.

- a) Median average for Bear Ages = \_\_\_\_\_ Months
  - b) Median average for Bear Head Length = \_\_\_\_\_ Inches
  - c) Median average for Bear Head Width = \_\_\_\_\_ Inches
  - d) Median average for Bear Neck Circumference = \_\_\_\_\_ Inches
  - e) Median average for Bear Length = \_\_\_\_\_ Inches
  - f) Median average for Bear Chest Size = \_\_\_\_\_ Inches
  - g) Median average for Bear Weight = \_\_\_\_\_ Pounds
- 



## Section 5C – Spread and Typical Values for Skewed Data, Quartiles, Interquartile Range (IQR), and the Five Number Summary

We have now seen that when data is not normal, we should use the median average as our measure of center and average.

The median is actually a type of quartile. Quartile analysis is an important part of understanding skewed.

**Definition of Quartiles:** The quartiles are three numbers that break the data into four equal groups. Think of them as three fences that separate the data into quarters when the data is in order.

**First Quartile (Q1):** This statistic is also called the 25<sup>th</sup> percentile and is the number that approximately 25% of the data is less than and 75% of the data is greater than.

**Second Quartile (Q2):** This statistic is also called the Median or the 50<sup>th</sup> percentile and is the number that approximately 50% of the data is less than and 50% of the data is greater than.

**Third Quartile (Q3):** This statistic is also called the 75<sup>th</sup> percentile and is the number that approximately 75% of the data is less than and 25% of the data is greater than.

Remember, when data set is skewed or not normal, we should not use the standard deviation to measure spread. So what measure of spread should we use for skewed data? In normal (bell-shaped) data, typical values are closer to the center. The empirical rule implies that for bell shaped data, about 68% is typical. In skewed data, the data is more spread out with less values being typical. For skewed data, we look for the middle 50% of the data for typical values. This is called the interquartile range.

**Interquartile Range (IQR):** The interquartile range is how far typical values are from each other in a skewed data set. IQR is the length between the middle 50% of the data values and is calculated by subtracting the third quartile (Q3) minus the first quartile (Q1).

Interquartile range formula:  $IQR = Q3 - Q1$

### Center and Spread Rule for Skewed Right, Skewed Left, or Non-normal Data

Center (Average) = Median (2<sup>nd</sup> Quartile)

Spread = Interquartile Range (IQR)

Typical Values = Between the 1<sup>st</sup> Quartile (Q1) and 3<sup>rd</sup> Quartile (Q3)

### **The Five Number Summary**

A common way to summarize skewed or non-normal quantitative data is by listing the following five statistics in this order. The five numbers are often referred to as the “five number summary”.

**Five Number Summary: Minimum Value, 1<sup>st</sup> Quartile (Q1), Median (Q2), 3<sup>rd</sup> Quartile (Q3), Maximum Value**

### Example 1

Different statistics programs sometimes give slightly different values for the quartiles. People sometimes overemphasize these differences. The key is to remember the idea, finding three fences that separate the data into four equal groups. Each quarter should have approximately the same number of values in that group.

Let us calculate the three quartiles and the interquartile range (IQR) for the following data set.



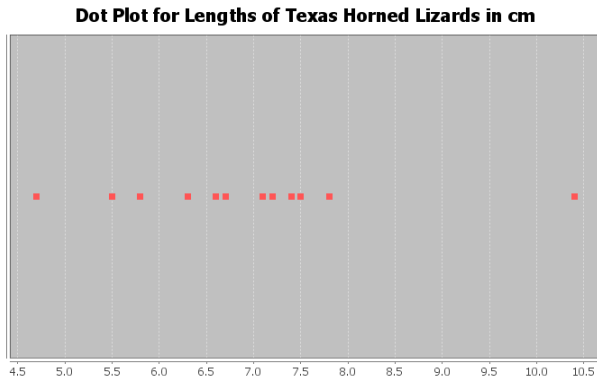
This material is from *Introduction to Data Analysis*, 2<sup>nd</sup> edition by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a “CC-BY” [Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by/4.0/) – 3/17/2021

### Texas Horned Lizard Lengths in Centimeters (cm)

6.7 , 10.4 , 7.8 , 4.7 , 5.5 , 5.8 , 7.2 , 7.5 , 7.1 , 6.3 , 6.6 , 7.4

When analyzing quantitative data, always determine the shape first.

Because of the one unusually large lizard of 10.4 cm, this data is probably skewed right. It is difficult to tell the shape of small data sets. Here is a dot plot of the data.



It does look like much of the data is bunched up on the left and the one unusual value on the right gives a longer tail to the right. So this data is skewed right. For skewed data, we should use the median for the average, IQR for the spread, and typical values in between the 1<sup>st</sup> and 3<sup>rd</sup> quartiles.

To calculate quartiles, we need to put the numbers in order first.

Texas Horned Lizard Length Data in order:

4.7 , 5.5 , 5.8 , 6.3 , 6.6 , 6.7 , 7.1 , 7.2 , 7.4 , 7.5 , 7.8 , 10.4

Now that the data is in order, think about quartering the data. Since there are 12 values in the data set, we should have  $12 / 4 = 3$  values in each quarter. Therefore, the quartiles should be placed between every three numbers.

4.7 , 5.5 , 5.8		6.3 , 6.6 , 6.7		7.1 , 7.2 , 7.4		7.5 , 7.8 , 10.4
Q1		Q2		Q3		

Therefore, Q1 should be half way between 5.8 and 6.3

$$Q1 = (5.8 + 6.3) / 2 = 6.05$$

1<sup>st</sup> Quartile Sentence: About 25% of the horned lizards had a length less than 6.05 cm.

Q2 (median) should be half way between 6.7 and 7.1

$$Q2 (\text{median}) = (6.7 + 7.1) / 2 = 6.9$$

2<sup>nd</sup> Quartile (Median) Sentence: About 50% of the horned lizards had a length less than 6.9 cm. The second quartile is also the median average for skewed data. So the average length of the horned lizards is also 6.9 cm.

Q3 should be half way between 7.4 and 7.5

$$Q3 = (7.4 + 7.5) / 2 = 7.45$$



This material is from *Introduction to Data Analysis*, 2<sup>nd</sup> edition by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a “CC-BY” [Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by/4.0/) – 3/17/2021

3<sup>rd</sup> Quartile Sentence: About 75% of the horned lizards had a length less than 7.45 cm.

This is how to think about quartiles. Notice we did not need a formula or fancy procedure to find the quartiles. We just needed to separate the data into four groups.

What about the interquartile range (IQR) for this data set?

$$\text{IQR} = Q3 - Q1 = 7.45 - 6.05 = 1.40 \text{ cm}$$

Interquartile Range Sentence: Typical Horned Lizard Lengths (middle 50%) were within 1.40 cm from each other.

#### Horned Lizard Data Summary

Average lizard length = 6.9 inches

Spread = 1.40 cm

Typical Lizard Lengths = Between 6.05 cm and 7.45 cm.

Horned Lizard Length Five Number Summary: 4.7 cm, 6.05 cm, 6.9 cm, 7.45 cm, 10.4 cm

(Minimum Value, 1<sup>st</sup> Quartile (Q1), Median (Q2), 3<sup>rd</sup> Quartile (Q3), Maximum Value)

Notice the five number summary shows the smallest and largest values in the data set, as well as the average (6.9 cm) and the 1<sup>st</sup> and 3<sup>rd</sup> quartiles that typical values fall in between.

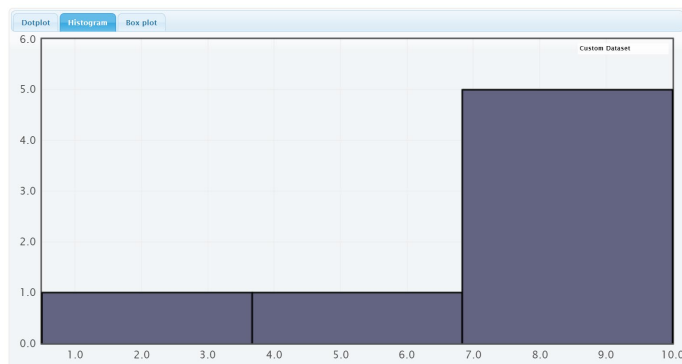
#### Example 2

There is some debate about how to calculate quartiles when the frequency is not divisible by four. This is especially true when there is an odd number of data values. Remember if there is an odd number of data values, the median (Q2) is an actual number in the data set. Since the median (Q2) is a value in the data set, it should be included in the top AND bottom halves of the data.

Let us look at the tips in dollars left at a bar. Here is the data in order.

\$0.50 , \$5.00 , \$7.50 , \$8.00 , \$8.50 , \$9.00 , \$10.00

Let's start by looking at the shape of this data. Here is a histogram created with StatKey. Notice the data is skewed left. Hence we should use the median (Q2) as the average, the IQR as the spread, and typical values should be in between Q1 and Q3.



Quartile Calculations: To find the quartiles, always start by finding the median (second quartile). The data is already in order from smallest to largest. In the previous section, we learned that since the number of values is odd, Q2 would be the middle number of \$8.00.

2<sup>nd</sup> Quartile (Median) Sentence: About 50% of the bar tips were less than \$8.00. The second quartile is also the median average for skewed data. So the average bar tip was \$8.00.

To find the first quartile (Q1), find the median of the bottom half of the data. Since the median is an actual value in the data, we will include it in our bottom half.

Bottom Half of Data: \$0.50 , \$5.00 , \$7.50 , \$8.00

Since we are including the median in the bottom half of the data, then there would be four numbers. The median of \$0.50 , \$5.00 , \$7.50 and \$8.00 would be just half way between \$5.00 and \$7.50 since this is an even number of values and there are two numbers in the middle.

$$Q1 = \text{median of bottom half of the data} = \frac{(5.0+7.5)}{2} = \$6.25$$

1<sup>st</sup> Quartile Sentence: About 25% of the bar tips were less than \$6.25.

To find the third quartile (Q3), find the median of the top half of the data. Since the median is an actual value in the data, we will include it in the top half of the data.

Top Half of the Data: \$8.00 , \$8.50 , \$9.00 , \$10.00

If we include the median in the top half, then there would be four numbers. The median of \$8.00 , \$8.50 , \$9.00 and \$10.00 would be half way between \$8.50 and \$9.00 since this is an even number of values and there are two numbers in the middle.

$$Q3 = \text{median of top half of the data} = \frac{(8.5+9.0)}{2} = \$8.75$$

3<sup>rd</sup> Quartile Sentence: About 75% of the bar tabs were less than \$8.75.

What about the interquartile range (IQR) for this data set?

$$IQR = Q3 - Q1 = \$8.75 - \$6.25 = \$2.50$$

Interquartile Range Sentence: Typical bar tips (middle 50%) were within \$2.50 from each other.

Bar Tips Data Summary

Average Bar Tip = \$8.00

Spread = \$2.50

Typical Bar Tips = Between \$6.25 and \$8.75.

Here is the summary statistics printout for the bar tip data from StatKey. Notice StatKey calculated the same median (Q2), the same 1<sup>st</sup> quartile (Q1) and the same 3<sup>rd</sup> Quartile (Q3). However, StatKey did not calculate the Interquartile Range (IQR). We would need to use the formula  $IQR = Q3 - Q1$  to calculate it.



## Summary Statistics

Statistic	Value
Sample Size	7
Mean	6.929
Standard Deviation	3.233
Minimum	0.5
Q <sub>1</sub>	6.250
Median	8.000
Q <sub>3</sub>	8.750
Maximum	10

**Bar Tip Five Number Summary:** \$0.50 , \$6.25 , \$8.00 , \$8.75 , \$10.00

(Minimum Value, 1<sup>st</sup> Quartile (Q1), Median (Q2), 3<sup>rd</sup> Quartile (Q3), Maximum Value)

Notice again that the five number summary shows the smallest and largest values in the data set, as well as the average (\$8.00) and the 1<sup>st</sup> and 3<sup>rd</sup> quartiles that typical values fall in between.

### Take away

Some computer programs have slight differences in how the quartiles are calculated. One program might give the 1<sup>st</sup> quartile as 2.5 mm and another program might give 2.6 mm. This difference in how quartiles are calculated is not something to dwell on. In a data set with ten thousand numbers, the quartiles will be about the same no matter what program you are using. In small data sets like the previous example, there can be some discrepancy, but it is not something to worry about.

Again, the key is to explain the meaning of statistics like median, Q1, Q3 and IQR. Use technology to calculate the statistics. Take whatever value the program gives and use it. It matters very little if Q1 came out to be 78.4 degrees Fahrenheit or 78.3 degrees Fahrenheit. The important thing when analyzing skewed quantitative data, is to be able to explain that the average is the median (Q2), the spread is IQR, typical values are between Q1 and Q3, approximately 25% of the values in the data were less than Q1, and approximately 75% of data values were less than Q3 and the average is Q2 (median).

### Calculating quartiles and IQR with technology

In large data sets, it is virtually impossible to calculate quartiles or any statistic for that matter with a calculator or by hand. We are now living in the age of “big data” where data sets often have hundreds of thousands of values or even millions of values. Surely, we cannot calculate the graphs and statistics we need from big data with a calculator. That is why it is so vital for data analysts to learn how to use statistics software like StatKey.

Here are the steps to calculating graphs and quantitative statistics with StatKey.

**StatKey Directions:** Go to [www.lock5stat.com](http://www.lock5stat.com). Click on “One Quantitative Variable” under the “Descriptive Statistics and Graphs” menu. Click on “Edit Data”. Copy and paste in the column of quantitative data you want to analyze. If the data has a title, check the box that says “Data has a header row”. If the data does NOT has a title, do NOT check the box that says “Data has a header row”. Do NOT check the box that says “First column is an identifier”. (You would only check the “identifier” box if there is a word next to every number in the data set.) Push “OK”. You should see the median (average), 1<sup>st</sup> Quartile, and 3<sup>rd</sup> Quartile in the list of “Summary Statistics” on the top right of the page.

NOTE: You will need to calculate the interquartile range (IQR) by subtracting the 3<sup>rd</sup> quartile in StatKey minus the 1<sup>st</sup> quartile in StatKey. (*Interquartile Range Formula:  $IQR = Q3 - Q1$* )



## Problem Set Section 5C

Directions: Put each of the following data sets in order from smallest to largest. Calculate the median (Q2), the first quartile (Q1), the third quartile (Q3) and the Interquartile Range ( $IQR = Q3 - Q1$ ).

Give the five number summary for the data set (*Minimum, Q1, Median, Q3, Maximum*).

1. { 5 , 7 , 8 , 8 , 9 , 11 , 14 , 16 , 17 , 19 , 21 , 25 , 26 , 29 , 31 , 33 , 36 }
  - a) Calculate the Median (Q2).
  - b) Calculate the 1<sup>st</sup> Quartile (Q1).
  - c) Calculate the 3<sup>rd</sup> Quartile (Q3).
  - d) Calculate the Interquartile Range  $IQR = Q3 - Q1$
  - e) Give the five number summary in order. (Minimum, Q1, Median, Q3, Maximum)
  
2. { 2.1 , 3.8 , 5.1 , 6.9 , 7.2 , 10.4 , 11.3 , 14.7 , 15.1 , 16.0 }
  - a) Calculate the Median (Q2).
  - b) Calculate the 1<sup>st</sup> Quartile (Q1).
  - c) Calculate the 3<sup>rd</sup> Quartile (Q3).
  - d) Calculate the Interquartile Range  $IQR = Q3 - Q1$
  - e) Give the five number summary in order. (Minimum, Q1, Median, Q3, Maximum)
  
3. { 31 , 34 , 41 , 52 , 68 , 71 , 79 , 83 , 88 , 90 , 103 }
  - a) Calculate the Median (Q2).
  - b) Calculate the 1<sup>st</sup> Quartile (Q1).
  - c) Calculate the 3<sup>rd</sup> Quartile (Q3).
  - d) Calculate the Interquartile Range  $IQR = Q3 - Q1$
  - e) Give the five number summary in order. (Minimum, Q1, Median, Q3, Maximum)
  
4. { 150 , 152 , 154 , 155 , 157 , 159 , 163 , 164 , 165 }
  - a) Calculate the Median (Q2).
  - b) Calculate the 1<sup>st</sup> Quartile (Q1).
  - c) Calculate the 3<sup>rd</sup> Quartile (Q3).
  - d) Calculate the Interquartile Range  $IQR = Q3 - Q1$
  - e) Give the five number summary in order. (Minimum, Q1, Median, Q3, Maximum)
  
5. { 7.5 , 2.3 , 4.6 , 1.9 , 2.8 , 9.4 , 8.3 , 6.1 }
  - a) Calculate the Median (Q2).
  - b) Calculate the 1<sup>st</sup> Quartile (Q1).
  - c) Calculate the 3<sup>rd</sup> Quartile (Q3).
  - d) Calculate the Interquartile Range  $IQR = Q3 - Q1$
  - e) Give the five number summary in order. (Minimum, Q1, Median, Q3, Maximum)



6. { 21 , 29 , 23 , 26 , 25 , 19 , 28 , 31 , 32 , 20 , 18 }

- a) Calculate the Median (Q2).
- b) Calculate the 1<sup>st</sup> Quartile (Q1).
- c) Calculate the 3<sup>rd</sup> Quartile (Q3).
- d) Calculate the Interquartile Range  $IQR = Q3 - Q1$
- e) Give the five number summary in order. (Minimum, Q1, Median, Q3, Maximum)

Directions: Use StatKey and the bear data to calculate minimum, maximum, median, Q1, Q3, and IQR for the following data sets. Then give the “five number summary” for each data set. Go to [www.lock5stat.com](http://www.lock5stat.com). Click on “One Quantitative Variable” under the “Descriptive Statistics and Graphs” menu. Click on “Edit Data”. Copy and paste in the column of quantitative data you want to analyze. If the data has a title, check the box that says “Data has a header row”. If the data does NOT have a title, do NOT check the box that says “Data has a header row”. Do NOT check the box that says “First column is an identifier”. (You would only check the “identifier” box if there is a word next to every number in the data set.) Push “OK”. You should see the median (average), 1<sup>st</sup> Quartile, and 3<sup>rd</sup> Quartile in the list of “Summary Statistics” on the top right of the page.

NOTE: You will need to calculate the interquartile range (IQR) by subtracting the 3<sup>rd</sup> quartile in StatKey minus the 1<sup>st</sup> quartile in StatKey. (*Interquartile Range Formula:  $IQR = Q3 - Q1$* )

7. Bear Ages in Months

- a) Calculate the Median (Q2).
- b) Calculate the 1<sup>st</sup> Quartile (Q1).
- c) Calculate the 3<sup>rd</sup> Quartile (Q3).
- d) Calculate the Interquartile Range  $IQR = Q3 - Q1$
- e) Give the five number summary in order. (Minimum, Q1, Median, Q3, Maximum)

8. Bear Head Length in Inches

- a) Calculate the Median (Q2).
- b) Calculate the 1<sup>st</sup> Quartile (Q1).
- c) Calculate the 3<sup>rd</sup> Quartile (Q3).
- d) Calculate the Interquartile Range  $IQR = Q3 - Q1$
- e) Give the five number summary in order. (Minimum, Q1, Median, Q3, Maximum)

9. Bear Head Width in Inches

- a) Calculate the Median (Q2).
- b) Calculate the 1<sup>st</sup> Quartile (Q1).
- c) Calculate the 3<sup>rd</sup> Quartile (Q3).
- d) Calculate the Interquartile Range  $IQR = Q3 - Q1$
- e) Give the five number summary in order. (Minimum, Q1, Median, Q3, Maximum)

10. Bear Neck Circumference in Inches

- a) Calculate the Median (Q2).
- b) Calculate the 1<sup>st</sup> Quartile (Q1).
- c) Calculate the 3<sup>rd</sup> Quartile (Q3).
- d) Calculate the Interquartile Range  $IQR = Q3 - Q1$
- e) Give the five number summary in order. (Minimum, Q1, Median, Q3, Maximum)





11. Bear Length in Inches

- a) Calculate the Median (Q2).
- b) Calculate the 1<sup>st</sup> Quartile (Q1).
- c) Calculate the 3<sup>rd</sup> Quartile (Q3).
- d) Calculate the Interquartile Range  $IQR = Q3 - Q1$
- e) Give the five number summary in order. (Minimum, Q1, Median, Q3, Maximum)

12. Bear Chest Size in Inches

- a) Calculate the Median (Q2).
- b) Calculate the 1<sup>st</sup> Quartile (Q1).
- c) Calculate the 3<sup>rd</sup> Quartile (Q3).
- d) Calculate the Interquartile Range  $IQR = Q3 - Q1$
- e) Give the five number summary in order. (Minimum, Q1, Median, Q3, Maximum)

13. Bear Weight in Pounds

- a) Calculate the Median (Q2).
  - b) Calculate the 1<sup>st</sup> Quartile (Q1).
  - c) Calculate the 3<sup>rd</sup> Quartile (Q3).
  - d) Calculate the Interquartile Range  $IQR = Q3 - Q1$
  - e) Give the five number summary in order. (Minimum, Q1, Median, Q3, Maximum)
- 



## Section 5D – Box Plots and Finding Unusual Values for Skewed Data

So far, we have seen that when a data set is skewed right, skewed left, or not normal, we should use the median as our center and average and the interquartile range (IQR) for the spread. We also learned that typical values will make up the middle 50% of data values and fall between the 1<sup>st</sup> and 3<sup>rd</sup> quartiles. What about finding unusual values (outliers) for skewed data sets?

### Unusual Values

For bell shaped data sets, unusual values (outliers) are more than two standard deviations from the mean, but skewed data involves more extreme values and is more spread out. It therefore has a different rule for finding unusual values (outliers). Here are the unusual cutoff values for skewed or non-normal data.

*Unusually High Cutoff for Skewed Data:*  $Q3 + (1.5 \times \text{IQR})$

*Unusually Low Cutoff for Skewed Data:*  $Q1 - (1.5 \times \text{IQR})$

The good news is that the typical and unusual values for skewed data are summarized nicely with a box plot. The box plot is a fabulous graph to look at when your data is skewed right, skewed left or not normal.

I like to call the unusual cutoff values the “Box and a Half Rule”, since  $1.5 \times \text{IQR}$  represents the length of a box and a half. So any values in the skewed data that is a box and half from the box are considered unusual (outlier).

### Introduction to Box Plots

Let us look at how box plots work. Remember to use technology when you create a box plot. No statistician, data analyst, or data scientist creates graphs by hand, especially with big data sets.

Let us look at an example where we do make the box plot by hand, just so we can understand the process.

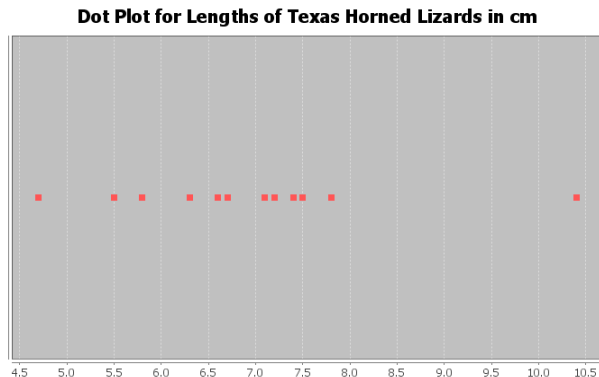
#### Example 1

Let us look at the Texas Horned Lizard data and create a box plot for the data.

Texas Horned Lizard Length Data in order:

4.7 , 5.5 , 5.8 , 6.3 , 6.6 , 6.7 , 7.1 , 7.2 , 7.4 , 7.5 , 7.8 , 10.4

A dot plot of the data indicated a skewed right shape. Therefore, this works nicely for a box plot.



In the last section, we calculated the three quartiles and the interquartile range for this data.

4.7 , 5.5 , 5.8 | 6.3 , 6.6 , 6.7 | 7.1 , 7.2 , 7.4 | 7.5 , 7.8 , 10.4



Q1

Q2

Q3

Q1 should be half way between 5.8 and 6.3

$$Q1 = (5.8 + 6.3) / 2 = 6.05$$

Q2 (median average) should be half way between 6.7 and 7.1

$$Q2 \text{ (median)} = (6.7 + 7.1) / 2 = 6.9$$

Q3 should be half way between 7.4 and 7.5

$$Q3 = (7.4 + 7.5) / 2 = 7.45$$

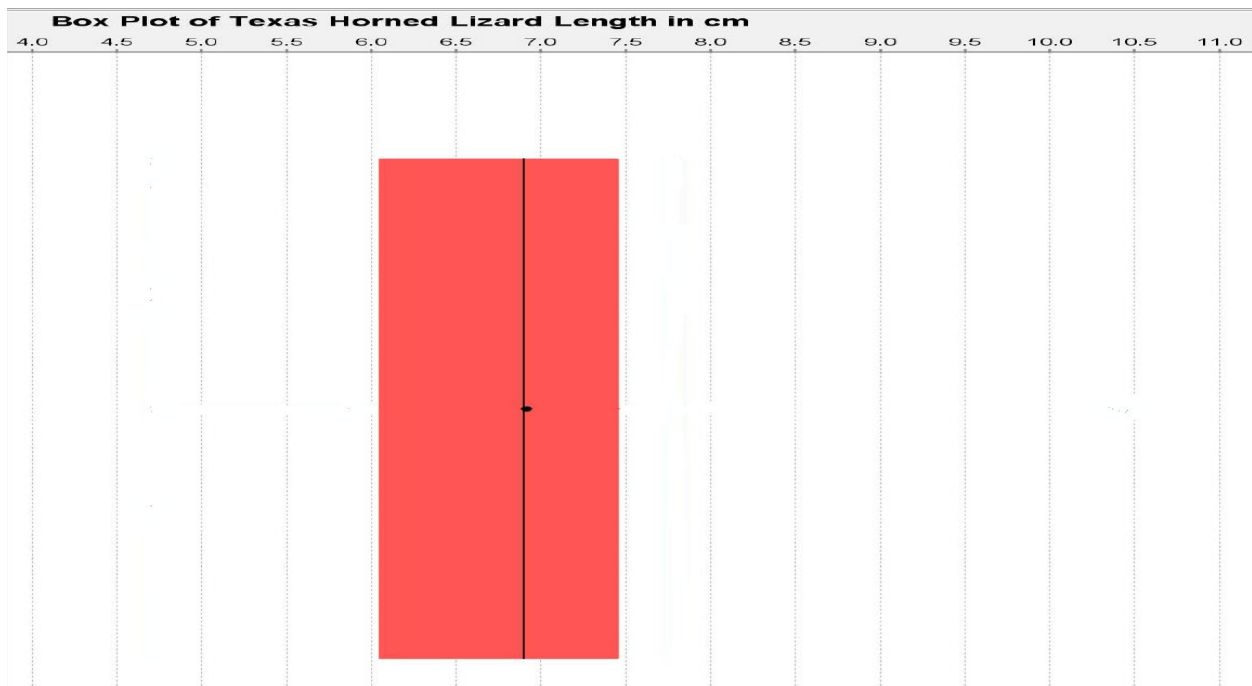
$$IQR = Q3 - Q1 = 7.45 - 6.05 = 1.40 \text{ cm}$$

Typical Values: Since this is skewed data, typical values will fall in between Q1 and Q3. So typical horned lizards in this data set have a length between 6.05 cm and 7.45 cm.

*Note: Q1 and Q3 do not accurately represent typical values in bell shaped data. You would need to use the standard deviation and the mean in that case.*

#### Making the box plot

Start by drawing an even number line that goes from the smallest and largest values in the data set. Then draw a box from Q1 to Q3. Draw a line in the box at the median average (Q2).



Now we need to calculate the unusual cutoff fences to determine if there are any unusual values (outliers) in the data set. To calculate the outliers, we will need to find the distance of a box and a half ( $1.5 \times IQR$ ). In this data  $1.5 \times IQR$



$= 1.5 \times 1.40 = 2.1$ . So any data values that are 2.1 or higher from Q3 are high outliers (unusually high values). Also any data values that are 2.1 or lower from Q1 are low outliers (unusually low values).

*Unusually High Cutoff for Skewed Data:*

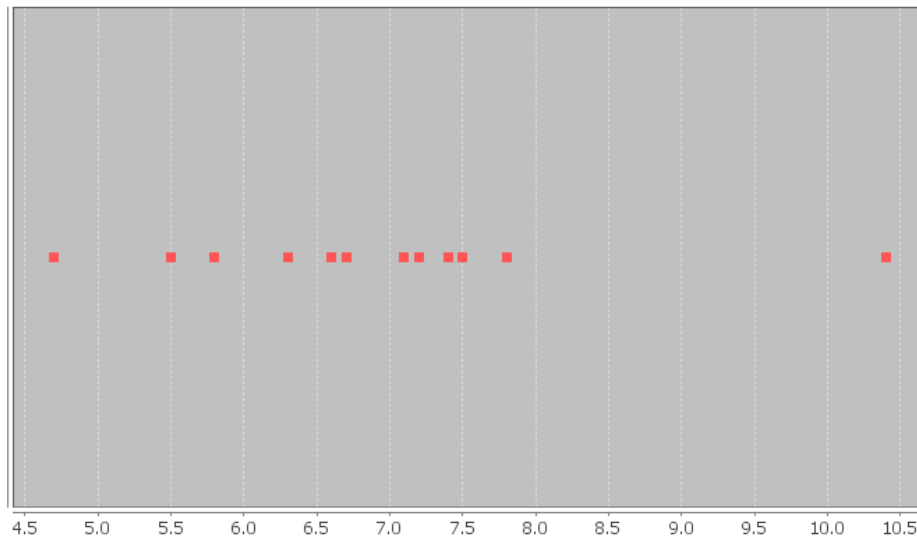
$$Q3 + (1.5 \times IQR) \approx 7.45 + (1.5 \times 1.40) \approx 7.45 + 2.1 \approx 9.55$$

*Unusually Low Cutoff for Skewed Data:*

$$Q1 - (1.5 \times IQR) \approx 6.05 - (1.5 \times 1.40) \approx 6.05 - 2.1 \approx 3.95$$

Let us look at the dot plot again and see if there are any numbers that are 3.95 or lower. We can also look to see if there are any numbers that are 9.55 or higher.

**Dot Plot for Lengths of Texas Horned Lizards in cm**



Notice there are no values in the data set that are 3.95 cm or below. That means there are no unusually low values in the data set.

There is one value in the data set that is 9.55 cm or higher. It is the maximum value of the data set 10.4 cm. Therefore, 10.4 cm is an unusually high value (high outlier) in the data set. We need to designate that value as an outlier (unusual). Some computer programs draw their outliers with a circle, some draw it with a triangle, and some draw it with a star. I will draw it with a triangle.



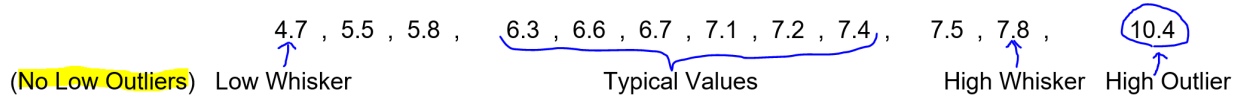


The box plot is also called the “box and whisker plot”. Now we need to determine where to draw the whiskers. The whiskers are drawn to the highest and lowest numbers in the data set that are not outliers (not unusual). Be careful. The whiskers are not drawn to the unusual cutoff fences. They must be drawn to numbers that are actually in the data set and are not outliers.

There was no unusually low value in the data set. Therefore, the low whisker on the left should be drawn to the smallest number in the data set, which is 4.7 cm.

There was an unusually high value (outlier) at 10.4 cm. That means we cannot draw the whisker to that value. We must choose a new maximum value in the data set that is not an outlier. Looking at the dot plot, we see that the next biggest number in the data set was 7.8 cm. That is 9.55 cm or below so it is not unusual. We will draw the high whisker (on the right) to 7.8 cm since that is the largest number in the data set that is not an outlier (not unusual).

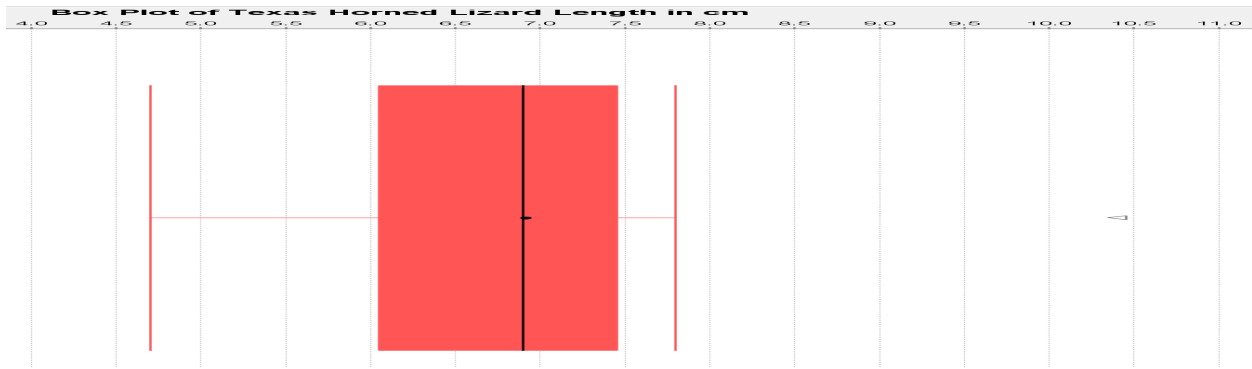
Texas Horned Lizard Length Data in order:



**Note:** Notice that there are values in the data that are neither typical nor an outlier. Some students make the mistake of thinking that if a data value is not typical, it is unusual. That is NOT true. There are many values in data sets that are not typical and not outliers.

Putting this information into our Box Plot, gives us the following graph.





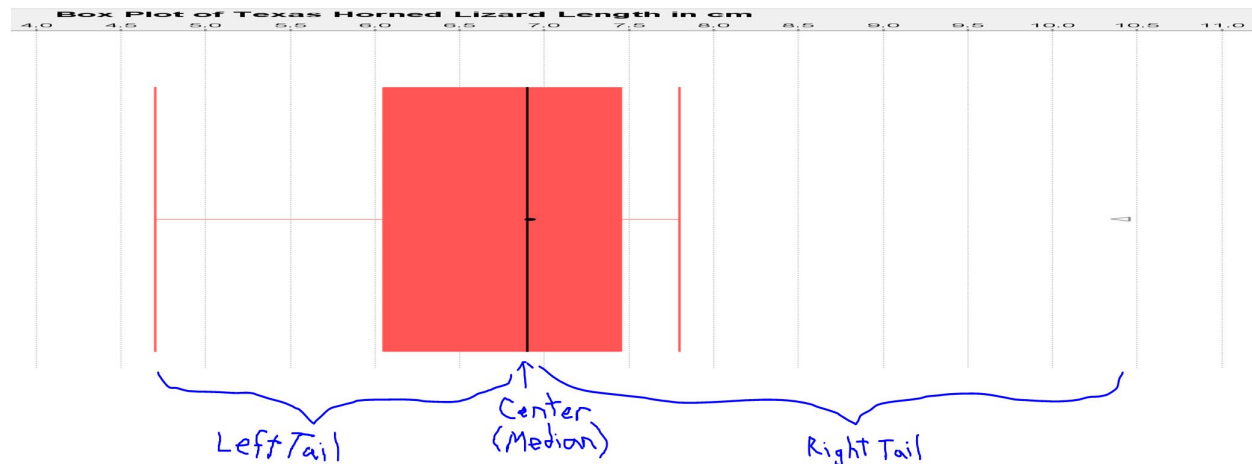
The whiskers probably get their name because they kind of look like cat whiskers. This is our complete box and whisker plot or box plot for short.

We can learn a lot by looking a box plot like this.

- The line in the box shows the median average.
- The edges of the box show the 1<sup>st</sup> and 3<sup>rd</sup> quartiles. Typical values will be in between them.
- The whiskers show the values that are neither typical, nor an outlier. The end of the left whisker shows the smallest number in the data set that is not an outlier. The end of the right whisker shows the largest number in the data set that was not an outlier.
- Stars, circles or triangles outside the whiskers are outliers (unusually low and high values).

#### Determining shape from a box plot

It is usually best to look at a histogram or dot plot when determining shape. Remember, a box plot is really a graph of the quartiles and outliers. However, since the median is the center and the line inside the box, we can at least compare the tails. In the box plot, we can look at the distance from the median to the largest value in the data and the distance from the median to the smallest value in the data. Remember this data did not have any low outliers so the smallest value is found at the end of the low whisker. Since the right tail is longer than the left tail, this is likely to be a skewed right data set.



#### Creating Box Plots with Technology

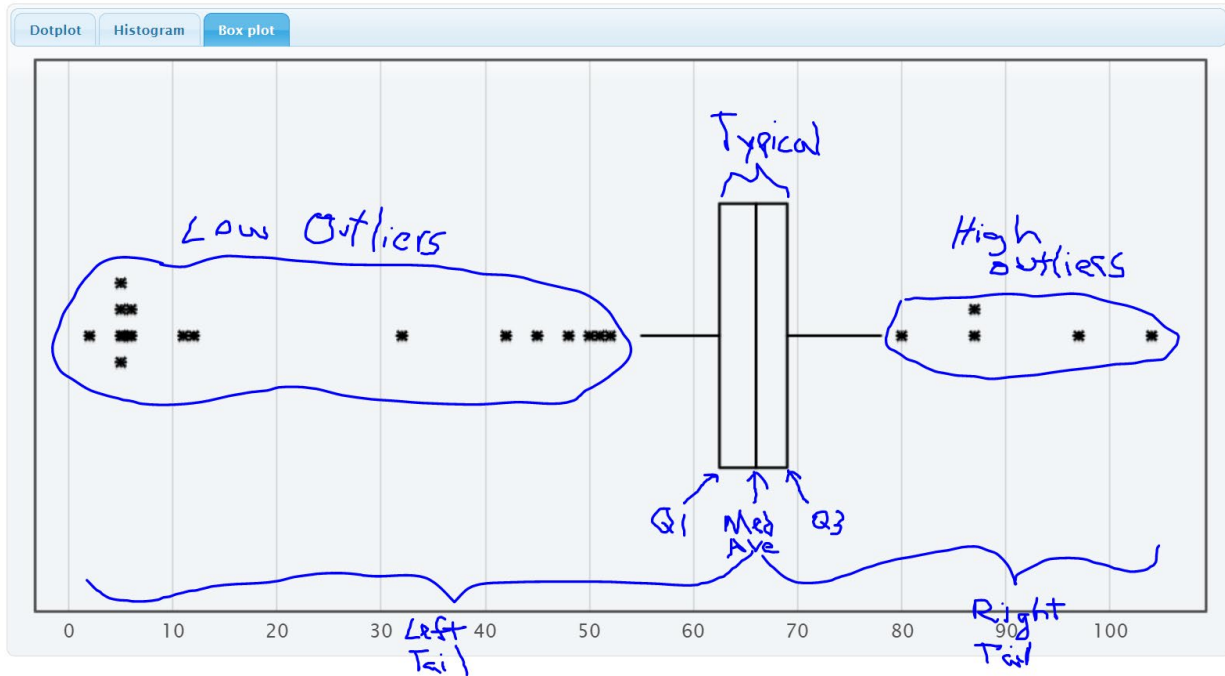
Let us look at how to create box plots with StatKey.



**Making a box plot in StatKey:** Go to [www.lock5stat.com](http://www.lock5stat.com). Click on "One Quantitative Variable" under the "Descriptive Statistics and Graphs" menu. Click on "Edit Data". Copy and paste in the column of quantitative data you want to analyze. If the data has a title, check the box that says "Data has a header row". If the data does NOT have a title, do NOT check the box that says "Data has a header row". Do NOT check the box that says "First column is an identifier". (You would only check the "identifier" box if there is a word next to every number in the data.) Now push "OK". At the top left of the graph, click on the "box plot" tab.

**Example**

In the spring 2016 semester, we asked math 140 statistics students how tall they were in inches. We created a box plot using StatKey. It is good to refer the summary statistics to see the actual value of the median and quartiles. Also if you hold your cursor over the stars (outliers), StatKey will tell you the value of the outlier.



**Summary Statistics**

Statistic	Value
Sample Size	357
Mean	64.390
Standard Deviation	12.040
Minimum	2
Q <sub>1</sub>	62.500
Median	66.000
Q <sub>3</sub>	69.000
Maximum	104



We can see a lot of information from this graph.

- The left tail is longer than the right tail so the data is likely to be skewed left.
- Since it is skewed, we should use the median and quartiles instead of the mean and standard deviation.
- The median average height of the stat students was 66 inches ( $5\frac{1}{2}$  feet).
- Typical heights fell between Q1 (62.5 inches) and Q3 (69 inches).
- The width of the box is the spread (IQR =  $69 - 62.5 = 6.5$  inches). So typical statistics students have a height within 6.5 inches of each other.
- The low outlier cutoff is  $Q1 - (1.5 \times IQR) = 62.5 - (1.5 \times 6.5) = 62.5 - 9.75 = 52.75$  inches. So any height below 52.75 inches would be considered a low outlier (unusually low). There are 18 low outliers ranging from 2 inches to 52 inches. Putting the data in order, we can identify the outliers in the excel spreadsheet below. Some students that said they were only a few inches tall. Maybe they thought the question was in feet.
- The high outlier cutoff is  $Q3 + (1.5 \times IQR) = 69 + (1.5 \times 6.5) = 69 + 9.75 = 78.75$  inches. So any height above 78.75 inches would be considered a high outlier (unusually high). There are five high outliers ranging from 80 inches to 104 inches. Putting the data in order, we can identify the outliers in the excel spreadsheet below.  
Some of these might also be a miscalculation. 80 inches (6 ft 8 in) is possible but 104 inches is 8 ft 8 inches. There is probably no stat student that tall.
- The whiskers indicate that there are some students whose heights are neither typical not unusual.

Excel spreadsheet showing low outliers

2	Low Outlier	
5	Low Outlier	
5	Low Outlier	
5	Low Outlier	
5	Low Outlier	
5.2	Low Outlier	
5.7	Low Outlier	
6	Low Outlier	
6	Low Outlier	
11	Low Outlier	
12	Low Outlier	
32	Low Outlier	
42	Low Outlier	
45	Low Outlier	
48	Low Outlier	
50	Low Outlier	
51	Low Outlier	
52	Low Outlier	
	Low Outlier Cutoff = 52.75	
55		
56		
56		

Excel Spreadsheet showing high outliers

76		
77		
77		
77.5		
78		
	High Outlier Cutoff = 78.75	
80	High Outlier	
87	High Outlier	
87	High Outlier	
97	High Outlier	
104	High Outlier	





### Summary Paragraph

In our previous chapter, we saw that we can write a summary paragraph with all the information about a quantitative data set. For normal data, our average, spread, typical values and outliers were calculated with the mean average and standard deviation. For data that is skewed left, skewed right, or not normal, the mean and standard deviation are not accurate. So we will use the median average, quartiles and interquartile range.

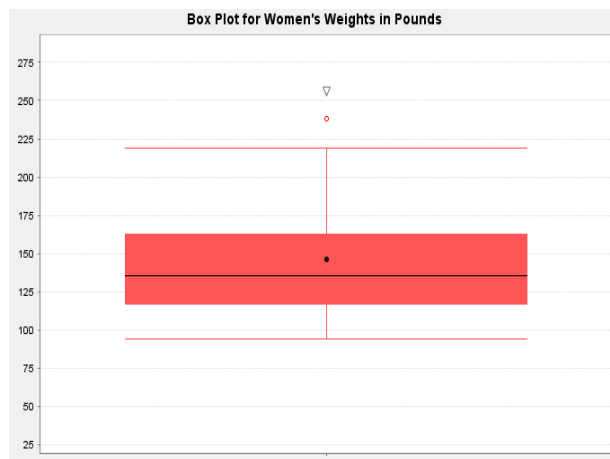
Here is the summary paragraph for the height of COC stat students in spring 2016. Notice since the data is skewed left we used the median average and IQR.

### Summary Paragraph:

*The data measured the heights in inches of 357 math 140 stat students in the spring 2016 semester. The shape of this quantitative data is skewed left. The median average height of the stat students was 66 inches ( $5\frac{1}{2}$  feet). The spread of the data was 6.5 inches (IQR). So typical statistics students have a height within 6.5 inches of each other. Typical heights fell between Q1 (62.5 inches) and Q3 (69 inches). There are 18 low outliers ranging from 2 inches to 52 inches. Some students that said they were only a few inches tall. These data values are obvious mistakes. There are five high outliers ranging from 80 inches to 104 inches. Some of these might also be a miscalculation. 80 inches (6 ft 8 in) is possible but 104 inches is 8 ft 8 inches. There is probably no stat student that tall.*

### Vertical Box Plots

Box plots can also be drawn vertically. Here is an example. Now high outliers are at the top and low outliers are at the bottom.

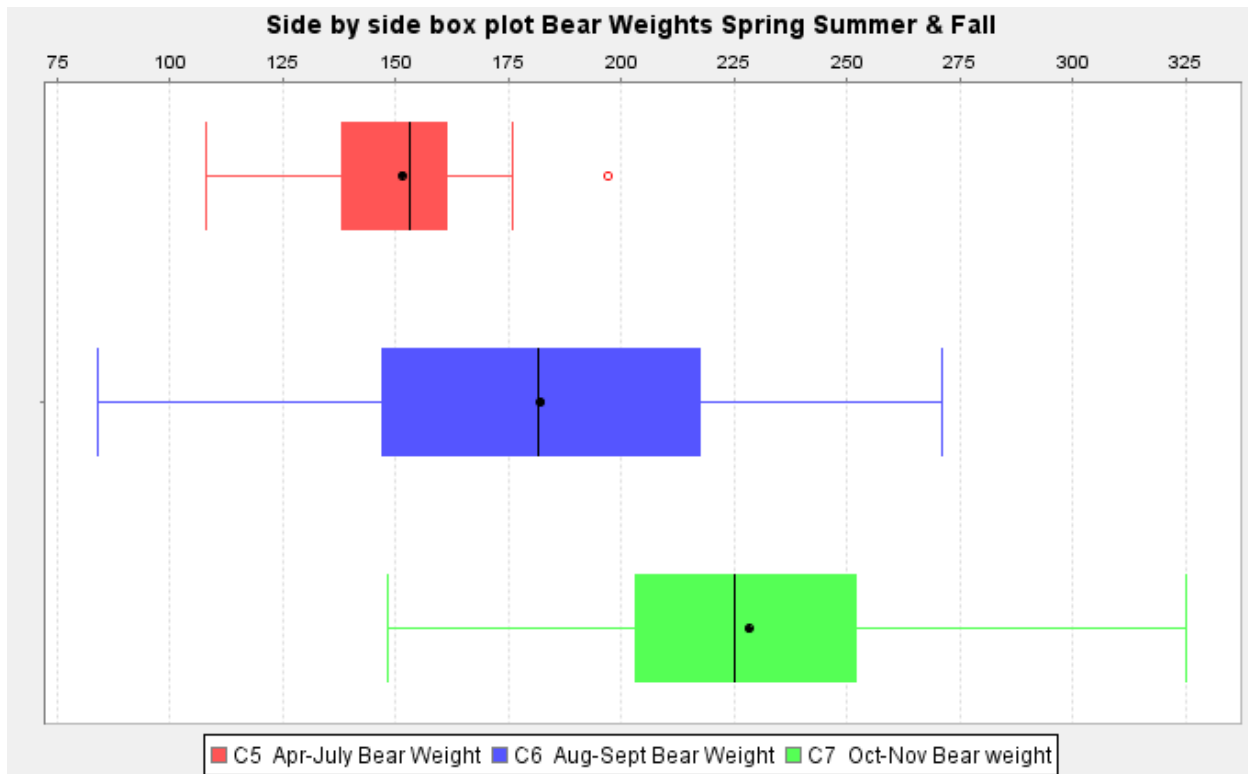


### Interpreting Box Plots

Remember to use technology to create graphs and find statistics. The important part is being able to interpret what the graph and statistics are telling us.

Box plots are often used to compare quantitative data from different groups. We call this kind of graph a “side by side” box plots. These graphs can be drawn horizontally or vertically. The following example was found from the North American black bear data. The bear weights were separated into three groups depending on what time of the year the measurements were taken (Spring, Summer or Fall). The box on top is describing bears measured in spring (April – July), the box in the middle is describing bears measured in summer (August – September) and the box on the bottom is describing bears measured in fall (October – November).





Graphs like this give us a lot of information.

Which group of bears had the highest average weight and what was the highest average weight?

*Notice the lines inside the box are the medians, which are very accurate measures of center. The group whose median line is farthest to the right is the fall bears (October-November). The bears measured in fall had the highest average weight. The line in the box looks like it falls around 225 pounds. So the average weight of the North American black bears measured in fall was 225 pounds. The average weight of the bears measured in spring was about 155 pounds and the average weight of bears measured in summer was about 180 pounds.*

Which group of bears had the most typical spread (variability) in their weights? *Typical spread (IQR) is the length of your box. So which group had the longest box? We can see it is the middle group of bears measured in summer (August – September). The bears measured in summer had the most variability in their weights. A typical bear measured in summer could have a weight from 145 pounds to 215 pounds.*



## Problem Set Section 5D

(#1-3) Directions: The median, 1<sup>st</sup> quartile and 3<sup>rd</sup> quartile are given for the following data sets. Calculate the IQR and outlier fences. Then identify the outliers, answer the questions, and draw a box plot for the data on a piece of paper. The data sets are already in order.

1. { 4 , 5 , 19 , 20 , 21 , 22 , 23 , 24 , 26 , 27 , 28 , 29 , 30 , 32 , 33 , 51 }

Median = 25

Q1 = 20.5

Q3 = 29.5

- Calculate the IQR =  $Q3 - Q1$
  - Calculate the high outlier fence  $Q3 + (1.5 \times IQR)$
  - Calculate the low outlier fence  $Q1 - (1.5 \times IQR)$
  - List all of the high outliers. (Data values larger than the high outlier fence.)
  - List all of the low outliers. (Data values smaller than the low outlier fence.)
  - What is the largest value in the data set that is NOT an outlier.  
(This is where the right whisker will go.)
  - What is the smallest value in the data set that is NOT an outlier.  
(This is where the left whisker will go.)
  - Draw the box plot including whiskers, outlier fences and outliers.
2. { 23 , 31 , 32 , 33 , 34 , 35 , 36 , 37 , 55 }

Median = 34

Q1 = 32

Q3 = 36

- Calculate the IQR =  $Q3 - Q1$
- Calculate the high outlier fence  $Q3 + (1.5 \times IQR)$
- Calculate the low outlier fence  $Q1 - (1.5 \times IQR)$
- List all of the high outliers. (Data values larger than the high outlier fence.)
- List all of the low outliers. (Data values smaller than the low outlier fence.)
- What is the largest value in the data set that is NOT an outlier.  
(This is where the right whisker will go.)
- What is the smallest value in the data set that is NOT an outlier.  
(This is where the left whisker will go.)
- Draw the box plot including whiskers, outlier fences and outliers.



3. { 8.4 , 9.6 , 10.8 , 10.9 , 11.0 , 11.2 , 11.3 , 11.4 , 11.6 , 11.7 , 12.9 , 13.1 }

Median = 11.25

Q1 = 10.85

Q3 = 11.65

- a) Calculate the IQR =  $Q3 - Q1$
- b) Calculate the high outlier fence  $Q3 + (1.5 \times IQR)$
- c) Calculate the low outlier fence  $Q1 - (1.5 \times IQR)$
- d) List all of the high outliers. (Data values larger than the high outlier fence.)
- e) List all of the low outliers. (Data values smaller than the low outlier fence.)
- f) What is the largest value in the data set that is NOT an outlier.  
(This is where the right whisker will go.)
- g) What is the smallest value in the data set that is NOT an outlier.  
(This is where the left whisker will go.)
- h) Draw the box plot including whiskers, outlier fences and outliers.

(#4-7) Directions: Use StatKey and the Math 075 Survey Data Fall 2015 to create Box Plots for the following data sets. Draw a rough sketch of the box plot on a piece of paper or save the box plot on a word document.

Go to [www.lock5stat.com](http://www.lock5stat.com). Click on "One Quantitative Variable" under the "Descriptive Statistics and Graphs" menu. Click on "Edit Data". Copy and paste in the column of quantitative data you want to analyze. If the data has a title, check the box that says "Data has a header row". If the data does NOT has a title, do NOT check the box that says "Data has a header row". Do NOT check the box that says "First column is an identifier". (You would only check the "identifier" box if there is a word next to every number in the data.) Now push "OK". At the top left of the graph, click on the "box plot" tab.

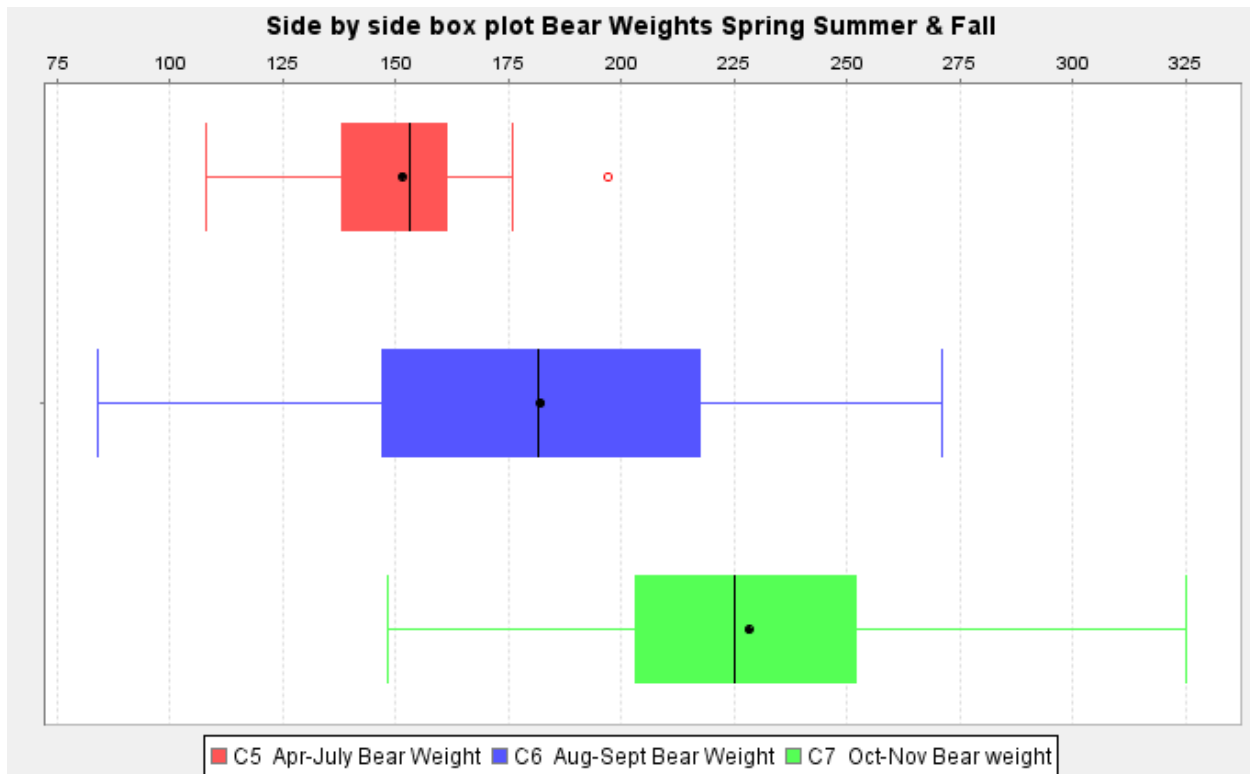
4. Age of Math 075 students in Fall 2015 (Column "C").
5. Work hours per week for Math 075 students in Fall 2015 (Column "J").
6. Exercise hours per week for Math 075 students in Fall 2015. (Column "L")
7. Commute time to campus in minutes for Math 075 students in Fall 2015. (Column "N")

Directions: Let us look again at the side-by-side box plot describing the weights of bears measured at different times of the year. The box on top is describing bears measured in spring (April – July), the box in the middle is describing



This material is from *Introduction to Data Analysis*, 2<sup>nd</sup> edition by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a "CC-BY" [Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by/4.0/) – 3/17/2021

bears measured in summer (August – September) and the box on the bottom is describing bears measured in fall (October – November).



8. Which group of bears had the lowest median average weight? Estimate the lowest median average weight from the graph for the three groups?
  
  9. Which group of bears had the lowest typical spread (IQR) in their weights? (*These are the bears whose weights were most consistent.*) Estimate the smallest IQR from the graph? Estimate the two values that typical values fall in between (Q1 and Q3) for the smallest spread group.
  
  10. Were there any outliers (unusual bear weights) in any of the three groups (yes or no)? If so, what group had an unusual value (outlier)? Was it a high outlier or a low outlier? Estimate the bear weight or weights from the graph that were considered unusual.
- 



## Section 5E – Measures of Center, Spread and Position

**Statistics:** Numbers calculated from sample data in order to understand the characteristics of the data.

Though the mean, median, standard deviation and IQR are used most often to analyze quantitative data, there are many different types of statistics that can also be used to dig deeper into the data. We will not be covering these statistics in depth, but it is good to at least have an idea of what they measure.

Memorize the following definitions so that you can explain these statistics if needed. You should also know if the statistic is a measure of center, spread or position.

- Measures of center (*mean, median, mode and midrange*) are types of averages.
- Measures of spread (*standard deviation, variance, range, and interquartile range*) measure variability or how much the data is spread out.
- Measures of position (*min, max, 1<sup>st</sup> quartile (Q1) and 3<sup>rd</sup> quartile (Q3)*) are statistics that we often use to identify where a data value falls compared to these positions.

### Measures of Center

**Mean Average:** The balancing point in terms of distances. The measure of center or average used when a quantitative data set is bell shaped (normal). The mean average is calculated by adding all of the data values and then dividing that sum by how many numbers are in the data.

**Median Average:** The center of the data in terms of order. Also called the second quartile (Q2) or the 50<sup>th</sup> percentile. Approximately 50% of the data will be less than the median and 50% will be above the median. This is the measure of center or average used when a data set is skewed (not bell shaped).

**Mode:** The number that occurs most often in a data set. Data sets may have no mode, one mode, or multiple modes. It is also sometimes used in bimodal or multimodal data.

**Midrange:** A quick measure of center that is usually not very accurate, but can be calculated quickly without a computer. The midrange lies half way between the smallest and largest values in the data.

$$\text{Midrange} = (\text{Max} + \text{Min}) \div 2$$

### Measures of Spread

**Standard Deviation:** How far typical values are from the mean in a normal (bell shaped) data set. It is the most accurate measure of spread for normal quantitative data. If you add and subtract the mean and standard deviation, you get two numbers that typical values in a bell shaped data set fall in between. It can also be used to find unusual values in bell shaped data. The standard deviation should not be used unless the data is bell shaped.

**Variance:** The standard deviation squared. A measure of spread used in ANOVA testing. Only accurate when the data is normal (bell shaped).

**Range:** The distance between the max and the min. All the data values are within this amount of each other. Range is a quick measure of spread that is not very accurate. It is based on unusual values and does not measure typical spread in the data set. It can be calculated quickly without a computer. The range is calculated by subtracting the maximum value in the data minus the minimum value in the data. (Range = Max – Min)

**Interquartile range (IQR):** How far typical values are from each other in a skewed data set. Measures the length of the middle 50% of the data. It is the most accurate measure of spread for skewed data sets. Interquartile range should not be used when data is bell shaped. Interquartile range is calculated by subtracting the 3<sup>rd</sup> quartile minus the 1<sup>st</sup> quartile. (IQR = Q3 – Q1)

### Measures of Position

**Minimum:** The smallest number in the data set.

**Maximum:** The largest number in the data set.

**First Quartile (Q1):** The number that approximately 25% of the data is less than and 75% of the data is greater than. Used for finding typical values for skewed data sets.



**Third Quartile (Q3):** The number that approximately 75% of the data is less than and 25% of the data is greater than. Used for finding typical values for skewed data sets.

### Total Frequency or Sample Size (n)

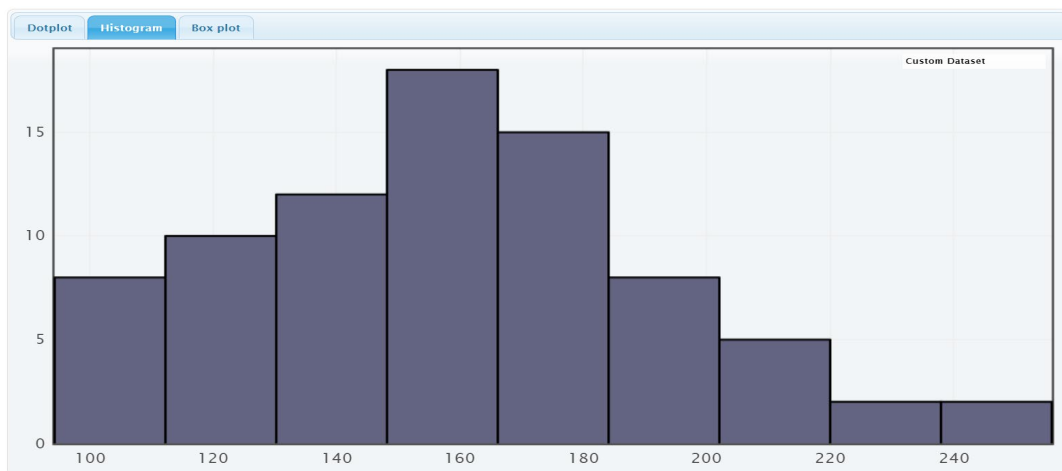
The total frequency or sample size of a data set (n) is not a measure of center, spread or position, but is important bit of information. It tells us how many numbers are in the data set.

### Example

Here is a StatKey printout of the weights in pounds of 80 randomly selected adults. Explain each of the statistics listed above in context.

### Summary Statistics

Statistic	Value
Sample Size	80
Mean	159.385
Standard Deviation	34.877
Minimum	94.3
Q <sub>1</sub>	135.000
Median	161.000
Q <sub>3</sub>	179.600
Maximum	255.9



Here is a similar printout from another computer program called Statcato. Notice it is similar to StatKey but has more statistics listed.

**Descriptive Statistics**

Variable	Mean	Standard Deviation	Variance
Weight (Lbs)	159.385	34.877	1216.397

Variable	Q1	Median	Q3	IQR	Mode	N for mode
Weight (Lbs)	135.0	161.0	179.75	44.75	161.9, 135.0, 156.3	2

Variable	Min	Max
Weight (Lbs)	94.3	255.9

Variable	N total
Weight (Lbs)	80

**Mean Average Weights:** The measure of center or average weight for these adults that balances the distances is 159.385 pounds. This average is only accurate if the data is normal (bell shaped).

**Median Average Weights:** The measure of center or average when the data values are put in order is 161 pounds. So approximately 50% of the adults weighed less than 161 pounds and approximately 50% of the adults weighed more than 161 pounds.

**Mode:** StatKey does not calculate the mode. We do see it on the Statcato printout though. The weights that appear most often in a data set are 135.0 pounds, 156.3 pounds and 161.9 pounds. Looking at “N for mode” we see that these three weights all appeared twice in the data set. This data set would be considered multimodal since it has three modes. Notice that two of the modes were close to the median (center) of the data.

**Midrange:** The midrange is not listed in either printout, but can be easily calculated.  
 $\text{Midrange} = (\text{Max} + \text{Min}) \div 2 = (255.9 + 94.3) \div 2 = (350.2) \div 2 = 175.1$  pounds. The midrange average weight was 175.1 pounds. Notice it is not close to the actual center (median) of the data, so not a very accurate average.

**Measures of Spread**

**Standard Deviation:** Typical values are within 34.877 pounds of each other. The standard deviation is only accurate if the data is bell shaped.

**Variance:** StatKey does not list the variance, but it can be easily calculated by squaring the standard deviation.  
 $\text{Variance} = 34.877 \times 34.877 \approx 1216.405$ . The variance or standard deviation squared is approximately 1216.4 square pounds. The variance is only accurate when the data is normal (bell shaped). *(Notice the Statcato printout lists the variance as 1216.397. Statcato is more accurate in this case because it keeps more decimal places. StatKey rounds the standard deviation to the thousandths place, so when we use it to calculate, we have a little bit of a rounding error.)*

**Range:** StatKey does not list the range but it can be easily calculated.  $\text{Range} = \text{Max} - \text{Min} = 255.9 - 94.3 = 161.6$  pounds. The distance between the max and the min for the weight data is 161.6 pounds. All the data values are within 161.6 pounds of each other.

**Interquartile range (IQR):** StatKey does not list the interquartile range (IQR) but it can be easily calculated.  
 $\text{IQR} = \text{Q3} - \text{Q1} = 179.6 - 135 = 44.6$  pounds. Typical data values are within 44.6 pounds of each other. This would only be accurate if the data is not normal. *(Notice that Statcato lists the IQR as 44.75 pounds. Computer programs often have slight differences in the quartile calculations. They are close though, so either would be ok to use.)*





### Measures of Position

Minimum: The lightest adult in the data set was 94.3 pounds.

Maximum: The heaviest adult in the data set was 255.9 pounds.

First Quartile (Q1): Approximately 25% of the adults in the data weighed less than 135 pounds.

Third Quartile (Q3): Approximately 75% of the adults in the data weighed less than 179.6 pounds (StatKey).  
(Notice that Statcato lists Q3 as 179.75 pounds. Computer programs often have slight differences in the quartile calculations. They are close though, so either would be ok to use.)

Total Frequency or Sample Size (n): Weight data was collected from a total of 80 adults. (Notice StatKey lists this statistic as “sample size” and Statcato lists it as “N total”).

---



Problem Set Section 5E

1. For each of the following statistics, classify it as a measure of center, spread or position.

- a) Q1
- b) Mean
- c) Variance
- d) Midrange
- e) Standard Deviation
- f) Minimum
- g) Q3
- h) Mode
- i) IQR
- j) Median
- k) Range
- l) Maximum

2. The following statistics were created from some weekly salary data in dollars from people living in Victoria, Australia. Write a sentence or two explaining the meaning of each of these statistics in context.

**Descriptive Statistics**

Variable	Mean	Standard Deviation	Variance
Victoria Salary	1149.050	516.553	266826.719

Variable	Q1	Median	Q3	IQR	Mode	N for mode
Victoria Salary	703.45	1015.74	1496.11	792.660	1011	2

Variable	Min	Max	Range
Victoria Salary	371.57	2396.28	2024.710

Variable	N total
Victoria Salary	35



## Chapter 5 Review Sheet

Here is a list of important ideas in this chapter.

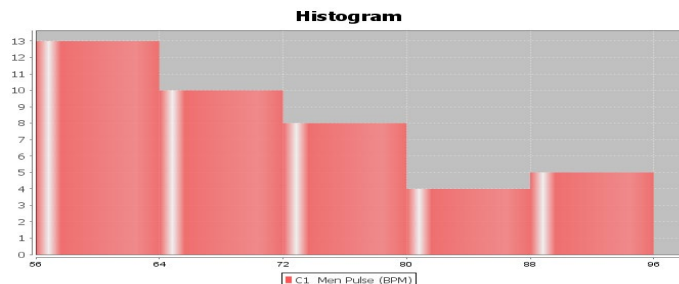
- Be able to distinguish between categorical data and quantitative (numerical measurement) data.
- Be able to create histograms and dot plots with technology and find the shape of a quantitative data set.
- Be able to find the five number summary (minimum, Q1, median, Q3, Maximum) with a calculator and with technology. Also find the interquartile range (IQR) and the total frequency (N).
- Write sentences to explain Q1, Median, Q3 and IQR.
- The interquartile range (IQR) tells us the maximum distance that typical values are from each other in a skewed data set. It measures the spread for the middle 50% and is the most accurate spread for skewed data sets.
- The first quartile Q1 is a divider that about 25% of the data values are less than and about 75% of the data values are greater than.
- The third quartile Q3 is a divider that about 75% of the data values are less than and about 25% of the data values are greater than.
- A center gives an average value for the data set is usually close to the highest bar or bars in the histogram.
- If a data set is skewed, we should use the median average as our measure of center and our average for the data set.
- A measure of spread or variability tells us how spread out the data set is. The more spread out the data is, the less consistent the data is and the harder it is to predict. A small amount of spread tells us that the data is more consistent and easier to predict.
- If a data set is skewed, we should use the interquartile range (IQR) as our measure of spread for the data set. If a data set is bell shaped, then we should not use the IQR.
- For Skewed Data:  $Q1 \leq \text{Typical Values} \leq Q3$
- Unusually High Cutoff for Skewed Data:  $Q3 + (1.5 \times \text{IQR})$  (Automatically calculated in a box plot)
- Unusually Low Cutoff for Skewed Data:  $Q1 - (1.5 \times \text{IQR})$  (Automatically calculated in a box plot)
- Be able to read and use a box plot to understand quartiles and percentages and identify unusual values in the data set.
- Be able to write a summary report paragraph summarizing the key characteristics of a skewed quantitative data set.
- Be able to classify various statistics as a measure of center, spread or position.

### Problems Chapter 5 Review Sheet

Directions: Give the shape of each of the following graphs from the men's health data. Then decide what the best measure of center and spread would be. (Mean/standard deviation or median/IQR?)

1. Men's Pulse Rate in Beats per Minute (BPM)

Shape = \_\_\_\_\_ Mean/Stand Dev **OR** Median/IQR? \_\_\_\_\_

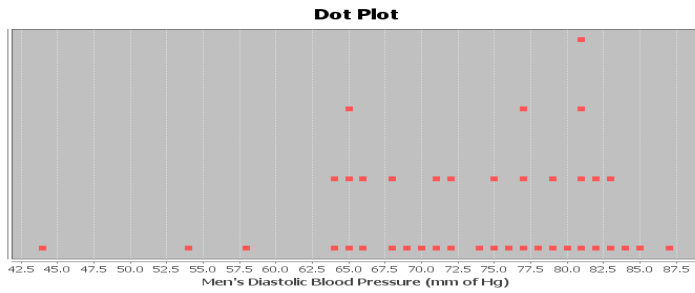


2. Men's Diastolic Blood Pressure in Millimeters of Mercury (mm of Hg)



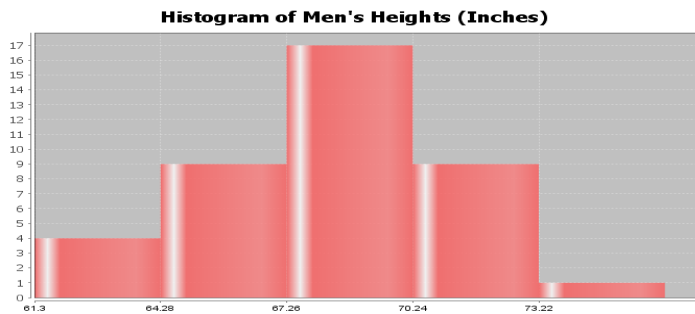
This material is from *Introduction to Data Analysis*, 2<sup>nd</sup> edition by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a “CC-BY” [Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by/4.0/) – 3/17/2021

Shape = \_\_\_\_\_ Mean/Stand Dev **OR** Median/IQR? \_\_\_\_\_



3. Men's Heights (inches)

Shape = \_\_\_\_\_ Mean/Stand Dev **OR** Median/IQR? \_\_\_\_\_



4. Calculate the Median, Q1, Q3 and IQR for the following data. work and put your answers in the spaces below.

The 16 numbers are already in order. Show

17 , 19 , 20 , 26 , 28 , 31 , 35 , 37 , 41 , 43 , 44 , 48 , 51 , 53 , 55 , 62

Median Average = \_\_\_\_\_

Q1 = \_\_\_\_\_

Q3 = \_\_\_\_\_

IQR = Q3-Q1 = \_\_\_\_\_

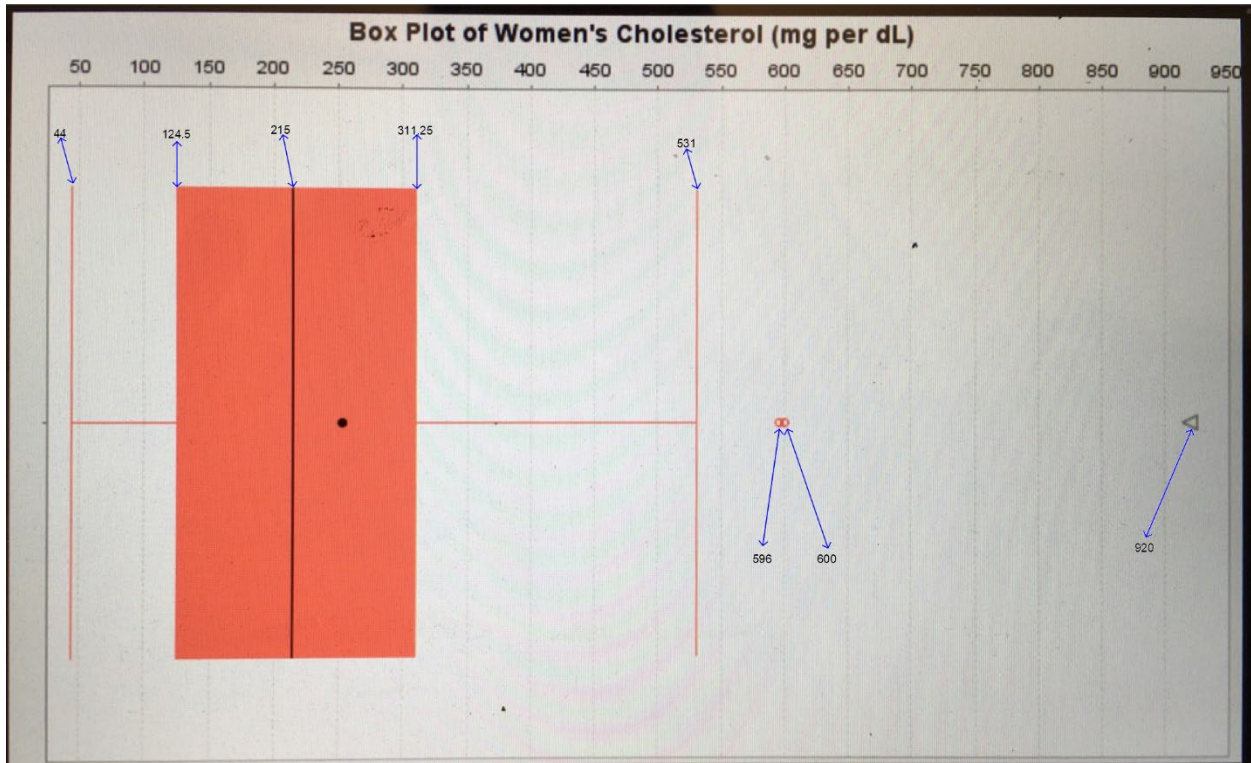
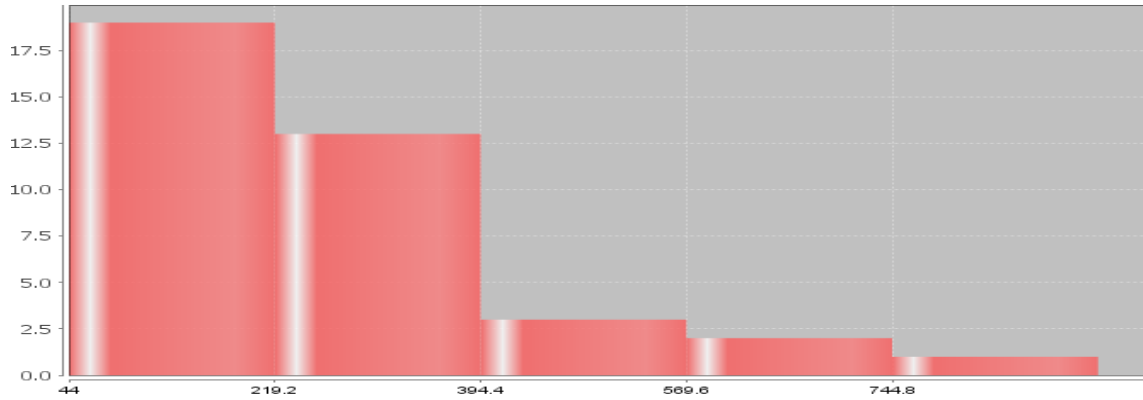
5. Interquartile Range (IQR) is an important measure of spread or variability in statistics. Give the basic definition of IQR.

6. How can we tell if we should use the median and IQR as our center and spread?



Look at the following Histogram, Box Plot and summary statistics of the women's cholesterol data and answer the following questions.

**Women's Cholesterol in mg per dL**



	Q1	Median	Q3	IQR	Min	Max	N total
Women Cholesterol (mg per dL)	124.5	215.0	311.25	186.75	44.0	920.0	38

7. What is this data measuring? \_\_\_\_\_

8. What are the units for the data set? \_\_\_\_\_



This material is from *Introduction to Data Analysis*, 2<sup>nd</sup> edition by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a "CC-BY" [Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by/4.0/) – 3/17/2021

9. What is the shape of the data set? \_\_\_\_\_
10. How many numbers are in the data set? \_\_\_\_\_
11. Are the median and IQR accurate for this data? (Yes or No) \_\_\_\_\_
12. What is the average cholesterol for these women? (Give a number) (No calculation needed)  
Average Cholesterol = \_\_\_\_\_
13. How far are typical values in the data set from each other? (Give a number) (No calculation needed)  
Average distance typical values are from each other = \_\_\_\_\_
14. Find two numbers that typical values fall in between and put your answer below. (No calculation needed)  
\_\_\_\_\_  $\leq$  typical cholesterol for these women  $\leq$  \_\_\_\_\_
15. Are there any unusually low values (low outliers) in the data set (yes or no)? \_\_\_\_\_
16. Are there any unusually high values (high outliers) in the data set (yes or no)? \_\_\_\_\_
17. List all unusual values (outliers) in the data set.  
Give the actual numbers, not a cutoff point. \_\_\_\_\_

(For #18-22, refer to the boxplot.)

18. What percent of these women had a cholesterol below 311.25? \_\_\_\_\_
19. What percent of these women had a cholesterol below 124.5? \_\_\_\_\_
20. What percent of these women had a cholesterol higher than 215? \_\_\_\_\_
21. What was the largest value in the data set that was not an  
outlier (not unusual)? \_\_\_\_\_
22. True or False? There were more numbers in the data set greater  
than 311.25 than there were numbers in the data set less than 124.5.
- 



## Chapter 5 Project Skewed Data Analysis

**Online Class Directions:** This will be an individual project. Each student will analyze one quantitative data set from the math 075-survey data fall 2015, create a poster summarizing their findings, and present the poster to other students in the class.

Each student will pick one of the following data sets from the math 075 survey data fall 2015 to analyze: Hours work per week, Hours sleep per night, Hours of exercise per week, Number of Minutes to get to school, College GPA, Number of Units completed at COC, Average cell phone bill per month, Dollars spent on a meal when eat out, Number of times eat at restaurant or fast food per week, Number of U.S. states visited, Number of minutes spent on social media.

### The Individual Poster Should Have

- **First and Last Name of student**
- **Why is this data important or interesting to you?**
- **Go to [www.lock5stat.com](http://www.lock5stat.com) and open StatKey. Click on “one quantitative variable” under the “descriptive statistics and graphs” menu. Click on “edit data” and copy and paste your one column of quantitative data into Statkey. If you data has a title, click on “data has a header row”. Do NOT click the box the says data has identifier. Press OK.**
- **Click on histogram in StatKey, and pull the slider to “3 buckets”. Your histogram should have 3 bars. Sketch the histogram onto your poster.**
- **Click on “Box Plot” in StatKey and sketch the box plot onto your poster.**
- **Write down the Median, 1<sup>st</sup> Quartile, 3<sup>rd</sup> Quartile, SMin, Max and Sample Size onto your poster. You will see them under the “Sample Statistics” menu in StatKey.**
- **Calculate the Interquartile Range  $IQR = Q3 - Q1$**
- **What is the data measuring?**
- **What are the units?**
- **How many numbers are in the data set : sample size (n)**
- **What is the Shape? Look at your histogram.**
- **Write a sentence to explain the median.**
- **What is the average? (Use the median if data is skewed.)**
- **What is your spread for the data? (Use the Interquartile Range if data is skewed.)**
- **Write a sentence to explain the interquartile range.**
- **Find two numbers that typical values fall in between (Q1 and Q3)**
- **Calculate Unusually high cutoff:  $Q3 + (1.5 \times IQR)$**
- **List all unusually high values (high outliers) in the data set. Find these on the box plot. If there are none, say “No high outliers”.**
- **Calculate Unusually low cutoff:  $Q1 - (1.5 \times IQR)$**
- **List all unusually high values (high outliers) in the data set. Find these on the box plot. If there are none, say “No high outliers”.**
- **Estimate the largest value in the data set that is not an outlier. Look at the right whisker on the box plot. Does not have to be exact.**
- **Estimate the smallest value in the data set that is not an outlier. Look at the left whisker on the box plot. Does not have to be exact.**
- **Decorate Poster**

Now take a picture of your poster project and submit the picture to your instructor in Canvas.

After submitting the picture of the poster, go to the discussion menu in Canvas and complete the “Chapter 5 Project Discussion”. You will be discussing your findings with other students in the class.



**Face to face Class Directions:** *The class will be separated into groups. Each group is required to pick a “team name” for their group and analyze one skewed quantitative data set from the math 075-survey data fall 2015, create a poster summarizing their findings, and present the poster to other students in the class.*

*Each group will have a different topic and will pick one of the following data sets from the math 075 survey data fall 2015 to present it to their classmates: Hours work per week, Hours sleep per night, Hours of exercise per week, Number of Minutes to get to school, College GPA, Number of Units completed at COC, Average cell phone bill per month, Dollars spent on a meal when eat out, Number of times eat at restaurant or fast food per week, Number of U.S. states visited, Number of minutes spent on social media.*

#### The Individual Poster Should Have

- **First and Last Name of student**
- **Why is this data important or interesting to you?**
- **Go to [www.lock5stat.com](http://www.lock5stat.com) and open StatKey. Click on “one quantitative variable” under the “descriptive statistics and graphs” menu. Click on “edit data” and copy and paste your one column of quantitative data into StatKey. If you data has a title, click on “data has a header row”. Do NOT click the box the says data has identifier. Press OK.**
- **Click on histogram in StatKey, and pull the slider to “3 buckets”. Your histogram should have 3 bars. Sketch the histogram onto your poster.**
- **Click on “Box Plot” in StatKey and sketch the box plot onto your poster.**
- **Write down the Median, 1<sup>st</sup> Quartile, 3<sup>rd</sup> Quartile, SMin, Max and Sample Size onto your poster. You will see them under the “Sample Statistics” menu in StatKey.**
- **Calculate the Interquartile Range  $IQR = Q3 - Q1$**
- **What is the data measuring?**
- **What are the units?**
- **How many numbers are in the data set : sample size (n)**
- **What is the Shape? Look at your histogram.**
- **Write a sentence to explain the median.**
- **What is the average? (Use the median if data is skewed.)**
- **What is your spread for the data? (Use the Interquartile Range if data is skewed.)**
- **Write a sentence to explain the interquartile range.**
- **Find two numbers that typical values fall in between (Q1 and Q3)**
- **Calculate Unusually high cutoff:  $Q3 + (1.5 \times IQR)$**
- **List all unusually high values (high outliers) in the data set. Find these on the box plot. If there are none, say “No high outliers”.**
- **Calculate Unusually low cutoff:  $Q1 - (1.5 \times IQR)$**
- **List all unusually high values (high outliers) in the data set. Find these on the box plot. If there are none, say “No high outliers”.**
- **Estimate the largest value in the data set that is not an outlier. Look at the right whisker on the box plot. Does not have to be exact.**
- **Estimate the smallest value in the data set that is not an outlier. Look at the left whisker on the box plot. Does not have to be exact.**
- **Decorate Poster**

#### Presentation

*Make sure each person on the team understands the poster and can present your findings. Bring your poster to a designated presentation area in the classroom and hang or tape your poster to a wall. One person at a time will present the poster. We will then rotate so that each member of the team gets to present. Everyone else will listen to presentations and give feedback.*

---





## Chapter 6 – Linear Quantitative Relationships

**Introduction:** In previous sections, we have looked at how to analyze a categorical data set. Then we looked at relationships between categorical data sets. In chapters 4 and 5, we looked at how to analyze quantitative data. It follows that now we are ready to look at relationships between quantitative data sets.

There are many different types of quantitative relationships that statisticians study. We will focus on the most common in this chapter, which is the study of linear relationships between quantitative variables.

**Algebra requirement:** Algebra classes study the subject of lines. However, they do not study lines the way statisticians and data scientists study lines. As with most things, statistics is the study of world around us with real data and real applications. For example, algebra classes may calculate slope between two points. A statistician studies the slope as an average rate of change between thousands or even millions of points based on real data. The slope calculation is much more complicated. An algebra class may find the equation of a line between two points. A statistician studies linear prediction formulas created from thousands of points, uses those formulas to predict world climate changes, and studies the accuracy of those predictions with residual analysis. While a basic understanding of slope and lines is helpful, realize that the calculations will be much more complicated. The study of linear quantitative relationships in statistics is extremely different from algebra.

**Technology:** Many of the calculations in this chapter are extremely difficult to calculate by hand with a calculator. As with many statistics and graphs, we prefer to use a computer software like StatKey to calculate. Then we can focus on understanding the meaning behind these statistics and graphs and what they tell us about the world around us.

### Terminology

**Correlation:** Statistical analysis that determines if there is a relationship between two different quantitative variables.

**Regression:** Statistical analysis that involves finding the line or model that best fits a quantitative relationship, using the model to make predictions, and analyzing error in those predictions.

**Scatter Plot:** A graph that shows the x-axis, y-axis and points at all of the ordered pairs in the data.

**Slope ( $b_1$ ):** The average amount of increase or decrease in the y-variable for every one-unit increase in the x-variable.

**Y-Intercept ( $b_0$ ):** The predicted y-value when the x-value is zero.

**Regression Line ( $\hat{y} = b_0 + b_1x$ ):** Also called the “Line of Least Squares” or the “Line of Best Fit”. This line will be represented by a linear equation, but realize it is not a line between two points. It is the line that best fits many points.

---



## Section 6A – Introduction to Quantitative Relationships, Explanatory and Response Variables, Scatterplots with Technology

Remember quantitative data is numerical measurement data, not categories. The numbers in the data set should measure something. They often have units and we should be able to take an average in context.

In this section, we will be focusing on two different quantitative variables with different units. It is much easier to compare the average salary in thousands of dollars from people in Arizona to the average salary in thousands of dollars from people in New Mexico. The two data sets have the same units and can be compared directly. For example, we can determine if the average salary of the people from Arizona is higher or lower than the average salary from the people from New Mexico.

When the units are different, you cannot just compare the centers or spreads. It becomes a much more complicated process. If you look at countries around the world and study the relationship between their unemployment rates and their national debts in millions of dollars, you cannot compare the national debt in millions of dollars to the unemployment rate percentage directly. They are completely different things.

So how do we analyze the relationships between two different quantitative variables? We will start by assigning one variable to be the explanatory variable and one variable to be the response variable.

### Explanatory and Response Variables

In algebra classes, we are often given an X and a Y variable and asked to plot a couple points. In statistics, we know it is not so simple. In statistics, we often call the X variable the “explanatory variable” or “independent variable”. We call the Y variable the “response variable” or “dependent variable”. I prefer explanatory and response because the terms independent and dependent can be confusing to students when they study the subject of independence. Real quantitative relationship analysis requires some serious thought about which variable should be the explanatory variable (X) and which variable should be the response variable (Y).

### Guidelines for choosing the explanatory (X) and the response (Y)

1. The response variable should respond.

Often business analysis involves studying the costs or profits of company over a period of several months or years. Should we assign the costs to be the explanatory variable (X) or the response variable (Y)? What about the time in months? Think of it this way. Does time respond to the costs of the company? Probably not. That does not sound right. Do the costs respond to time? That may be true. Whichever variable responds to the other should probably be your response variable. In this case, I should assign the time (months) as my explanatory variable (X) and the costs (thousands of dollars) as my response variable (Y).

2. The response variable should be the focus of your study or the variable you may want to make predictions about.

Let us look at the example of the unemployment rates and national debts of various countries. Those variables may relate to each other. In other words either variable could be the responses variable. In that case, pick the variable you are most interested in to be the response variable (Y). I was studying unemployment rates in various countries and wanted to see if the national debt was related to unemployment. I was also interested in trying to predict unemployment rates with my prediction equation. Since the focus of my study was unemployment and I wanted to eventually make predictions about unemployment, I let the unemployment rates be my response variable (Y). Therefore, my explanatory variable (X) will be the national debts.

### Ordered Pairs

Once you have chosen which variable is X and which variable is Y, you will need to find ordered pair data. Ordered pair data pairs an X in the first data set with a Y value in the second data set. There needs to be some kind of relationship between them. For example, I do not want to pair 20 random unemployment rates with 50 national debts. First there needs to be the same number of X and Y values. Computer programs will give error messages if the frequency N for one quantitative variable is not the same as the frequency N for the other data set. If I want to study the relationship between national debt and unemployment rates, I do not want to pair any national debt with any unemployment rate. I want to collect the data together. The national debt and unemployment rate should come from the same country and hopefully the same year. I went from country to country and looked up their estimated national



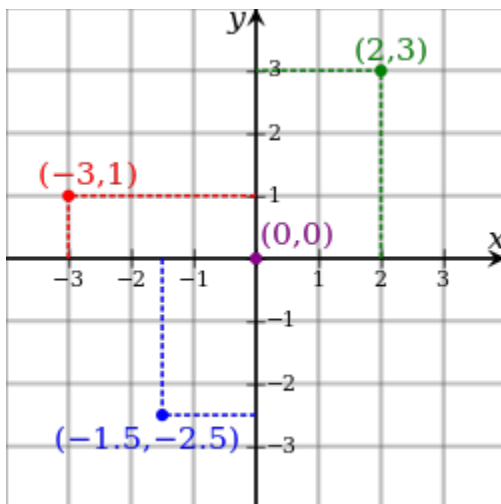
debt and their unemployment rate at the same time (August 2017). This is difficult to do. Getting data is difficult job. Websites, articles and various sources often disagree with one another. Therefore, these values are just approximations and may not be perfectly accurate.

Country	August 2017 National Debt (Billions of U.S. Dollars)	Unemployment Rate (%)
France	2472.9	9.6
Mexico	474.5	3.5
U.S.A	19873.5	
Japan	9094.9	3.06
Canada	826.5	
Australia	406.0	
United Kingdom	2279.8	

You can now write an ordered pair. They are often written in the form  $(X, Y)$ . So for Mexico's data I would write the ordered pair as  $(474.5 \text{ Billion } \$, 3.5\% \text{ unemployment})$  and Japan's data as  $(9094.9 \text{ Billion } \$, 3.06\% \text{ unemployment})$ . Notice the first number in the pair describes the explanatory variable (X) and is often called the "X coordinate". The second number in the ordered pair describes the response variable (Y) and is often called the "Y coordinate".

### Rectangular Coordinate System

Once you have your ordered pairs, you can graph them. To graph quantitative variables with different units, we will need both an X-axis and a Y-axis.



Notice to graph the point  $(2, 3)$  we find 2 on the X-axis and 3 on the Y-axis and then the point would be where they meet. Notice that to make the ordered pair, a rectangle is created. That is why this system of graphing with X and Y-axes is often called the "rectangular coordinate system".



The key is the units though. Always pay close attention to the units for your explanatory and response variables. For example, the x-axis could be describing temperature in degrees Fahrenheit and the y-axis could be describing profits in thousands of dollars. So ( 2 , 3 ) is really describing the ordered pair (2 degrees Fahrenheit , 3 thousand dollars) and ( -3 , 1) is really describing the ordered pair (-3 degrees Fahrenheit , 1 thousand dollars).

### Scatterplots

There are many types of graphs statisticians look at when studying relationships between quantitative variables with different units. The most important graph though is the scatterplot. We said in the last couple of chapters that the first step when analyzing quantitative data is to find the shape. The scatterplot is a graph of the ordered pairs on the rectangular coordinate system. This graph shows the shape of the quantitative relationship.

We should again use technology to create a scatterplot. Once you have collected your ordered pair data and chosen which column will be the explanatory (X) and which will be the response (Y), you can create a scatterplot with any statistics software.

#### How to create a Scatterplot with StatKey:

- Open the data. Then open a new spreadsheet and paste the two quantitative data sets next to each other side by side. It is customary to have the explanatory column (X) on the left and the response column (Y) on the right. Then copy the two columns together.
- Now we will go to [www.lock5stat.com](http://www.lock5stat.com) and click on “StatKey”. Under the “Descriptive Statistics and Graphs” menu, click on “Two Quantitative Variables”. Click on “Edit Data” at the top. Push Control A on your keyboard to highlight old data and then push “delete” on your keyboard to delete all old data in the edit data field. Then paste the two columns of quantitative data into the “Edit Data” field. If your data has a title, click the box that says “Data has header row”. If your data does not have a title, do NOT check the box that says “Data has header row”. Then press OK. The graph you see is the scatterplot.

**Note:** Your X variable should be on the horizontal axis and the Y variable should be on the vertical axis. If the X and Y variables are backward in the graph, simply click the “switch variables” button. It is also nice to check the “show regression line” box. The regression line is the line that best fits the points in the scatterplot. StatKey will also give us some statistics to help understand the relationship. These statistics we will explore in future sections.

**Note:** The titles of the quantitative columns of data can be problematic sometimes if they are too long. If StatKey gives an error message, it may be a problem with the title. If that is the case, you may need to either shorten the title or completely delete the title.

#### Example 1

Let us look at the health data again. Statistics analysis always starts with a question, even if it is a question in your own mind. My first question was to see if there is a relationship between the weight of a man and his cholesterol.

Step 1: First notice that the health data does have ordered pair data containing the weight and cholesterol of the same 40 men. Having ordered pair data is vital to studying quantitative relationships. I cannot take a weight of one man and pair it with the cholesterol of a different man. The weight and cholesterol need to come from the same man. We opened the health data and found the columns for men’s weight and men’s cholesterol.

AD	AE	AF	AG	AH	AI	AJ	AK
Men Age (years)	Men Ht (in)	Men Wt (Lbs)	Men Waist (cm)	Men Pulse (BPM)	Men Syst BP (mm of Hg)	Men Diast BP (mm of Hg)	Men Cholesterol (mg per deciliter)
58	70.8	169.1	90.6	68	125	78	522
22	66.2	144.2	78.1	64	107	54	127
32	71.7	179.3	96.5	88	126	81	740
31	68.7	175.8	87.7	72	110	68	49
28	67.6	152.6	87.1	64	110	66	230
46	69.2	166.8	92.4	72	107	83	316



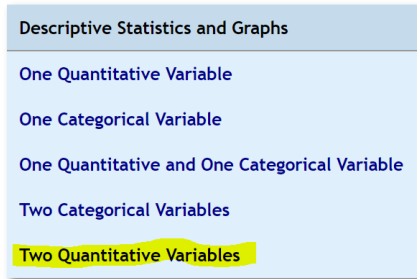
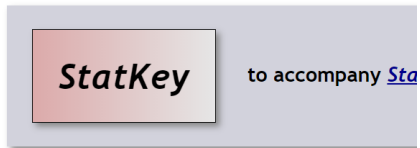
Step 2: The next step is to choose which variable is to be the explanatory (X) and which variable should be the response variable (Y). The variable you want to predict should be the response variable (Y). I want to maybe predict cholesterol levels based on a man's weight, so I will make cholesterol my response (Y). Therefore, I will make the weight the explanatory variable (X).

Step 3: Now we need to open a new excel spread sheet and copy and paste the men's weight and men's cholesterol data into two columns next to each other. It is common to put the X variable on the left and the Y variable on the right.

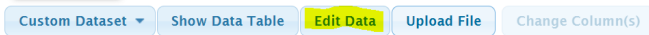
	A	B
1	Men Wt (Lbs)	Men Cholesterol (mg per deciliter)
2	169.1	522
3	144.2	127
4	179.3	740
5	175.8	49
6	152.6	230
7	166.8	316
8	135	590
9	201.5	466
10	175.2	121
11	139	578
12	156.3	78
13	186.6	265
14	191.1	250
15	151.3	265
16	209.4	273
17	237.1	272
18	176.7	972
19	220.6	75
20	166.1	138
21	137.4	139
22	164.2	638
23	162.4	613
24	151.8	762
25	144.1	303
26	204.6	690
27	193.8	31
28	172.9	189
29	161.9	957
30	174.8	339
31	169.8	416
32	213.3	120
33	198	702
34	173.3	1252
35	214.5	288
36	137.1	176
37	119.5	277
38	189.1	649
39	164.7	113
40	170.1	656
41	151	172



Step 4: Now we will go to [www.lock5stat.com](http://www.lock5stat.com) and click on “StatKey”. Under the “Descriptive Statistics and Graphs” menu, click on “Two Quantitative Variables”. Click on “Edit Data” at the top. Push Control A on your keyboard and delete all old data in the edit data field. Then paste the two columns of quantitative data into the “Edit Data” field. Since these columns of data have titles, we will click the box that says “Data has header row”. Then press OK.



**StatKey** Descriptive Statistics for Two Quantitative Variables



Edit data

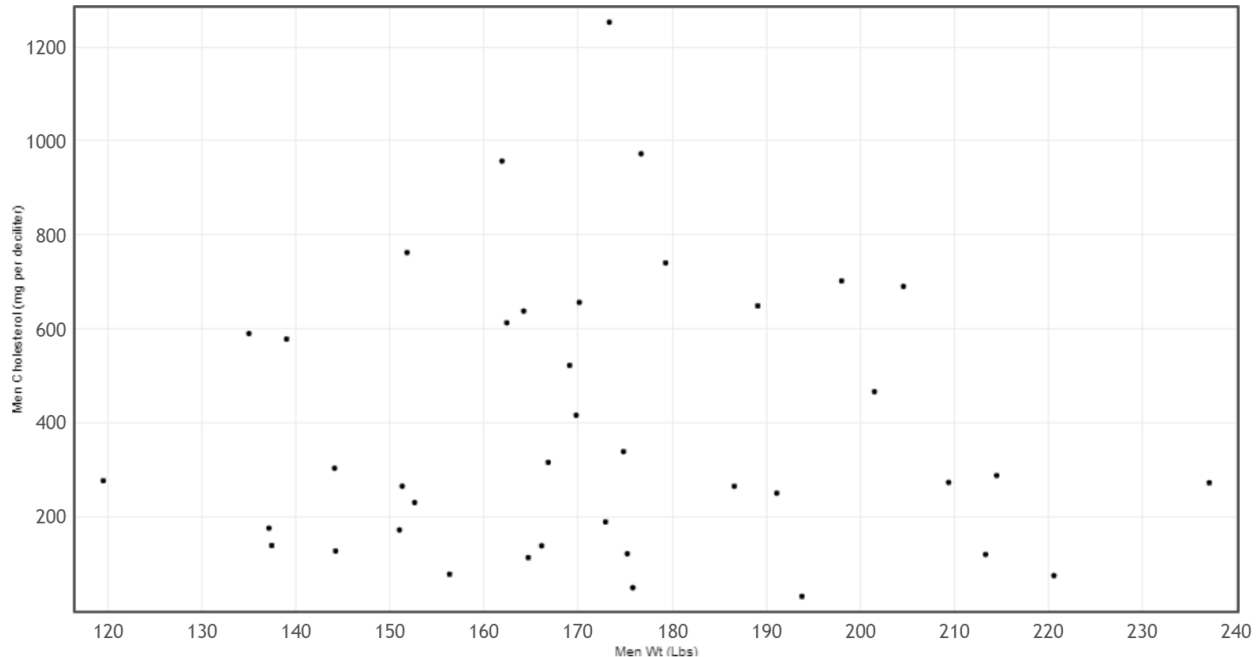
Men Wt (Lbs)	Men Cholesterol (mg per deciliter)
169.1	522
144.2	127
179.3	740
175.8	49
152.6	230
166.8	316
135	590
201.5	466
175.2	121
139	578
156.3	78
186.6	265
191.1	250
151.3	265
209.4	273
237.1	272
176.7	972
220.6	75
166.1	138

Data has header row

Manually edit the values above or paste a tab or comma separated file into the box and click Ok. The file must have only two columns

Ok





### Shape of a scatterplot

When looking at the shape of a scatterplot, you have to forget about what you learned in Algebra classes. In algebra classes, we often start with a linear or curved function and find ordered pairs that lie on that line or curve. That is not how real data works in statistics. The dots will rarely go through a line or a curve. In some ways, statistics analysis is the opposite of algebra. Instead of focusing on a linear and curved function and then finding ordered pairs, in statistics, we start with a graph of all the real data ordered pairs and then find the line or curve that best fits all of the data. This can be a difficult process.

**Key Question:** Start by asking yourself a simple question. The points will not lie on a line or a curve, but can I imagine a line or a curve that the points could be relatively close to? That is the key. You have to take all of the points into account.

In the men's weight / men's cholesterol example, the points seem very scattered all over with no obvious pattern. Do not force a computer program like StatKey to draw a best-fit line or curve if there is no pattern in the data. The line or curve the computer draws will not be very accurate, since the points will not follow any particular pattern. This scatterplot tells that there is hardly any relationship between the weight of these men and the cholesterol of these men. Sometimes lighter men had a high cholesterol. Sometimes lighter men had a low cholesterol. Sometimes heavier men had a high cholesterol. Sometimes heavier men had a low cholesterol. Sometimes we refer to this as "no correlation", "no relationship" or "no association".

### Example 2

Let us look at another example from the health data. This time I wanted to look at the weight of the men (in pounds) and the waist size of the men (in centimeters).

Step 1: Do I have ordered pair data? Yes. The health data contained the weights and waist sizes of the same 40 men.

Step 2: Pick which variable is the explanatory (X) and the response (Y). I was interested in predicting the weight of a man from his waist size. Remember the variable you want to predict and are most interested in should be your response (Y). So I picked the men's waist size (in cm) to be the explanatory variable (X) and the men's weight (in pounds) to be the response variable (Y).



A	B
Men Waist (cm)	Men Wt (Lbs)
90.6	169.1
78.1	144.2
96.5	179.3
87.7	175.8
87.1	152.6
92.4	166.8
78.8	135
103.3	201.5
89.1	175.2
82.5	139
86.7	156.3
103.3	186.6
91.8	191.1
75.6	151.3
105.5	209.4
108.7	237.1
104	176.7
103	220.6
91.3	166.1
75.2	137.4
87.7	164.2
77	162.4
85	151.8
79.6	144.1
103.8	204.6
103	193.8
97.1	172.9
86.9	161.9
88	174.8
91.5	169.8
102.9	213.3
93.1	198
98.9	173.3
107.5	214.5
81.6	137.1
75.7	119.5
95	189.1
91.1	164.7
94.9	170.1
79.9	151



This material is from *Introduction to Data Analysis*, 2<sup>nd</sup> edition by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a “CC-BY” [Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by/4.0/) – 3/17/2021



Step 3: Make a scatterplot with technology. We will use StatKey and follow the same steps as in example 1.

Edit data ✕

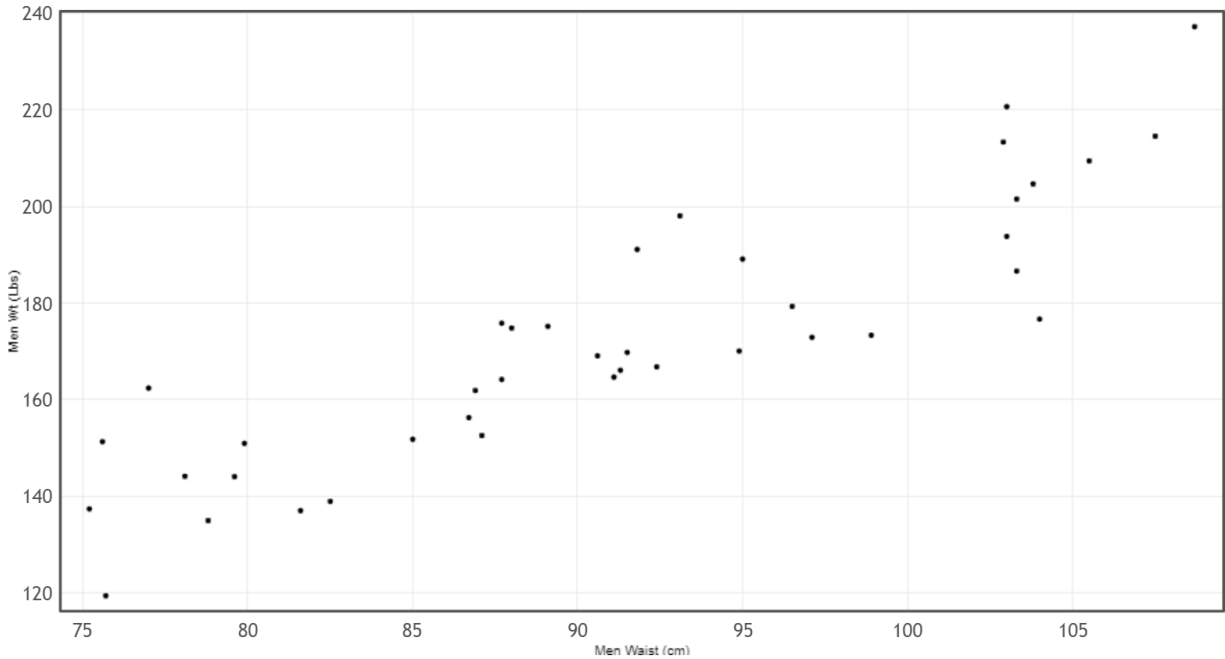
Men Waist (cm)	Men Wt (Lbs)
90.6	169.1
78.1	144.2
96.5	179.3
87.7	175.8
87.1	152.6
92.4	166.8
78.8	135
103.3	201.5
89.1	175.2
82.5	139
86.7	156.3
103.3	186.6
91.8	191.1
75.6	151.3
105.5	209.4
108.7	237.1
104	176.7
103	220.6
91.3	166.1

Data has header row

Manually edit the values above or paste a tab or comma separated file into the box and click Ok. The file must have only two columns

Ok

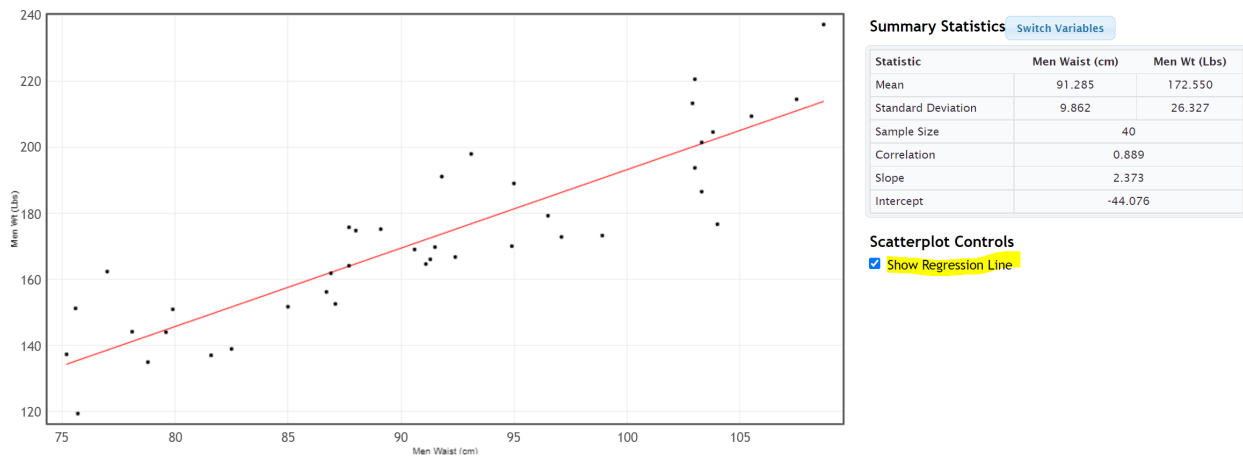




Step 4: Interpret the shape of the scatterplot.

Remember the points will not lie on a line or a curve. The key question is are they close to a line or a curve?

This scatterplot shows a distinct linear pattern. The points look like they are close to a line going up from left to right. This is often called a “positive linear relationship” or a “positive correlation”. In fact we can have the computer find the line of best fit. This is called the “regression line”. By clicking on “Show Regression Line” we see the line of best fit drawn. Notice the points do not go through the line, but they are close to the line. Also notice the line goes up from left to right.

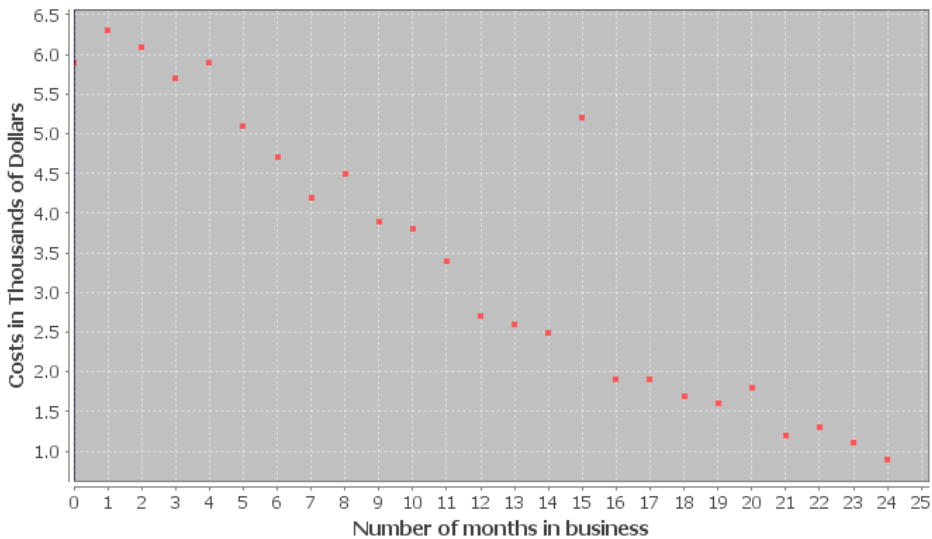


### Example 3

To be run well, businesses often require a large amount of statistical analysis. Here is a scatterplot made from data describing the monthly costs of running a company. The data describes the month (X) and the costs (Y) in thousands of dollars. Notice month 0 is the initial cost of starting up the company.



### Scatterplot of number of months and costs of company



Notice most of the points do seem to be close to a line. They are following a linear pattern. In fact, the linear pattern seems to be going down from left to right. We often call this a “negative linear relationship” or a “negative correlation”.

**Unusual Value (Outlier):** Notice there appears to be a point that does not follow the pattern. The company had an unusually high cost in the 15<sup>th</sup> month of operation. You may want to check with the company to determine if this was a mistake in the data. Maybe the cost was supposed to be 2.5 thousand dollars instead of 5.2 thousand dollars. However, this outlier was not a mistake. The company had some equipment break down and had to replace it. When studying quantitative relationships with scatterplots, it is important to look for these unusual values (outliers). Notice this point does not seem to follow the negative linear pattern.

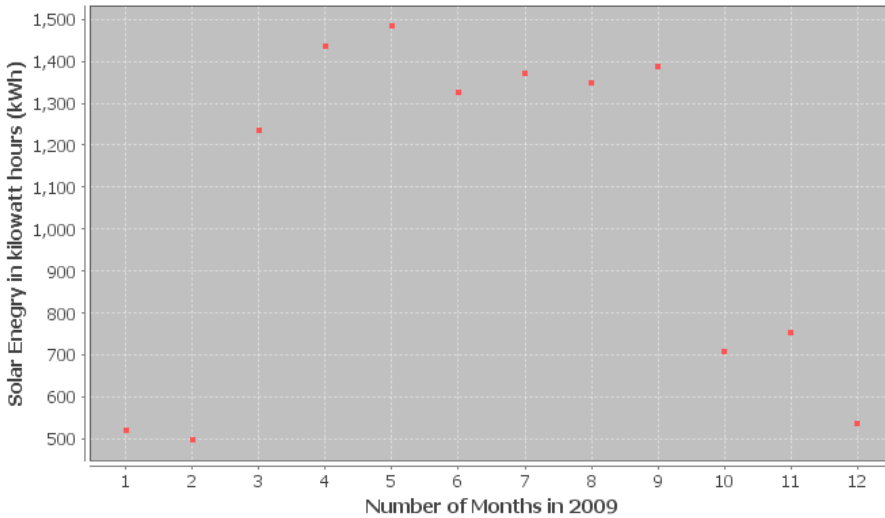
Aside from the outlier, this graph is good news for the company. As the months increase, the costs of the company seem to be decreasing dramatically.

#### Example 4

A college started a solar energy program. Here is a scatterplot describing their data in 2009. The explanatory variable (X) was the # months in 2009 and the response variable was the solar energy generated from that month in kilowatt-hours (kWh).

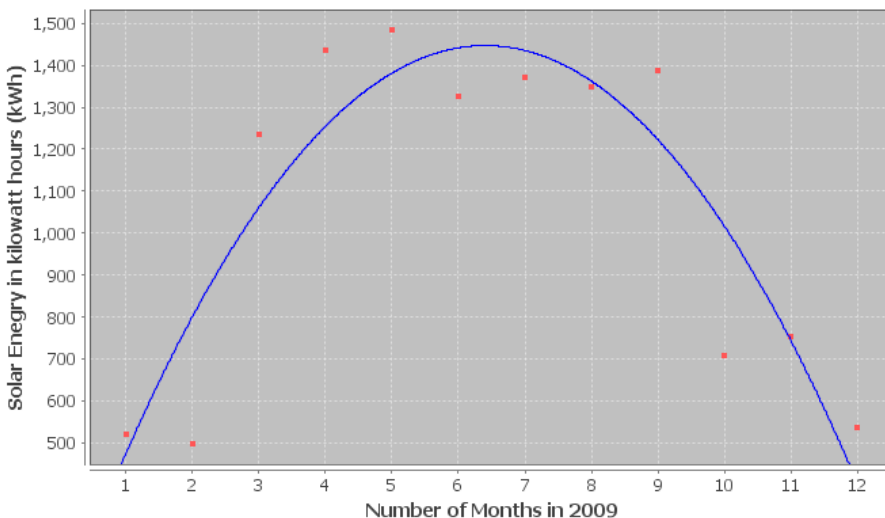


### Scatterplot of number of months and solar energy



Notice the points in the scatterplot do not seem to be close to a line, but there seems to be some relationship. It seems to follow a curve of some kind. You would have to use your imagination some, but can you draw a curve that might fit this data? Here is a possible curve.

### Scatterplot of number of months and solar energy



Notice the curve seems to fit this data pretty well. The points are pretty close to the curve. You may remember from previous algebra classes that this “U” shaped curve is called a “parabola” or a “quadratic curve”. So this scatterplot indicates that there is no linear relationship, but there is a quadratic relationship between time in months and solar energy.

#### Scatterplots and Shapes

We have seen in this section that to analyze two quantitative data sets with different units, we have to find ordered pair data and chose one variable to be the explanatory variable (X) and the other variable to be the response variable (Y). We can then create a scatterplot to see if there is a relationship between the variables. We saw various possibilities for these quantitative relationships.

- **No relationship at all:** Points in scatterplot are spread out all over and do not seem to be close to any line or curve.



- **Positive Correlation** (Positive Linear Relationship): Points in the scatterplot seem to be close to a line that is going up from left to right (increasing). This is sometimes called a “direct relationship” in mathematics. As the X variable increases, the Y variable also increases. Look out for points that do not seem to fit the linear pattern (outliers).
- **Negative Correlation** (Negative Linear Relationship): Points in the scatterplot seem to be close to a line that is going down from left to right (decreasing). This is sometimes called an “indirect relationship” or “inverse relationship” in mathematics. As the X variable increases, the Y variable decreases. Look out for points that do not seem to fit the linear pattern (outliers).
- **Curved Relationship** (Non-linear Relationship): Points in the scatterplot do not seem to follow any line, but do seem to be close to a curve. There are many different types of curves possible when looking at data. Look out for points that do not seem to fit the curve pattern (outliers).

Note: The use of the word “correlation” denotes a linear relationship between two quantitative variables. If you have a relationship between categorical variables or between a categorical and quantitative variable, we usually refer to that as an “association” or a “relationship”.

Note: Correlation does not imply causation. It is wrong to state that one variable causes another just because they are related.

---

## Problem Set Section 6A

Directions: Open the health data from Canvas or from [www.matt-teachout.org](http://www.matt-teachout.org). Use the indicated columns of data to create scatterplots with StatKey. Save the scatterplot on a word document or make a general sketch of the graph on a sheet of paper and answer the questions.

How to create a Scatterplot with StatKey:

- Open the data. Then open a new spreadsheet and paste the two quantitative data sets next to each other side by side. It is customary to have the explanatory column (X) on the left and the response column (Y) on the right. Then copy the two columns together.
- Now we will go to [www.lock5stat.com](http://www.lock5stat.com) and click on “StatKey”. Under the “Descriptive Statistics and Graphs” menu, click on “Two Quantitative Variables”. Click on “Edit Data” at the top. Push Control A on your keyboard to highlight old data and then push “delete” on your keyboard to delete all old data in the edit data field. Then paste the two columns of quantitative data into the “Edit Data” field. If your data has a title, click the box that says “Data has header row”. If your data does not have a title, do NOT check the box that says “Data has header row”. Then press OK. The graph you see is the scatterplot. If your scatterplot has the X and Y variables backward, simply click the “Switch Variables” button in StatKey.

1. Explore the relationship between a woman’s weight and height.

- Which variable did you chose to be the explanatory variable?
- Which variable did you chose to be the response variable?
- Create a Scatterplot with Statcato. Label the x and y axes and give the graph a title. Save it on a word document or make a rough sketch of it on a piece of paper.
- Look at the scatterplot. Does it look like the variables have a linear pattern, curved pattern, or no relationship at all?
- Are there any outliers that do not seem to fit the pattern? Hold your cursor over the point in StatKey and estimate the x and y coordinate for the outliers.

2. Explore the relationship between a man’s weight and height.

- Which variable did you chose to be the explanatory variable?



- b) Which variable did you chose to be the response variable?
  - c) Create a Scatterplot with Statcato. Label the x and y axes and give the graph a title. Save it on a word document or make a rough sketch of it on a piece of paper.
  - d) Look at the scatterplot. Does it look like the variables have a linear pattern, curved pattern, or no relationship at all?
  - e) Are there any outliers that do not seem to fit the pattern? Hold your cursor over the point in StatKey and estimate the x and y coordinate for the outliers.
3. Explore the relationship between a woman's cholesterol and age.
- a) Which variable did you chose to be the explanatory variable?
  - b) Which variable did you chose to be the response variable?
  - c) Create a Scatterplot with Statcato. Label the x and y axes and give the graph a title. Save it on a word document or make a rough sketch of it on a piece of paper.
  - d) Look at the scatterplot. Does it look like the variables have a linear pattern, curved pattern, or no relationship at all?
  - e) Are there any outliers that do not seem to fit the pattern? Hold your cursor over the point in StatKey and estimate the x and y coordinate for the outliers.
4. Explore the relationship between a man's cholesterol and age.
- a) Which variable did you chose to be the explanatory variable?
  - b) Which variable did you chose to be the response variable?
  - c) Create a Scatterplot with Statcato. Label the x and y axes and give the graph a title. Save it on a word document or make a rough sketch of it on a piece of paper.
  - d) Look at the scatterplot. Does it look like the variables have a linear pattern, curved pattern, or no relationship at all?
  - e) Are there any outliers that do not seem to fit the pattern? Hold your cursor over the point in StatKey and estimate the x and y coordinate for the outliers.
5. Explore the relationship between a woman's weight and body mass index (BMI).
- a) Which variable did you chose to be the explanatory variable?
  - b) Which variable did you chose to be the response variable?
  - c) Create a Scatterplot with Statcato. Label the x and y axes and give the graph a title. Save it on a word document or make a rough sketch of it on a piece of paper.
  - d) Look at the scatterplot. Does it look like the variables have a linear pattern, curved pattern, or no relationship at all?
  - e) Are there any outliers that do not seem to fit the pattern? Hold your cursor over the point in StatKey and estimate the x and y coordinate for the outliers.
6. Explore the relationship between a man's weight and body mass index (BMI).
- a) Which variable did you chose to be the explanatory variable?
  - b) Which variable did you chose to be the response variable?
  - c) Create a Scatterplot with Statcato. Label the x and y axes and give the graph a title. Save it on a word document or make a rough sketch of it on a piece of paper.
  - d) Look at the scatterplot. Does it look like the variables have a linear pattern, curved pattern, or no relationship at all?
  - e) Are there any outliers that do not seem to fit the pattern? Hold your cursor over the point in StatKey and estimate the x and y coordinate for the outliers.



7. Explore the relationship between a woman's systolic blood pressure and her diastolic blood pressure.
- Which variable did you chose to be the explanatory variable?
  - Which variable did you chose to be the response variable?
  - Create a Scatterplot with Statcato. Label the x and y axes and give the graph a title. Save it on a word document or make a rough sketch of it on a piece of paper.
  - Look at the scatterplot. Does it look like the variables have a linear pattern, curved pattern, or no relationship at all?
  - Are there any outliers that do not seem to fit the pattern? Hold your cursor over the point in StatKey and estimate the x and y coordinate for the outliers.
8. Explore the relationship between a man's systolic blood pressure and his diastolic blood pressure.
- Which variable did you chose to be the explanatory variable?
  - Which variable did you chose to be the response variable?
  - Create a Scatterplot with Statcato. Label the x and y axes and give the graph a title. Save it on a word document or make a rough sketch of it on a piece of paper.
  - Look at the scatterplot. Does it look like the variables have a linear pattern, curved pattern, or no relationship at all?
  - Are there any outliers that do not seem to fit the pattern? Hold your cursor over the point in StatKey and estimate the x and y coordinate for the outliers.
9. Explore the relationship between the length of a man's leg and the circumference of his wrist.
- Which variable did you chose to be the explanatory variable?
  - Which variable did you chose to be the response variable?
  - Create a Scatterplot with Statcato. Label the x and y axes and give the graph a title. Save it on a word document or make a rough sketch of it on a piece of paper.
  - Look at the scatterplot. Does it look like the variables have a linear pattern, curved pattern, or no relationship at all?
  - Are there any outliers that do not seem to fit the pattern? Hold your cursor over the point in StatKey and estimate the x and y coordinate for the outliers.
10. Explore the relationship between a woman's pulse and her cholesterol level.
- Which variable did you chose to be the explanatory variable?
  - Which variable did you chose to be the response variable?
  - Create a Scatterplot with Statcato. Label the x and y axes and give the graph a title. Save it on a word document or make a rough sketch of it on a piece of paper.
  - Look at the scatterplot. Does it look like the variables have a linear pattern, curved pattern, or no relationship at all?
  - Are there any outliers that do not seem to fit the pattern? Hold your cursor over the point in StatKey and estimate the x and y coordinate for the outliers.
- 



## Section 6B – Strength and Direction of Linear Relationships and the Correlation Coefficient “r”

We may be able to see if a scatterplot has a linear relationship, but it is hard to quantify how much of a linear relationship it has. Sometimes the scale can make it look like the points are not close to a line, when indeed they are. We need a way to measure the linear relationship.

Fortunately, there are ways statisticians measure the strength of a linear relationship. One of the statistics that measures quantitative linear relationships is the correlation coefficient “r”.

**Definition of the correlation coefficient (r):** The correlation coefficient “r” is a number between -1 and +1 that describes the strength and direction of the linear relationship. “r” values can tell us if the linear relationship is strong, moderate or weak, or does not exist. It can tell us if the linear relationship is positive (linear pattern going up from left to right) or negative (linear pattern going down from left to right).

### Interpreting the correlation coefficient “r”

Step 1: Look at a scatterplot.

Always start by looking at a scatterplot. Have an idea of what the scatterplot looks like before you try to find and interpret the correlation coefficient.

Step 2: Calculate “r”

Once you have seen the scatterplot, use a statistics software like StatKey to calculate the correlation coefficient “r”.  
Warning: The correlation coefficient “r” is extremely difficult and time consuming to calculate. No data analyst or statistician calculates “r” with a formula and calculator, especially for big data sets. Always use a computer software program to do the difficult calculation and then focus on being able to interpret and explain the meaning of the correlation coefficient.

### Creating the Correlation Coefficient “r” with StatKey:

- To put the data into StatKey, you will want to open a fresh excel spreadsheet and paste the two quantitative data sets next to each other side by side. Then copy the two columns together.
- Now we will go to [www.lock5stat.com](http://www.lock5stat.com) and click on “StatKey”. Under the “Descriptive Statistics and Graphs” menu, click on “Two Quantitative Variables”. Click on “Edit Data” at the top. Push Control A on your keyboard to highlight old data and then push “delete” on your keyboard to delete all old data in the edit data field. Then paste the two columns of quantitative data into the “Edit Data” field. If your data has a title, click the box that says “Data has header row”. If your data does not have a title, do NOT check the box that says “Data has header row”. Then press OK.
- You will see the correlation coefficient “r” under “Summary Statistics”. Look next to “Correlation”.

### Summary Statistics Switch Variables

Statistic	Men Waist (cm)	Men Wt (Lbs)
Mean	91.285	172.550
Standard Deviation	9.862	26.327
Sample Size	40	
Correlation	0.889 = r	
Slope	2.373	
Intercept	-44.076	



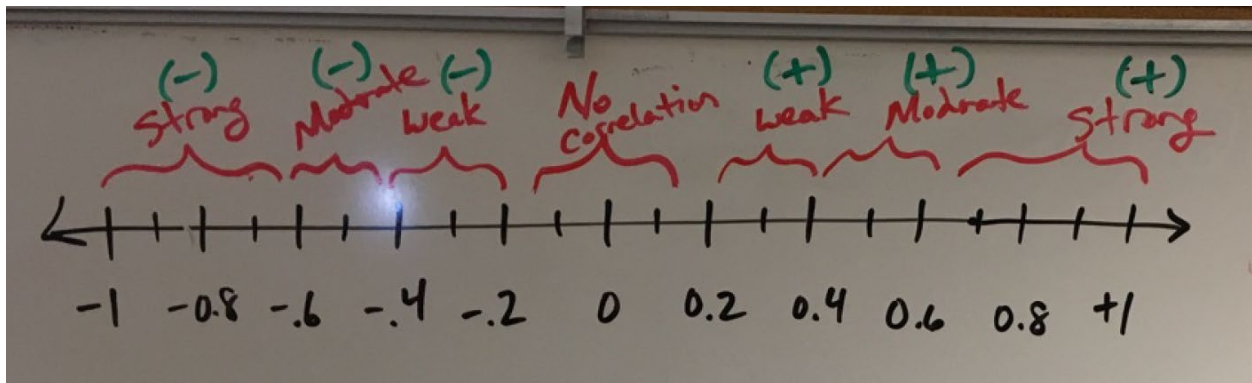


### Interpret what “r” is telling us about the quantitative relationship

Let us see what the “r” value is telling us about the linear relationship. Correlation coefficients are difficult to read, but here are some general guidelines. These are not “set in stone” rules. The number of points in the data can make a difference in the interpretation of the correlation coefficient.

#### Notes about “r”

- r close to +1: This tells us that there is a strong positive correlation. Strong in the sense that the points are close to the regression line and positive means that the regression line is going up from left to right (increasing).
- r close to -1: This tells us that there is a strong negative correlation. Strong in the sense that the points are close to the regression line and negative means that the regression line is going down from left to right (decreasing).
- r close to 0: This tells us that there is no linear relationship between the variables.
- r values in between  $\pm 0.6$  to  $\pm 1.0$  are usually pretty strong. Again, the negative tells us the line is going down from left to right and the positive tells us the line is going up from left to right. The sign does not tell us the strength of the relationship.
- r values in between  $\pm 0.4$  to  $\pm 0.5$  are usually moderate in strength. This means there is a linear relationship, but it is not necessarily strong or weak. It is more in the middle.
- r values in between  $\pm 0.2$  to  $\pm 0.3$  are usually pretty weak. This means there is a linear relationship, but it is very weak.
- r values in between 0 to  $\pm 0.1$  usually tell us there is no linear relationship between the variables. Be careful of the signs when you get an “r” value close to zero. For example, an “r” value of  $-0.044$  does not mean there is a negative linear relationship. Remember the “r” value usually needs to be around  $-0.2$  to even be considered weak.
- The correlation coefficient will be the same if the X and Y are switched. The calculation for “r” does not depend on which variable is X or Y.
- I always find it is helpful to keep the following number line in mind when interpreting a correlation coefficient.



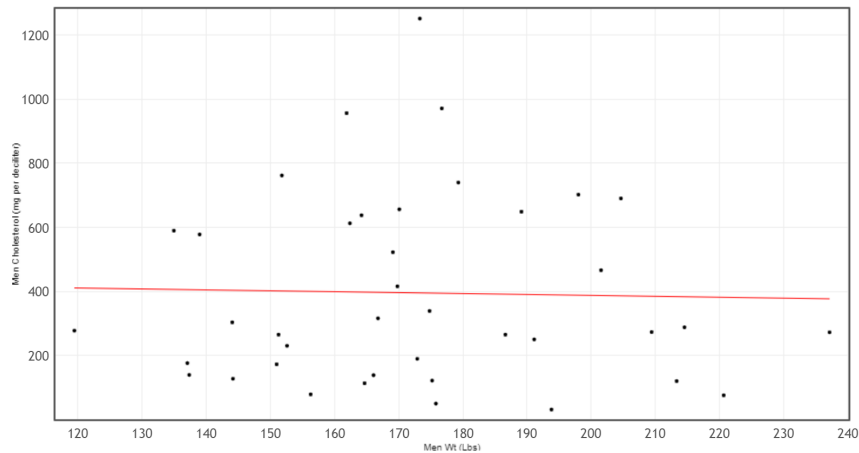
**Note:** The question is often asked, what is the strength and direction of the linear relationship (correlation)? Remember the strength (strong, moderate, weak, none) is asking how close the points are to the line. The direction is asking if the line is going up or down from left to right (increasing or decreasing).

#### Example 1

In the previous section, we looked at men’s weight and cholesterol from the health data. I wanted to see if there is a relationship between the weight of a man and his cholesterol.

Since I was most interested in the cholesterol, I let the cholesterol be the response variable (Y) and the weight be the explanatory variable (X). We then used StatKey to create the following scatterplot and summary statistics.





Summary Statistics [Switch Variables](#)

Statistic	Men Wt (Lbs)	Men Cholesterol (mg per deciliter)
Mean	172.550	395.225
Standard Deviation	26.327	292.412
Sample Size		40
Correlation		-0.026
Slope		-0.288
Intercept		444.879

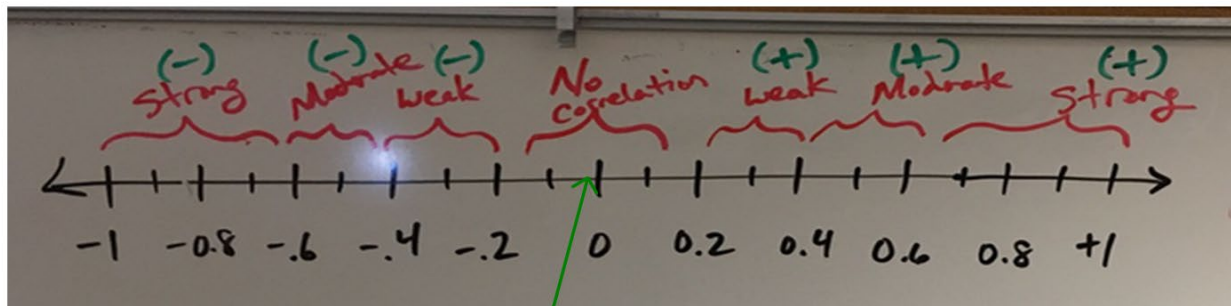
Scatterplot Controls

Show Regression Line

In this scatterplot, the points seem very scattered all over with no obvious pattern. The points do not seem to follow the regression line. Visually we think that there may be no linear relationship. Does the correlation coefficient confirm this suspicion?

You will find the correlation coefficient “r” under “Summary Statistics”. Look for the number next to “Correlation”. We see that the correlation coefficient  $r = -0.026$

Interpretation: So what does this statistic of  $r = -0.026$  tell us? Looking at the number line, we see that the r value of  $-0.026$  though negative is extremely close to zero on the number line. This does not tell us there is negative correlation. This statistics agrees with what we said earlier when we looked at the scatterplot. There seems to be no linear relationship (no correlation) between the weight and cholesterol of these men.



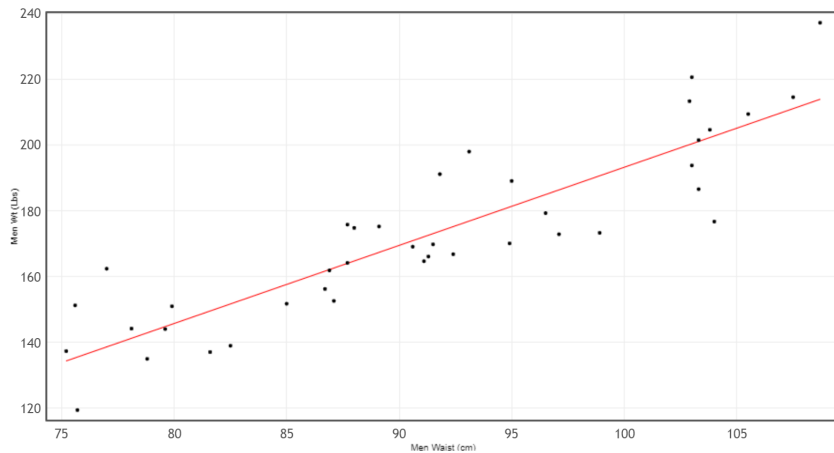
$r = -0.026$

Example 2

In the last section, we also looked at the relationship between the weight of the men (in pounds) and the waist size of the men (in centimeters). I was interested in predicting the weight of a man from his waist size, so I picked the men's waist size (in cm) to be the explanatory variable (X) and the men's weight (in pounds) to be the response variable (Y).

We used StatKey to find the following scatterplot and correlation coefficient “r”.





Summary Statistics [Switch Variables](#)

Statistic	Men Waist (cm)	Men Wt (Lbs)
Mean	91.285	172.550
Standard Deviation	9.862	26.327
Sample Size	40	
Correlation	0.889	
Slope	2.373	
Intercept	-44.076	

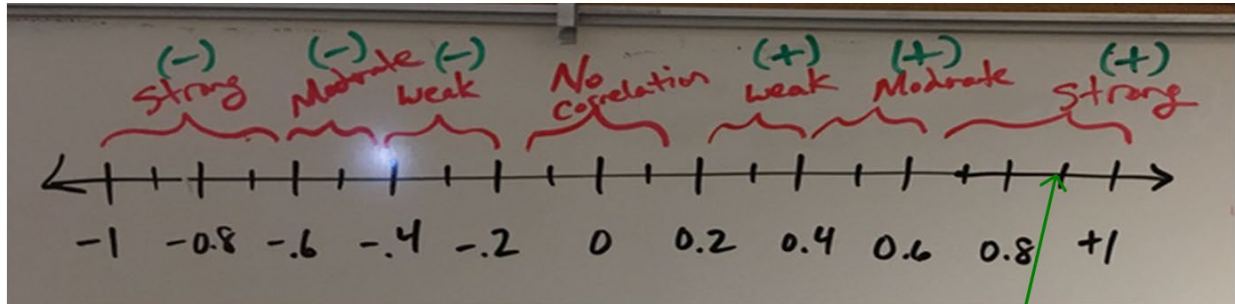
Scatterplot Controls  
 Show Regression Line

The points in the scatterplot seem to be close to the regression line and the line is going up from left to right. The scatterplot shows a “positive linear relationship” or a “positive correlation”, but how strong is this relationship? To determine this we can look at the correlation coefficient “r”.

The correlation coefficient came out to be 0.889. I like to put a positive sign in front of the correlation coefficient since 0.889 really means +0.889 and the sign of the correlation coefficient is important to the interpretation.

Interpretation: “Strong, Positive Correlation”

Look again at the correlation coefficient number line. Notice that +0.889 is a number very close to +1. That means that this correlation coefficient is telling us that there is a strong positive correlation between the waist size of a man (in cm) and his weight (in pounds). Therefore, it again confirms what our eyes were telling us when we looked at the scatterplot. The points seem to be close to a line (strong) and that line is going up from left to right (positive).



$r = +0.889$

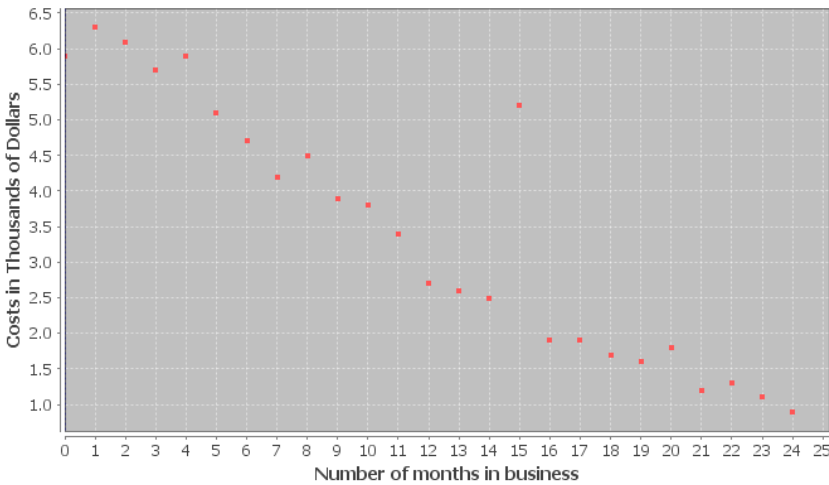
**Important Note: Correlation does not imply causation.** Just because there is a strong correlation between the waist size and weight of these 40 men, it does NOT imply that waist size CAUSES a man to have a certain weight. There are other factors involved. In order to prove cause and effect, the data must be collected with experimental design and the confounding variables must be controlled. Neither is the case in this example.

### Example 3

In the last section, we also looked at an example with an outlier. The data describes the number of months in business (X) and the company costs (Y) in thousands of dollars. The scatterplot shown below shows that most of the data follows a negative linear pattern, but month 15 had a higher cost than expected and did not seem to follow the pattern. The computer also gave us the correlation coefficient “r”. Let’s see if we can interpret it.



### Scatterplot of number of months and costs of company



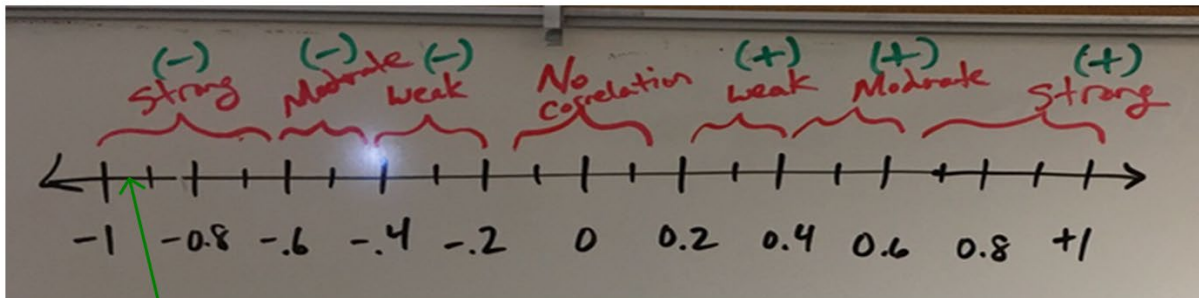
Correlation Coefficient

$r = -0.9409$

When a scatterplot shows an unusual point (outlier), it is often asked, “How influential is that outlier?” In other words, is the outlier doing a lot of damage to the overall relationship? Outliers can make a big difference to the strength of the relationship. Correlation coefficients can show a weak relationship with the outlier, but a very strong relationship without the outlier. When this happens, we call this an “influential outlier”.

Let us use the correlation coefficient “r” to shed some light on this relationship. Is the outlier influential? It seems like the relationship should be pretty strong, but how is the outlier effecting the overall strength of the relationship?

Interpretation: Look at the correlation coefficient number line again. The correlation coefficient of  $r = -0.9409$  is very close to  $-1$ . That means that despite the outlier, the correlation is still very strong. This tells us that the outlier is not very influential. The overall interpretation is still strong and negative. Therefore, there is a strong negative correlation between the number of months in business and the costs of the company in thousands of dollars.



$r = -0.9409$

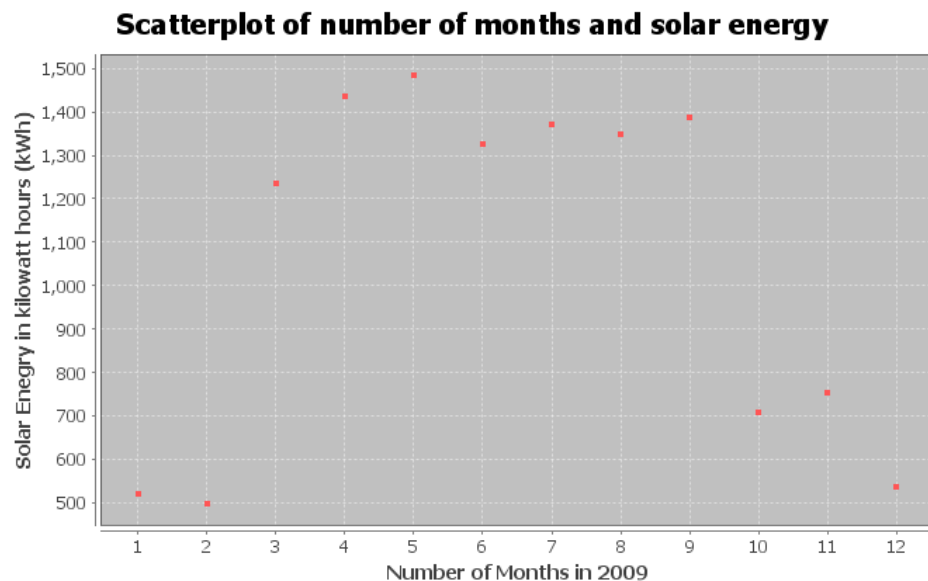
**Important Note: Correlation does not imply causation.** Just because there is a strong correlation between time in months and costs for this business, it does NOT imply that time CAUSES the company to have certain costs. There are other factors involved. In order to prove cause and effect, the data must be collected with experimental design and the confounding variables must be controlled. Neither is the case in this example.



#### Example 4

Curved relationships can be tricky. Remember if you do have a curved relationship, you really want to ask the computer to draw a curve that best fits the data, not a line. Some students get into trouble because they look at the  $r$  value for a line and try to apply it to a curve. When you ask a program to calculate “ $r$ ”, you are asking the computer how close your points are to a line, not a curve!

In the last section, we looked at a scatterplot that showed a curved pattern. The explanatory variable ( $X$ ) was the # months in 2009 and the response variable was the solar energy generated from that month in kilowatt-hours (kWh).



Notice the points in the scatterplot do not seem to be close to a line, but there seems to be a curved relationship.

Test Statistic

$r$       $-0.0568$

This is why it is so important to look at scatterplot before interpreting the correlation coefficient. We need to know what shape we are dealing with. This correlation coefficient is very close to 0, meaning that there is no linear relationship. If a student looked at just the  $r$  value without looking at a scatterplot, they may incorrectly think that there is no relationship at all between time (months) and solar energy. There is actually a strong relationship, but it is just not linear.

Interpretation: “Strong Curved Relationship”

It is important to know the shape before calculating statistics. Calculating statistics for a linear relationship when it is really curved can be very misleading. We will learn in our next chapter how to analyze curved relationships.

---

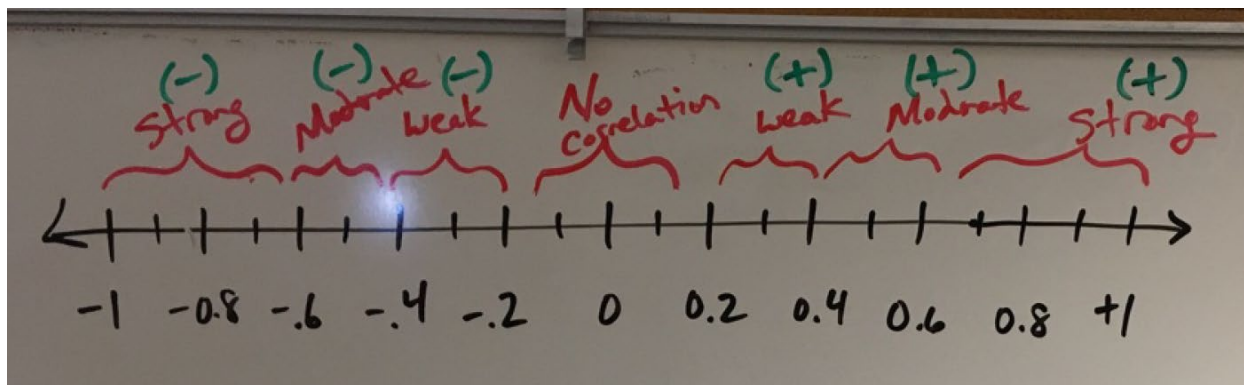


## Problem Set Section 6B

Directions: Open the “Health Data” from Canvas or from [www.matt-teachout.org](http://www.matt-teachout.org). Use the indicated columns of data to create scatterplots with StatKey and calculate the correlation coefficient “r”.

How to create a Scatterplot and calculate the correlation coefficient “r” with StatKey:

- Open the data. Then open a new spreadsheet and paste the two quantitative data sets next to each other side by side. It is customary to have the explanatory column (X) on the left and the response column (Y) on the right. Then copy the two columns together.
- Now we will go to [www.lock5stat.com](http://www.lock5stat.com) and click on “StatKey”. Under the “Descriptive Statistics and Graphs” menu, click on “Two Quantitative Variables”. Click on “Edit Data” at the top. Push Control A on your keyboard to highlight old data and then push “delete” on your keyboard to delete all old data in the edit data field. Then paste the two columns of quantitative data into the “Edit Data” field. If your data has a title, click the box that says “Data has header row”. If your data does not have a title, do NOT check the box that says “Data has header row”. Then press OK. The graph you see is the scatterplot. If your scatterplot has the X and Y variables backward, simply click the “Switch Variables” button in StatKey. The correlation coefficient “r” is listed “Summary Statistics” where it says “Correlation”.



1. Explore the relationship between a woman’s weight and height.
  - f) What is the correlation coefficient “r”.
  - g) Use the scatterplot and the correlation coefficient “r” to classify the relationship as one of the following: strong positive correlation, strong negative correlation, moderate positive correlation, moderate negative correlation, weak positive correlation, weak negative correlation, or no linear relationship.
2. Explore the relationship between a man’s weight and height.
  - a) What is the correlation coefficient “r”.
  - b) Use the scatterplot and the correlation coefficient “r” to classify the relationship as one of the following: strong positive correlation, strong negative correlation, moderate positive correlation, moderate negative correlation, weak positive correlation, weak negative correlation, or no linear relationship.
3. Explore the relationship between a woman’s cholesterol and age.
  - f) What is the correlation coefficient “r”.
  - g) Use the scatterplot and the correlation coefficient “r” to classify the relationship as one of the following: strong positive correlation, strong negative correlation, moderate positive correlation, moderate negative correlation, weak positive correlation, weak negative correlation, or no linear relationship.

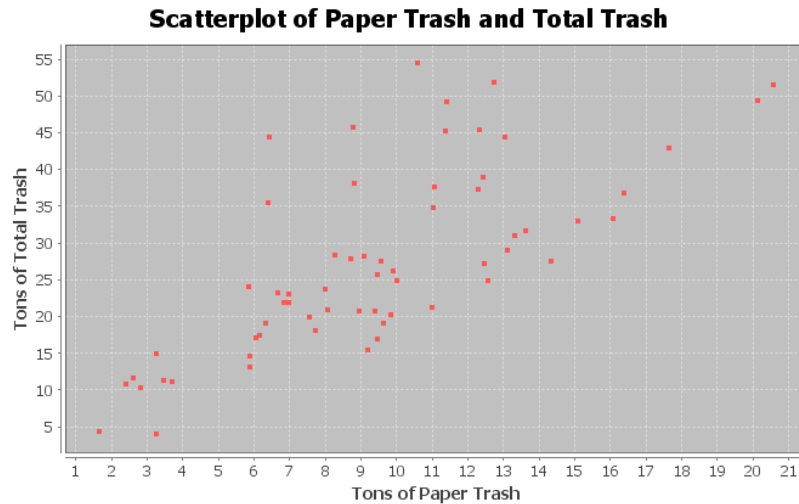


4. Explore the relationship between a man's cholesterol and age.
  - a) What is the correlation coefficient "r".
  - b) Use the scatterplot and the correlation coefficient "r" to classify the relationship as one of the following: strong positive correlation, strong negative correlation, moderate positive correlation, moderate negative correlation, weak positive correlation, weak negative correlation, or no linear relationship.
5. Explore the relationship between a woman's weight and body mass index (BMI).
  - a) What is the correlation coefficient "r".
  - b) Use the scatterplot and the correlation coefficient "r" to classify the relationship as one of the following: strong positive correlation, strong negative correlation, moderate positive correlation, moderate negative correlation, weak positive correlation, weak negative correlation, or no linear relationship.
6. Explore the relationship between a man's weight and body mass index (BMI).
  - a) What is the correlation coefficient "r".
  - b) Use the scatterplot and the correlation coefficient "r" to classify the relationship as one of the following: strong positive correlation, strong negative correlation, moderate positive correlation, moderate negative correlation, weak positive correlation, weak negative correlation, or no linear relationship.
7. Explore the relationship between a woman's systolic blood pressure and her diastolic blood pressure.
  - a) What is the correlation coefficient "r".
  - b) Use the scatterplot and the correlation coefficient "r" to classify the relationship as one of the following: strong positive correlation, strong negative correlation, moderate positive correlation, moderate negative correlation, weak positive correlation, weak negative correlation, or no linear relationship.
8. Explore the relationship between a man's systolic blood pressure and his diastolic blood pressure.
  - a) What is the correlation coefficient "r".
  - b) Use the scatterplot and the correlation coefficient "r" to classify the relationship as one of the following: strong positive correlation, strong negative correlation, moderate positive correlation, moderate negative correlation, weak positive correlation, weak negative correlation, or no linear relationship.
9. Explore the relationship between the length of a man's leg and the circumference of his wrist.
  - a) What is the correlation coefficient "r".
  - b) Use the scatterplot and the correlation coefficient "r" to classify the relationship as one of the following: strong positive correlation, strong negative correlation, moderate positive correlation, moderate negative correlation, weak positive correlation, weak negative correlation, or no linear relationship.
10. Explore the relationship between a woman's pulse and her cholesterol level.
  - a) What is the correlation coefficient "r".
  - b) Use the scatterplot and the correlation coefficient "r" to classify the relationship as one of the following: strong positive correlation, strong negative correlation, moderate positive correlation, moderate negative correlation, weak positive correlation, weak negative correlation, or no linear relationship.

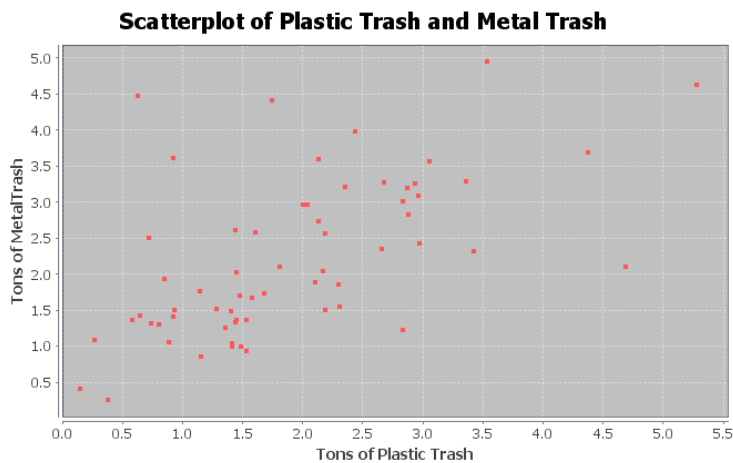


(#11-17) Directions: Use the given scatterplot and correlation coefficient  $r$  to determine the strength and direction of the correlation.

11. The x variable is describing the number of tons of paper trash and the y variable is the number of tons of total trash. ( $r = 0.7287$ ) Use the scatterplot and the correlation coefficient " $r$ " to classify the relationship as one of the following: strong positive correlation, strong negative correlation, moderate positive correlation, moderate negative correlation, weak positive correlation, weak negative correlation, or no linear relationship.



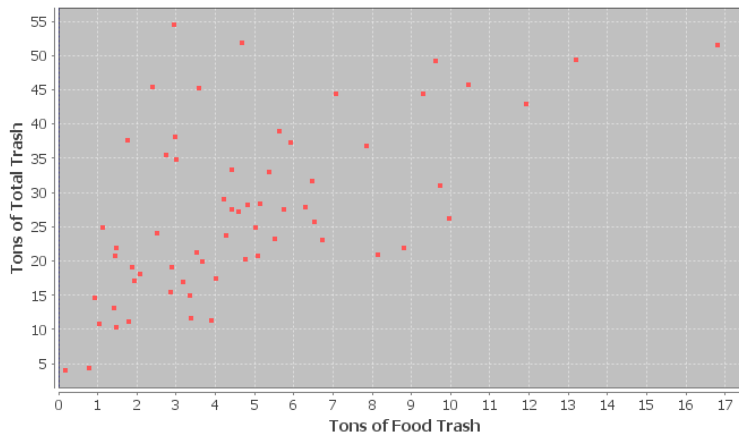
12. The x variable is describing the number of tons of plastic trash and the y variable is the number of tons of metal trash. ( $r = 0.5862$ ) Use the scatterplot and the correlation coefficient " $r$ " to classify the relationship as one of the following: strong positive correlation, strong negative correlation, moderate positive correlation, moderate negative correlation, weak positive correlation, weak negative correlation, or no linear relationship.





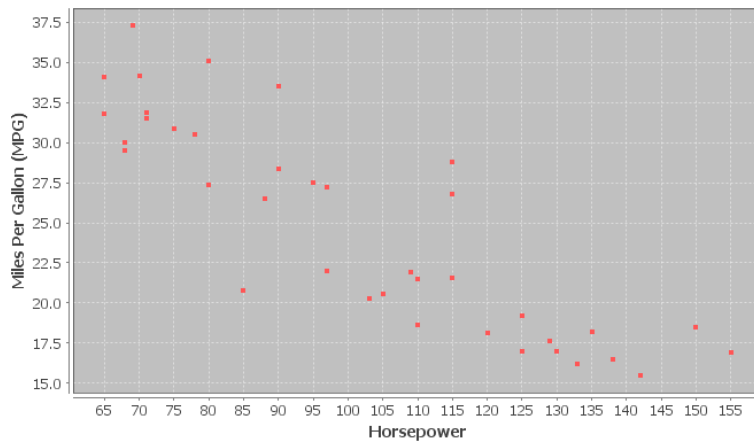
13. The x variable is describing the number of tons of food trash and the y variable is the number of tons of total trash. ( $r = 0.5833$ ) Use the scatterplot and the correlation coefficient “r” to classify the relationship as one of the following: strong positive correlation, strong negative correlation, moderate positive correlation, moderate negative correlation, weak positive correlation, weak negative correlation, or no linear relationship.

**Scatterplot of Food and Total Trash**

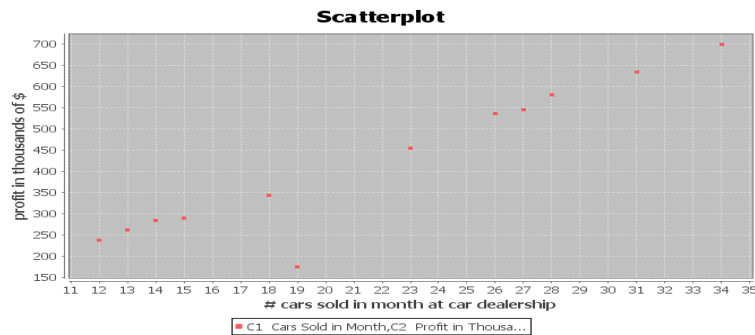


14. The x variable is describing the horsepower of an automobile and the y variable is describing the miles per gallon. ( $r = -0.8713$ ) Use the scatterplot and the correlation coefficient “r” to classify the relationship as one of the following: strong positive correlation, strong negative correlation, moderate positive correlation, moderate negative correlation, weak positive correlation, weak negative correlation, or no linear relationship.

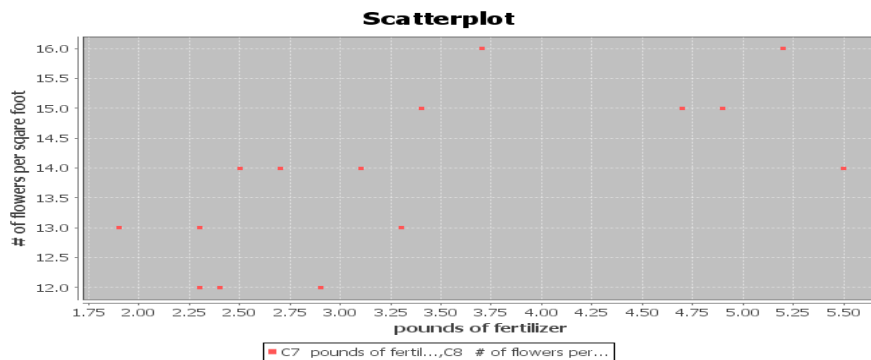
**Scatterplot of Car Horsepower and MPG**



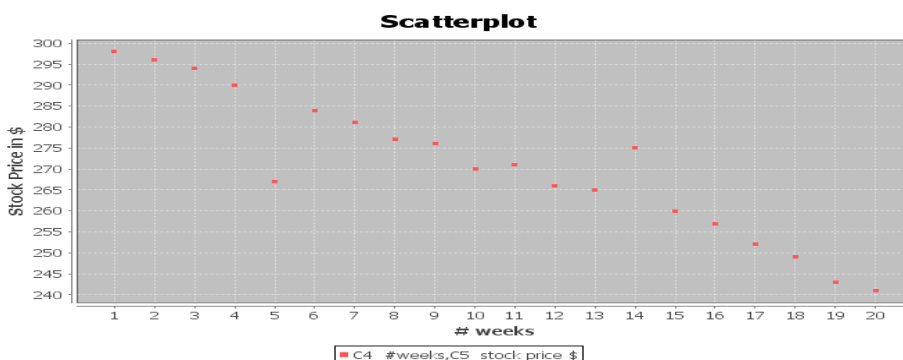
15. The x variable is the number of cars sold and the y variable is the total profit in thousands of dollars. ( $r = 0.9404$ ) Use the scatterplot and the correlation coefficient “r” to classify the relationship as one of the following: strong positive correlation, strong negative correlation, moderate positive correlation, moderate negative correlation, weak positive correlation, weak negative correlation, or no linear relationship.



16. The x variable is the number of pounds of fertilizer used and the y variable is the number of flowers per square foot. ( $r = 0.6727$ ) Use the scatterplot and the correlation coefficient “r” to classify the relationship as one of the following: strong positive correlation, strong negative correlation, moderate positive correlation, moderate negative correlation, weak positive correlation, weak negative correlation, or no linear relationship.



17. The x variable is the week and the y variable is the stock price in dollars. ( $r = -0.9429$ ) Use the scatterplot and the correlation coefficient “r” to classify the relationship as one of the following: strong positive correlation, strong negative correlation, moderate positive correlation, moderate negative correlation, weak positive correlation, weak negative correlation, or no linear relationship.



## Section 6C – Confounding Variables, r-squared, Correlation is not Causation, and Multivariable Studies

### Correlation is NOT Causation

There is a famous saying in statistics, “correlation is not causation”. We saw this when we looked at categorical relationships. Just because there is a relationship between two variables, does not mean that one variable causes the other.

Why? Why doesn't a relationship imply causation?

The real reason is confounding variables. Confounding variables (also called “lurking variables”) are other variables that might be related to the response variable other than the explanatory variable you are looking at. It helps to look at an example.

In the previous section, we found that there is a strong positive linear relationship (correlation) between the waist size and weight of forty men in the health data. Does that mean that the waist size of a man causes them to have a certain weight? No, it does not. The weight of a man is influenced by many factors other than just his waist size. Can you think of any?

Height of the man  
Genetics (How tall are his parents?)  
Amount of Exercise  
Quality of his diet  
Body Mass Index  
Amount of Muscle  
Amount of Fat

These are called confounding variables. Many variables influence a man's weight other than just his waist size. That is why it is wrong to say things like, “a small waist size causes a man to not weigh very much”. (Athletes often have a lot of muscle mass and may weigh a lot, but have a small waist size.)

### The Coefficient of Determination ( $r^2$ )

Suppose we are studying the weight of a man and what to know which variables have the strongest relationship with weight. An important statistic often used in studies like this is the “coefficient of determination” ( $r^2$ ). The coefficient of the determination ( $r^2$ ) is calculated by squaring the  $r$  – value (squaring the correlation coefficient).

**Definition of the Coefficient of Determination ( $r^2$ ):** The percentage of variability in the response variable (Y) that can be explained by the relationship with the explanatory variable (X).

### Notes about r-squared

- R-squared is often given in correlation and regression printouts. StatKey however does not provide r-squared in its printout. It is easy to calculate though if you have the correlation coefficient “ $r$ ”. Just push the square button on a calculator or multiply the “ $r$ ” value by itself.
- R-squared is always positive. Remember when you square a number (even a negative number) the result will be positive. R-squared is never negative.
- R-squared is a proportion that can be converted into a percentage. Make sure to take the r-squared value in the computer and multiply it by 100% to convert it into a percentage.
- Do not convert the correlation coefficient  $r$  into a percentage. “ $r$ ” is not a percentage. It is a decimal number between  $-1$  and  $+1$ .
- The higher the r-squared percentage is, the stronger the relationship. The lower the r-squared percentage is, the weaker the relationship.
- R-squared can be calculated for lines and curves. R-squared is a great statistic for quantitative relationship studies because it can be calculated for linear relationships and for curved relationships. (As long as you tell the computer what relationship to calculate.)



### Looking at r and r-squared

Note: The following table is not exact. R and r-squared interpretations can differ depending on the data and how many points you have. These are just general guidelines.

	Correlation Coefficient (r)	Coefficient of Determination (r-squared)
No Correlation	$\approx 0 \rightarrow \pm 0.19$	$\approx 0 \rightarrow 3\%$
Weak Correlation	$\approx \pm 0.2 \rightarrow \pm 0.39$	$\approx 4\% \rightarrow 15\%$
Moderate Correlation	$\approx \pm 0.4 \rightarrow \pm 0.59$	$\approx 16\% \rightarrow 35\%$
Strong Correlation	$\approx \pm 0.6 \rightarrow \pm 1.0$	$\approx 36\% \rightarrow 100\%$

### Calculating the Coefficient of Determination ( $r^2$ ) with StatKey:

- To put the data into StatKey, you will want to open a fresh excel spreadsheet and paste the two quantitative data sets next to each other side by side. Then copy the two columns together.
- Now we will go to [www.lock5stat.com](http://www.lock5stat.com) and click on "StatKey". Under the "Descriptive Statistics and Graphs" menu, click on "Two Quantitative Variables". Click on "Edit Data" at the top. Push Control A on your keyboard to highlight old data and then push "delete" on your keyboard to delete all old data in the edit data field. Then paste the two columns of quantitative data into the "Edit Data" field. If your data has a title, click the box that says "Data has header row". If your data does not have a title, do NOT check the box that says "Data has header row". Then press OK.
- You will see the correlation coefficient "r" under "Summary Statistics". Look next to "Correlation".
- Square the r-value by either multiplying "r" by itself or using the square button on a calculator.

### Example 1

In our last section, we used the health data to calculate the correlation coefficient "r" with StatKey. This helped us see that there was a strong positive correlation between the waist size and weight of these men. The computer calculated that  $r = 0.889$ . From this, we can calculate r-squared.

Coefficient of Determination (r-squared) =  $0.889 \times 0.889 = 0.790321 \approx 79.0\%$ .

Sentence explaining r-squared in context: 79.0% of the variability in the men's weights can be explained by the linear relationship with the men's waist size.

Interpretation: This percentage is very high indicated this is an extremely strong linear relationship. Waist size is a good explanatory variable for predicting weight. However this does not indicate that waist size causes weight to increase (correlation is not causation). There are many confounding variables involved.

### Summary Statistics Switch Variables

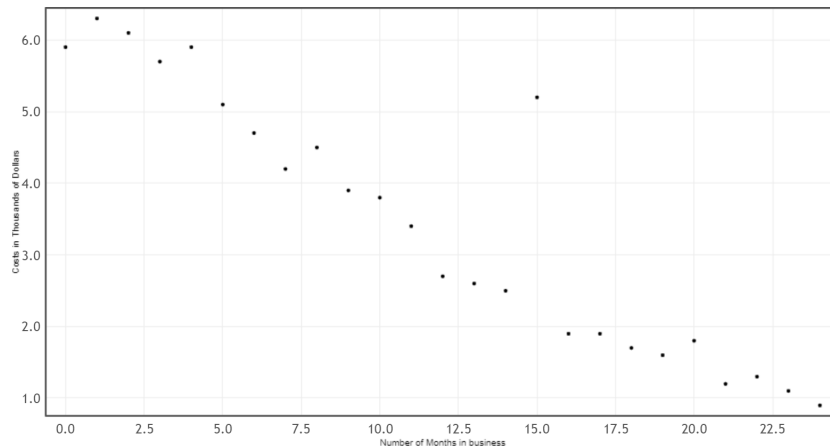
Statistic	Men Waist (cm)	Men Wt (Lbs)
Mean	91.285	172.550
Standard Deviation	9.862	26.327
Sample Size	40	
Correlation	0.889 = r	
Slope	2.373	
Intercept	-44.076	



## Example 2

Let's analyze the following time and cost data for a company. Use StatKey to calculate the correlation coefficient ( $r$ ) and then use " $r$ " to calculate the coefficient of determination ( $r$ -squared).

Number of Months in business	Costs in Thousands of Dollars
0	5.9
1	6.3
2	6.1
3	5.7
4	5.9
5	5.1
6	4.7
7	4.2
8	4.5
9	3.9
10	3.8
11	3.4
12	2.7
13	2.6
14	2.5
15	5.2
16	1.9
17	1.9
18	1.7
19	1.6
20	1.8
21	1.2
22	1.3
23	1.1
24	0.9



### Summary Statistics [Switch Variables](#)

Statistic	Number of Months in business	Costs in Thousands of Dollars
Mean	12.000	3.436
Standard Deviation	7.360	1.820
Sample Size	25	
Correlation	-0.941	
Slope	-0.233	
Intercept	6.227	

### Scatterplot Controls

Show Regression Line

Coefficient of Determination ( $r$ -squared) =  $(-0.941) \times (-0.941) = +0.885481 \approx 88.5\%$ .

Sentence explaining  $r$ -squared in context: 88.5% of the variability in costs for the company can be explained by the linear relationship with time in months.

Interpretation: This percentage is very high indicated this is an extremely strong linear relationship. Time is a good explanatory variable for predicting costs. However this does not indicate that time causes costs to decrease (correlation is not causation). There are many confounding variables involved.



### Example 3: Multiple Variable Quantitative Relationship Studies

What variables have the strongest relationship with the weight of a man? (What variables are most important to study when looking at men's weight?) This kind of study is sometimes called "multiple regression".

The health data has several variables that we might look at, but which ones have the strongest relationships with men's weight? This is actually not as difficult as it seems. We will need to choose a different explanatory variable (X) each time and then use a statistics software program to calculate r-squared for each variable with the men's weight.

Response Variable: Weight of Men (in pounds)

What variables that might be related to the weight of the men?

For this example, we will focus on the variables in the health data and on linear relationships only. Linear relationships are the most common type of quantitative relationship statisticians study.

Men's Age (years)  
Men's Height (inches)  
Men's Waist Size (cm)  
Men's Pulse (beats per minute)  
Men's Systolic Blood Pressure (mm of Hg)  
Men's Diastolic Blood Pressure (mm of Hg)  
Men's Cholesterol (mg per deciliter)  
Men's Body Mass Index (BMI) (kg per m<sup>2</sup>)  
Men's Leg Length (inches)  
Men's Elbow Circumference (Inches)  
Men's Wrist Circumference (Inches)  
Men's Arm Length (Inches)

Letting each of these variables be the explanatory variable, we can calculate the r-squared value for each. Remember we need to keep the men's weight as the response variable though, since that is the variable we are studying.

Men's Age (years) and Weight (pounds): r-squared = 0.0815  $\approx$  8.2%

Men's Height (inches) and Weight (pounds): r-squared = 0.2727  $\approx$  27.3%

Men's Waist Size (cm) and Weight (pounds): r-squared = 0.7902  $\approx$  79.0%

Men's Pulse (beats per minute) and Weight (pounds): r-squared = 0.0031  $\approx$  0.31%

Men's Systolic Blood Pressure (mm of Hg) and Weight (pounds): r-squared = 0.1240  $\approx$  12.4%

Men's Diastolic Blood Pressure (mm of Hg) and Weight (pounds): r-squared = 0.1503  $\approx$  15.0%

Men's Cholesterol (mg per deciliter) and Weight (pounds): r-squared = 0.0007  $\approx$  0.07%

Men's Body Mass Index (BMI) (kg per m<sup>2</sup>) and Weight (pounds): r-squared = 0.6395  $\approx$  64.0%

Men's Leg Length (inches) and Weight (pounds): r-squared = 0.1380  $\approx$  13.8%

Men's Elbow Circumference (Inches) and Weight (pounds): r-squared = 0.4034  $\approx$  40.3%

Men's Wrist Circumference (Inches) and Weight (pounds): r-squared = 0.2696  $\approx$  27.0%



This material is from *Introduction to Data Analysis*, 2<sup>nd</sup> edition by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a "CC-BY" [Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by/4.0/) – 3/17/2021

Men's Arm Length (Inches) and Weight (pounds):  $r\text{-squared} = 0.6750 \approx 67.5\%$

Multiple Regression Interpretation:

So which variables had the strongest relationship with the weight of the men?

Height (27.3%), Waist Size (79.0%), Body Mass Index (64.0%), Elbow Circumference (40.3%), Wrist Circumference (27.0%) and Arm Length (67.5%) all showed a linear relationship to the weight of the men. Waist size, body mass index and arm length showed very strong linear relationships, but all of these variables showed some correlation and we should think about all of them if we want to study men's weights.

What about the other variables?

Pulse (0.31%) and cholesterol (0.07%) showed no relationship at all. (Notice their r-squared values are very close to zero.)

Surprisingly, the following variables showed a weak linear relationship with the men's weight.

Age (8.2%), Systolic Blood Pressure (12.4%), Diastolic Blood Pressure (15.0%),  
Leg Length (13.8%),

---



## Problem Set Section 6C

(#1-10) Answer the following questions to explain each of the following r-squared values. In each of these examples the weight was the response variable (Y).

1. Men's Waist Size (cm) and Weight (pounds):  $r\text{-squared} = 0.7902$ 
  - a) Convert the given r-squared value into a percentage by multiplying by 100 and adding the % sign.
  - b) Write the r-squared definition sentence in context to explain the r-squared in this problem.
  - c) Does the r-squared indicate that there is a weak correlation, moderate correlation, strong correlation or no correlation between the variables.
  
2. Men's Pulse (beats per minute) and Weight (pounds):  $r\text{-squared} = 0.0031$ 
  - a) Convert the given r-squared value into a percentage by multiplying by 100 and adding the % sign.
  - b) Write the r-squared definition sentence in context to explain the r-squared in this problem.
  - c) Does the r-squared indicate that there is a weak correlation, moderate correlation, strong correlation or no correlation between the variables.
  
3. Men's Systolic Blood Pressure (mm of Hg) and Weight (pounds):  $r\text{-squared} = 0.1240$ 
  - a) Convert the given r-squared value into a percentage by multiplying by 100 and adding the % sign.
  - b) Write the r-squared definition sentence in context to explain the r-squared in this problem.
  - c) Does the r-squared indicate that there is a weak correlation, moderate correlation, strong correlation or no correlation between the variables.
  
4. Men's Diastolic Blood Pressure (mm of Hg) and Weight (pounds):  $r\text{-squared} = 0.1503$ 
  - a) Convert the given r-squared value into a percentage by multiplying by 100 and adding the % sign.
  - b) Write the r-squared definition sentence in context to explain the r-squared in this problem.
  - c) Does the r-squared indicate that there is a weak correlation, moderate correlation, strong correlation or no correlation between the variables.
  
5. Men's Cholesterol (mg per deciliter) and Weight (pounds):  $r\text{-squared} = 0.0007$ 
  - a) Convert the given r-squared value into a percentage by multiplying by 100 and adding the % sign.
  - b) Write the r-squared definition sentence in context to explain the r-squared in this problem.
  - c) Does the r-squared indicate that there is a weak correlation, moderate correlation, strong correlation or no correlation between the variables.
  
6. Men's Body Mass Index (BMI) ( $\text{kg per m}^2$ ) and Weight (pounds):  $r\text{-squared} = 0.6395$ 
  - a) Convert the given r-squared value into a percentage by multiplying by 100 and adding the % sign.
  - b) Write the r-squared definition sentence in context to explain the r-squared in this problem.
  - c) Does the r-squared indicate that there is a weak correlation, moderate correlation, strong correlation or no correlation between the variables.

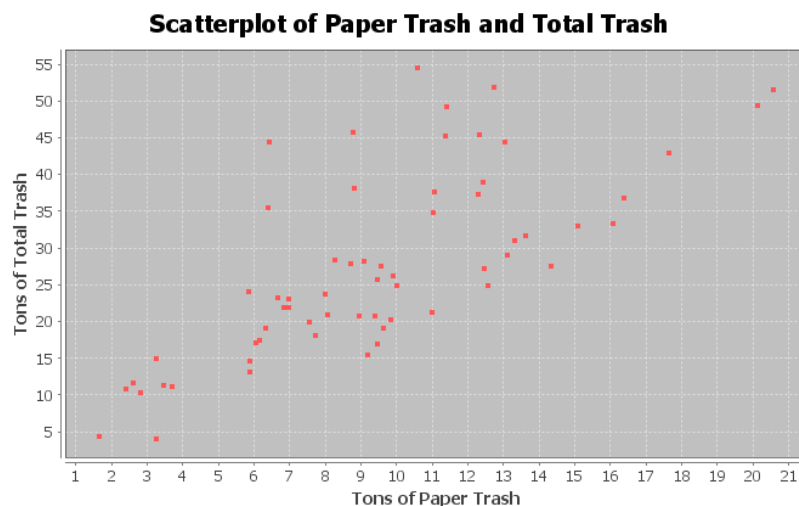




7. Men's Leg Length (inches) and Weight (pounds):  $r\text{-squared} = 0.1380$
- Convert the given  $r\text{-squared}$  value into a percentage by multiplying by 100 and adding the % sign.
  - Write the  $r\text{-squared}$  definition sentence in context to explain the  $r\text{-squared}$  in this problem.
  - Does the  $r\text{-squared}$  indicate that there is a weak correlation, moderate correlation, strong correlation or no correlation between the variables.
8. Men's Elbow Circumference (Inches) and Weight (pounds):  $r\text{-squared} = 0.4034$
- Convert the given  $r\text{-squared}$  value into a percentage by multiplying by 100 and adding the % sign.
  - Write the  $r\text{-squared}$  definition sentence in context to explain the  $r\text{-squared}$  in this problem.
  - Does the  $r\text{-squared}$  indicate that there is a weak correlation, moderate correlation, strong correlation or no correlation between the variables.
9. Men's Wrist Circumference (Inches) and Weight (pounds):  $r\text{-squared} = 0.2696$
- Convert the given  $r\text{-squared}$  value into a percentage by multiplying by 100 and adding the % sign.
  - Write the  $r\text{-squared}$  definition sentence in context to explain the  $r\text{-squared}$  in this problem.
  - Does the  $r\text{-squared}$  indicate that there is a weak correlation, moderate correlation, strong correlation or no correlation between the variables.
10. Men's Arm Length (Inches) and Weight (pounds):  $r\text{-squared} = 0.6750$
- Convert the given  $r\text{-squared}$  value into a percentage by multiplying by 100 and adding the % sign.
  - Write the  $r\text{-squared}$  definition sentence in context to explain the  $r\text{-squared}$  in this problem.
  - Does the  $r\text{-squared}$  indicate that there is a weak correlation, moderate correlation, strong correlation or no correlation between the variables.

(#11-17) Directions: Use the given scatterplots and correlation coefficients " $r$ " to answer the following questions for each problem.

11. The  $x$  variable is describing the number of tons of paper trash and the  $y$  variable is the number of tons of total trash. ( $r = 0.7287$ )

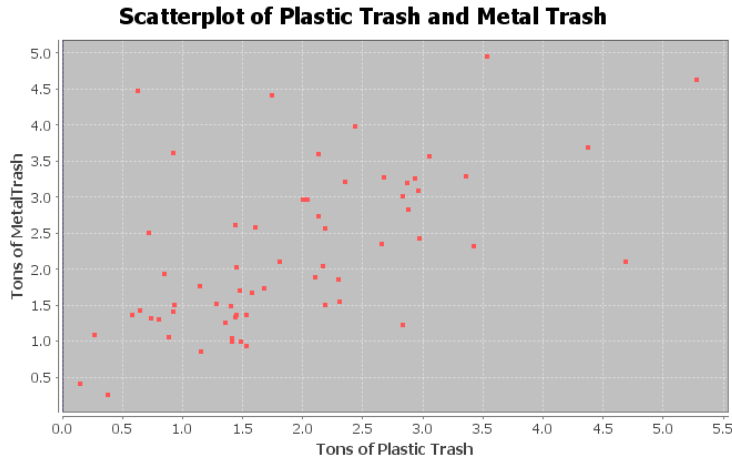


- Find the value of  $r\text{-squared}$  by squaring the correlation coefficient " $r$ " with your calculator or by multiplying  $r \times r$ .



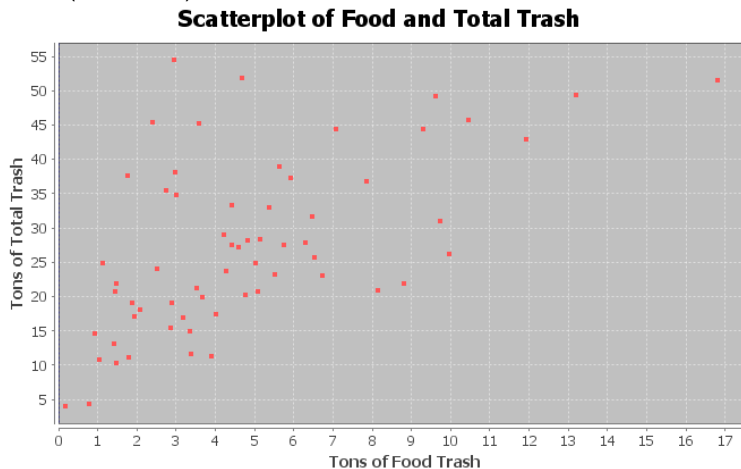
- b) Convert r-squared in part (a) into a percentage by multiplying by 100 and adding the % sign.
- c) Write the r-squared definition sentence in context to explain the r-squared in this problem.
- d) List other possible confounding variables that may also account for the variability in y.
- e) Since this data did not use experimental design, there are confounding variables that are not controlled. So is it ok to say that the X variable causes the Y? Why or why not?

12. The x variable is describing the number of tons of plastic trash and the y variable is the number of tons of metal trash. ( $r = 0.5862$ )



- a) Find the value of r-squared by squaring the correlation coefficient “r” with your calculator or by multiplying  $r \times r$ .
- b) Convert r-squared in part (a) into a percentage by multiplying by 100 and adding the % sign.
- c) Write the r-squared definition sentence in context to explain the r-squared in this problem.
- d) List other possible confounding variables that may also account for the variability in y.
- e) Since this data did not use experimental design, there are confounding variables that are not controlled. So is it ok to say that the X variable causes the Y? Why or why not?

13. The x variable is describing the number of tons of food trash and the y variable is the number of tons of total trash. ( $r = 0.5833$ )

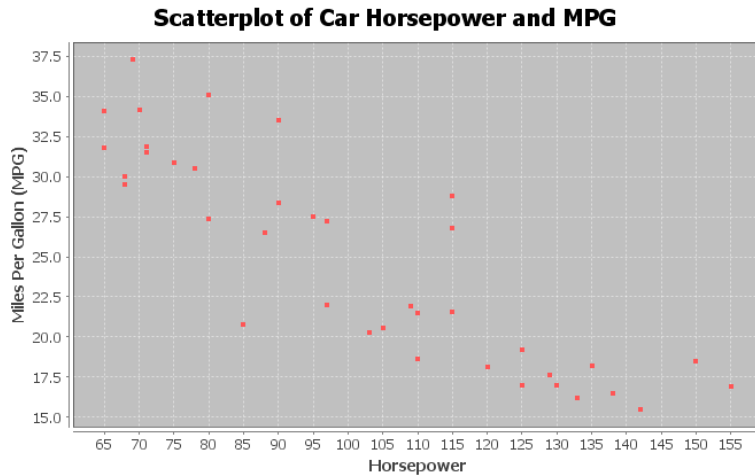


- a) Find the value of r-squared by squaring the correlation coefficient “r” with your calculator or by multiplying  $r \times r$ .
- b) Convert r-squared in part (a) into a percentage by multiplying by 100 and adding the % sign.



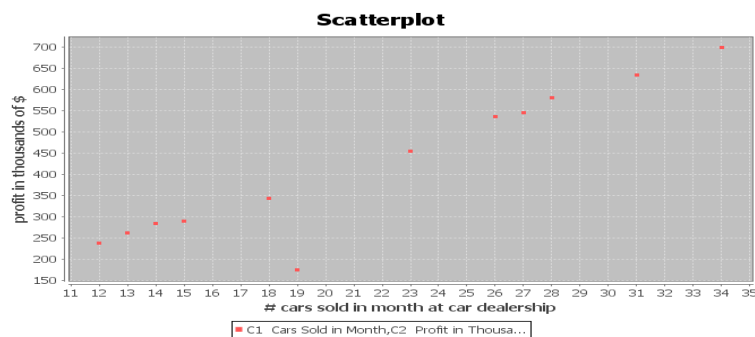
- c) Write the r-squared definition sentence in context to explain the r-squared in this problem.
- d) List other possible confounding variables that may also account for the variability in y.
- e) Since this data did not use experimental design, there are confounding variables that are not controlled. So is it ok to say that the X variable causes the Y? Why or why not?

14. The x variable is describing the horsepower of an automobile and the y variable is describing the miles per gallon. ( $r = -0.8713$ )



- a) Find the value of r-squared by squaring the correlation coefficient “r” with your calculator or by multiplying  $r \times r$ .
- b) Convert r-squared in part (a) into a percentage by multiplying by 100 and adding the % sign.
- c) Write the r-squared definition sentence in context to explain the r-squared in this problem.
- d) List other possible confounding variables that may also account for the variability in y.
- e) Since this data did not use experimental design, there are confounding variables that are not controlled. So is it ok to say that the X variable causes the Y? Why or why not?

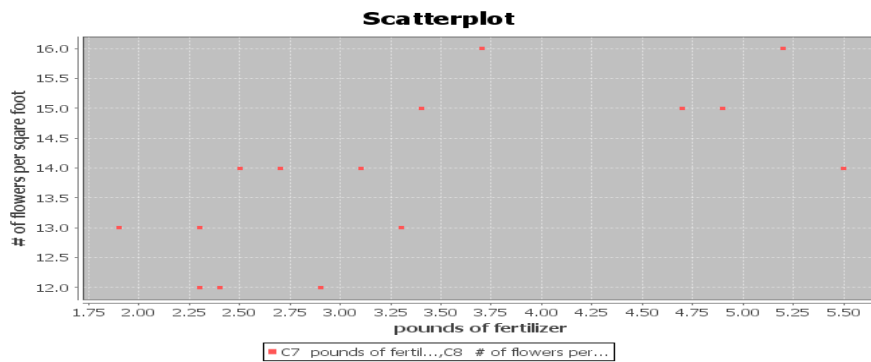
15. The x variable is the number of cars sold and the y variable is the total profit in thousands of dollars. ( $r = 0.9404$ )



- a) Find the value of r-squared by squaring the correlation coefficient “r” with your calculator or by multiplying  $r \times r$ .
- b) Convert r-squared in part (a) into a percentage by multiplying by 100 and adding the % sign.
- c) Write the r-squared definition sentence in context to explain the r-squared in this problem.
- d) List other possible confounding variables that may also account for the variability in y.
- e) Since this data did not use experimental design, there are confounding variables that are not controlled. So is it ok to say that the X variable causes the Y? Why or why not?

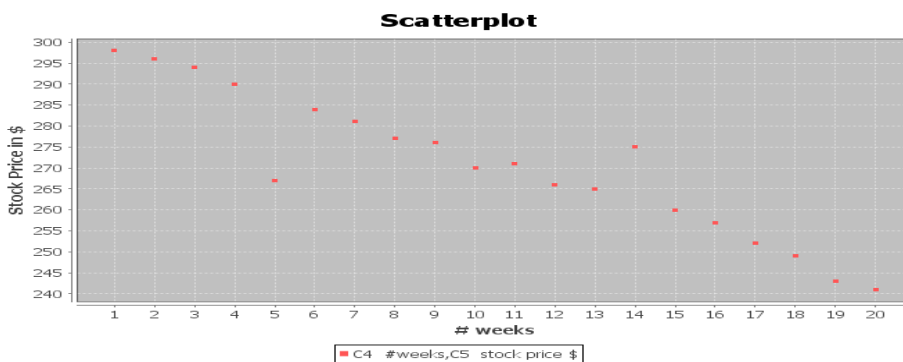


16. The x variable is the number of pounds of fertilizer used and the y variable is the number of flowers per square foot. ( $r = 0.6727$ )



- Find the value of r-squared by squaring the correlation coefficient “r” with your calculator or by multiplying  $r \times r$ .
- Convert r-squared in part (a) into a percentage by multiplying by 100 and adding the % sign.
- Write the r-squared definition sentence in context to explain the r-squared in this problem.
- List other possible confounding variables that may also account for the variability in y.
- Since this data did not use experimental design, there are confounding variables that are not controlled. So is it ok to say that the X variable causes the Y? Why or why not?

17. The x variable is the week and the y variable is the stock price in dollars. ( $r = -0.9429$ )



- Find the value of r-squared by squaring the correlation coefficient “r” with your calculator or by multiplying  $r \times r$ .
- Convert r-squared in part (a) into a percentage by multiplying by 100 and adding the % sign.
- Write the r-squared definition sentence in context to explain the r-squared in this problem.
- List other possible confounding variables that may also account for the variability in y.
- Since this data did not use experimental design, there are confounding variables that are not controlled. So is it ok to say that the X variable causes the Y? Why or why not?



## Section 6D – Best Fit Regression Line with Technology, Slope and Y-intercept Interpretation

We said that one of the main differences lines in algebra classes and lines in statistics is the number of points. Algebra talks about the equation of a line between two points, while in statistics we talk about the line that best fit thousands or even millions of points.

In this section, we will discuss how to find a line that minimizes the distance between itself and thousands or millions of points in a scatterplot. As you can imagine, it is much more complicated than finding a line between two points.

This line of best fit is often called the “regression line”.

**Definition of the “regression line”:** This line best fits all the points in the scatterplot by minimizing the vertical distance between itself and all of the points in the scatterplot. It is also called the “line of best fit” or the “line of least squares”. It also is sometimes called a “prediction line” because if the two quantitative data sets have correlation, then the regression line can become a formula for making predictions.

Note: If there is no linear relationship (no correlation), then we should not use the regression line to make predictions.

Calculating the slope and y-intercept of the regression line with StatKey:

- To put the data into StatKey, you will want to open a fresh excel spreadsheet and paste the two quantitative data sets next to each other side by side. Then copy the two columns together.
- Now we will go to [www.lock5stat.com](http://www.lock5stat.com) and click on “StatKey”. Under the “Descriptive Statistics and Graphs” menu, click on “Two Quantitative Variables”. Click on “Edit Data” at the top. Push Control A on your keyboard to highlight old data and then push “delete” on your keyboard to delete all old data in the edit data field. Then paste the two columns of quantitative data into the “Edit Data” field. If your data has a title, click the box that says “Data has header row”. If your data does not have a title, do NOT check the box that says “Data has header row”. Then press OK.
- You will see the slope and y-intercept of the regression line under “Summary Statistics”. Look next to “Slope” and “Intercept”.
- To see the regression line on the scatterplot, check the box that says “Show Regression Line”.

### Calculating the Regression Line

As with most statistics, the regression line from the points in your scatterplot is very complex and time-consuming calculation. It is best to calculate the line with a statistics software program like StatKey. We will show the process of how the line is calculated though.

To calculate the slope ( $b_1$ ) and the y-intercept ( $b_0$ ) of the regression line, StatKey will calculate five different statistics, each of which is a very time consuming calculation. However, if you have these statistics already calculated, you can use them to get the regression line with a couple formulas.

$\bar{x}$  : mean average of the explanatory data set (X)

$\bar{y}$  : mean average of the response data set (Y)

$S_x$  : standard deviation of the explanatory data set (X)

$S_y$  : standard deviation of the response data set (Y)

$r$  : correlation coefficient between X and Y

Recall that the equation of a line is made up of two values, the slope of the line and the y-intercept. If you can find the slope and the y-intercept, you can write the formula for the equation of the line.

Equation of the Regression Line:



This material is from *Introduction to Data Analysis*, 2<sup>nd</sup> edition by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a “CC-BY” [Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by/4.0/) – 3/17/2021

$$\hat{y} = (\text{Y-intercept}) + (\text{slope}) x$$

$$\hat{y} = b_0 + b_1 x$$

Notice that the order of the slope and Y-intercept are backwards from algebra classes you may have seen. Algebra usually writes the equation with the slope first. In statistics, we prefer to write the Y-intercept first. The reason why is that Y-intercepts are usually initial values and the slope is how much the variables change after this initial Y-intercept value. Therefore, it makes sense that the initial value comes first in the formula. Just remember, whether you are looking at a line from algebra or statistics, the number in front of the "X" is the slope.

### Calculate the Slope ( $b_1$ )

We start by calculating the slope of the regression line. How do you find the slope that best fits thousands of data points? Start with the definition of slope. Slope is defined as the rate of change between the Y variable and X variable. In algebra, they often define slope as "rise over run" or "change in Y / change in X". It is easy to measure this change when you have two points, but how do you measure this change when you have thousands of points? Think of change in Y as the variability in Y and change in X as the variability in X. So we need a measure of the variability (spread) of X and Y. In regression line calculations, we use the standard deviation as our measure of spread.

$$\text{Slope} \approx \frac{\text{Variability in Y}}{\text{Variability in X}} \approx \frac{\text{Standard Deviation of all the Y values } (S_y)}{\text{Standard Deviation of all the X values } (S_x)}$$

Now there are two problems with leaving the formula like this. The first is that standard deviation is a distance calculation and is always positive. If all we do is divide the standard deviations, it will be impossible to get a negative slope (which happen all the time). The second problem is we need to take into account the strength of the correlation. It turns out that both of these problems can be solved by multiplying this ratio by the correlation coefficient. Remember the correlation coefficient measure the strength and direction (negative or positive) of the linear relationship.

$$\text{Best Fit Slope} = r \left( \frac{S_y}{S_x} \right) = \frac{r (S_y)}{S_x}$$

### Calculate the Y-intercept ( $b_0$ )

If you recall from your algebra classes, you can calculate the Y-intercept if you know the slope and point on the line. This is true for statistics as well, but what point should we use? There may be thousands of points in a scatterplot and the regression line does not have to go through any of them. The regression line gets close as possible to all of them.

Point on the regression line: It turns out the point we want to use for the regression line calculation is not any of the points in the scatterplot. Remember we want the line to go through the center of the spread of points. The mean average is a measure of spread, so we like to use the ordered pair (mean of X, mean of Y) to calculate the Y-intercept. Hence, to calculate the Y-intercept for the best-fit line, we will use the mean of all the X values, the mean of the Y values, and the best-fit slope.

$\bar{x}$  : mean average of the explanatory data set (X)

$\bar{y}$  : mean average of the response data set (Y)

Best Fit Y-intercept = (mean of Y values) – (slope) (mean of X values)

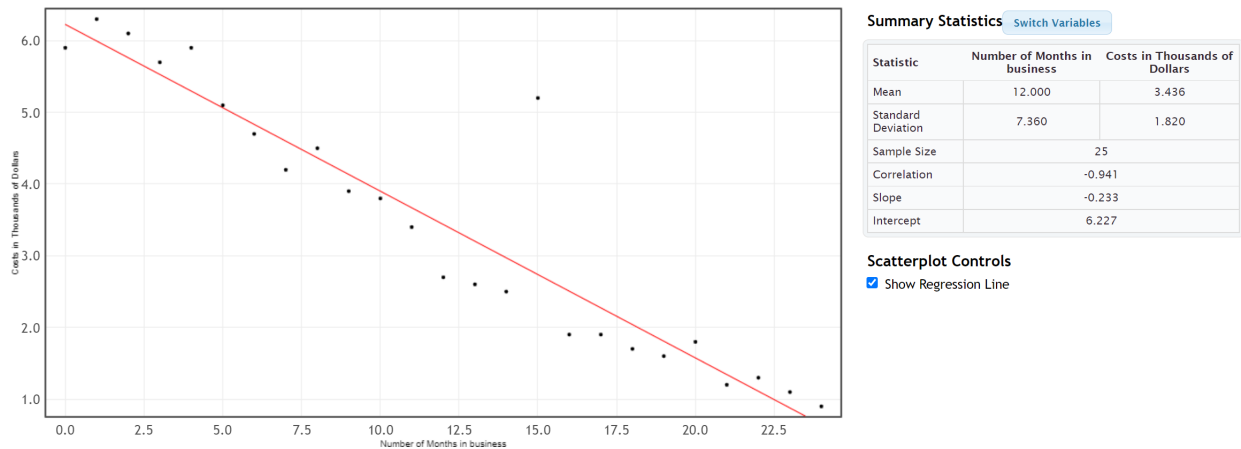


$$\text{Best Fit Y-intercept} = \bar{y} - \text{slope} (\bar{x})$$

When calculating the Y-intercept, we need to multiply the slope times the mean average of the explanatory data set (X). Then subtract the answer from the mean average of the response data set (Y).

### Example 1

Let us look again at an example from the last section. We looked at some data that gave the number of months a company had been in business and their costs in thousands of dollars. We found that there was an outlier, but that overall there was a strong negative correlation. This is important. Always check to see if there is correlation, before using the regression line. If there is no correlation, then the regression line is not accurate. Here is the scatterplot and statistics from StatKey.



The scatter plot and correlation coefficient ( $r = -0.941$ ) indicated that this data has a strong negative correlation. Notice the slope and y-intercept have already been calculated. Since there is a strong correlation, the slope and y-intercept are pretty accurate. Let's discuss how the computer calculated the slope and y-intercept.

Explanatory Variable (X): Number of months the company is in business

Response Variable (Y): Company costs in thousands of dollars

We will need the statistics listed above if we are going to calculate the slope and y-intercept for the regression line.

$\bar{x}$  : mean average of the explanatory data set (X)

$\bar{y}$  : mean average of the response data set (Y)

$S_x$  : standard deviation of the explanatory data set (X)

$S_y$  : standard deviation of the response data set (Y)

$r$  : correlation coefficient between X and Y



These are listed in the StatKey printout.

## Summary Statistics [Switch Variables](#)

Statistic	Number of Months in business	Costs in Thousands of Dollars
Mean	12.000 = $\bar{x}$	3.436 = $\bar{y}$
Standard Deviation	7.360 = $s_x$	1.820 = $s_y$
Sample Size	25	
Correlation	-0.941 = $r$	
Slope	-0.233 = $b_1$	
Intercept	6.227 = $b_0$	

To calculate the slope of the best fit regression line, we will take the correlation coefficient “r”, multiply by the standard deviation of the Y values (costs) and then divide by the standard deviation of the X values (months).

$$\text{Best Fit Slope } (b_1) = \frac{r \times s_y}{s_x} = \frac{-0.941 \times 1.820}{7.360} \approx -0.232692934 \approx -0.233 \text{ (Same as StatKey)}$$

Now that we have the slope ( $b_1$ ), we can use the mean averages to calculate the best fit Y-intercept ( $b_0$ ). It is better to use the

$$\text{Best Fit Yintercept } (b_0) = \bar{y} - (\text{slope} \times \bar{x}) = 3.436 - (-0.232692934 \times 12.000) = 3.436 - (-2.792315217) = 3.436 + (+2.792315217) \approx 6.228 \text{ (Close to what StatKey gave. Computer programs will always be more accurate than hand calculations since they keep more decimal places and round less.)}$$

Now that we know the best-fit slope ( $b_1$ ) and Y-intercept ( $b_0$ ), we can write the equation of the regression line.

Equation of the Regression Line:

$$\hat{y} = (\text{Y-intercept}) + (\text{slope}) x$$

$$\hat{y} = 6.227 + (-0.233) x$$

### Interpreting the Slope and Y-intercept

Remember, calculating a regression line is not enough. We need to know what the slope and Y-intercept tell us about the relationship between the real data variables. We need to understand these statistics and be able to explain them to others.

#### Example 1 (Interpretation)

Let us see if we can explain what the slope and the Y-intercept for the time/cost data. This information may be very important for the company.





Summary Statistics [Switch Variables](#)

Statistic	Number of Months in business	Costs in Thousands of Dollars
Mean	12.000 = $\bar{x}$	3.436 = $\bar{y}$
Standard Deviation	7.360 = $S_x$	1.820 = $S_y$
Sample Size	25	
Correlation	-0.941 = $r$	
Slope	-0.233 = $b_1$	
Intercept	6.227 = $b_0$	

Interpreting the slope

To interpret the slope, you have to remember that the slope measures the change in Y divided by the change in X. In other words, you cannot interpret the slope without thinking of it as a fraction and including the units.

**Definition of Slope of the Regression Line:** A rate of change that measures the average increase or decrease in the Y variable per 1 unit of the X variable.

The slope was -0.233. A good way to think of this decimal as a fraction is to put it over +1. Then put the units of Y in the numerator and the units of X in the denominator.

$$\text{Slope} = -0.233 = \frac{-0.233 \text{ thousand dollars}}{+1 \text{ month}}$$

First, recognize what the slope is not saying. The slope is not saying that the company had a cost of -0.2326 thousand dollars in the first month. (*In fact, the company had a cost of over six thousand dollars in its first month.*)

So what is the slope telling us? You also have to remember that slope is a rate of change (increase or a decrease). If it is negative, it is an average decrease and if it is positive, it is an average increase.

Since this slope was negative, it is a decrease. The slope is telling us that monthly costs are decreasing about 0.233 thousand dollars per month on average. Looking at the units, we can also explain it this way: Monthly costs are decreasing about \$233 per month on average. (*Notice we did not say that the costs were decreasing -0.233 or decreasing -\$233. The negative is described by the word "decrease".*)

**Sentence Explaining the Slope in Context:** Monthly costs for this company are decreasing about \$233 per month on average.

Interpretation of Y-intercept

Remember the slope is the number in front of the X. The Y-intercept is the initial number in the regression line formula that is by itself. So the Y-intercept for the cost data is  $b_0 = 6.227$ .

**Definition of Y-intercept of the Regression Line:** The predicted Y-value when X is zero.

Summary Statistics [Switch Variables](#)

Statistic	Number of Months in business	Costs in Thousands of Dollars
Mean	12.000 = $\bar{x}$	3.436 = $\bar{y}$
Standard Deviation	7.360 = $S_x$	1.820 = $S_y$
Sample Size	25	
Correlation	-0.941 = $r$	
Slope	-0.233 = $b_1$	
Intercept	6.227 = $b_0$	



A Y-intercept is the predicted Y value when X is zero. Do not forget to include the units. Therefore, the Y-intercept of 6.227 really represents the ordered pair (0 months, 6.227 thousand dollars).

**Sentence Explaining the Y-intercept in Context:** At the start of the company (month 0), we predict there was an average initial cost of approximately 6.227 thousand dollars (or \$6,227).

*Note about Y-intercept Interpretations: The regression line is meant to apply to the X and Y values in the two data sets. Zero is often not represented in the X values of many quantitative data sets. When zero is not in the scope of the X values, the Y-intercept will not make a whole lot of sense. It is still an important number in the formula if our predictions are to be accurate, but the formula may not be designed to plug in zero for X.*

*Important Note about Shape: We have seen in this section that the mean and standard deviations of the two quantitative variables are used to calculate the regression line. Remember that mean averages and standard deviations are only accurate if the data is bell shaped. Therefore, our regression line is not very accurate when the data is not bell shaped. We will see in the next section that to verify the shape requirement we will look at a special histogram called the “histogram of the residuals” to check this bell shaped requirement.*

---



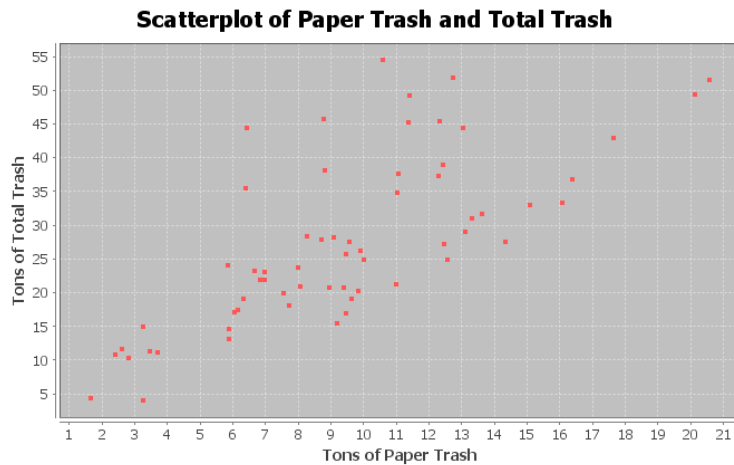
## Problem Set Section 6D

(#1-6) Use the regression line formulas below and a calculator to calculate the slope, Y-intercept, and equation of the regression line for the following ordered pair data. The correlation coefficient ( $r$ ), mean averages and standard deviations for both X and Y variables have been provided.

- Slope of the Regression Line = (multiply correlation coefficient  $r$  times the standard deviation of Y values)  $\div$  standard deviation of X values
- Y-intercept of the Regression Line (*multiply before you subtract.*)  
= (mean of Y values) – (multiply slope times mean of X values)
- Equation of the Regression Line (*plug in the slope and Y-intercept but leave the X and Y in the formula*)

$$\hat{y} = (Y\text{-intercept}) + (\text{slope}) x$$

1.



$$r = 0.7287$$

	Mean	StDev
(x)Paper Trash	9.428	4.168
(y)Total Trash	27.44	12.46

Slope = \_\_\_\_\_

Sentence Explaining the Slope:

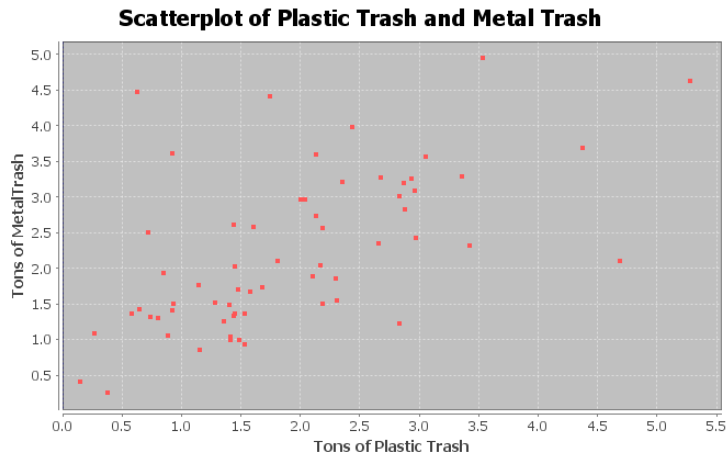
Y – Intercept = \_\_\_\_\_

Sentence Explaining the Y-intercept:

Equation of the Regression Line:  $Y = \text{_____} + \text{_____} X$



2.



$r = 0.5862$

	Mean	StDev
(x) Plastic Trash	1.911	1.065
(y) Metal Trash	2.218	1.091

Slope = \_\_\_\_\_

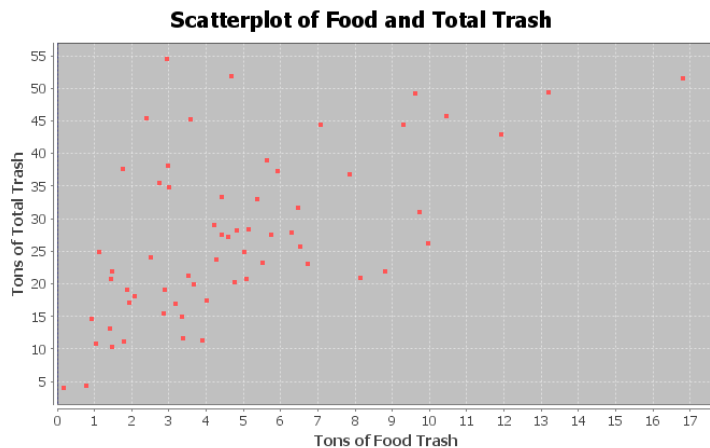
Sentence Explaining the Slope:

Y – Intercept = \_\_\_\_\_

Sentence Explaining the Y-intercept:

Equation of the Regression Line:  $Y = \text{_____} + \text{_____} X$

3.



$r = 0.5833$

	Mean	StDev
(x)Food Trash	4.816	3.297
(y)Total Trash	27.44	12.46

Slope = \_\_\_\_\_

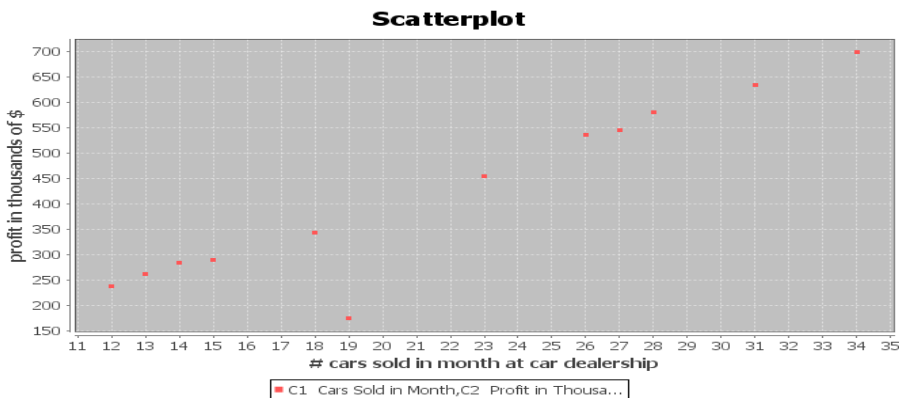
Sentence Explaining the Slope:

Y – Intercept = \_\_\_\_\_

Sentence Explaining the Y-intercept:

Equation of the Regression Line:  $Y = \text{_____} + \text{_____} X$

4. This data describes the relationship between the number of cars sold and total profit in thousands of dollars.



$r = 0.9404$

	Mean	Standard Deviation
(x) Cars Sold in Month	21.667	7.512
(y) Profit in Thousands of Dollars	420.25	175.615

Slope = \_\_\_\_\_

Sentence Explaining the Slope:

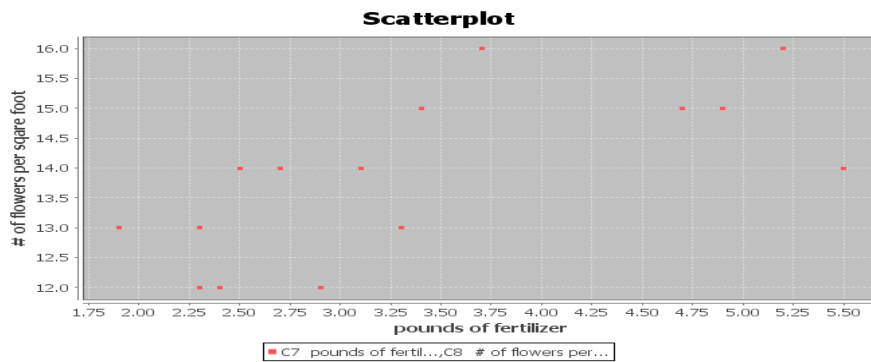
Y – Intercept = \_\_\_\_\_

Sentence Explaining the Y-intercept:

Equation of the Regression Line:  $Y = \text{_____} + \text{_____} X$



5. This data describes the relationship between the number of pounds of fertilizer used and the number of flowers per square foot.



$r = 0.6727$

Variable	Mean	Standard Deviation
(x) pounds of fertilizer used	3.387	1.165
(y) # of flowers per sq. ft.	13.867	1.356

Slope = \_\_\_\_\_

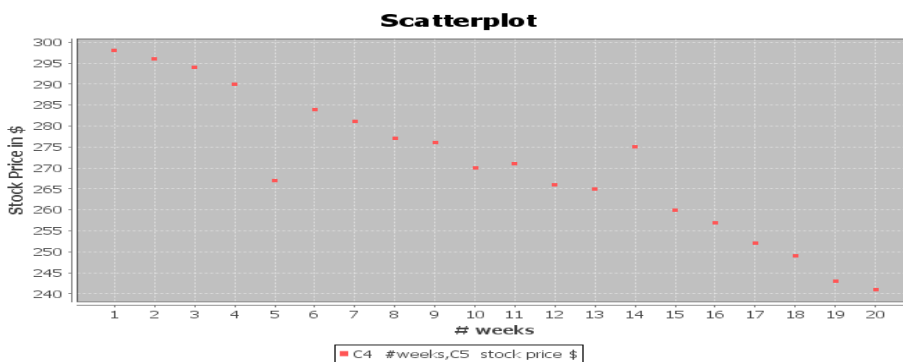
Sentence Explaining the Slope:

Y – Intercept = \_\_\_\_\_

Sentence Explaining the Y-intercept:

Equation of the Regression Line:  $Y = \text{_____} + \text{_____} X$

6. The following data describes the Price of a stock over the first 20 weeks of this year.



$r = -0.9429$



Variable	Mean	Standard Deviation
(x) #weeks	10.5	5.916
(y) stock price \$	270.6	17.031

Slope = \_\_\_\_\_

Sentence Explaining the Slope:

Y – Intercept = \_\_\_\_\_

Sentence Explaining the Y-intercept:

Equation of the Regression Line:  $Y = \text{_____} + \text{_____} X$

(#7-12) Directions: For the following problems, use the indicated data and StatKey to calculate the correlation coefficient “r” and the slope and y-intercept of the regression line. Then answer the questions.

Calculating the slope and y-intercept of the regression line with StatKey:

- To put the data into StatKey, you will want to open a fresh excel spreadsheet and paste the two quantitative data sets next to each other side by side. Then copy the two columns together.
- Now we will go to [www.lock5stat.com](http://www.lock5stat.com) and click on “StatKey”. Under the “Descriptive Statistics and Graphs” menu, click on “Two Quantitative Variables”. Click on “Edit Data” at the top. Push Control A on your keyboard to highlight old data and then push “delete” on your keyboard to delete all old data in the edit data field. Then paste the two columns of quantitative data into the “Edit Data” field. If your data has a title, click the box that says “Data has header row”. If your data does not have a title, do NOT check the box that says “Data has header row”. Then press OK.
- You will see the slope and y-intercept of the regression line under “Summary Statistics”. Look next to “Slope” and “Intercept”.
- The correlation coefficient “r” will be listed under “Summary Statistics”. Look under “Correlation”.
- To see the regression line on the scatterplot, check the box that says “Show Regression Line”.

7. Open the “Cigarette Data” from Canvas or from [www.matt-teachout.org](http://www.matt-teachout.org). Explore the relationship between mg of nicotine and mg of tar in cigarettes.

- Let nicotine be the explanatory variable and tar be the response variable. Create a scatter plot with the regression line and find the correlation coefficient in order to verify correlation between the variables. You do NOT need to save or copy the scatterplot. Give the r-value and describe the strength and direction of the linear relationship. (*This tells us how well the line fits the data.*)
- What is the Y-intercept of the regression line? Write a sentence interpreting the meaning of the Y-intercept using the units of the explanatory and response variable.
- What is the slope of the regression line? Write a sentence interpreting the meaning of the slope using the units of the explanatory and response variable.
- What is the equation of the regression line?



8. Open the “Cigarette Data” from Canvas or from [www.matt-teachout.org](http://www.matt-teachout.org). Explore the relationship between mg of nicotine and the carbon monoxide (CO) (parts per million PPM) in cigarettes.

a) Let nicotine be the explanatory variable and CO be the response variable. Create a scatter plot with the regression line and find the correlation coefficient in order to verify correlation between the variables. You do NOT need to save or copy the scatterplot. Give the r-value and describe the strength and direction of the linear relationship. (*This tells us how well the line fits the data.*)

b) What is the Y-intercept of the regression line? Write a sentence interpreting the meaning of the Y-intercept using the units of the explanatory and response variable.

c) What is the slope of the regression line? Write a sentence interpreting the meaning of the slope using the units of the explanatory and response variable.

d) What is the equation of the regression line?

9. Open the “Health Data” from Canvas or from [www.matt-teachout.org](http://www.matt-teachout.org). Explore the relationship between a woman’s waist size in cm and her body mass index (BMI) in  $\text{kg per m}^2$ .

a) Let waist size be the explanatory variable and body mass index (BMI) be the response variable. Create a scatter plot with the regression line and find the correlation coefficient in order to verify correlation between the variables. You do NOT need to save or copy the scatterplot. Give the r-value and describe the strength and direction of the linear relationship. (*This tells us how well the line fits the data.*)

b) What is the Y-intercept of the regression line? Write a sentence interpreting the meaning of the Y-intercept using the units of the explanatory and response variable.

c) What is the slope of the regression line? Write a sentence interpreting the meaning of the slope using the units of the explanatory and response variable.

d) What is the equation of the regression line?

10. Open the “Health Data” from Canvas or from [www.matt-teachout.org](http://www.matt-teachout.org). Explore the relationship between a woman’s systolic blood pressure and her diastolic blood pressure.

a) Let systolic blood pressure be the explanatory variable and diastolic blood pressure be the response variable. Create a scatter plot with the regression line and find the correlation coefficient in order to verify correlation between the variables. You do NOT need to save or copy the scatterplot. Give the r-value and describe the strength and direction of the linear relationship. (*This tells us how well the line fits the data.*)

b) What is the Y-intercept of the regression line? Write a sentence interpreting the meaning of the Y-intercept using the units of the explanatory and response variable.

c) What is the slope of the regression line? Write a sentence interpreting the meaning of the slope using the units of the explanatory and response variable.

d) What is the equation of the regression line?

11. Open the “Bear Data” from Canvas or from [www.matt-teachout.org](http://www.matt-teachout.org). Explore the relationship between the length of a bear in inches and the weight of the bear in pounds.

a) Let length be the explanatory variable and weight be the response variable. Create a scatter plot with the regression line and find the correlation coefficient in order to verify correlation between the variables. You do NOT need to save or copy the scatterplot. Give the r-value and describe the strength and direction of the linear relationship. (*This tells us how well the line fits the data.*)

b) What is the Y-intercept of the regression line? Write a sentence interpreting the meaning of the Y-intercept using the units of the explanatory and response variable.





- c) What is the slope of the regression line? Write a sentence interpreting the meaning of the slope using the units of the explanatory and response variable.
- d) What is the equation of the regression line?

12. Open the “Bear Data” from Canvas or from [www.matt-teachout.org](http://www.matt-teachout.org). Explore the relationship between the head length of a bear in inches and the head width of the bear in inches.

- a) Let the head width be the explanatory variable and head length be the response variable. Create a scatter plot with the regression line and find the correlation coefficient in order to verify correlation between the variables. You do NOT need to save or copy the scatterplot. Give the r-value and describe the strength and direction of the linear relationship. (*This tells us how well the line fits the data.*)
  - b) What is the Y-intercept of the regression line? Write a sentence interpreting the meaning of the Y-intercept using the units of the explanatory and response variable.
  - c) What is the slope of the regression line? Write a sentence interpreting the meaning of the slope using the units of the explanatory and response variable.
  - d) What is the equation of the regression line?
- 



## Section 6E – Residuals, Standard Deviation of the Residual Errors ( $S_e$ ), Residual Plots and Histogram of the Residuals

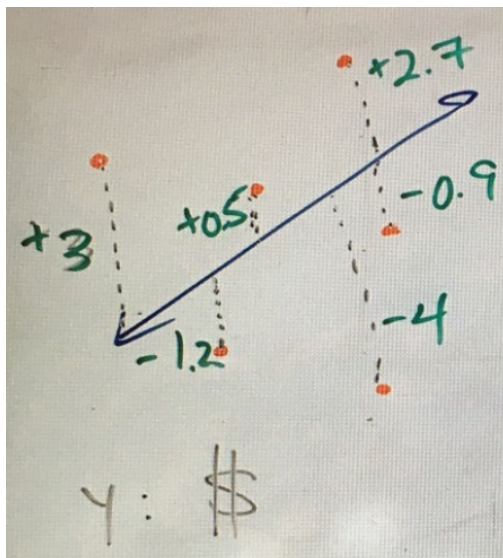
**Note about section 6E:** StatKey does not calculate residual plots, histogram of the residuals, or the standard deviation of the residual errors. We will focus on interpreting residuals in this section. Students will not be asked to calculate them with computer software.

Statisticians often look deeper when they study quantitative relationships. One topic that is often explored is the study of “residuals”.

**Definition of Residual:** A “residual” or “residual error” is a measure of the vertical distance that each point in the scatterplot is above or below the line. It measures the difference between predicted Y values from the regression line and the actual Y values in the response data. If the residual is positive, then the point is above the line. If the residual is negative then the point is below the line.

### Notes about Residuals

- If the residual is positive, then the point is above the line. If the residual is negative then the point is below the line.
- The residual measures vertical distance to the line.
- The units of the residual are always the same as the response variable (Y).
- Since the regression line itself is sometimes used to make predictions, the residuals measure the amount of prediction error for each X value in the explanatory data. That is why residuals are often called “residual errors”.
- If the residual is positive, then the point is above the regression line. The line is where the predicted values are. Therefore, a positive residual tells us the line be beneath the actual point meaning that the prediction for that particular x value is too low.
- If the residual is negative, then the point is below the regression line. The line is where the predicted values are. Therefore, a negative residual tells us the line be above the actual point meaning that the prediction for that particular x value is too high.



The picture above shows the idea of residuals. In this example, the response variable was in dollars so all of the residual errors are in dollars. A point that has a residual of +3 means the point was 3 dollars above the regression line. A point that has a residual of -4 means that the point was 4 dollars below the regression line. What does this tell us about the predicted values on the line itself? The +3 residual tells us the point is 3 dollars above the line and



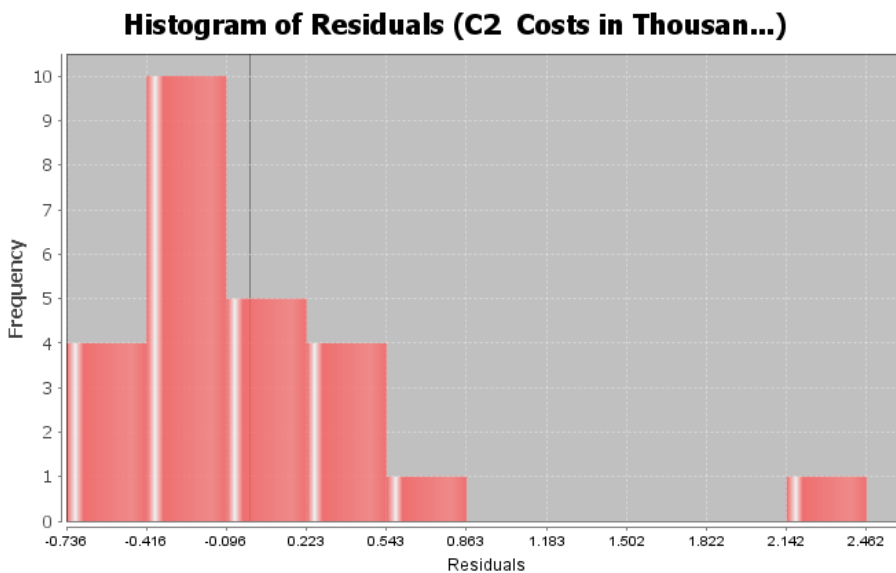
that the line is 3 dollars below the point. That means the predicted value is 3 dollars too low. The -4 residual tells us the point is 4 dollars below the line and that the line is 4 dollars above the point. That means the predicted value is 4 dollars too high.

### Histogram of the Residuals

We saw at the end of the last section that since the regression line is based on the mean and standard deviations of the quantitative data sets, we need to check if the data is bell shaped. What we really need to check is to see if the “residuals” are bell shaped. To that end, a graph that is often looked at is the “histogram of the residuals”.

#### Example 1

In this chapter, we have been looking at the months in business and cost data in thousands of dollars. Using the steps above, I made the following histogram of the residuals for the month and cost data.



### Interpreting a Histogram of the Residual Errors

There are two things we like to check when we look at the histogram of the residual errors. As we have said before, we want to check to see if the data is close to bell shaped. It does not have to be a perfect bell shape, but it should not be radically skewed. The other factor is that that we want the center to be close to zero. The dark vertical line in the histogram is a marker for zero. An easy way to check this requirement is to see if the zero line is close to the highest bar. When the residuals are not bell shaped or if the graph is not centered at zero, then our regression line will not be as accurate as we think.

#### Two Requirements to Check when looking at a histogram of the residual errors:

1. The histogram should be close to bell shaped and not radically skewed.
2. The histogram should be centered close to zero. The zero line should coincide with the highest bar.

Interpretation of the histogram of the residuals: In the histogram shown above describing the residuals for the cost data, we see that the graph is not very bell shaped. In fact, it is skewed right. We also see that the zero line is a little off from the highest bar. This tells us the graph is not centered at zero as well as we would like. This would be a red flag for a statistician or data analyst that the formula for the regression line is not as accurate as we think.

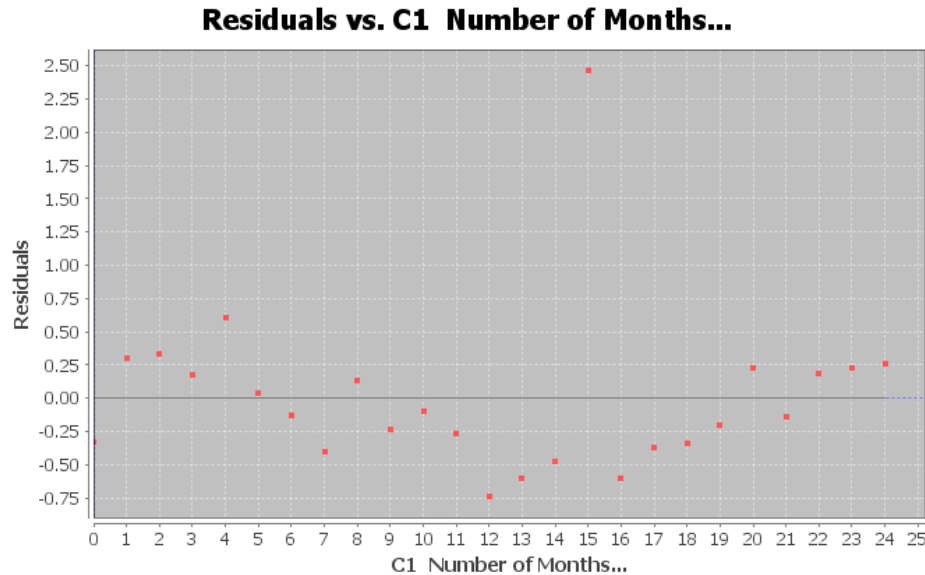
As with most things in statistics, there is a lot of grey area. In this last example, the zero line was not dramatically off from the highest hill.



### Residual Plot versus the X Variable

Another graph that statisticians often look at is the “residual plot”. A residual plot is a graph of the residuals showing how far each point is from the regression line. Residuals are positive if the point is above the regression line and negative if the point is below the regression line. Therefore, in the residual plot, you will see negative and positive numbers. The horizontal zero line represents the regression line itself. Though there are many more advanced types of residual plots, we will focus on the “residual plot versus the X variable”. This graph shows the residuals with the original explanatory variable as the X-axis.

Here is a residual plot versus the X variable for the month and cost data again.



Let us see if we can understand what we are looking at. The horizontal line at zero represents the regression line itself. Notice the vertical scale on the left is no longer the same as the scatterplot. It has positive numbers above zero and negative numbers below zero. The units for the vertical access are still the same as Y variable (costs in thousands of dollars), but now it is showing the residual (how far each point is above or below the regression line). The X-axis is the same as the scatterplot, showing the number of months the company has been in business.

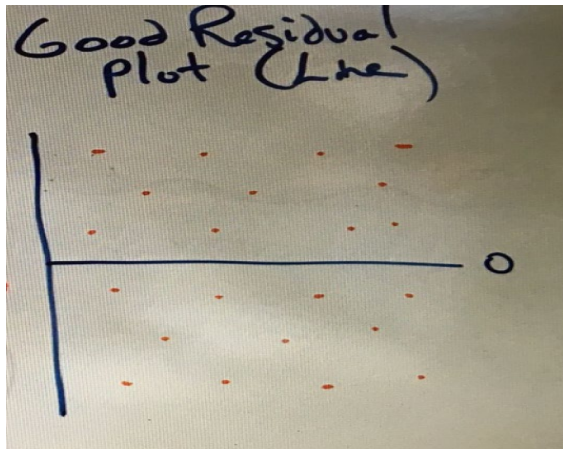
### Interpreting a Residual Plot

There are many things statisticians look at when they study residual plots. We will focus on two in particular. The first is to look and see if the points are evenly spread out from zero line. We do not want to see a “V” or “fan” shape in the residual plot.

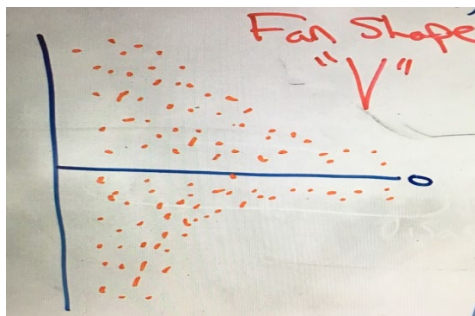
Good Residual Plot (Evenly Spread Out)



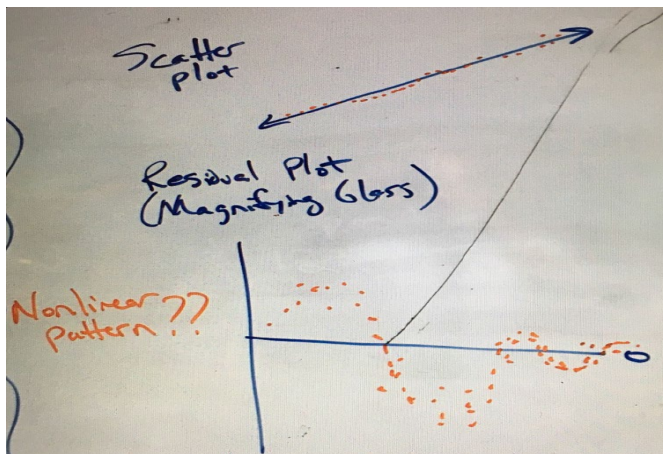
This material is from *Introduction to Data Analysis*, 2<sup>nd</sup> edition by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a “CC-BY” [Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by/4.0/) – 3/17/2021



Bad Residual Plot ("V" Shape, Not evenly spread out)



You can also look for curved (nonlinear) patterns in the residual plot. I like to think of residual plots as a magnifying glass. Points in a scatterplot can often look so small that it is difficult to see patterns in the data other than just the line. You can see the distances really well in a residual plot, which I find makes it easier to see curved relationships.



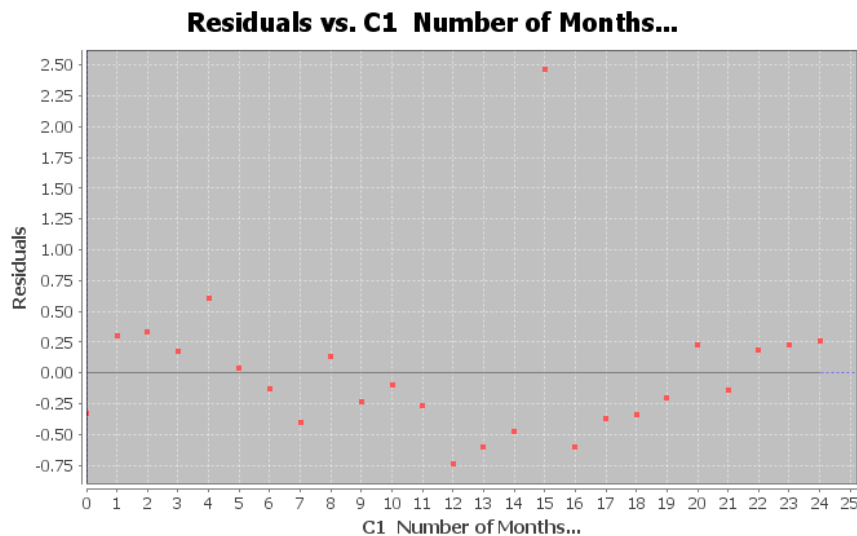
Note about terminology: The term "nonlinear" can refer to a curved pattern in the data, but can be misleading. Statisticians include many curved patterns under the heading of "linear regression" since they focus on the study of transformations from curves to lines.



### Interpreting a Residual Plot

- The points should be evenly spread out from the zero line. You should not see a “V” shape.
- Look for curved patterns in the data that may not have been apparent in the scatterplot. If we see a curved pattern, we may want to use some type of regression curve, instead of the regression line.

Interpretation of the residual plot: Here is the residual plot again for the month and cost data. Let us see if we can see any key features from this graph.



Notice we can see the outlier at 15 months, which is close to 2.5 thousand higher than what the regression line might predict. You may wish to set that outlier apart. It was an unusual event where the company had to purchase replacement equipment.

*Note: You have to be careful judging outliers from residual plots. The magnifying glass of a residual plot tends to make a lot of the ordered pairs look like outliers. I prefer to judge possible outliers with the scatterplot and correlation coefficient  $r$ .*

Note the curved “U” shaped pattern in the residual plot. This is an indication that a quadratic curve (parabola) may be a better fit for the data than the line was.

How about the even spread requirement? If we take the outlier out of the data. The rest of the dots have a pretty even spread from the zero line. We do not see any “V” shape.

### Standard Deviation of the Residual Errors (Se)

Recall that the standard deviation of a single quantitative data set is a statistic that tells us how far typical values in the data set are from the mean in bell shaped (normal) data. Standard deviation is used in many different contexts in statistics. In regression theory, we can calculate the how far typical points are from the line or curve. We call this the “standard deviation of the residuals” or the “standard deviation of the residual errors”.

Here is the standard deviation of the residuals for the month and cost data.

Standard Deviation of the Residual Errors ( $S_e$ ) = 0.6295

### Interpreting the Standard Deviation of the Residual Errors

The standard deviation of the residual errors tells us two important things. Like the standard deviation for a single data set, the standard deviation for the residuals tells us how far typical points are from the regression line on average. Think of it as a measure of the average distance from the line. If there is correlation, we can use the regression line as a formula to make predictions. Therefore, the standard deviation gives us a measure of the



This material is from *Introduction to Data Analysis*, 2<sup>nd</sup> edition by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a “CC-BY” [Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by/4.0/) – 3/17/2021

average prediction error. If we use the regression line to make a prediction, then the standard deviation of the residuals tells us how far off the prediction might be on average.

1. The average distance that the points are from the regression line.
2. The average amount of prediction error.

#### Notes about Standard Deviation of the Residuals

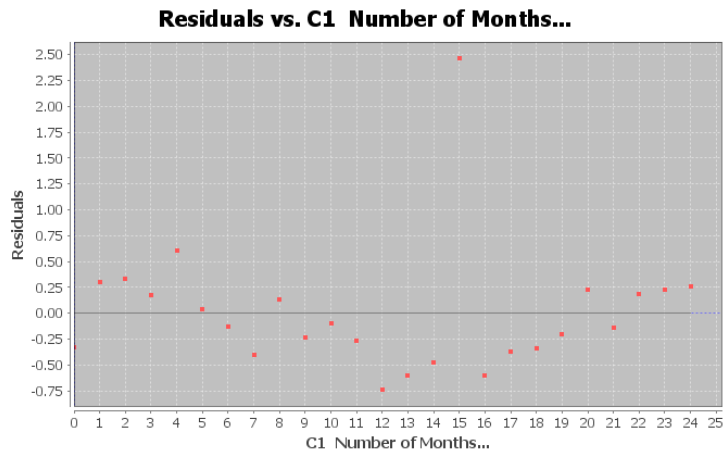
- *Standard deviation is a measure of spread for bell shaped (normal) data sets. If the histogram of the residuals is not bell shaped, then the standard deviation is not as accurate.*
- *Do not confuse the standard deviation of the X values, the standard deviation of the Y values and the standard deviation of the residual errors. These are three different standard deviations that measure different things. The standard deviation of the Y values is how far typical values are on average from the mean of the Y values (response variable). The standard deviation of the X values is how far typical values are on average from the mean of the X values (explanatory variable). The standard deviation of the residual errors measures how far typical points in the scatterplot are on average from the regression line.*
- *Terminology: Some shorten the name for the “Standard Deviation of the Residual Errors” to just “Standard Error”. This can be confusing, because the term “standard error” is used to describe the standard deviation of a sampling distribution.*

Example 1 continued

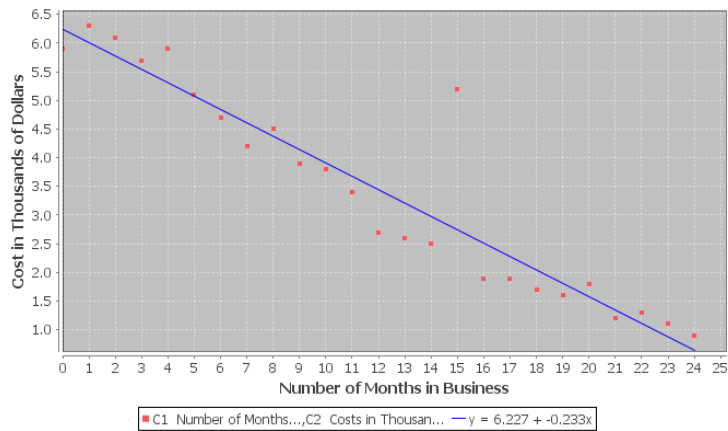
Here again is the standard deviation of the residuals for the month and cost data.

Standard error of estimate = 0.6295

What does this tell us about the data? It is good to look at the scatterplot and residual plot to give us a visual.



**Scatterplot of Month and Cost Data**



The standard deviation of the residuals always has the same units as the Y variable. If you look at the vertical axis on the scatterplot, you can see that the Y variable is the cost in thousands of dollars. So the standard deviation = 0.6295 thousand dollars (or \$629.50).

The standard deviation tells us two important things in this example. The first is that the points in the scatterplot are on average about 0.6295 thousand dollars from the line. The second is that the average prediction error for this regression line will be about 0.6295 thousand dollars (or \$629.50). If we use the regression line to make a prediction in the scope of the X values, our prediction could be off by about \$629.50 on average.

Remember that the accuracy of the regression line and the standard deviation of the residuals is tied to the residual plot being bell shaped (normal). In this example, the histogram of the residuals was skewed right. This tells us that the regression line and standard deviation will not be quite as accurate.

---



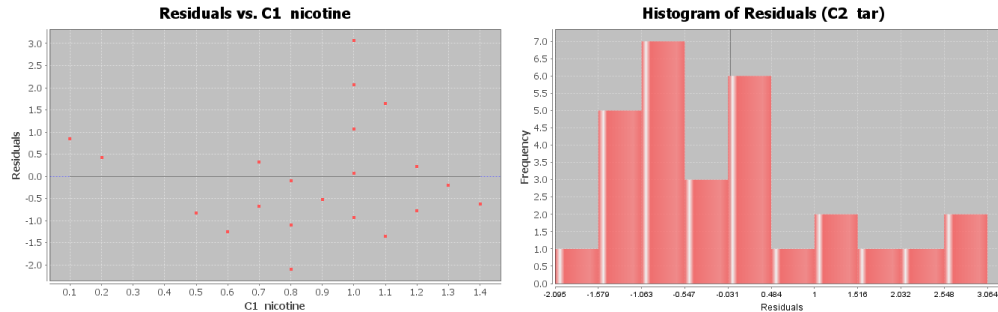


## Problem Set Section 6E

**Note about the problems in section 6E: StatKey does not calculate residual plots, histogram of the residuals, or the standard deviation of the residual errors. We will focus on interpreting residuals in this section. Students will not be asked to calculate them with computer software.**

1. Mg of nicotine is the explanatory variable and mg of tar is the response variable. Use the given residual plot versus the x-variable, histogram of the residuals, and the standard deviation of the residual errors ( $S_e$ ) to answer the following questions.

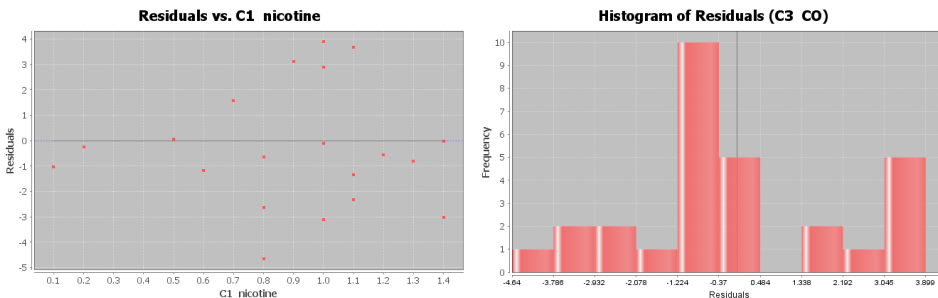
$$S_e = 1.2984 \text{ mg of tar}$$



- Is the histogram bell shaped? If not, what shape is it? Is the histogram centered close to zero?
- Was the points in the residual plot evenly spaced out or was there a sideways "V" shape? Did you see a curved pattern in the residual plot? If so, describe the shape of the curved pattern ("U" shaped or "S" shaped for example).
- Interpret the standard deviation in context by describing the two meanings of the standard deviation.

2. Mg of nicotine is the explanatory variable and mg of carbon monoxide (CO) is the response variable. Use the given residual plot versus the x-variable, histogram of the residuals, and the standard deviation of the residual errors ( $S_e$ ) to answer the following questions.

$$S_e = 2.2961 \text{ PPM}$$

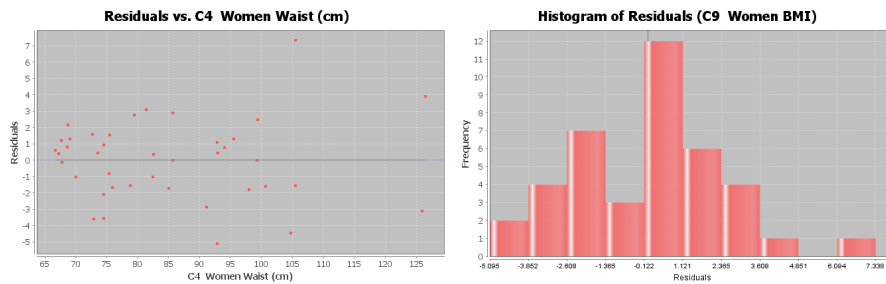


- Is the histogram bell shaped? If not, what shape is it? Is the histogram centered close to zero?
- Was the points in the residual plot evenly spaced out or was there a sideways "V" shape? Did you see a curved pattern in the residual plot? If so, describe the shape of the curved pattern ("U" shaped or "S" shaped for example).
- Interpret the standard deviation in context by describing the two meanings of the standard deviation.



3. Women's waist size is the explanatory variable and women's body mass index (BMI) is the response variable. Use the given residual plot versus the x-variable, histogram of the residuals, and the standard deviation of the residual errors ( $S_e$ ) to answer the following questions.

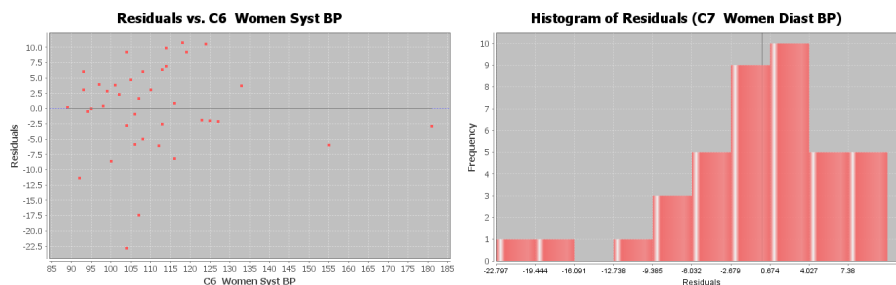
$$S_e = 2.4761 \text{ kg/m}^2$$



- Is the histogram bell shaped? If not, what shape is it? Is the histogram centered close to zero?
- Was the points in the residual plot evenly spaced out or was there a sideways "V" shape? Did you see a curved pattern in the residual plot? If so, describe the shape of the curved pattern ("U" shaped or "S" shaped for example).
- Interpret the standard deviation in context by describing the two meanings of the standard deviation.

4. Women's systolic blood pressure is the explanatory variable and women's diastolic blood pressure is the response variable. Use the given residual plot versus the x-variable, histogram of the residuals, and the standard deviation of the residual errors ( $S_e$ ) to answer the following questions.

$$S_e = 7.2912 \text{ mm of Hg}$$

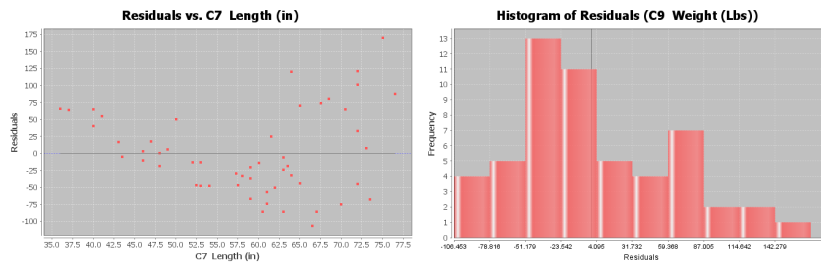


- Is the histogram bell shaped? If not, what shape is it? Is the histogram centered close to zero?
- Was the points in the residual plot evenly spaced out or was there a sideways "V" shape? Did you see a curved pattern in the residual plot? If so, describe the shape of the curved pattern ("U" shaped or "S" shaped for example).
- Interpret the standard deviation in context by describing the two meanings of the standard deviation.



5. Bear length in inches is the explanatory variable and bear weight in pounds the response variable. Use the given residual plot versus the x-variable, histogram of the residuals, and the standard deviation of the residual errors ( $S_e$ ) to answer the following questions.

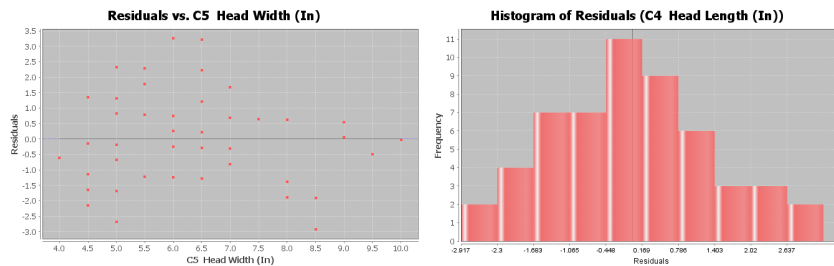
$$S_e = 61.8272 \text{ pounds}$$



- Is the histogram bell shaped? If not, what shape is it? Is the histogram centered close to zero?
- Was the points in the residual plot evenly spaced out or was there a sideways "V" shape? Did you see a curved pattern in the residual plot? If so, describe the shape of the curved pattern ("U" shaped or "S" shaped for example).
- Interpret the standard deviation in context by describing the two meanings of the standard deviation.

6. Bear head width in inches is the explanatory variable and bear head length in inches is the response variable. Use the given residual plot versus the x-variable, histogram of the residuals, and the standard deviation of the residual errors ( $S_e$ ) to answer the following questions.

$$S_e = 1.4231 \text{ inches}$$



- Is the histogram bell shaped? If not, what shape is it? Is the histogram centered close to zero?
- Was the points in the residual plot evenly spaced out or was there a sideways "V" shape? Did you see a curved pattern in the residual plot? If so, describe the shape of the curved pattern ("U" shaped or "S" shaped for example).
- Interpret the standard deviation in context by describing the two meanings of the standard deviation.



## Section 6F – Predictions, Scope of the X-values, Extrapolation, and Using the Standard Deviation of the Residual Errors

**Note about section 6F:** StatKey can calculate the scatterplot, the correlation coefficient, and the slope and y-intercept of the regression line. We expect students to be able to use StatKey to calculate these. However, StatKey does not calculate the standard deviation of the residual errors. We will focus on interpreting the standard deviation of the residual errors and students will not be asked to calculate it.

In this chapter, we have seen that we can use scatterplots,  $r$ , and  $r$ -squared to analyze and measure linear relationships (correlation) between two different quantitative variables. We also found that if there is a linear relationship, we could find the line of best fit (regression line).

If there is a correlation between the variables, then we can use the regression line to predict  $Y$  values. In this section, we will look at how to make predictions and guidelines for interpreting those predictions.

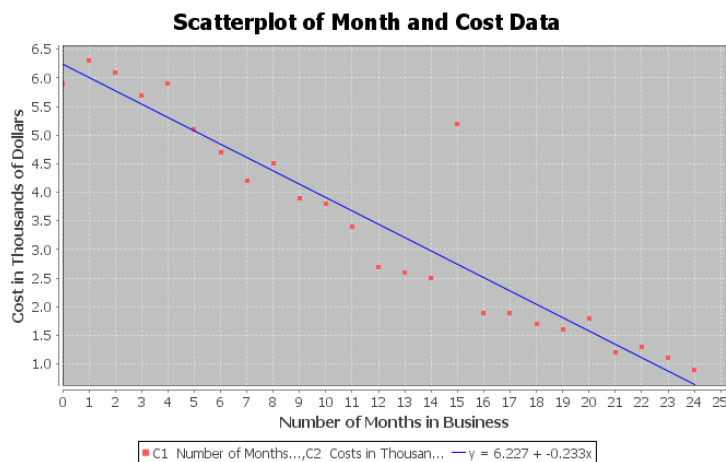
### Notes about Making Predictions with the Regression Line

#### 1. There should be some correlation between the variables.

If there is no linear relationship between the variables, then the regression line does not fit the data, meaning our predictions will not be accurate. Always check correlation with the scatterplot and the correlation coefficient " $r$ " before using the regression line to make a prediction.

#### 2. The scope of the X values and "Extrapolation"

In general, we like to use  $X$  values in the scope of the data. What do we mean by this? Look at the following scatterplot of the month and cost data.



Notice that the  $X$ -axis of the scatterplot represents the number of months the company has been in business. The  $X$  values go from 0 to 24 months. This is called the "scope of the  $X$  values" or the "scope of the data". Recall that this regression line fits the data really well and had a strong linear relationship with a high  $r$ -squared value. This tells us the regression line should be pretty accurate for predicting costs. The thing to remember is that this regression line and the standard deviation of the residuals (prediction error) was based on the  $X$  values of 0 to 24 months. Plugging in  $X$  values between 0 and 24 will give pretty accurate predictions.

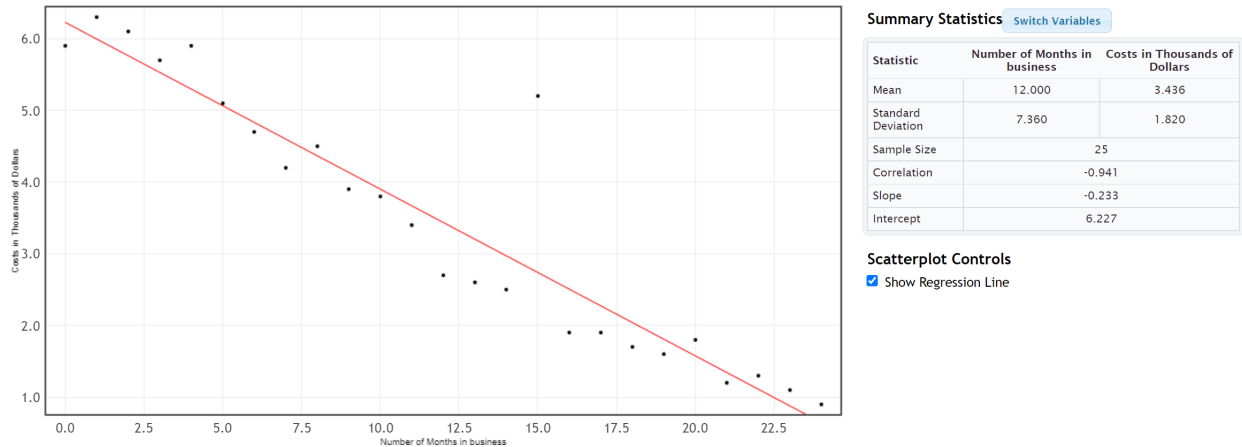
If you plug in values outside of the scope, you are using the formula for something it was not designed for. Many people love to use regression to make predictions about the future. Remember, there is no guarantee that the data will follow this pattern into the future. Also the standard deviation no longer applies to predictions outside the scope, so we cannot measure how much prediction error we may have.



**Definition of “Extrapolation”:** Plugging in an X value outside the scope of the data into a regression line or curve in order to make a prediction. The predictions outside the scope may have a significant increase in error and we will be unable to measure the error.

Extrapolating outside the scope of the X values can lead to large errors in your prediction. Also, remember the standard deviation prediction error only applies if your X value is in the scope of the data.

For example, let us look at scatterplot and statistics for the time (months) and costs (thousands of dollars) for a company.



The average decrease in costs per month (slope) is 0.233 thousand dollars. At this rate, the line will begin to predict negative costs for the company, something that is very unlikely to happen. In fact, if we extrapolate and plug in 28 months for X into the formula, we would get a prediction of about -0.3 thousand dollars (negative \$300). The costs of the company will likely not drop to negative \$300.

This problem with extrapolation is a good reason why we need to recalculate with new data every few years.

I will not say never extrapolate. Many data analysts extrapolate. People are always interested in what the data tells us about the future. I would say if you do extrapolate, do not extrapolate too much (excessive extrapolation). In the last example, we may like to extrapolate a little and predict the costs in month 25, but I would not use this equation to predict the costs in month 48.

Keep the scope of the X values in mind whenever you are making predictions with a regression line. Remember if you do extrapolate, proceed with caution. You may be telling someone a predicted Y value that is very wrong.

*Note about the Y-intercept: The Y-intercept of the regression line is the predicted Y value when X is zero. Therefore, the Y intercept is the prediction you would get if you plug in zero for X in the formula. In the last section, we said that the Y-intercept often does not make sense when we try to explain it. The reason for this is that sometime zero is an extrapolation. Zero may not be in the scope of the X values. In other words, if zero is not in the scope, then the formula was never designed for you to plug in zero for X. Therefore, it is not surprising that the Y-intercept value does not make much sense in context. Extrapolations can give very unusual answers sometimes.*

### 3. Making the Prediction

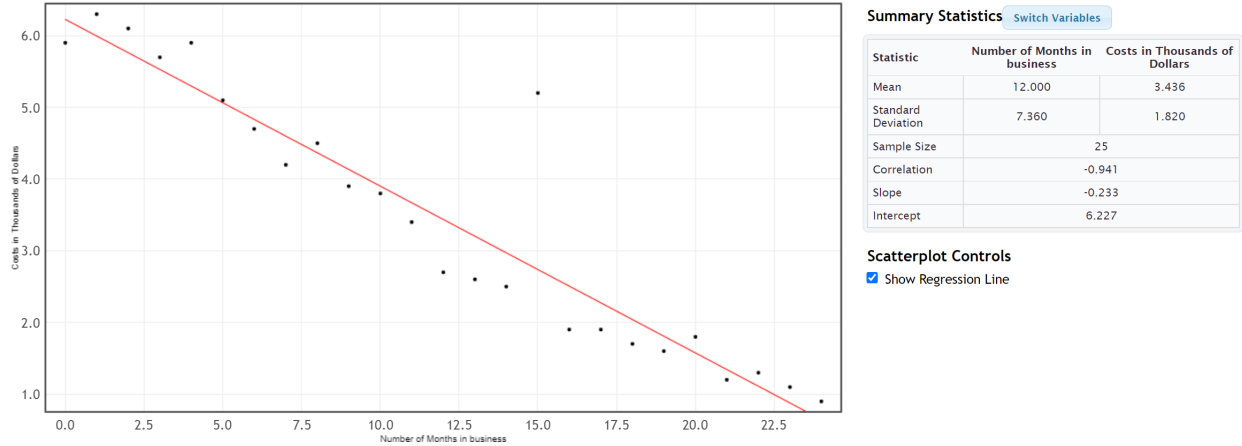
To make the prediction, plug in an X value into the regression line equation for X and use order of operations and a calculator to calculate the corresponding Y value. This Y value is the prediction. Many statistics software programs have prediction functions where it will calculate the prediction automatically. Unfortunately, StatKey does not have the prediction function at this time.

Make sure to follow the order of operations when you make your prediction. Multiply the X value by the slope first. Then add the Y-intercept. Be careful of negative number mistakes.



## Example 1

Let us look at scatterplot and statistics for the the time (months) and costs (thousands of dollars) for a company. Recall that in month 15, the company had an unusually high cost due to having to buy some replacement parts. Use the regression line to predict the average costs of the company in month 15 if they had not had to replace those parts.



First notice that the regression line does fit the data in scatterplot and the correlation coefficient  $r$  is close to  $-1$ . This means there is a strong (negative) correlation and the regression line formula is likely to be more accurate. First we need to plug in the slope and y-intercept into the regression line formula.

$$\hat{y} = \text{Y-intercept} + (\text{Slope}) x$$

$$\hat{y} = 6.227 + (-0.233) x$$

This formula can now be used to make predictions about  $y$ . The symbol  $\hat{y}$  means “predicted  $y$  value”. When we plug in a number for  $x$  and find  $\hat{y}$ , we are making a prediction based on data.

Remember, the  $X$  value is the month and  $Y$  value is the costs in thousands of dollars. Always keep your units in mind. To make the prediction, plug in 15 for  $X$  into the regression line formula. Remember to follow the order of operations and multiply first before adding. Be careful of making a calculation error with the negative numbers.

$$\hat{y} = 6.227 + (-0.233) x$$

$$\hat{y} = 6.227 + (-0.233) (15)$$

$$\hat{y} = 6.227 + (-3.495)$$

$$\hat{y} = +2.732 \text{ thousand dollars}$$

Therefore, if the company had not had to replace those parts, we predict their average costs would have been about 2.732 thousand dollars (\$2,732) in their 15<sup>th</sup> month in business.

*How much error could there be in that prediction?* 0.6295 thousand dollars (\$629.50)

Remember, the standard deviation of the residual errors tells us how much prediction error we have.

Standard Deviation of the Residual Errors = 0.6295

This is the average prediction error for any prediction in the scope of the  $X$  values. Month 15 was in the scope of the  $X$  values. The units of the standard deviation of the residual errors is the same as the predicted  $Y$  value (thousands of dollars).

So we predict that in month 15, the companies costs would have been about \$2732. This prediction could be off by about \$629.50 on average.



This material is from *Introduction to Data Analysis*, 2<sup>nd</sup> edition by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a “CC-BY” [Creative Commons Attribution 4.0 International License](#) – 3/17/2021

### Example 2

Can we use the month and cost regression line from Example 1 to predict the costs in month 50?

No. This is an extrapolation. Our prediction may be very off. Also, our standard deviation of the residual errors would not be accurate. We are out of the scope of the data.

---

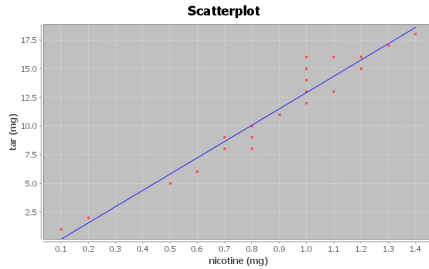


This material is from *Introduction to Data Analysis*, 2<sup>nd</sup> edition by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a “CC-BY” [Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by/4.0/) – 3/17/2021

## Problem Set Section 6F

*Directions: You will not need to use StatKey for these problems. Graphs and statistics were already calculated for you. Use the given scatterplot, correlation coefficient ( $r$ ), standard deviation of residual errors, and the equation of the regression line to answer the following questions.*

1. Explore the relationship between mg of nicotine and mg of tar in cigarettes. Let nicotine be the explanatory variable and tar be the response variable.

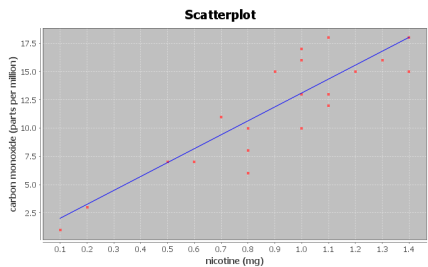


Correlation Coefficient  $r = 0.9614$

Regression Line Equation:  $Y = -1.2713 + 14.2076 X$

$s_e = 1.2984$  mg of tar

- Is there correlation between the variables? Explain why.
  - How well does the regression line fit the data? Do you think the regression line equation will be suitable to make predictions with? Why or why not?
  - What is the scope of the  $x$  values of the data? Does zero fall in that scope? What does that tell us about the  $y$ -intercept?
  - Provided the regression line is suitable for making predictions, predict the amount of tar we can expect to have in a cigarette that has 0.8 mg of nicotine. How far off on average could our prediction be?
  - One company is working on a cigarette with 4.75 mg of nicotine in it. Would it be ok to predict the amount of tar for this new cigarette? Why or why not?
  - Why do you think it is important that people know how much tar is in cigarettes?
2. Explore the relationship between mg of nicotine and carbon monoxide (CO) in part per million (ppm) in cigarettes. Let nicotine be the explanatory variable and CO be the response variable.



Correlation Coefficient  $r = 0.8633$

Regression Line Equation:  $Y = 0.7950 + 12.3057 X$

$s_e = 2.2961$  parts per million (ppm)

- Is there correlation between the variables? Explain why.

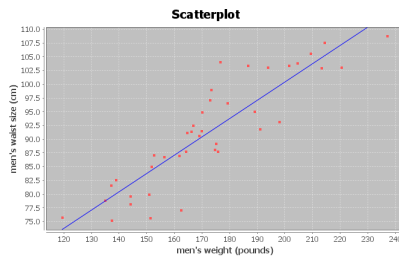


This material is from *Introduction to Data Analysis*, 2<sup>nd</sup> edition by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a “CC-BY” [Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by/4.0/) – 3/17/2021



- b) How well does the regression line fit the data? Do you think the regression line equation will be suitable to make predictions with? Why or why not?
- c) What is the scope of the x values of the data? Does zero fall in that scope? What does that tell us about the y-intercept?
- d) Provided the regression line is suitable for making predictions, predict the amount of Carbon Monoxide (CO) we can expect to have in a cigarette that has 1.2 mg of nicotine. How far off on average could our prediction be?
- e) One company is working on a cigarette with 4.75 mg of nicotine in it. Would it be ok to predict the amount of carbon monoxide released from this new cigarette? Why or why not?
- f) Why do you think it is important that people know how much carbon monoxide is released when a cigarette is smoked?

3. Explore the relationship between a man's weight in pounds and his waist size in inches. Let weight be the explanatory variable and waist size be the response variable.



Correlation Coefficient  $r = 0.8889$

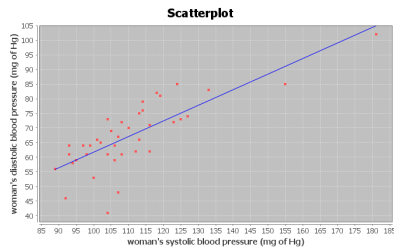
Regression Line Equation:  $Y = 33.8291 + 0.3330 X$

$s_e = 4.5763$  cm

- a) Is there correlation between the variables? Explain why.
- b) How well does the regression line fit the data? Do you think the regression line equation will be suitable to make predictions with? Why or why not?
- c) What is the scope of the x values of the data? Does zero fall in that scope? What does that tell us about the y-intercept?
- d) Provided the regression line is suitable for making predictions, predict the waist size of a man that weighs 200 pounds. How far off on average could our prediction be?
- e) Should we use the regression line equation to predict the waist size of a man that weighs 400 pounds? Why or why not?
- f) Can you think of any confounding variables that might influence waist size other than weight?



4. Explore the relationship between a woman's systolic blood pressure (mm of Hg) and her diastolic blood pressure (mm of Hg). Let systolic blood pressure be the explanatory variable and diastolic blood pressure be the response variable.



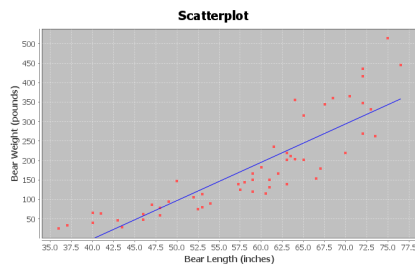
Correlation Coefficient  $r = 0.7854$

Regression Line Equation:  $Y = 8.3079 + 0.5335 X$

$s_e = 7.2912$  mm of Hg

- Is there correlation between the variables? Explain why.
- How well does the regression line fit the data? Do you think the regression line equation will be suitable to make predictions with? Why or why not?
- What is the scope of the x values of the data? Does zero fall in that scope? What does that tell us about the y-intercept?
- Provided the regression line is suitable for making predictions, predict the diastolic blood pressure of a person with a systolic blood pressure of 135. How far off on average could our prediction be?
- One woman with hypertension has a systolic blood pressure of 240. Would the regression line equation give an accurate prediction of her diastolic blood pressure? Why or why not?

5. Explore the relationship between the length of a bear in inches and the weight of the bear in pounds. Let length be the explanatory variable and weight be the response variable.



Correlation Coefficient  $r = 0.8644$

Regression Line Equation:  $Y = -393.8391 + 9.8390 X$

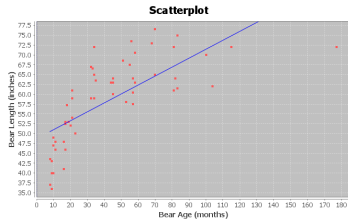
$s_e = 61.8272$  pounds

- Is there correlation between the variables? Explain why.
- How well does the regression line fit the data? Do you think the regression line equation will be suitable to make predictions with? Why or why not?
- What is the scope of the x values of the data? Does zero fall in that scope? What does that tell us about the y-intercept?



- d) Provided the regression line is suitable for making predictions, predict the weight of a bear that is 72 inches long. How far off on average could our prediction be?
- e) A young bear is only 18 inches long. Should we use the regression line equation to estimate the weight of this young bear? Why or why not?
- f) Why do you think it would be useful to a researcher to be able to estimate the weight of a bear in the wild by measuring its length?

6. Explore the relationship between the age of a bear in months and the length of the bear in inches. Let age be the explanatory variable and length be the response variable.



Correlation Coefficient  $r = 0.7188$

Regression Line Equation:  $Y = 48.6903 + 0.2281 X$

$s_e = 7.5109$  inches

- a) Is there correlation between the variables? Explain why.
- b) How well does the regression line fit the data? Do you think the regression line equation will be suitable to make predictions with? Why or why not?
- c) What is the scope of the x values of the data? Does zero fall in that scope? What does that tell us about the y-intercept?
- d) Provided the regression line is suitable for making predictions, predict the length of a bear that is 120 months old (10 years old). How far off on average could our prediction be?
- e) Will this formula give accurate predictions of the length of newborn bears? Why or why not?
- 



## Chapter 6 Review Sheet

In this chapter, we looked at finding a linear relationship (correlation) between two different quantitative variables with different units.

- When analyzing two different quantitative data sets, start by choosing one variable to be the response variable (Y) and the other to be the explanatory variable (X). In general, the response variable (Y) should respond to the explanatory variable (X). If both variables respond, choose the variable you are more interested in and want to make predictions about to be the response variable (Y).
  - The scatterplot and correlation coefficient “r” can tell us the strength of the linear relationship (strong, moderate, weak, or none) and the direction of the linear relationship (positive or negative). If r is close to +1, it is strong positive correlation. If r is close to -1, it is strong negative correlation. If r is close to zero, there is no correlation.
  - R-squared is the percentage of variability in the y variable that can be explained by the relationship with the x variable. The higher the percentage, the stronger the relationship.
  - Correlation is not causation. There are many other confounding variables that might influence the response variable other than the explanatory variable being studied.
  - The regression line is the line that best fits the data and minimizes the vertical distances from all the points in the scatterplot to the line.
  - Slope is the increase or decrease in the Y variables for every 1-unit increase in the X variable.
  - Y-intercept is the predicted Y value when X is zero.
  - The standard deviation of the residual errors (Se) gives the average vertical distance that the points are from the line. It also tells us the average prediction error.
  - Residuals are the vertical distance that each point is above or below the line. Points above the line have a positive residual. Points below the line have a negative residual.
  - A histogram of the residuals should be bell shaped (normal) and centered close to zero.
  - A residual plot versus the x variable should be evenly spread out from the zero line and not fan shaped (not “V” shaped).
  - To make a prediction with the regression line, first determine if the line fits the data. You should not make predictions when there is no correlation. If there is correlation, plug in the X value you want to predict in for X in the formula and solve for Y. The X value you plug in should be in the scope of the X values. Plugging in X values out of the scope is called “extrapolation” and can result in huge prediction errors.
- 

## Problems for Chapter 6 Review Sheet

1. Define each of the following.

- a) Explanatory variable
- b) Response variable
- c) Correlation Coefficient “r”
- d) r-squared
- e) slope
- f) y-intercept
- g) residual
- h) standard deviation of the residual errors.

2. When doing a correlation study with two quantitative variables, explain how we can tell which variable should be the explanatory variable (x) and the response variable (y).

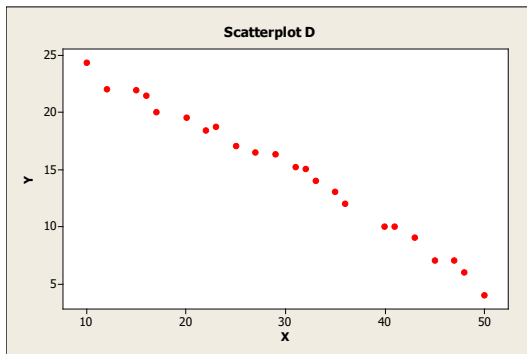
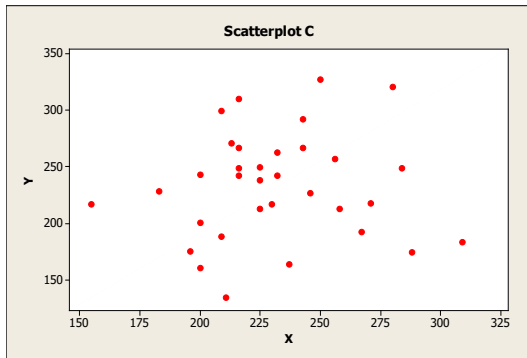


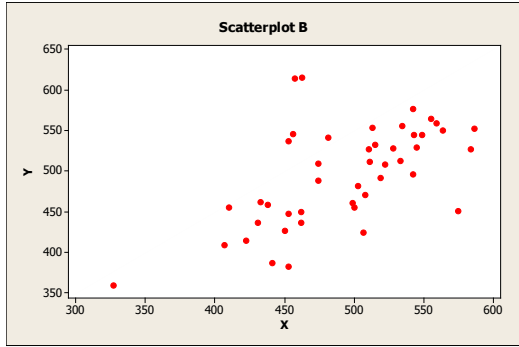
3. Sanvi is a medical student studying sleep patterns and migraine headaches. She is hoping to use the hours someone sleeps in order to predict the number of migraines. She went to a clinic that specializes in treating people who suffer with headaches and recorded some data. When a patient with migraines came into the clinic, Sanvi recorded the average number of migraines they have each month and how many hours per night the person sleeps on average.

- Which variable should be the explanatory variable (x)?  
(Number of migraines or hours of sleep)
- Which variable should be the response variable (y)?  
(Number of migraines or hours of sleep)
- Sanvi is hoping to prove that lack of sleep causes migraines. If the data showed a strong correlation between migraines and sleep, would this prove that lack of sleep causes people to get migraines? Why or why not?

4. Match the correlation coefficients (r) with their scatterplots. (Each r value corresponds to only one graph.) For each graph describing the strength and direction of the linear relationship (correlation).

$r = 0.592$ ,  $r = -0.993$ ,  $r = 0.023$





5. Use the following formulas to compute the slope and y-intercept for the regression line.  
Show your work and Round your answers to the hundredths place.

$$(r = 0.819)$$

	Mean	Standard Deviation
(x) Explanatory Variable	$\bar{x} = 19.18$	$s_x = 3.83$
(y) Response Variable	$\bar{y} = 82.55$	$s_y = 11.64$

a)

$$\text{slope: } m = \frac{r \cdot s_y}{s_x}$$

*To calculate the slope: r times standard deviation of y values, then divide by standard deviation of x values)*

Slope = \_\_\_\_\_

b)

$$\text{y-intercept: } b = \bar{y} - m(\bar{x})$$

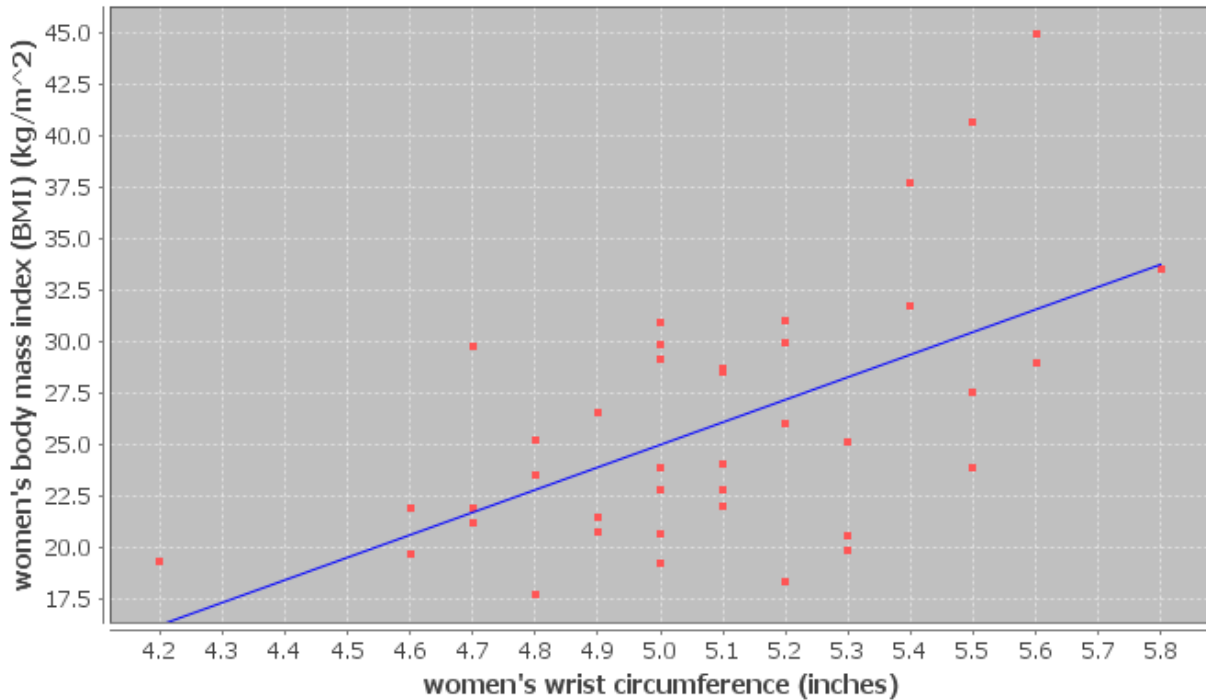
*To calculate the y – intercept (slope times the mean of x values, then subtract the answer from the mean of the y-values)*

Y – Intercept = \_\_\_\_\_



(#6-19) Let us look at the relationship between the wrist circumference of a woman (in inches) and her body mass index (BMI) in kg per meters squared. We used the health data and Statcato to create the following graphs and statistics. The explanatory variable (X) is the wrist circumference and the response variable (Y) is the body mass index.

### Scatterplot



x = Women Wrist (in)  
y = Women BMI (kg/m<sup>2</sup>)  
Sample size n = 40

#### Correlation

<b>r</b>	0.5870

#### Regression

Regression equation  $Y = b_0 + b_1x$

$b_0 = -29.7018$

$b_1 = 10.9407$

$r^2 = 0.3446$

Standard Deviation of the Residual Errors = 5.0568

6. Use the scatterplot and the correlation coefficient “r” to describe the strength and direction of the linear relationship.



7. Look at the scatterplot. Estimate the scope of the x-values and put your answer below. Just give approximate values.

\_\_\_\_\_  $\leq$  Wrist Circumference (Inches)  $\leq$  \_\_\_\_\_

8. What was the slope of the regression line? Write a sentence to explain the slope in context.

9. What was r-squared? Convert the r-squared value into a percentage.

10. Write a sentence to explain r-squared in context.

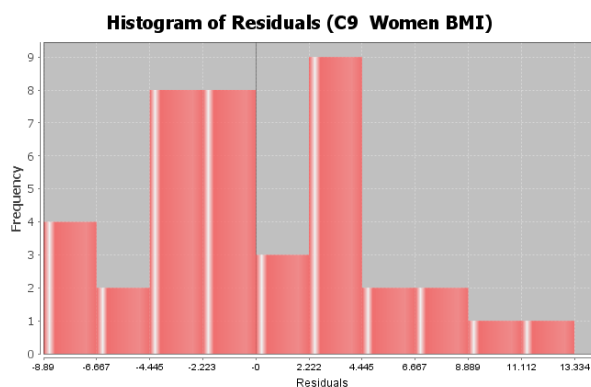
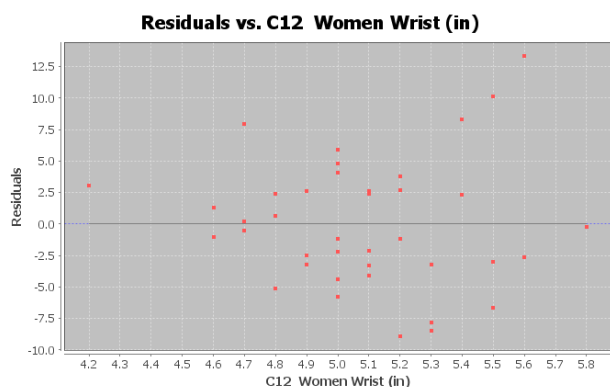
11. List three confounding variables that might influence the body mass index of a woman other than size of her wrist.

12. Does this study prove that the size of these women's wrist causes them to have a certain body mass index? (Yes or No) Explain why.

13. What was the standard deviation of the residual errors? What units does the standard deviation of the residuals have in this problem?

14. Explain the two meanings of the standard deviation of the residuals in the context of the wrist and BMI data.

Residual Plot and Histogram of the Residuals



15. Does the residual plot above on the left show a “V” shape or is it evenly spread out?

16. Does the residual plot above on the left show a curved pattern in the data?





17. Is the Histogram of the Residuals bell shaped? (Yes or No)

18. Is the histogram of the residuals centered close to zero? (Yes or No)

19. Use your calculator and the regression equation below to predict the body mass index for a woman with a wrist circumference of 4.5 inches. (Plug in 4.5 for X.) Show work.

$$Y = 48.802 - 8.367 X$$

20. In the previous problem, how far off could our BMI prediction be on average (prediction error)? (No calculation needed)

21. Will this formula give an accurate prediction for the body mass index for a female child with a wrist circumference of 3.1 inches? Why or why not?

---



## Chapter 6 Project

### Linear Quantitative Relationships

**Online Class Directions:** This will be an individual project. Each student is required to analyze a pair of quantitative data sets from the following topic list. You can find your two columns of data in the “Math 075 Project Data Correlation Regression”. You can find this data in Canvas or at [www.matt-teachout.org](http://www.matt-teachout.org). Put your two columns of data into StatKey to calculate, create a poster summarizing their findings, and email a picture of your poster to your instructor.

**Topics:** IQ/Brain Volume , MLB Runs Scored/Attendance , MLB Runs Allowed/Wins , Price Item/Customer Satisfaction , Meat/illness , CEO Golf Score/Stock Price , Swim time/Pulse , Boats/Manatee Deaths , Cost of Living/Aviation Pay , Poverty/BMI , Alcohol/Tobacco in England

- Pick one of the pairs of quantitative variables and pick which should be X and which should be Y. The poster should give the explanatory variable (x) and response variables (y) and the units for x and y.
- Why this topic was important or interesting to your group.
- Go to [www.lock5stat.com](http://www.lock5stat.com). Click on “Two quantitative variables” under the “descriptive statistics and graphs” menu. Click on “edit data” button in StatKey. Copy and paste the two columns together into StatKey. If your two columns have titles, click the box that says “data has a header row”. Push OK. Click the box that says “Show Regression Line”.
- Look at the scatterplot carefully. Make sure the x variable you picked is on the horizontal axis and the response variable (y) is on the vertical axis. If they are not push the “Switch Variables” button.
- Draw a rough sketch of the scatterplot on your poster. Label your axes and draw the regression line also.
- What is the correlation coefficient (r) for your data? You will find this where it says “Correlation” under “Summary Statistics”.
- Use the r-value to classify the linear relationship as one of the following: strong positive correlation, strong negative correlation, moderate positive correlation, moderate negative correlation, weak positive correlation, weak negative correlation, or no correlation.
- Square your r-value for multiply  $r \times r$  in order to calculate the coefficient of determination (r-squared). Put r-squared on your poster.
- Write a sentence describing its meaning of r-square in context.
- What is the slope of the regression line? You will find this where it says “Slope” under “Summary Statistics” in StatKey. Put the slope on your poster.
- Write a sentence describing the meaning of the slope in context.
- What is the Y-intercept of the regression line? You will find this where it says “Intercept” under “Summary Statistics” in StatKey. Put the Y-intercept on your poster.
- Write a sentence describing the meaning of the y-intercept in context.
- Put in the slope and Y-intercept into the regression line equation  $\hat{y} = (Y - intercept) + (Slope)X$ . Put the regression equation on your poster.
- Find the scope of the x-values? (Estimate two numbers on the X-axis that the points on the scatterplot are in between.)
- Pick any x-value in the scope. Plug in that x value into the regression line equation and predict the y value.
- Decorate your poster to spark interest.
- Now take a picture of your poster project and submit the picture to your instructor in Canvas.
- After submitting the picture of the poster to your instructor, go to the discussion menu in Canvas and complete the “Chapter 6 Project Discussion”. You will be discussing your findings with other students in the class.



**Face to face Class Directions:** The class will be separated into groups. Each group is required to pick a “team name” for their group and analyze a pair of quantitative data sets from the following topic list. You can find your two columns of data in the “Math 075 Project Data Correlation Regression”. You can find this data in Canvas or at [www.matt-teachout.org](http://www.matt-teachout.org). Put your two columns of data into StatKey to calculate, create a poster summarizing their findings, and present the poster to other students in the class.

**Topics:** IQ/Brain Volume , MLB Runs Scored/Attendance , MLB Runs Allowed/Wins , Price Item/Customer Satisfaction , Meat/illness , CEO Golf Score/Stock Price , Swim time/Pulse , Boats/Manatee Deaths , Cost of Living/Aviation Pay , Poverty/BMI , Alcohol/Tobacco in England

- Pick one of the pairs of quantitative variables and pick which should be X and which should be Y. The poster should give the explanatory variable (x) and response variables (y) and the units for x and y.
- Why this topic was important or interesting to your group.
- Go to [www.lock5stat.com](http://www.lock5stat.com). Click on “Two quantitative variables” under the “descriptive statistics and graphs” menu. Click on “edit data” button in StatKey. Copy and paste the two columns together into StatKey. If your two columns have titles, click the box that says “data has a header row”. Push OK. Click the box that says “Show Regression Line”.
- Look at the scatterplot carefully. Make sure the x variable you picked is on the horizontal axis and the response variable (y) is on the vertical axis. If they are not push the “Switch Variables” button.
- Draw a rough sketch of the scatterplot on your poster. Label your axes and draw the regression line also.
- What is the correlation coefficient (r) for your data? You will find this where it says “Correlation” under “Summary Statistics”.
- Use the r-value to classify the linear relationship as one of the following: strong positive correlation, strong negative correlation, moderate positive correlation, moderate negative correlation, weak positive correlation, weak negative correlation, or no correlation.
- Square your r-value for multiply  $r \times r$  in order to calculate the coefficient of determination (r-squared). Put r-squared on your poster.
- Write a sentence describing its meaning of r-square in context.
- What is the slope of the regression line? You will find this where it says “Slope” under “Summary Statistics” in StatKey. Put the slope on your poster.
- Write a sentence describing the meaning of the slope in context.
- What is the Y-intercept of the regression line? You will find this where it says “Intercept” under “Summary Statistics” in StatKey. Put the Y-intercept on your poster.
- Write a sentence describing the meaning of the y-intercept in context.
- Put in the slope and Y-intercept into the regression line equation  $\hat{y} = (Y - intercept) + (Slope)X$ . Put the regression equation on your poster.
- Find the scope of the x-values? (Estimate two numbers on the X-axis that the points on the scatterplot are in between.)
- Pick any x-value in the scope. Plug in that x value into the regression line equation and predict the y value.
- Decorate your poster to spark interest.

#### Presentation

Make sure each person on the team understands the poster and can present your findings. Bring your poster to a designated presentation area in the classroom and hang or tape your poster to a wall. One person at a time will present the poster. We will then rotate so that each member of the team gets to present. Everyone else will listen to presentations and give feedback.



## Chapter 7 – Curved Quantitative Relationships

**Introduction:** In the last unit, we saw that when looking for relationships between quantitative data sets it is often useful to create a scatter plot of the data. Remember that the data should be quantitative and in paired form. The paired data should be quantitative, which means that it should be a measurable quantity with defined units, not categories.

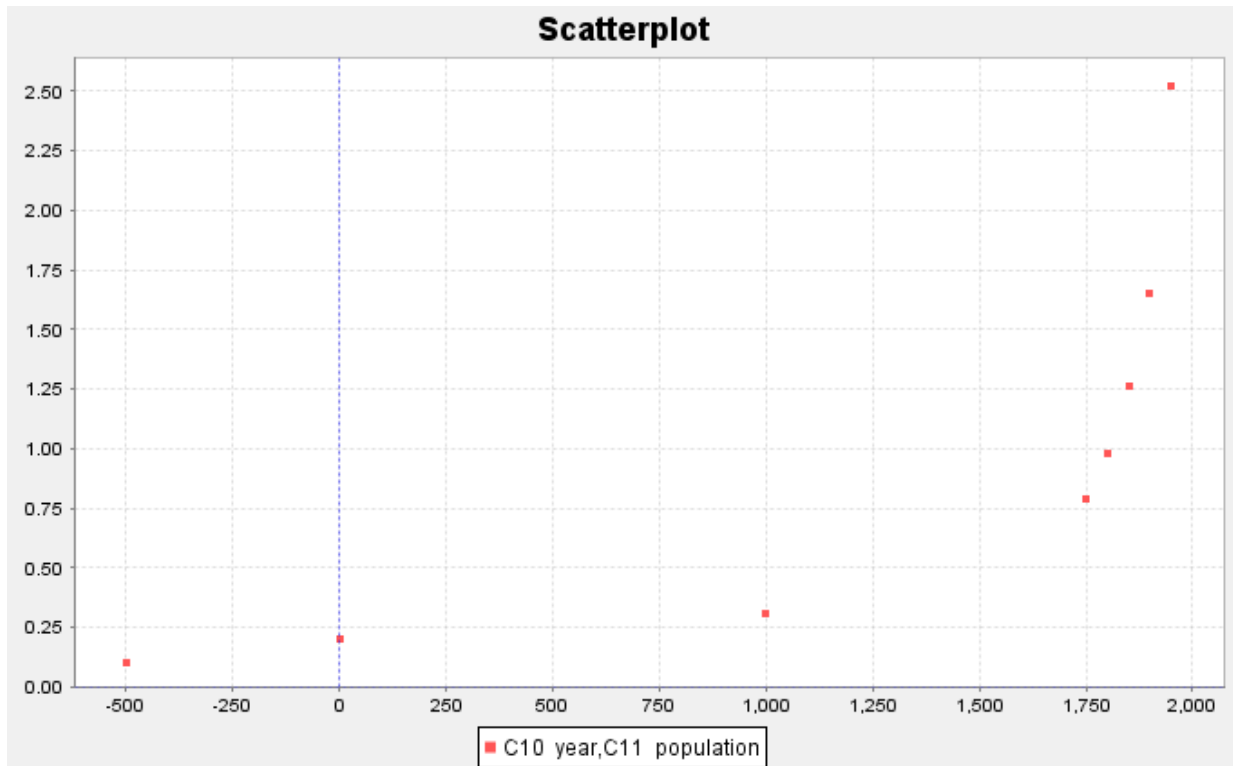
We do this by designating one data set to be the explanatory variable ( $x$ ) and one data set to be the response variable ( $y$ ). The choosing of the explanatory and response variables is very important. Remember to choose the response variable to be the one that might naturally respond to changes in the explanatory variable. Every case is different and in some paired data, either data set might be the response. The variable that you want to make a prediction of, should be the response variable.

**EXAMPLE:** Let us look at the following paired data set giving the year and the world population in that year. Note that “-500” means 500 B.C.

Year	World Population in Billions
-500	0.1
1	0.2
1000	0.31
1750	0.791
1800	0.978
1850	1.262
1900	1.65
1950	2.519

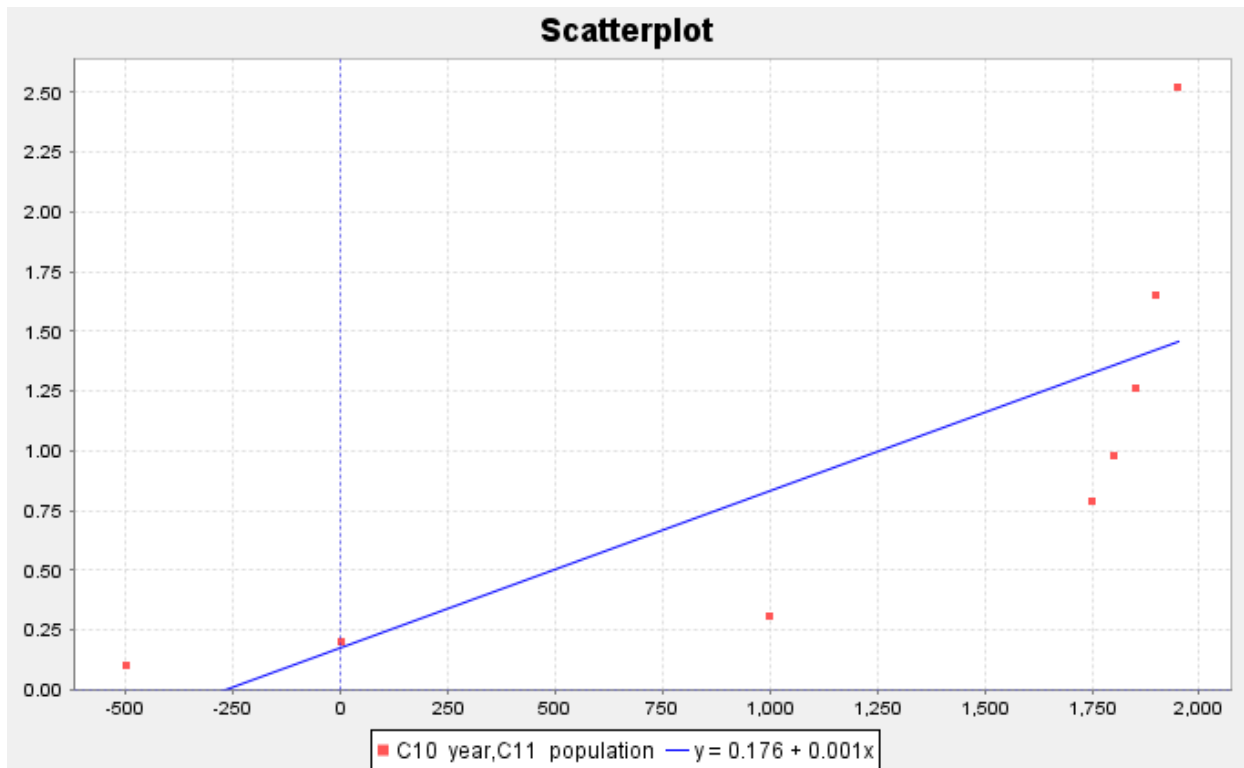
In thinking about this paired data, we wonder if there is a relationship, but which variable should be the explanatory and which should be the response? It seems logical that the population might change or respond to the year, so we will make the year be the explanatory variable ( $x$ ) and the world population be the response variable ( $y$ ). Plugging the data into a statistics software program like Statcato, we can generate the following scatterplot.





We notice right away that this graph does not have a linear shape. Using our statistics software, we can find our least squares regression line for the data and an *R-squared (linear)* = 0.5882 (58.8%). This tells us that approximately 58.8% of the variability in population can be explained by the linear relationship with time. The graph and the R-squared value confirm that this scatterplot does have some linear relationship (correlation), but it is not as strong as we might like. The statistics software also gives us the equation of the regression line  $\hat{y} = 0.1763 + 0.0007x$ , but it is clear that this line is not a good model for the data.





Let us look at this scatterplot again. Just because there does not seem to be a linear relationship, does not mean there is no relationship at all. In fact, the scatterplot shows a very strong curved relationship in the data. If our goal were to make predictions of what the world population will be, we would need to find a function that matches that curve.

Hence, it is useful for anyone studying data to have some knowledge of curved functions.

---

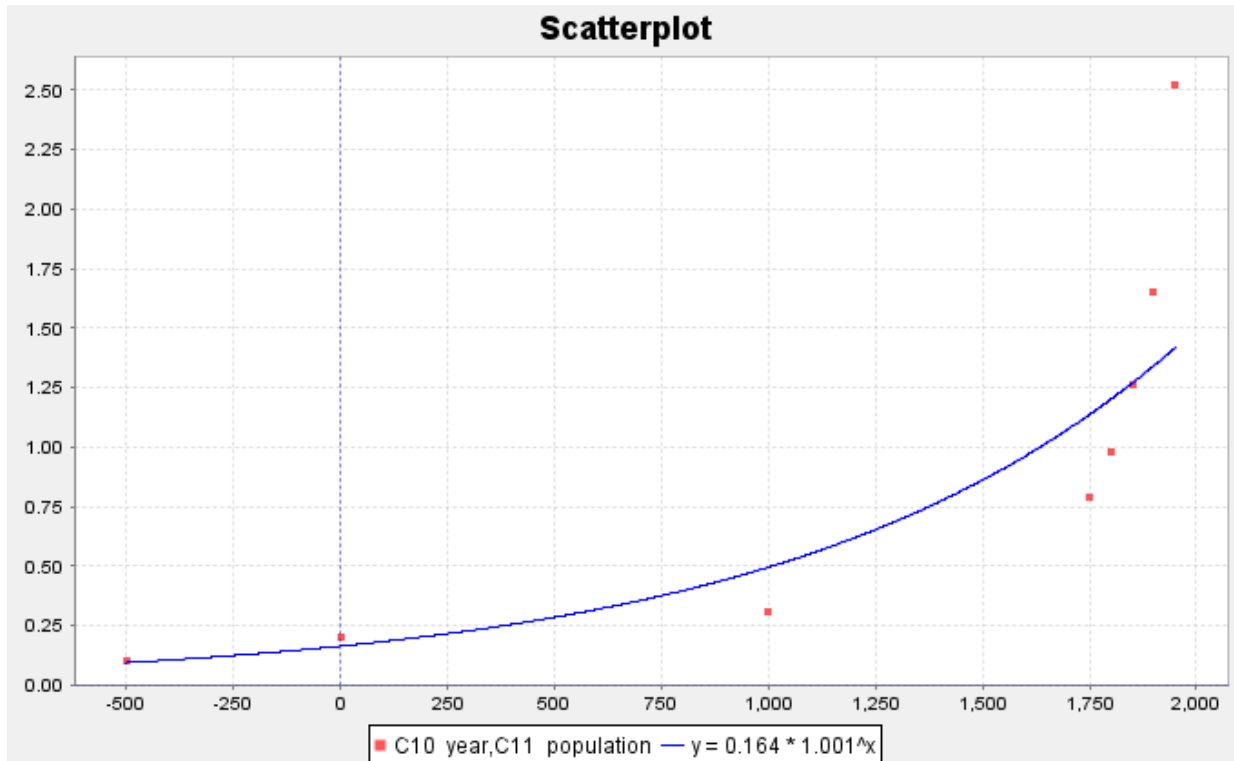


## Section 7A – Exponential Quantitative Relationships

Probably the most common type of curved pattern seen in data is the exponential curve. This pattern is seen in a variety of data analysis settings including population growth and decay, and compound interest.

Let us look again at our world population data. Using a statistics software program, we have the program plot an exponential curve over the scatter plot. Does the curve seem to fit the data better than the regression line we found in the introduction section? Not only have we found a better fit for the paired data, but the program also gives the equation for that exponential curve

$$\hat{y} = 0.16365(1.00111)^x.$$



We can see a few things from this graph. The first is the shape of an exponential curve. Do you see how the exponential curve has a backwards “L” shape to it? The curve increases as we go from left to right. This is very common with exponential curves, especially with population growth. That is why this pattern is called “exponential growth”. Exponential decay functions tend to have a regular “L” shape and decrease from left to right. Did you also notice how the exponential function does not cross the x-axis, but simply approaches it? This is also very common in exponential functions. If you think about it, the response variable (y) is describing the world population. Of course, the y values must be positive. If the y-value was zero or negative, I would not be here talking to you. The world population cannot be zero or a negative number. This tells us that the y-values of exponential curves can only be positive numbers.



What else can we learn from this graph? Did you notice that the curve tends to get close to the x-axis for the first years in the data set (500 BC, 1 AD and 1000 AD)? When a curve gets close to a line for certain x-values, we call this line an asymptote. In this graph, the x-axis is an asymptote. Did you also notice that the graph starts to get very big, very quickly? From 1000 AD to 1950 AD, the population has risen from approximately 310 million to over 2.5 billion people! That is an incredible increase if you think about it. Have you ever heard someone say the following? “That is growing so fast that it is growing exponentially.” That statement comes from the shape of the exponential growth function.

Now let us look at the equation of the exponential function  $\hat{y} = 0.16365(1.00111)^x$  that the software found for us.

Different software’s can write this formula differently. Do you notice how the x variable is actually the exponent in the equation? That is how “exponential” functions get their name. Recall that the number an exponent is attached to is called the “base”. In this equation, the number (1.00111) is the base. Notice also that 0.16365 is the number that the exponential expression (exponent and base) is multiplied by. This number is usually called the “initial value” or in this case the “initial population”. In this data set, our “initial” ordered pair was 500 BC. Unfortunately, this is not the initial value described in the equation. When we say our initial value, we mean the y-value when the x = 0. If you have studied any algebra, you may remember that this would be the y-intercept. Look at the graph of the exponential curve. Approximately, where does the curve cross the y-axis? Did you notice that the curve seems to cross the y-axis at 0.16365? This is the same initial value as given in the equation.

#### Assessing the fit of an exponential function

This is fine for Statcato to give us an exponential function, but how well does this curve really fit the data? We again see that the exponential curve does not fit the data perfectly, but it does seem to fit better than the line. If we are going to use this exponential function to maybe make predictions, then we need to have some way of assessing how well the curve fits the data.

#### R-Squared

One of the first things to look at when assessing the fit of a curve to a scatterplot is the “R-squared” value. Remember that “R” is the correlation coefficient. However, when we square R it gives the percent of variability in y that can be explained by the relationship with x. For our exponential function and the population growth data Statcato determined that  $R^2 \approx 0.9078$ . This tells us that approximately 90.8% of the variability in population can be explained by the exponential relationship with time. This is a very high percentage and indicates the exponential function fits pretty well.

Another use of R-squared is to determine which model is a better fit. For example, suppose I want to know if the exponential model is a better fit than the linear model. We can determine this by comparing the R-squared values.

Regression Line:  $R^2 \approx 0.5878$

Exponential Regression Curve:  $R^2 \approx 0.9078$

While only 58.8% of the variability in population can be explained by the linear relationship, almost 91% of the variability can be explained by the exponential relationship. The model with the higher R-squared is the better fit.

*Note: We prefer to use the simpler formula (linear) when possible. If there is a significant increase in the r-squared value for the curved function, we will use the curve. However, if the curve has only a slightly better r-squared, we prefer to use the simpler model. In the last example, the R-squared value for the exponential was 90.8%. This is significantly higher than the regression line’s R-squared value of 58.8%. Therefore, we would most definitely prefer the exponential model to the linear model. Let us suppose in a different problem, the R-squared value for a curve is 84% and the R-squared for the linear model is 83%, then we would stay with simpler model (linear) because there is not a significant increase.*





## Standard Deviation of the Residual Errors ( $S_e$ )

Let us look at one of the ordered pairs in our data set, say (1000 year, 0.31 billion people). Can you find the ordered pair on the curve with that same x-value (1000 AD)? If we plug in 1000 into the regression equation, we can get the y value on the curve. This is often called our predicted value  $\hat{y}$ . Let us calculate the predicted value for the year 1000 AD.

$$\hat{y} = 1.00111^{1000} \times 0.16365$$

$$\hat{y} = 0.496$$

So the regression line predicted that the population in the year 1000 AD would be approximately 0.496 billion people. The actual observed population in the year 1000 AD was 0.31 billion. So how much error was in our prediction? One way to measure error is through residuals. Recall that a residual is the difference between the observed ordered pair (y) and the predicted value ( $\hat{y}$ ) if the original x value is plugged into the function. Another way to explain the residual is that it is the vertical distance from the curve to the point. For example, for the year 1000 AD, the residual would be calculated as follows:

$$y - \hat{y} = 0.31 - 0.496$$

$$y - \hat{y} = -0.186$$

Notice this gives us a residual (error) of -0.186. This means that the ordered pair is 0.186 below the curve when x = 1000 AD. Let us now make a table of the residuals. For each x value in the data set, we will plug the x value into the regression curve from Statcato  $\hat{y} = 0.16365(1.00111)^x$ . This will give us our predicted  $\hat{y}$  values. Subtracting the actual y value minus the predicted  $\hat{y}$  value gives us the residual.

Notice that when the curve is too high, the residuals are negative and when the curve is too low, the residuals are positive. We still have the problem of assessing how well the curve fits the data set. One possibility would be to find the Standard Deviation of the Residual Errors ( $S_e$ ) as we did for lines. By squaring the residuals, we are able to eliminate the negative residuals. Now we add up the squares, divide by n-2 and take the square root of the answer. Recall that the Standard Deviation of the Residual Errors ( $S_e$ ) will give us how far on average the points are from the curve and will give us the average prediction error should we use the curve to make a prediction. Let us look at the calculation of the Standard Deviation ( $S_e$ ) below.

Year (x)	World Pop (y) in Billions	pred y from Exp curve	Residual Error	Residuals Squared
-500	0.1	0.093975847	0.006024153	0.0000362904
1	0.2	0.163831652	0.036168349	0.001308149
1000	0.31	0.496267158	-0.186267158	0.034695454
1750	0.791	1.140420931	-0.349420931	0.122094987
1800	0.978	1.205466528	-0.227466528	0.051741021
1850	1.262	1.274222097	-0.012222097	0.00014938
1900	1.65	1.346899241	0.303100759	0.09187007
1950	2.519	1.423721635	1.095278365	1.199634697

We first find the Sum of the Squares of the Residual Errors (SSE). Do not be confused. The SSE is not the standard deviation. SSE and  $S_e$  are completely different. In a sense, we need to use the sum of squares to get the standard deviation.



$$SSE = 0.0000362904 + 0.001308149 + \dots + 1.199634697$$

$$SSE \approx 1.501530049$$

Now we can use the standard deviation formula  $S_e = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{1.501530049}{8-2}} \approx 0.5002549$  to calculate the standard deviation of the residual errors.

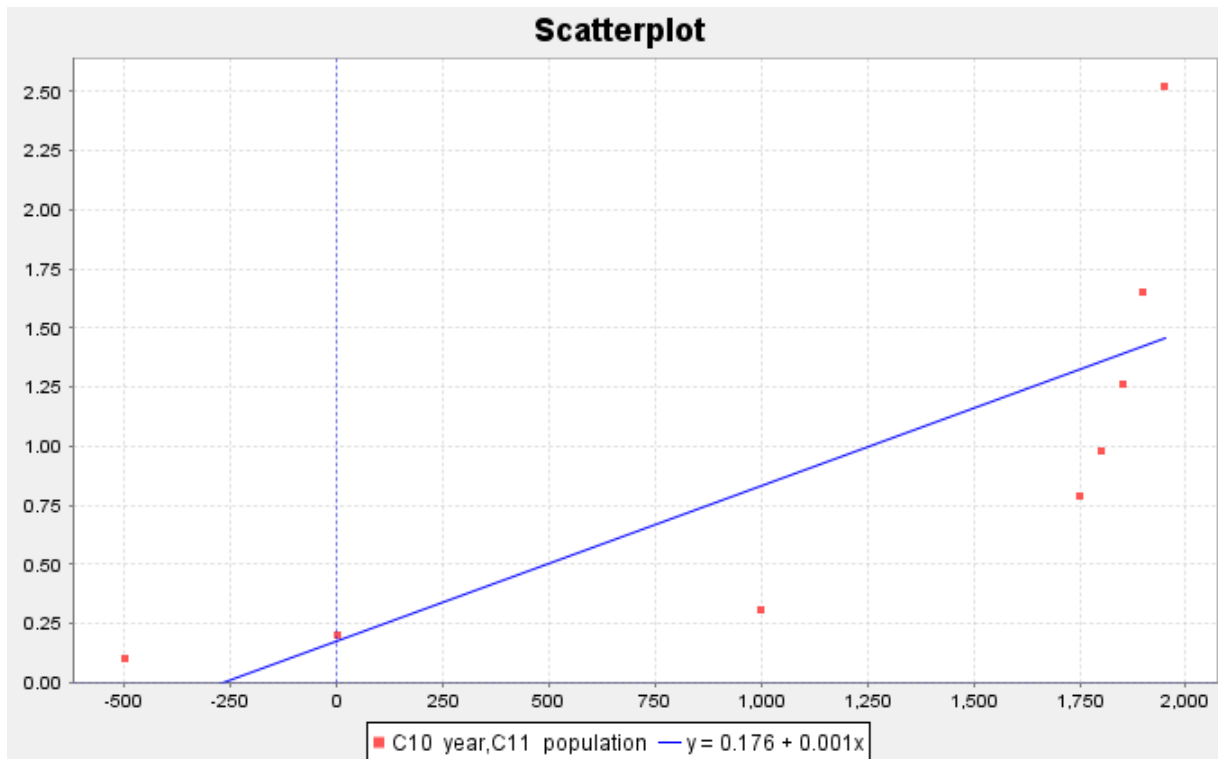
Remember were-as one data set has degrees of freedom  $n-1$ , ordered pair data has a degrees of freedom  $n-2$ . This is why we divide by  $n-2$  instead of  $n-1$ .

So  $S_e \approx 0.5$  billion. As with chapter 3, the standard deviation of the residuals tells us how well the data fits the regression curve. A regression curve tries to minimize this vertical distance. Therefore, for exponential curves  $\hat{y} = a(b)^x$ , the curve  $\hat{y} = 0.16365(1.00111)^x$  was the best fit. This again means that it minimized the vertical distance to the curve ( $S_e$ ). So no other function of the form  $\hat{y} = a(b)^x$  will have a smaller  $S_e$  than the function  $\hat{y} = 0.16365(1.00111)^x$ .

Sometimes we may want to know if one curve or line fits the data better than another does. The Standard Deviation of the Residual Errors can be used for this purpose. The curve that has the smallest Standard Deviation of the Residual Errors will be the one that fits the data best.

Let us explore this a little bit. We just found that the Standard Deviation for the Exponential Curve Residuals ( $S_e$ ) was about 0.500 billion. Earlier we said that we thought the exponential curve fit the data better than the regression line. Can we confirm what our eyes are telling us? Look at the scatterplot below. The software found that the regression line that best fits the population data was  $\hat{y} = 0.1763 + 0.0007x$  and calculated the Standard Deviation of the Residual Errors ( $S_e$ ).





First we plug in each year ( $x$ ) into the regression line  $\hat{y} = 0.1763 + 0.0007x$  and obtain our predicted  $\hat{y}$  values. Subtracting the observed population  $y$  values minus the predicted  $\hat{y}$  gives us the residuals. We had Statcato calculate the Standard Deviation this time.

For the regression line,  $S_e \approx 0.572$ . Notice this is larger than the standard deviation for the exponential curve (0.500). Since there is much less error when we use the exponential function versus the linear function, this implies that the exponential curve is a much better fit to this population data than the regression line.

**Key Idea:** A linear or curved model with a larger R-squared and smaller Standard Deviation of the Residual Errors gives evidence of a better fit.

**Note:** *The study of regression is broad and complicated branch of Statistics. It would be wrong to suggest that all of regression can be summarized into the highest R-squared and the lowest Standard Deviation. We often study many factors before deciding on a particular model. For example, the histogram of the residuals should be bell shaped and centered at zero. In addition, the residual plot verses the x-value should be evenly spread out. Another is that there should not be any significant outliers in the scatterplot. These are but a few.*

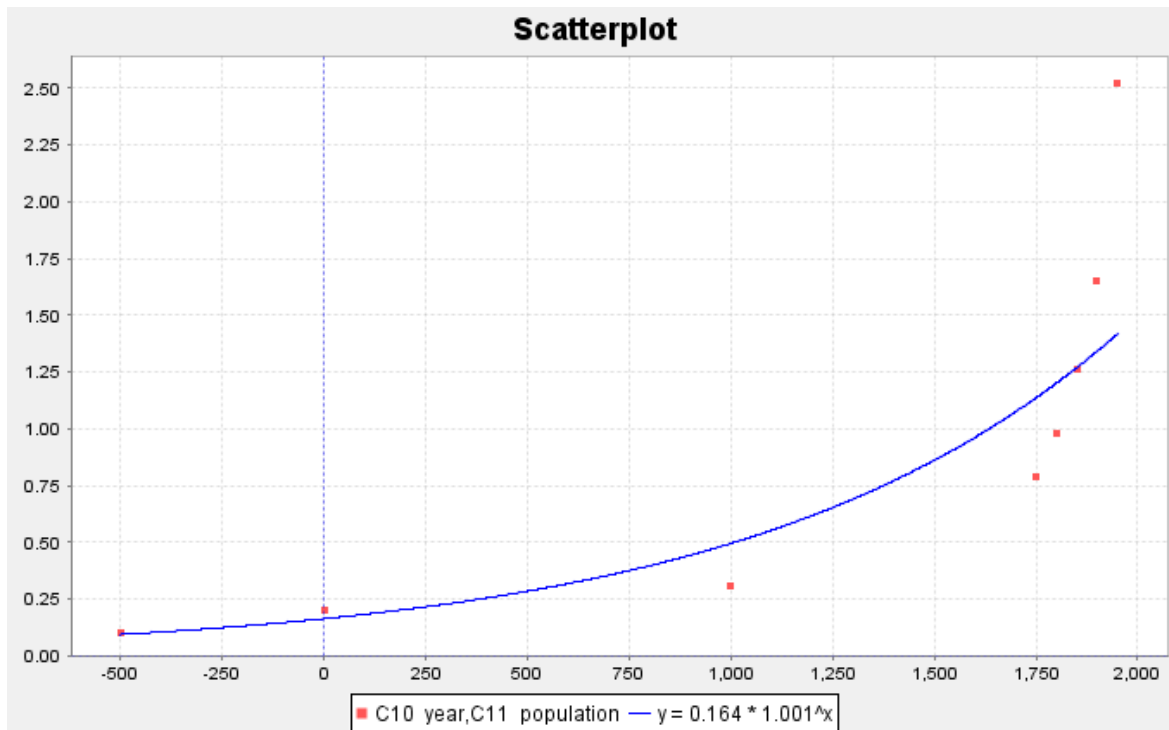
#### Making Predictions from an Exponential Function

Remember the goal of finding the exponential curve for the population data was to hopefully use it to make predictions. So now that we have assessed that the exponential curve does fit the population data reasonably well, let us use the function to make a prediction.

**Caution!!** Remember that we should only make predictions within the scope of the data. The x-values for our population paired data were between -500 (500 BC) and the year 1950. We should not try to make predictions outside of this range. If you recall making predictions out of the scope of the data is called Extrapolation. People that extrapolate tend to have a lot of error in their predictions because there is no guarantee that the data will follow the curve outside the scope of the data. Remember the Standard Deviation of the Residual Errors only applies in the scope of the data. Once you go outside the scope of the data, there is no telling how much error there may be.



So let us predict the world population in Billions for a given year. Let us look at the scatterplot below. Try to estimate the world population in the year 600 AD.



By plugging in 600 for x in our exponential curve, we can get our prediction. Remember to follow the order of operations and perform the exponent first, then multiply.

$$\hat{y} = 0.16365(1.00111)^{600}$$

$$\hat{y} \approx 0.3184$$

Hence in 600 AD, the exponential function predicts that the world population was 0.318 Billion (318 Million) people.

---

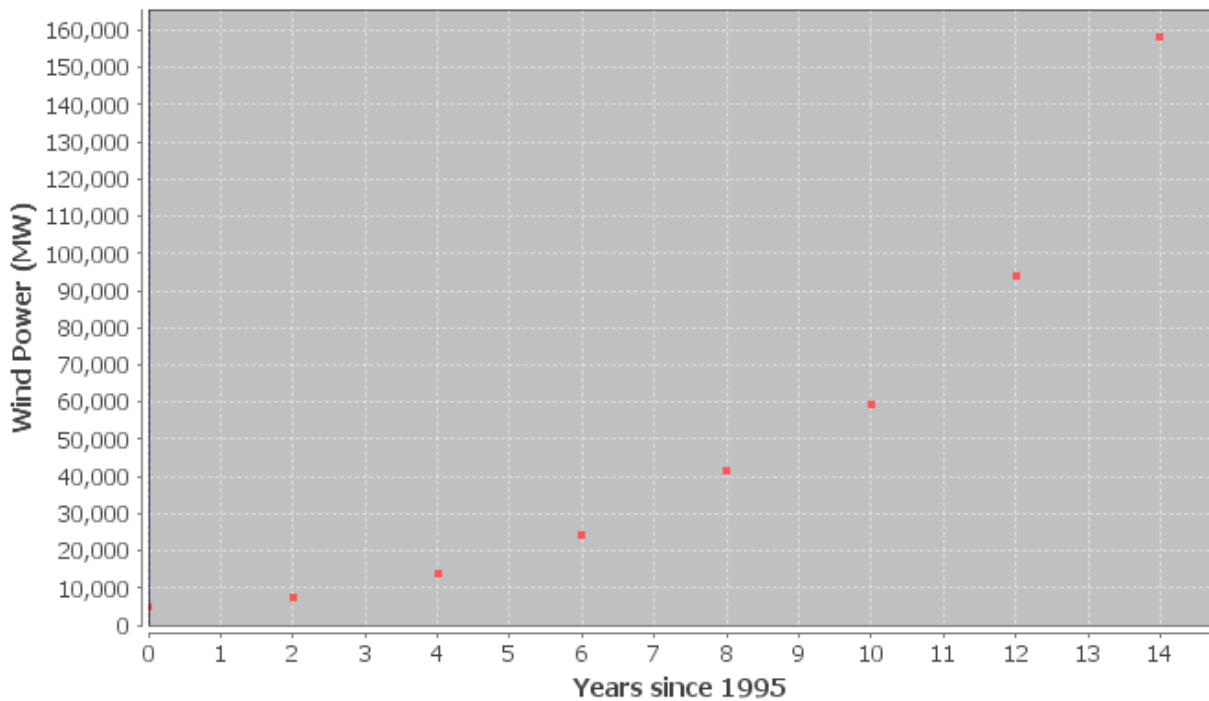


## Problem Set Section 7A

NOTE: You will not need Statcato or StatKey to do this assignment. You only need to analyze the graphs and statistics provided.

1. The following scatterplots and printouts describe the number of years since 1995 and the worldwide wind power capacity in MW (megawatts). The number of years is the explanatory variable (X) and the wind power (MW) is the response variable (Y).
  - a) Examine the following scatterplot of the ordered pairs. Does the scatterplot look like an exponential model might fit? If so, would it be exponential growth or exponential decay? What is the scope of the data (x values)? What years do the scope represent?

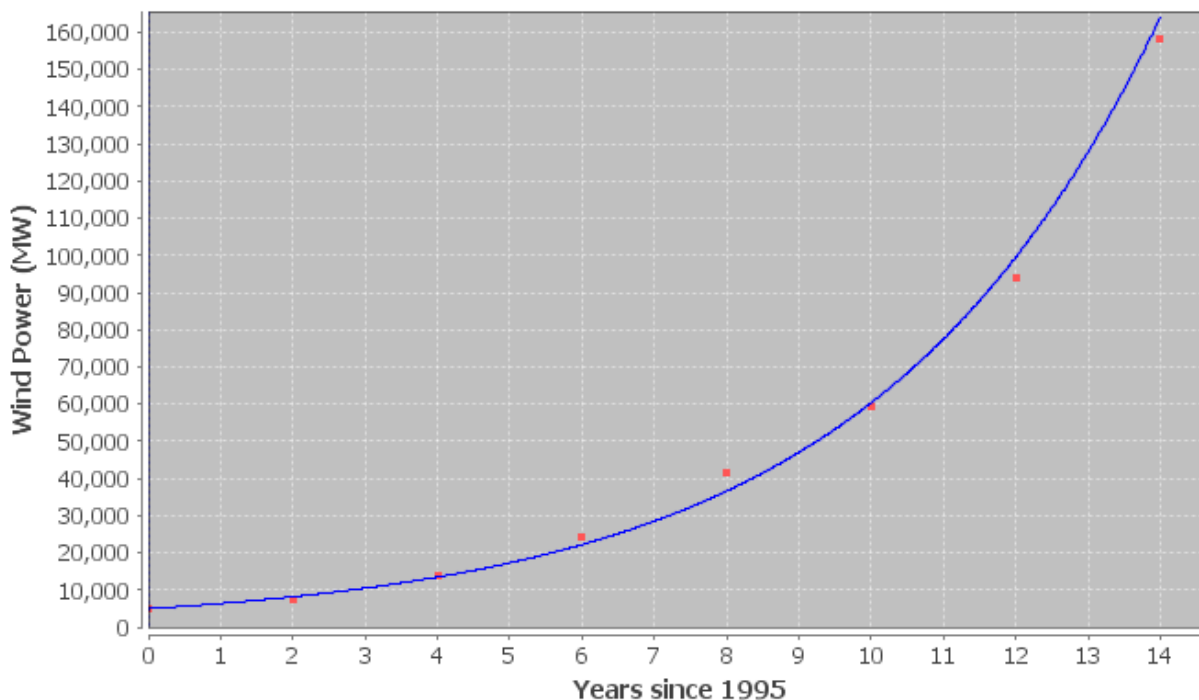
### Scatterplot



- b) The following scatterplot includes the exponential curve. Is this an exponential growth curve or an exponential decay curve? Do you think that the exponential curve fits the data well? Are the points close to the curve?



## Plot



The following Statcato printout describes the exponential relationship between the years since 1995 and the wind power (MW). Use the printout to answer the following questions.

### Non-Linear Modeling:

x (independent/explanatory variable): Years since 1995

y (dependent/response variable): Wind Power (MW)

**Exponential Model:  $y = a b^x$**

a = 4945.11427

b = 1.28424

Sample size = 8

Coefficient of determination  $r^2 = 0.9965$

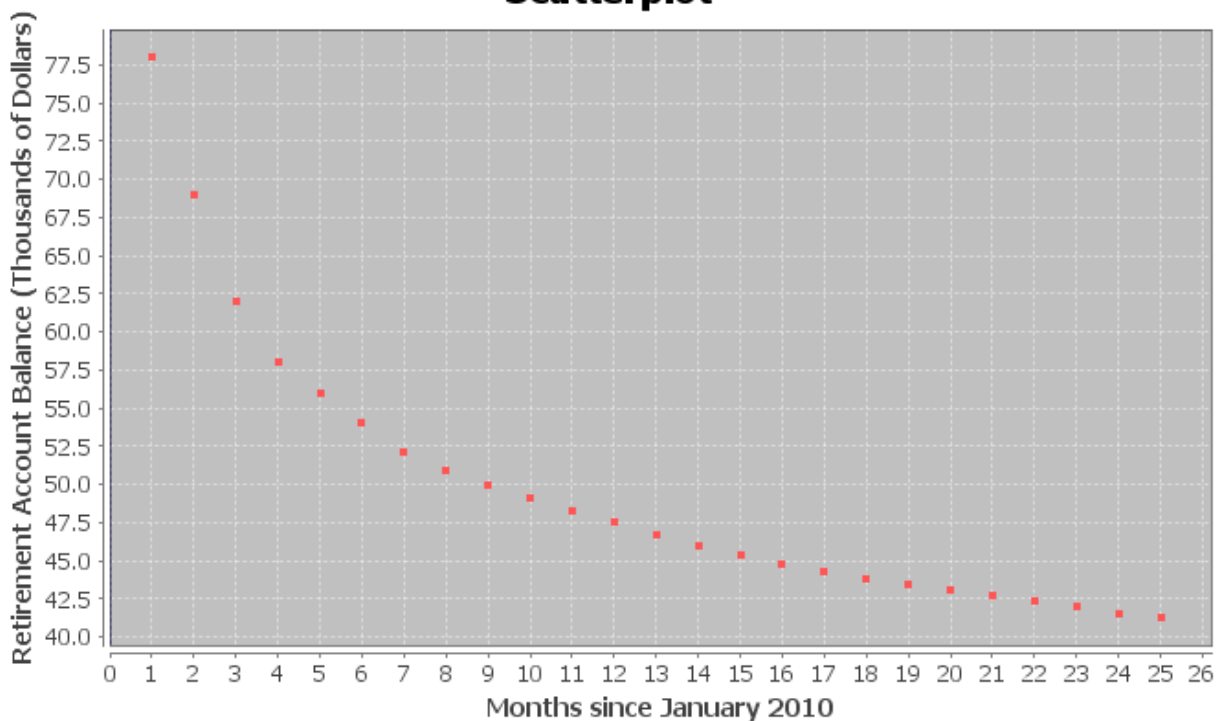
Standard Deviation of Residual Errors = 4039.2

- How many ordered pairs were in the data?
- Give equation of the exponential curve that Statcato found. What is the y-intercept? What is the base? Is the base greater than 1 or less than 1? What does that tell you about the shape of the exponential curve?
- What is the coefficient of determination ( $r^2$ )? Convert  $r^2$  into a percentage. Then write a sentence explaining the meaning of  $r^2$  in this context.
- Most computer programs have a problem calculating the standard deviation of the residuals errors for exponential curves. Statcato originally showed that the standard deviation for the exponential curve was 0.0788. This was wrong. The graph shows that the points are much farther from the curve than this. The real standard deviation is actually 4039.2 MW. Write two sentences explaining the meaning of the standard deviation of the residual errors 4039.2 MW in this context.



- g) Plug in 7 for “x” into the equation of the exponential curve and solve for “y” in order to predict the the Wind Power in 2002 (year 7)? How far off could this prediction be on average?
- h) Plug in 13 for “x” into the equation of the exponential curve and solve for “y” in order to predict the the Wind Power in 2008 (year 13)? How far off could this prediction be on average?
- i) Do you think it would be all right to use this model to predict the worldwide wind power in 2065 (year 70)? Why or why not? If a person did make this prediction, would it have the same prediction error as parts (g) and (h)?
2. The following scatterplots and printouts describe the number of months since January 2010 and the amount of money in a retirement account in thousands of dollars. The number of months is the explanatory variable (X) and the retirement account balance is the response variable (Y). The account started with \$78,000 in their account in 2010, and have been slowly making withdrawals for their living expenses.
- a) Examine the following scatterplot of the ordered pairs. Does the scatterplot look like an exponential model might fit? If so, would it be exponential growth or exponential decay? What is the scope of the data (x values)? What months do the scope represent?

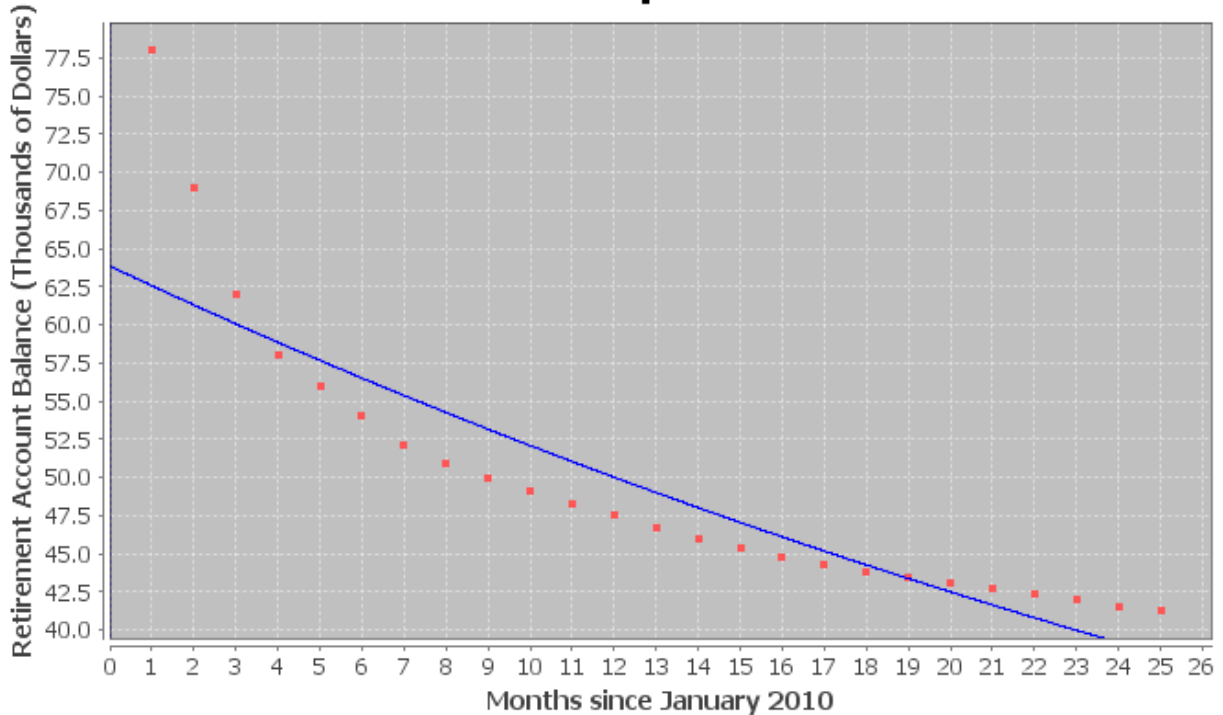
### Scatterplot



- b) The following scatterplot includes the exponential curve. Is this an exponential growth curve or an exponential decay curve? Do you think that the exponential curve fits the data well? Are the points close to the curve?



## Scatterplot



The following Statcato printout describes the exponential relationship between the months since January 2010 and the retirement account balance (in thousands of dollars). Use the printout to answer the following questions.

### Non-Linear Modeling:

x (independent/predictor variable): Months since January 2010

y (dependent/response variable): Retirement Account Balance (Thousands of \$)

**Exponential Model:  $y = a b^x$**

a = 63.85340

b = 0.97985

Sample size = 25

Coefficient of determination  $r^2 = 0.8337$

Standard Deviation of Residual Errors = 4.182

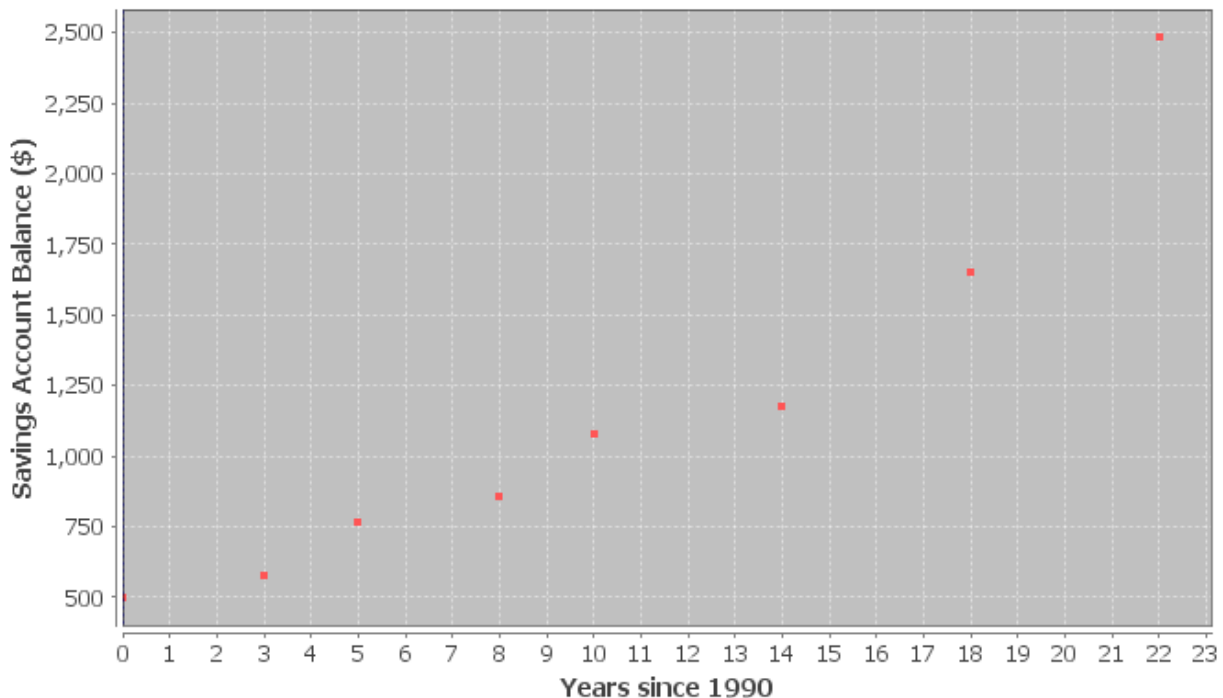
- How many ordered pairs were in the data?
- Give equation of the exponential curve that Statcato found. What is the y-intercept? What is the base? Is the base greater than 1 or less than 1? What does that tell you about the shape of the exponential curve?
- What is the coefficient of determination ( $r^2$ )? Convert  $r^2$  into a percentage. Then write a sentence explaining the meaning of  $r^2$  in this context.



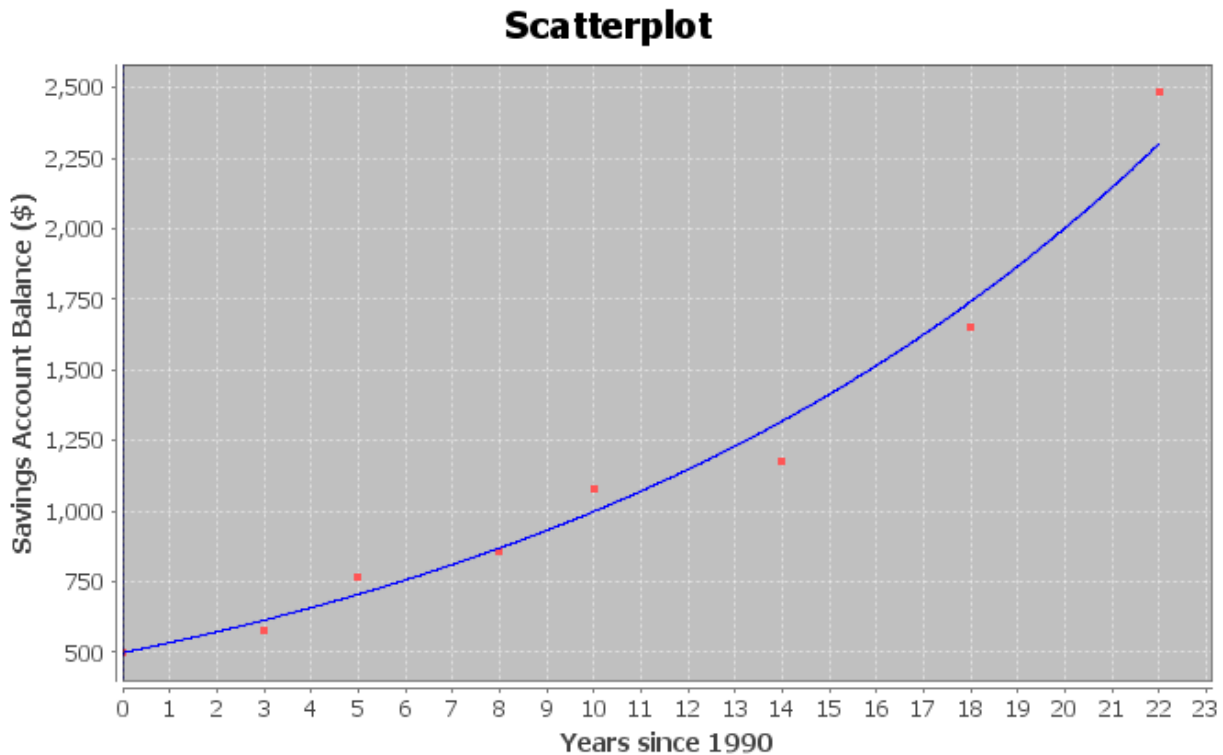


- f) Most computer programs have a problem calculating the standard deviation of the residuals errors for exponential curves. Statcato originally showed that the standard deviation for the exponential curve was 0.0684. This was wrong. The graph shows that the points are much farther from the curve than this. The real standard deviation is actually 4.182 thousand dollars. Write two sentences explaining the meaning of the standard deviation of the residual errors 4.182 thousand dollars in this context.
- g) Plug in 11.5 for “x” into the equation of the exponential curve and solve for “y” in order to predict the retirement account balance December 15<sup>th</sup> 2010 (month 11.5). How far off could this prediction be on average?
- h) Plug in 24.5 for “x” into the equation of the exponential curve and solve for “y” in order to predict the retirement account balance January 15<sup>th</sup> 2012 (month 24.5). How far off could this prediction be on average?
- i) Do you think it would be all right to extrapolate a lot and use this model to predict the retirement account at the start of 2050 (month 480)? Why or why not? If a person did make this prediction, would it have the same prediction error as parts (f) and (g)?
3. A person opened a savings account in 1990. The following scatterplots and printouts describe the number of years since 1990 and the corresponding savings account balance (in dollars). The number of years since 1990 is the explanatory variable (X) and the savings account balance (in dollars) is the response variable (Y).
- a) Examine the following scatterplot of the ordered pairs. Does the scatterplot look like an exponential model might fit? If so, would it be exponential growth or exponential decay? What is the scope of the data (x values)? What years do the scope represent?

### Scatterplot



- b) The following scatterplot includes the exponential curve. Is this an exponential growth curve or an exponential decay curve? Do you think that the exponential curve fits the data well? Are the points close to the curve?



The following Statcato printout describes the exponential relationship between the years since 1990 and the savings account balance (in dollars). Use the printout to answer the following questions.

**Non-Linear Modeling:**

x (independent/predictor variable): Years Since 1990

y (dependent/response variable): Amount in Savings Account (\$)

**Exponential Model:  $y = a b^x$**

a = 497.44019

b = 1.07211

Sample size = 8

Coefficient of determination  $r^2 = 0.9806$

Standard Deviation of Residual Errors = \$110.94

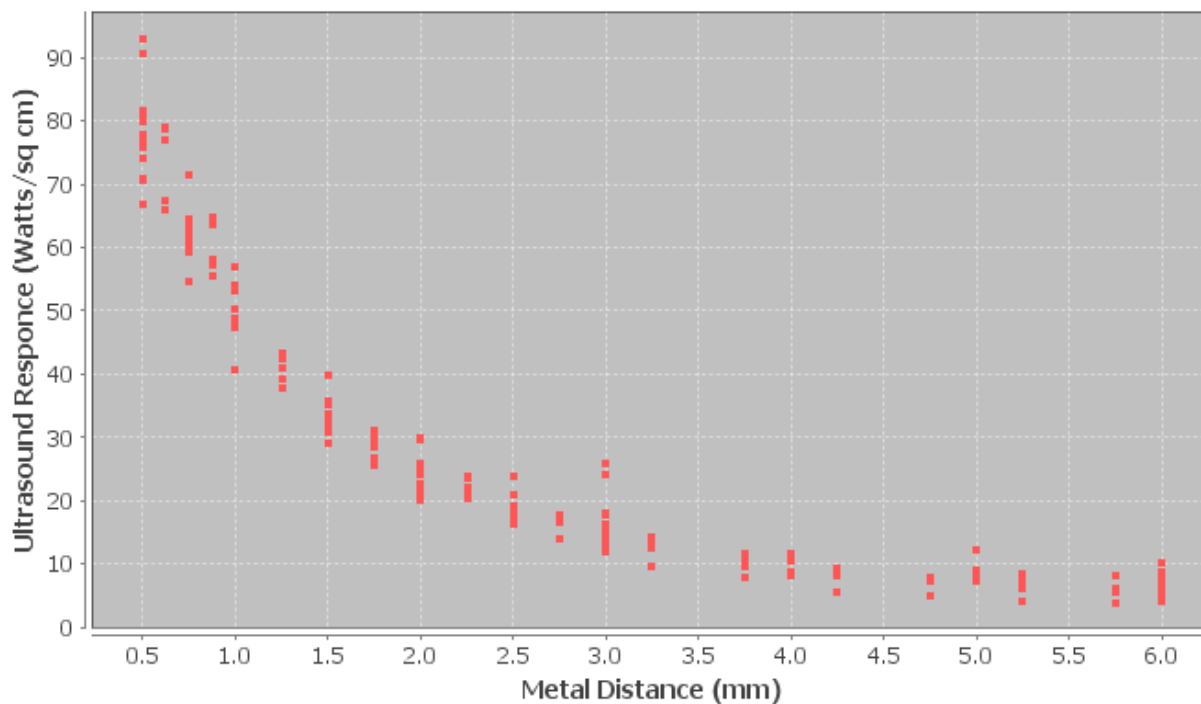
- c) How many ordered pairs were in the data?
- d) Give equation of the exponential curve that Statcato found. What is the y-intercept? What is the base? Is the base greater than 1 or less than 1? What does that tell you about the shape of the exponential curve?



- e) What is the coefficient of determination ( $r^2$ )? Convert  $r^2$  into a percentage. Then write a sentence explaining the meaning of  $r^2$  in this context.
- f) Most computer programs have a problem calculating the standard deviation of the residuals errors for exponential curves. Statcato originally showed that the standard deviation for the exponential curve was 0.0801. This was wrong. The graph shows that the points are much farther from the curve than this. The real standard deviation is actually \$110.94. Write two sentences explaining the meaning of the standard deviation of the residual errors \$110.94 in this context.
- g) Plug in 16 for “x” into the equation of the exponential curve and solve for “y” in order to predict the savings account balance in 2006 (year 16). How far off could this prediction be on average?
- h) Plug in 21 for “x” into the equation of the exponential curve and solve for “y” in order to predict the savings account balance in 2011 (year 21). How far off could this prediction be on average?
- i) Do you think it would be all right to extrapolate a lot and use this model to predict the savings account in 2040 (year 50)? Why or why not? If a person did make this prediction, would it have the same prediction error as parts (g) and (h)?
4. Ultrasound is used in a variety of applications. The following scatterplots and printouts describe the relationship between the metal distance in millimeters and the ultrasound response. Let the explanatory variable (X) be the metal distance in millimeters and the response variable (Y) be the ultrasound response in Watts per square centimeter.
- a) Examine the following scatterplot of the ordered pairs. Does the scatterplot look like an exponential model might fit? If so, would it be exponential growth or exponential decay? What is the scope of the data (x values)?



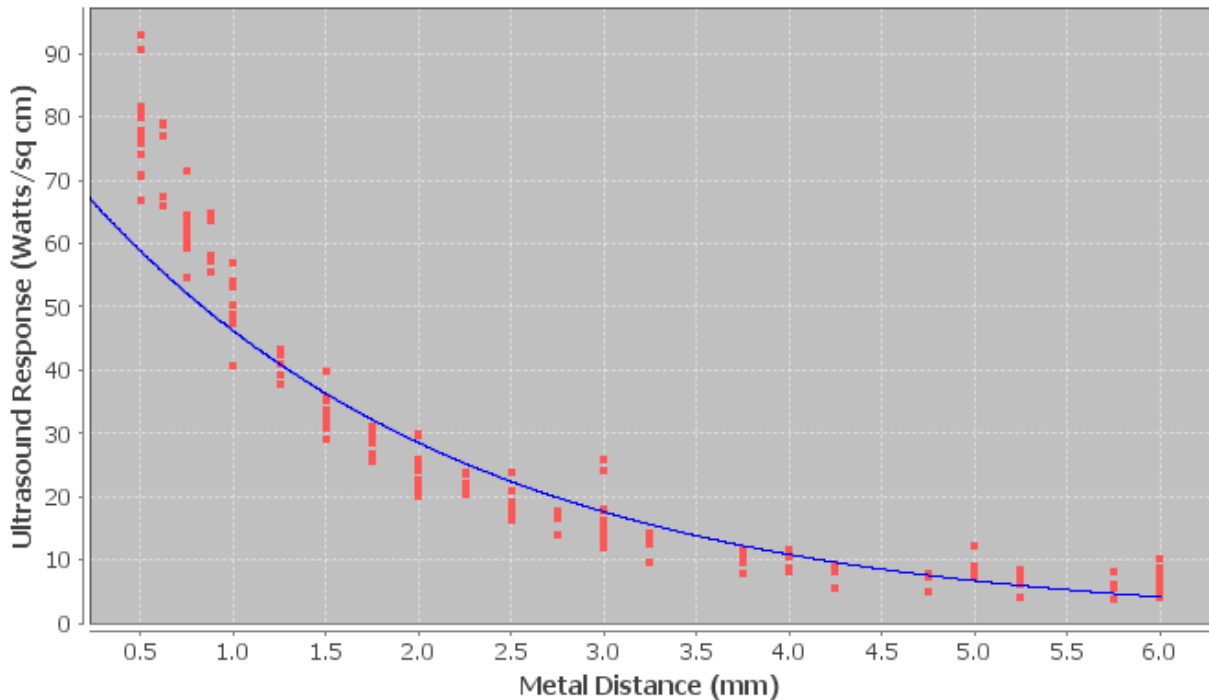
## Scatterplot



- b) The following Statcato scatterplot includes the exponential curve. Is this an exponential growth curve or an exponential decay curve? Do you think that the exponential curve fits the data well? Are the points close to the curve?



## Scatterplot



The following Statcato printout describes the exponential relationship between the metal distance in millimeters and the ultrasound response (Watts per square centimeter). Use the printout to answer the following questions.

### Non-Linear Modeling:

x (independent/predictor variable): Metal Distance (mm)

y (dependent/response variable): Ultrasonic Response (Watts/sq cm)

**Exponential Model:  $y = a b^x$**

a = 74.91083

b = 0.61685

Sample size = 214

Coefficient of determination  $r^2 = 0.9126$

Standard Deviation of Residual Errors = 8.239 Watts/sq cm

- How many ordered pairs were in the data?
- Give equation of the exponential curve that Statcato found. What is the y-intercept? What is the base? Is the base greater than 1 or less than 1? What does that tell you about the shape of the exponential curve?
- What is the coefficient of determination ( $r^2$ )? Convert  $r^2$  into a percentage. Then write a sentence explaining the meaning of  $r^2$  in this context.



- f) Most computer programs have a problem calculating the standard deviation of the residuals errors for exponential curves. Statcato originally said that the standard deviation for the exponential curve is 0.2466. This is wrong. The graph shows that the points are much farther from the curve than this. The real standard deviation is 8.239 Watts per square cm. Write two sentences explaining the meaning of the standard deviation of the residual errors 8.239 Watts/sq cm in this context.
- g) Plug in 2.83 for “x” into the equation of the exponential curve and solve for “y” in order to predict the ultrasonic response if the metal is 2.83 mm away. How far off could this prediction be on average?
- h) Plug in 4.51 for “x” into the equation of the exponential curve and solve for “y” in order to predict the ultrasonic response if the metal is 4.51 mm away. How far off could this prediction be on average?
- i) Do you think it would be all right to extrapolate and use this model to predict ultrasonic response if the metal is 12.75 mm away? Why or why not? If a person did make this prediction, would it have the same prediction error as parts (g) and (h)?
5. If a data set has zero or negative numbers in the data for response variable (Y), Statcato said “Error” and “Not Available”. The programs were unable to find an exponential function to fit the data. However, Statcato had no problem finding an exponential function when the data for the explanatory variable (X) is zero or negative. Explain why this happened.
- 



## Section 7B – Logarithmic Quantitative Relationships

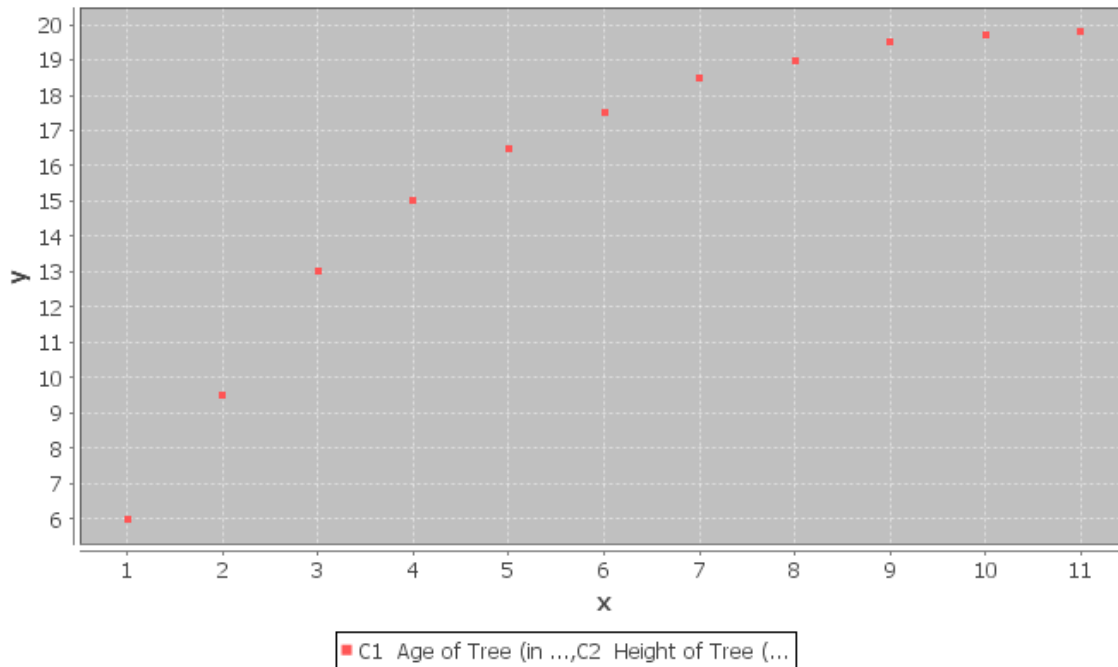
Let us examine another data set. The following data set gives the height of a tree in feet and the age of the tree in years.

Age of Tree (in years)	Height of Tree (in feet)
1	6.0
2	9.5
3	13.0
4	15.0
5	16.5
6	17.5
7	18.5
8	19.0
9	19.5
10	19.7
11	19.8

As with the population data, we wonder if there is a relationship between the age of the tree and the height of the tree, but which variable should be the explanatory and which should be the response? It seems logical that the height of the tree responds to its age, so we will make the year the explanatory variable ( $x$ ) and the height the response variable ( $y$ ). It also makes sense to make the height the response variable since we may want to predict the height of the tree from knowing the age of the tree. Plugging the data into a statistics software, we get the following scatterplot.

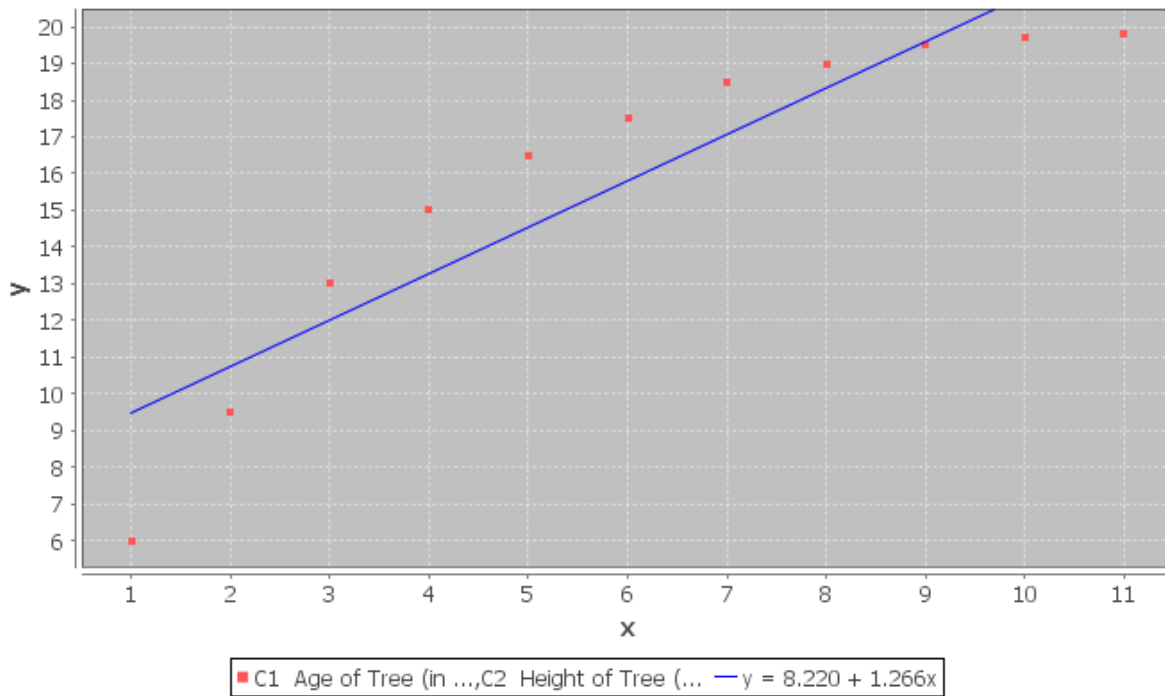


## Scatterplot



Let us start by seeing how well a line will fit the data. Creating a scatterplot with the regression line  $\hat{y} = 8.2200 + 1.2664x$  drawn and we see that the data fits the line reasonably well.

## Scatterplot

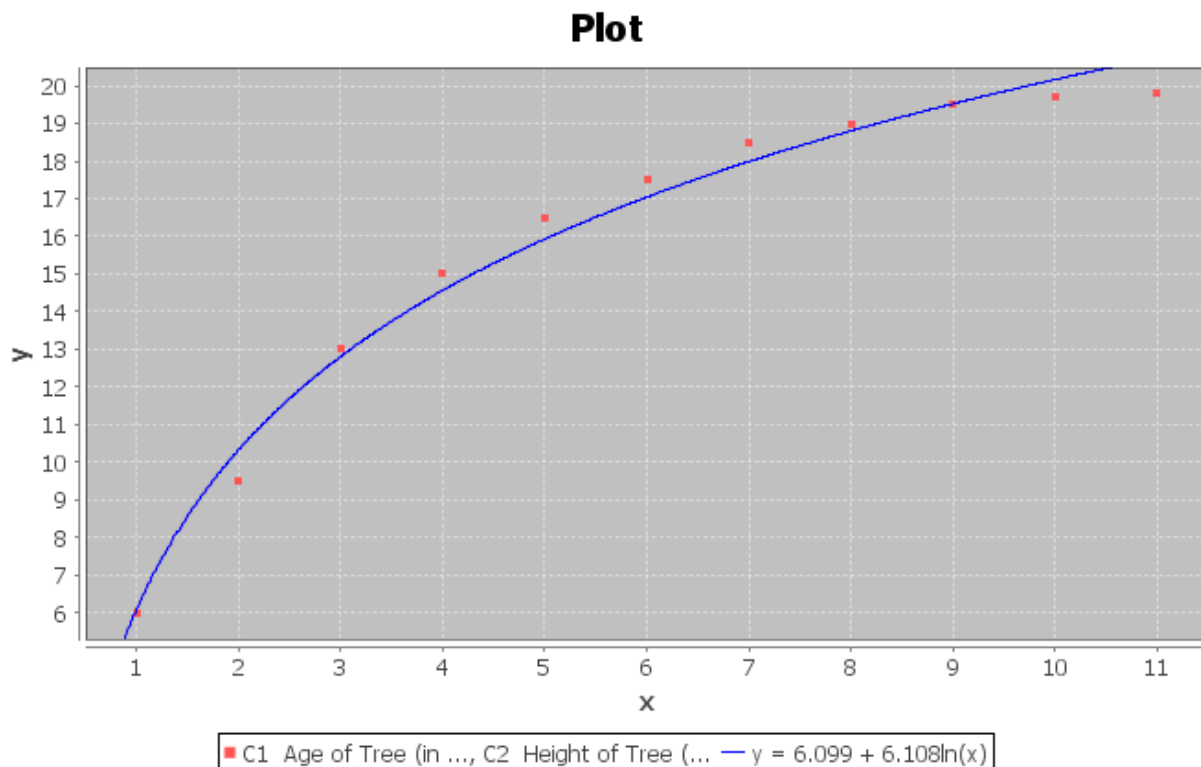




Let us see how well the line really fits. The software calculated the R-squared to be 0.8400 and the Standard Deviation of the Residual Errors to be 1.9325 feet. Therefore, we see that the regression line fits the data reasonably well. Points in the scatterplot are an average of 1.9325 feet from the regression line and predictions with the regression line formula will have an average error of 1.9325 feet.

The line does seem to fit reasonably, but would a curve fit the data even better? Do you notice how after 8 years, the trees start to approach a maximum height of about 20 feet. This causes the scatterplot to take on an upside down L shape. The curve seems to be increasing from left to right, but it is increasing very slowly. This is the shape of another type of curve, the Logarithmic curve. Logarithmic curves (or Log curves for short) have a shape that frequently occurs when we analyze data sets.

For example, we can find out how many years it will take money to grow in your bank account with a Log function. So let us try to graph a Log curve with statistics software that approximates this data set and see what happens.



We can see right away that the Log curve appears to fit the data better than the line. This software uses the Natural Log or (LN) for short. The function came out to be  $\hat{y} = 6.09934 + 6.10818LN(x)$ . The found the function in terms of the Natural Log because it is one of the few types of logarithms you can find on your calculator. Some programs use Log base 10 (LOG). After all, isn't the purpose again of finding this function to use it to predict the height of a tree? So how does logarithms work and in particular the Natural Logarithm function?

### About Logarithms

Logarithms are really the inverse of exponentials. Logs in fact are exponents. When you find the LN(8) for example, on your calculator you are finding an exponent on a particular base that when evaluated gives you an answer of 8. However, what is the base for the Natural Log function? The answer to that question is the number "e". "e" is an irrational number (infinite non-repeating decimal) that is approximately 2.718. Again, "e" is not exactly 2.718 but that is pretty close.



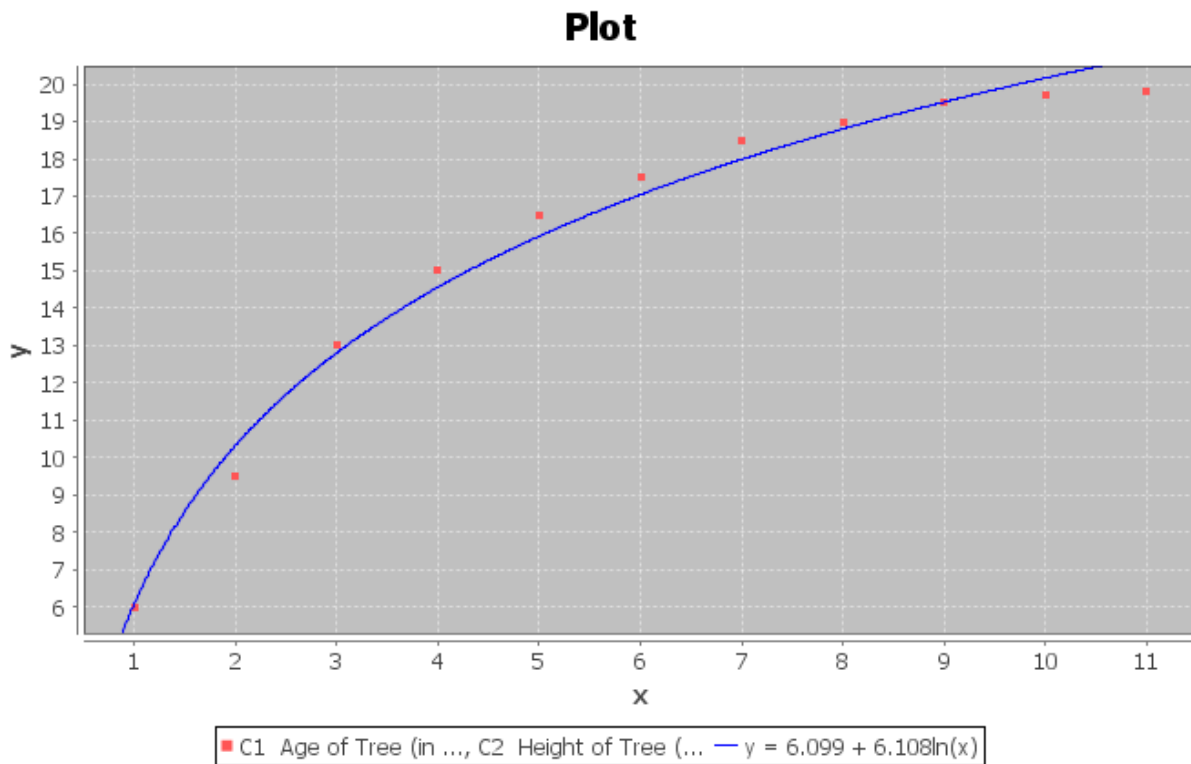
So let us see if we can understand this. When we find the LN(8) on our calculator we are really finding the following exponent ( $e^{??} = 8$ ). If we replace e with 2.718, we get  $2.718^{??} = 8$ . See if you can find the LN(8) on your calculator? Every calculator is different. For most calculators, you will push the “LN” key then the number 8 and then enter or =. For a few calculators, you may have to push the 8 first, then the LN key. You should have gotten an answer of 2.079. Therefore, this implies that  $e^{2.079} \approx 8$ .

Let us plug in some other numbers into the LN function. Find the LN(0) or LN(-5). What does your calculator tell you? It probably said “ERROR” or “UNDEFINED”. There is a reason for this. The number you plug into the LN function is equal to 2.718 to some power. 2.718 to some power will always be a positive number. Hence, we can only plug in positive numbers into the LN function. Do you remember the name for the values of x we are allowed to plug into a formula? You are right. It is called the Domain. So what is the Domain of the natural Log function?

Since we can only plug in positive numbers for x, our Domain is all positive numbers ( $x > 0$ ). This Domain is very common in most basic logarithms. This also implies that if we have negative numbers in our explanatory data set (x values) we should not use a Log function as our model.

### Assessing the fit of a Log function

Let us go back now to our tree data. Statcato found that the natural Log function that best fit the data was  $\hat{y} = 6.099 + 6.108LN(x)$ . Notice the distinctive upside down “L” shape with a slow growth.



But how well does this log curve fit the tree data? One way to measure this is with the standard deviation of the residual errors ( $S_e$ ). As we did with the standard deviation calculation in the line above, we will plug in all the ages (x values) into  $\hat{y} = 6.09934 + 6.10818LN(x)$  and get our predicted  $\hat{y}$  values. Let us try to calculate the predicted height  $\hat{y}$  for a tree that is 2 years old. Plugging in 2 for x in the natural Log equation gives the following.



$$\hat{y} = 6.09934 + 6.10818 \ln(2)$$

$$\hat{y} = 6.09934 + 6.10818 \times 0.69314718$$

$$\hat{y} = 6.09934 + 4.233867745$$

$$\hat{y} \approx 10.3 \text{ feet}$$

When doing the calculation on a calculator, be sure to follow the order of operations. So be sure to do the  $\ln(2)$  first, then the multiplication, and lastly the addition. How far off was this predicted value? Recall that a two-year-old tree in the data set had an actual height of 9.5 feet. By subtracting the actual  $y$  value minus the predicted  $\hat{y}$ , we get that  $y - \hat{y} = 9.5 - 10.33 \approx -0.83$ . If you remember, this number is often called a “residual” and tells us that the ordered pair in the data set  $(2, 9.5)$  was 0.83 feet below the natural Log curve. If we calculate all the predicted values  $\hat{y}$  and the residuals  $y - \hat{y}$ , we will get the following table.

Age of tree (years)	Height of tree (feet)	pred $y$	Residual
1	6	6.099	-0.099
2	9.5	10.33274	-0.83274
3	13	12.80932	0.190676
4	15	14.56649	0.433514
5	16.5	15.92945	0.570553
6	17.5	17.04307	0.456933
7	18.5	17.98462	0.515381
8	19	18.80023	0.199771
9	19.5	19.51965	-0.01965
10	19.7	20.16319	-0.46319
11	19.8	20.74534	-0.94534

Notice again that a positive residual means that the ordered pair was above the natural Log curve and a negative residual means that the ordered pair was below the natural Log curve. To calculate the Standard Deviation of the Residual Errors square all the residuals and add them. If you recall this is called the sum of squares. Now divide by  $n-2$  and take the square root. The statistics software calculated the Standard deviation for us and found it to be 0.5653 feet. Therefore, if we predict the height with the Natural Log curve, we will have an average error of 0.5653 feet.

This is much better than the standard deviation for the regression line of 1.9325 feet we calculated earlier. So not only is the natural log curve a much better fit, but if we use it to make predictions, we will have a much smaller average error. Recall also that the R-squared for the regression line was 0.84. The R-squared for the natural log curve is 0.9863 and is much better than the regression line. Only 84% of the variability in height can be explained by the regression line, while 98.6% of the variability can be explained by the natural log curve. This also re-enforces that the natural log is a much better overall fit.

### Making Predictions with the Log Equations

Since the natural Log equation was a good fit for the tree data. Let us see if we can use it to make predictions. Remember, we should only make predictions in the scope of the data. Since our  $x$  values were between 1 year and 11 years, we should only make predictions for  $1 \leq x \leq 11$ . If we make a prediction for an  $x$  value out of the scope of the data, we should expect more error in the prediction.



Use the natural Log equation  $\hat{y} = 6.099 + 6.108LN(x)$  to predict the height of a tree that is 10.5 years old. Plugging in 10.5 for x in the equation and using the order of operations to simplify we get the following:

$$\begin{aligned}\hat{y} &= 6.099 + 6.108LN(10.5) \\ &= 6.099 + 6.108 \times 2.351375 \\ &= 6.099 + 14.3622 \\ &\approx 20.5 \text{ ft}\end{aligned}$$

Therefore, we expect a tree that is 10.5 years old to be about 20.5 ft. Since we found earlier that the standard deviation of the residual errors was 0.5653 feet, we know that our prediction of 20.5 ft. could have an approximate error of 0.5653 feet.

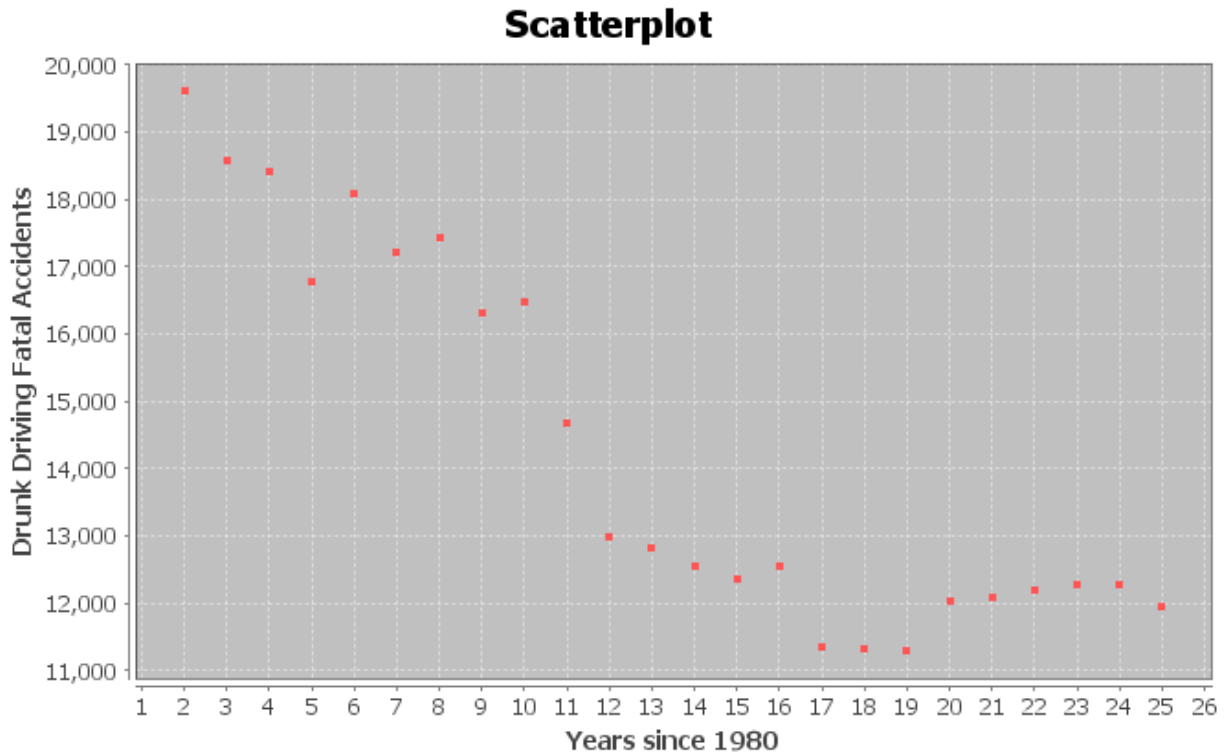
---



## Problem Set Section 7B

NOTE: You will not need Statcato or StatKey to do this assignment. You only need to analyze the graphs, printouts and statistics provided.

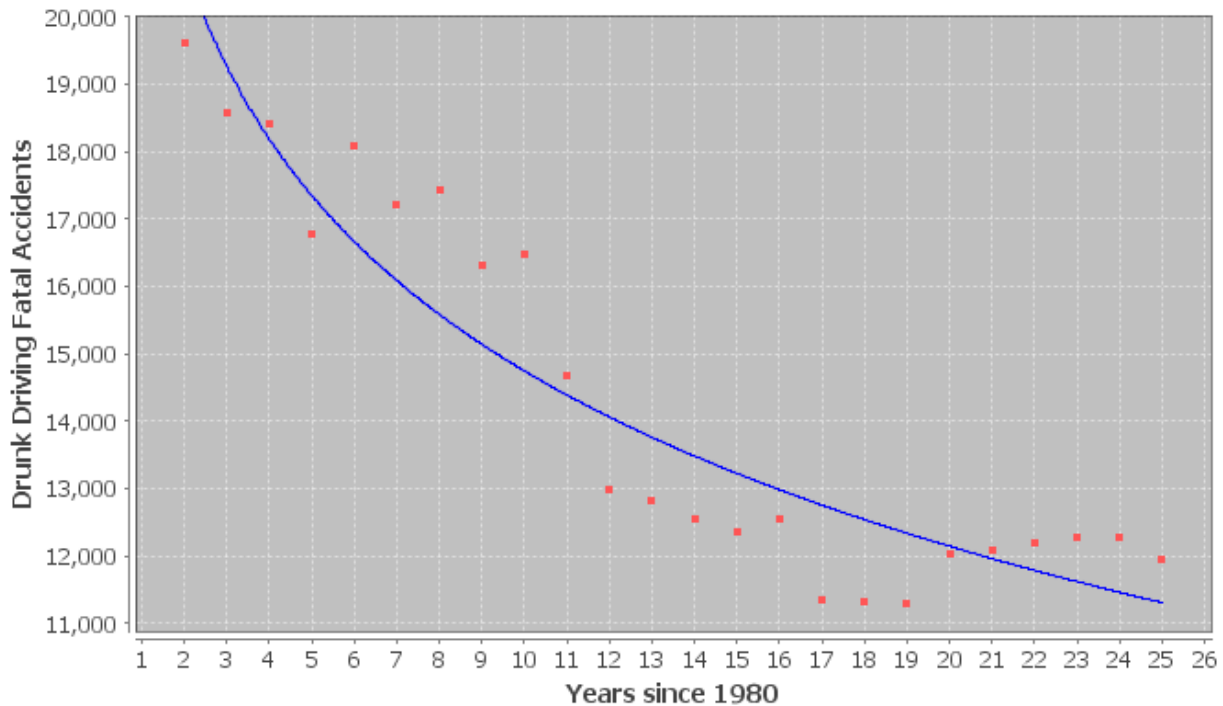
- The following scatterplots and printouts describe the number of years since 1980 and the number of drunk driving fatal accidents. The number of years be the explanatory variable (X) and the number of drunk driving fatal accidents be the response variable (Y).
  - The following graph shows the scatterplot for the data. Does the scatterplot look like a logarithmic model might fit? If so, would it be logarithmic growth or logarithmic decay? What is the scope of the data (x values)?



- The following graph shows the logarithmic curve that that best fits data. Do you think that the logarithmic curve fits the data well? Are the points close to the curve?



## Plot



The following Statcato printout describes the logarithmic relationship between the years since 1980 and the number of drunk driving fatal accidents. Use the printout to answer the following questions.

### Non-Linear Modeling:

x (explanatory variable): Years Since 1980

y (response variable): Drunk Driving Fatal Accidents

**Logarithmic Model:  $y = b_0 + b_1 \ln x$**

$b_0 = 23389.29331$

$b_1 = -3753.61219$

Sample size = 24

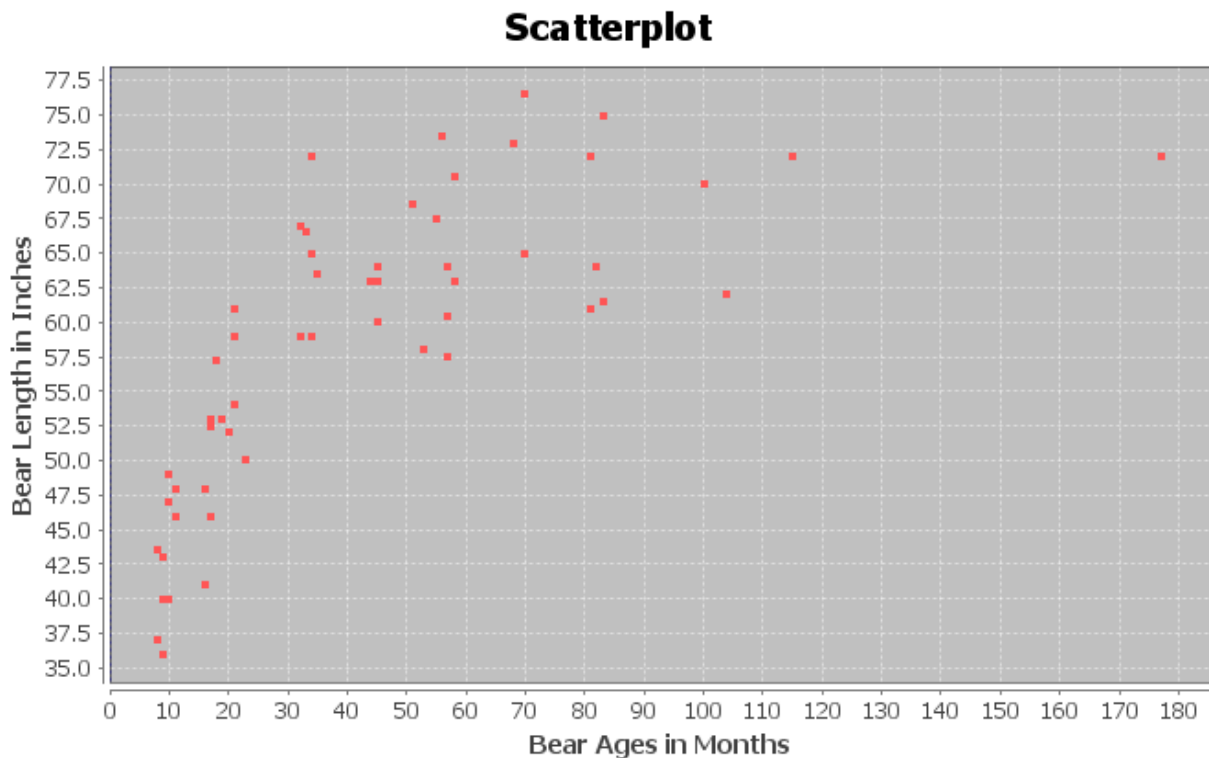
Coefficient of determination  $r^2 = 0.8688$

Standard Deviation of the Residual Errors = 1036.1736 fatal accidents

- How many ordered pairs were in the data?
- What is the coefficient of determination ( $r^2$ )? Write a sentence explaining the meaning of  $r^2$  in this context.
- The standard deviation of the residual errors in Statcato is accurate for log curves. What was the standard deviation of the residual errors? Write two sentences explaining the standard deviation in this context.
- Give the equation of the log curve that best fits the data.
- In the equation of the log curve, was the number in front of "ln x" positive or negative? What does that tell you about whether the curve is logarithmic growth or logarithmic decay? Does this agree with your answer in letter (a)?



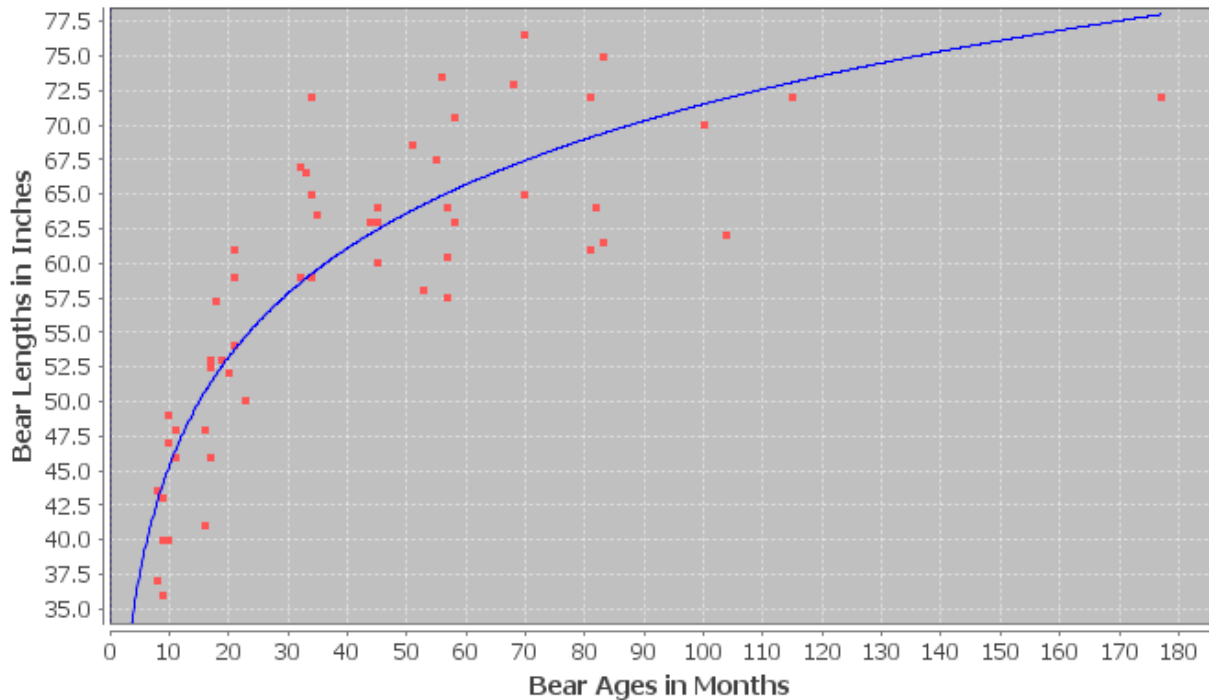
- h) Use the logarithmic equation to predict the number of fatal drunk driving accidents that may occur in year 12.5 (half way through the year 1992). How far off could this prediction be on average?
- i) Use the logarithmic equation to predict the number of fatal drunk driving accidents that may occur in year 23.75 (three quarters of the way through the year 2003). How far off could this prediction be on average?
- j) Do you think it would be all right to extrapolate and use this model to predict the number of fatal car accidents in year 70 (this would be the year 2050)? Why or why not? If a person did make this prediction, would it have the same prediction error as parts (h) and (i)?
2. The following scatterplots and printouts describe the age in months and the length in inches of American black bears. The age is the explanatory variable (X) and the length is the response variable (Y).
- a) The following graph shows the scatterplot for the data. Does the scatterplot look like a logarithmic model might fit? If so, would it be logarithmic growth or logarithmic decay? What is the scope of the data (x values)?



- b) The following graph shows the logarithmic curve that that best fits data. Do you think that the logarithmic curve fits the data well? Are the points close to the curve?



## Plot



The following Statcato printout describes the logarithmic relationship between the ages of bears in months and the length of bears in inches. Use the printout to answer the following questions.

### Non-Linear Modeling:

x (explanatory variable): Bear AGE (months)

y (response variable): Bear LENGTH (inches)

**Logarithmic Model:  $y = b_0 + b_1 \ln x$**

$b_0 = 19.12622$

$b_1 = 11.37504$

Sample size = 54

Coefficient of determination  $r^2 = 0.7539$

Standard Deviation of the Residual Errors = 5.3594 inches

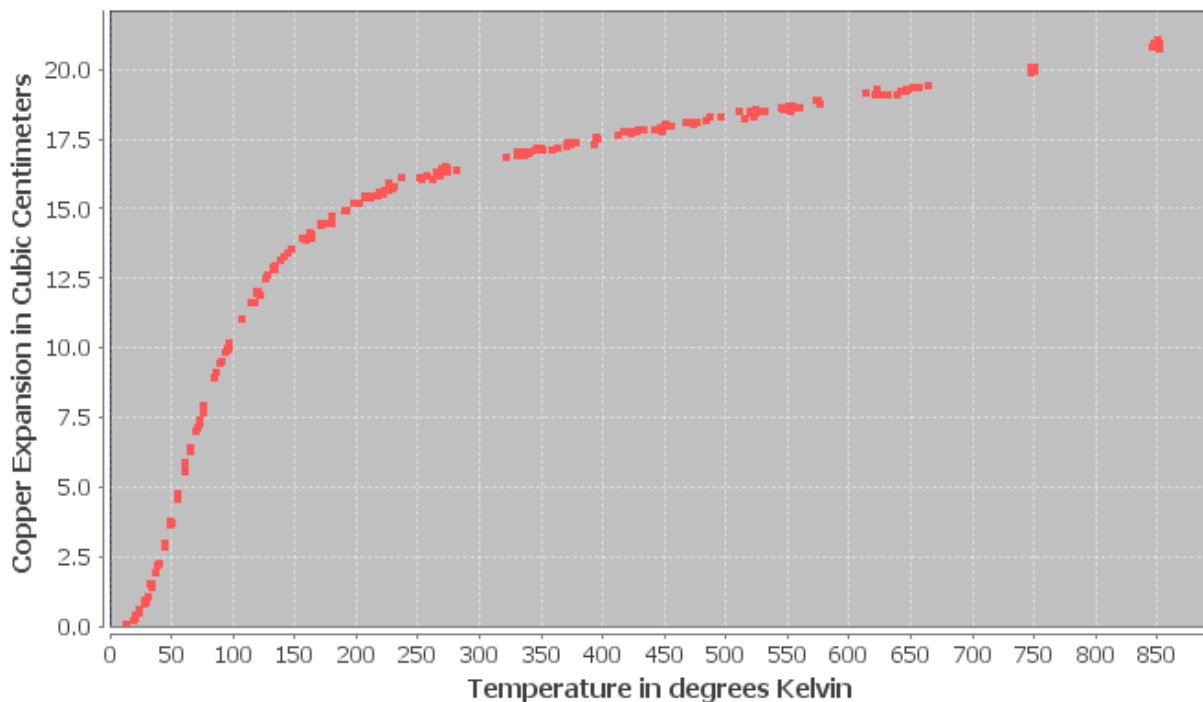
- How many ordered pairs were in the data?
- What is the coefficient of determination ( $r^2$ )? Write a sentence explaining the meaning of  $r^2$  in this context.
- The standard deviation of the residual errors in Statcato is accurate for log curves. What was the standard deviation of the residual errors? Write two sentences explaining the standard deviation in this context.
- Give the equation of the log curve that best fits the data.
- In the equation of the log curve, was the number in front of "ln x" positive or negative? What does that tell you about whether the curve is logarithmic growth or logarithmic decay? Does this agree with your answer in letter (a)?





- h) Use the logarithmic function to predict the length of a black bear that is 48 months old. How far off could this prediction be on average?
- i) Use the logarithmic function to predict the length of a black bear that is 120 months old. How far off could this prediction be on average?
- j) Do you think it would be all right to extrapolate and use this model to predict the length of a bear that is 600 months old? (This would be a 50 year old bear.) Why or why not? If a person did make this prediction, would it have the same prediction error as parts (h) and (i)?
3. The following scatterplots and printouts describe the temperature of copper in degrees Kelvin and how much the volume of the copper expands in cubic centimeters. The temperature is the explanatory variable (X) and the copper expansion is the response variable (Y).
- a) The following graph shows the scatterplot for the data. Does the scatterplot look like a logarithmic model might fit? If so, would it be logarithmic growth or logarithmic decay? What is the scope of the data (x values)?

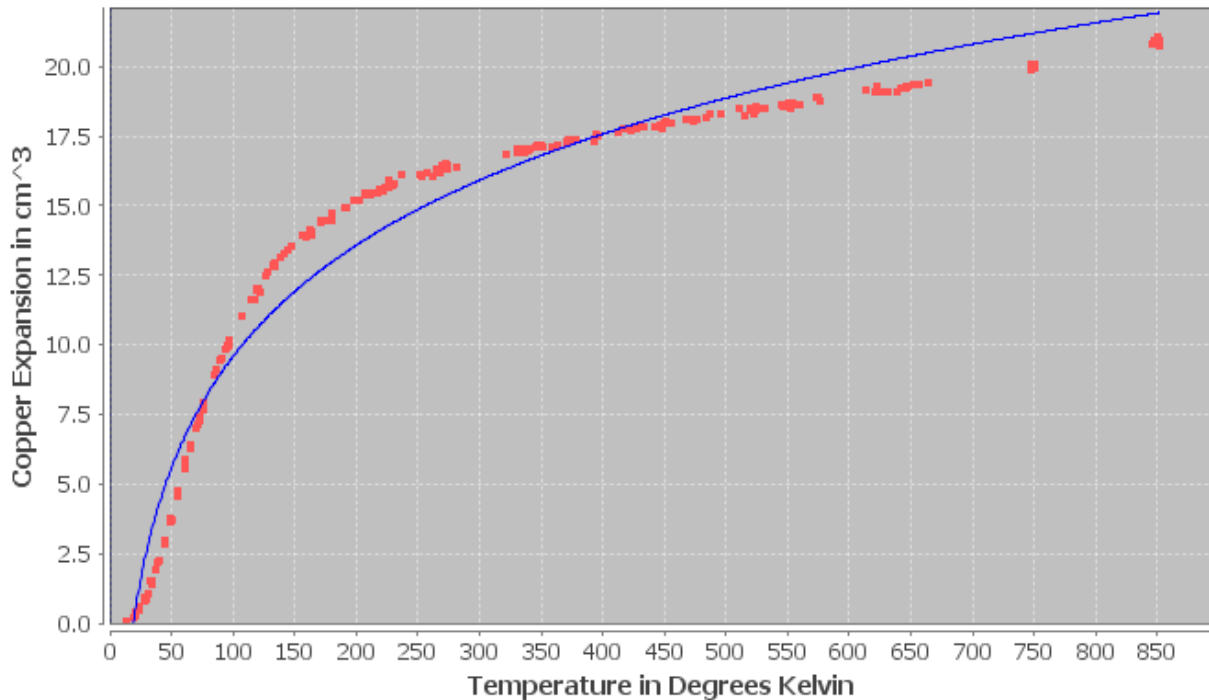
### Scatterplot



- b) The following graph shows the logarithmic curve that that best fits data. Do you think that the logarithmic curve fits the data well? Are the points close to the curve?



## Plot



The following Statcato printout describes the logarithmic relationship between the temperature in degrees Kelvin and the amount copper expands in cubic centimeters. Use the printout to answer the following questions.

### Non-Linear Modeling:

x (explanatory variable): Temperature (Kelvin)

y (response variable): Copper Expansion (cubic cm)

**Logarithmic Model:  $y = b_0 + b_1 \ln x$**

$b_0 = -16.99737$

$b_1 = 5.77086$

Sample size = 236

Coefficient of determination  $r^2 = 0.9628$

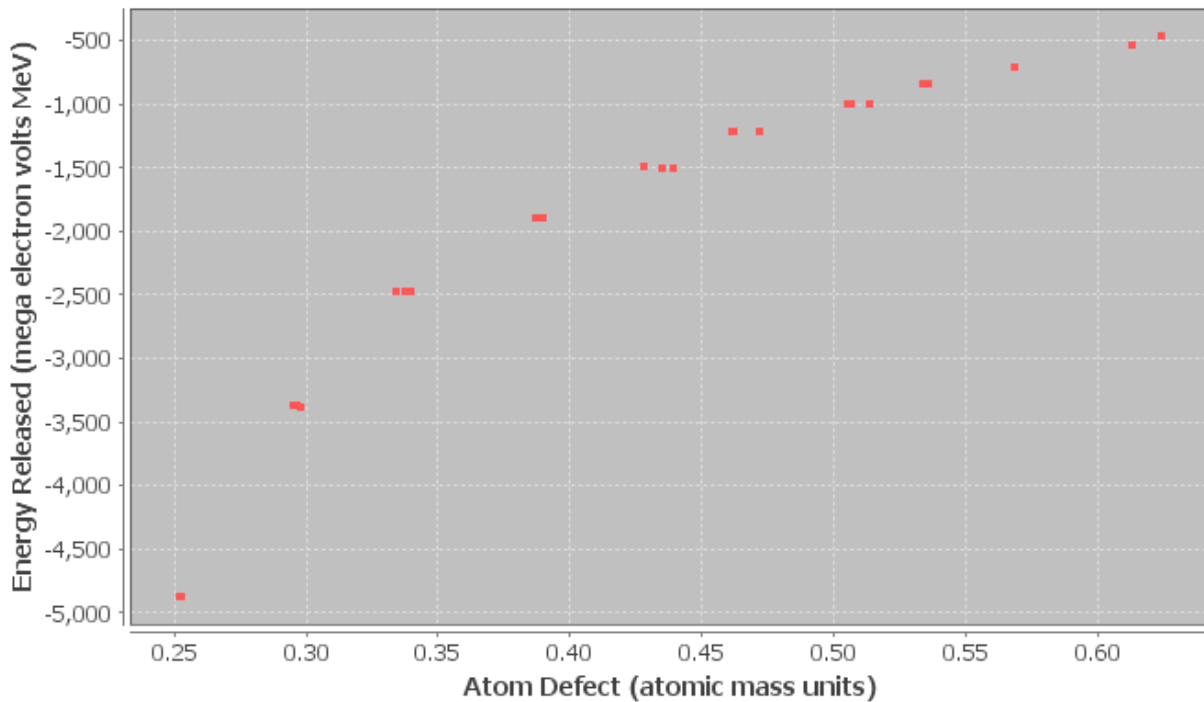
Standard Deviation of the Residual Errors = 1.1149

- How many ordered pairs were in the data?
- What is the coefficient of determination ( $r^2$ )? Write a sentence explaining the meaning of  $r^2$  in this context.
- The standard deviation of the residual errors in Statcato is accurate for log curves. What was the standard deviation of the residual errors? Write two sentences explaining the standard deviation in this context.
- Give the equation of the log curve that best fits the data.
- In the equation of the log curve, was the number in front of "ln x" positive or negative? What does that tell you about whether the curve is logarithmic growth or logarithmic decay? Does this agree with your answer in letter (a)?



- h) Use the logarithmic function to predict the amount of expansion when the temperature is 400 degrees Kelvin. How far off could this prediction be on average?
  - i) Use the logarithmic function to predict the amount of expansion when the temperature is 600 degrees Kelvin. How far off could this prediction be on average?
  - j) Do you think it would be all right to extrapolate and use this model to predict how much copper would expand when the temperature is 1000 degrees Kelvin? Why or why not? If a person did make this prediction, would it have the same prediction error as parts (h) and (i)?
4. The following scatterplots and printouts describe the defects of atoms and the amount of energy released in mega electron volts (MeV). The atomic defect is the explanatory variable (X) and the energy released is the response variable (Y). The release of energy numbers in the data are all negative, denoting the loss of electrons.
- a) The following graph shows the scatterplot for the data. Does the scatterplot look like a logarithmic model might fit? If so, would it be logarithmic growth or logarithmic decay? What is the scope of the data (x values)?

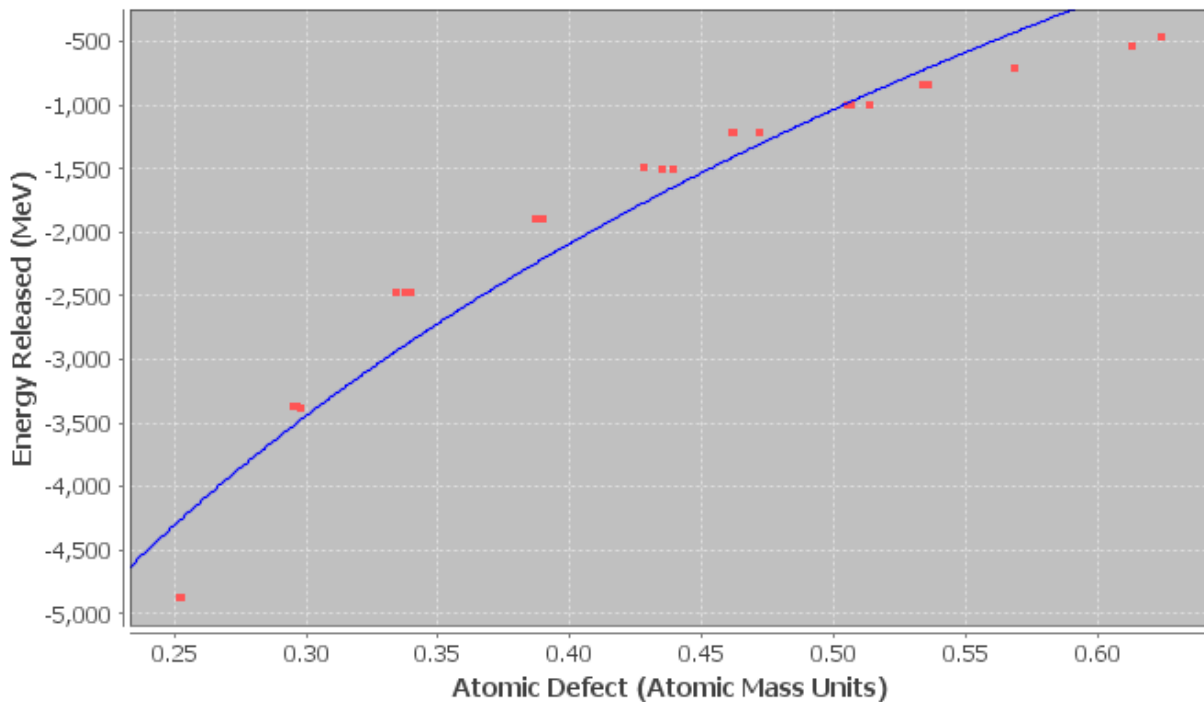
### Scatterplot



- b) The following graph shows the logarithmic curve that that best fits data. Do you think that the logarithmic curve fits the data well? Are the points close to the curve?



## Plot



The following Statcato printout describes the logarithmic relationship between the amount an atom defects and the energy released. Use the printout to answer the following questions.

### Non-Linear Modeling:

x (explanatory variable): Atomic Defect (atomic mass units)

y (response variable): Energy (MeV)

**Logarithmic Model:  $y = b_0 + b_1 \ln x$**

$b_0 = 2241.60751$

$b_1 = 4720.83078$

Sample size = 25

Coefficient of determination  $r^2 = 0.9410$

Standard Deviation of the Residual Errors = 341.8581

- How many ordered pairs were in the data?
- What is the coefficient of determination ( $r^2$ )? Write a sentence explaining the meaning of  $r^2$  in this context.
- The standard deviation of the residual errors in Statcato is accurate for log curves. What was the standard deviation of the residual errors? Write two sentences explaining the standard deviation in this context.
- Give the equation of the log curve that best fits the data.



- g) In the equation of the log curve, was the number in front of “ln x” positive or negative? What does that tell you about whether the curve is logarithmic growth or logarithmic decay? Does this agree with your answer in letter (a)?
- h) Use the logarithmic function to predict the amount of energy released if the atom has a defect of 0.37. How far off could this prediction be on average?
- i) Use the logarithmic function to predict the amount of energy released if the atom has a defect of 0.56. How far off could this prediction be on average?
- j) Do you think it would be all right to extrapolate and use this model to predict the energy released if the atom defect is 0.9? Why or why not? If a person did make this prediction, would it have the same prediction error as parts (h) and (i)?
5. Statcato was unable to find a log function when the explanatory variable (X) had data values that were either zero or negative. However, Statcato had no problem finding a log function when the response variable (Y) was zero or negative. Explain why this happened. What does this tell us about data sets that cannot be modeled with logarithmic models?
6. Draw an exponential growth curve, exponential decay curve, logarithmic growth curve, and logarithmic decay curve and discuss the key features of each curve. What is the relationship between logarithmic functions and exponential functions?
- 



## Section 7C – Quadratic Relationships

Another type of curve seen in scatterplots is the quadratic curve. Quadratic curves have a distinctive “U” shape. This U shape is often called a “parabola”. Because of their parabolic shape, quadratic curves are commonly used to map airplane flights or missile launches. A scatterplot does not have to have a U shape to use a quadratic curve. Quadratic curves can be used to model many different patterns in the scope of the x-values. Think of it as using a piece of the parabola instead of the whole thing. It only has to match in the scope of the x-values.

However, what makes a curve Quadratic? What does the equation look like? Whereas lines have the form  $\hat{y} = mx + b$ , exponential curves are known for the variable exponents, and log curves have “LN(x)” in the formula, Quadratic curves are known for their squared variables. The standard form for a quadratic function is  $\hat{y} = c + bx + ax^2$  where  $a$ ,  $b$  and  $c$  are real numbers.

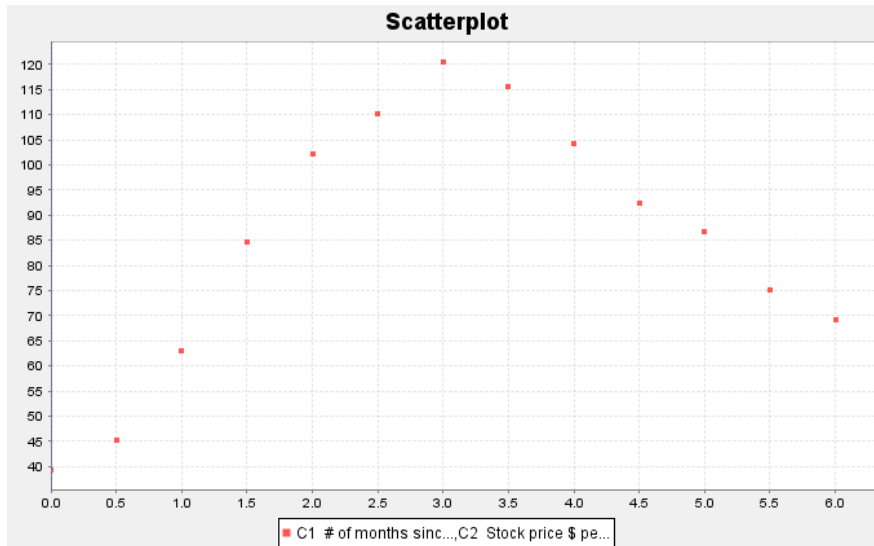
### Example 1

Let us look at the following data set. This gives the value of a stock over a 6-month period. Stocks are notorious for going up and down in value depending on the state of the economy at the time. We let the explanatory variable be the number of months since January 1st, and the response variable is the stock price per share in dollars.

# of months since January 1	Stock price \$ per share
0	39.4
0.5	45.35
1	62.91
1.5	84.71
2	102.31
2.5	110.19
3	120.4
3.5	115.63
4	104.23
4.5	92.5
5	86.61
5.5	75.12
6	69.29

If we make a scatterplot of this data, we get the following graph. Notice the distinctive upside down U shape.



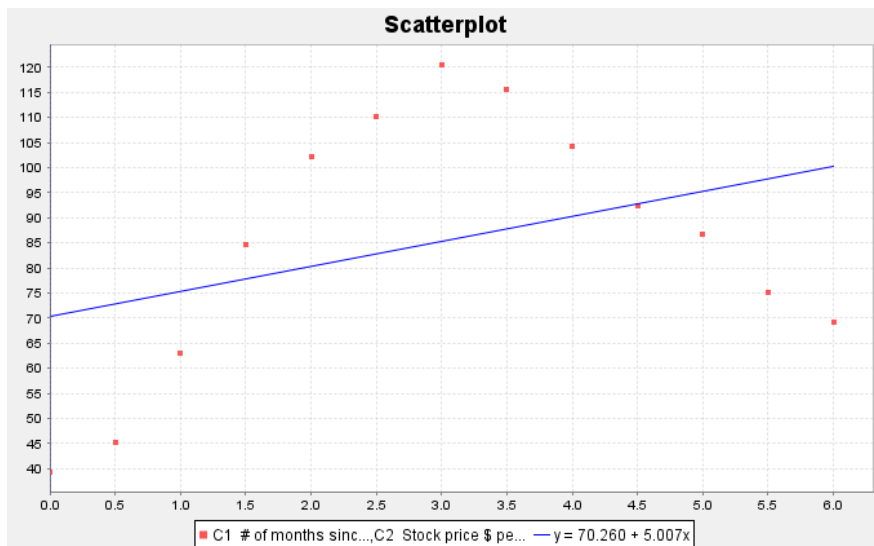


So far, we have discussed the linear, exponential, and log curves. Using statistics software, we found the least squares regression line and the best-fit exponential function for this data.

#### Scatterplot with regression line

Coefficient of determination  $r^2 = 0.1419$

Standard Deviation of Residual Errors = 25.0389



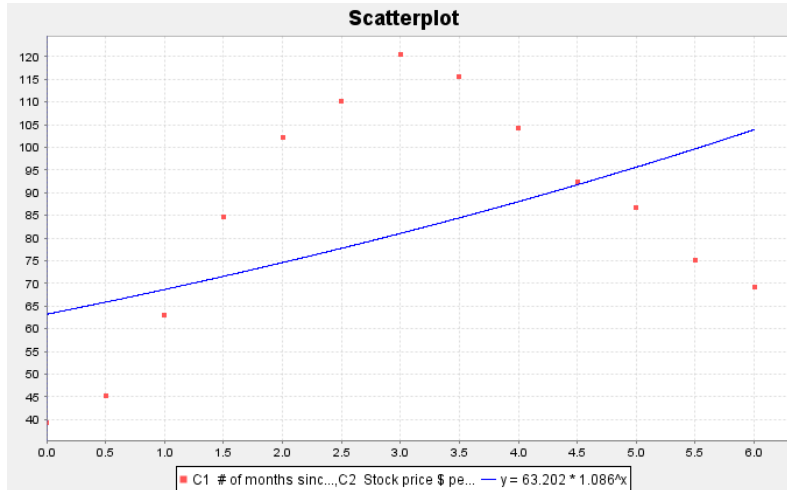
Notice the regression line does not fit the data very well. R-squared is rather low at 0.1419 (14.9%) and the Standard Deviation of the Residual Errors (prediction error) is rather high at \$25.03 per share. So if we use the regression line to predict stock prices we will have an average error or \$25.03. The line is not a good model for this data.



Scatterplot with exponential curve

Coefficient of determination  $r^2 = 0.2123$

Standard Deviation of the Residual Errors = \$26.425056



# months since January	Stock Price \$ per share	predicted Y	Residual	residual squared
0	39.4	63.20152	-23.80152	566.5123543
0.5	45.35	65.87436	-20.52435945	421.249331
1	62.91	68.66024	-5.750235282	33.0652058
1.5	84.71	71.56393	13.14607212	172.8192122
2	102.31	74.59042	27.7195802	768.3751263
2.5	110.19	77.7449	32.44509567	1052.684233
3	120.4	81.03279	39.36720564	1549.77688
3.5	115.63	84.45973	31.17026828	971.5856248
4	104.23	88.0316	16.19840319	262.3882659
4.5	92.5	91.75452	0.745481254	0.5557423
5	86.61	95.63489	-9.024885828	81.44856421
5.5	75.12	99.67936	-24.55935653	603.1619932
6	69.29	103.8949	-34.60487092	1197.497091
			Stand Dev Resid	26.42505633

Notice the exponential curve does not fit the data very well either. R-squared is rather low at 0.2123 (21.23%) and the Standard Deviation of the Residual Errors (prediction error) is rather high at \$26.43 per share. So if we use the exponential curve equation to predict stock prices we will have an average error or \$26.43. Neither the exponential curve nor the regression line are good fits for this data.

*Note: The exponential curve had a slightly higher r-squared than the regression line, but a higher standard deviation. A higher R-squared usually goes with a smaller prediction error. This is probably due to a round. As discussed in previous sections, the computer software has an error in the standard deviation of the residuals calculation. We had to calculate it with excel.*



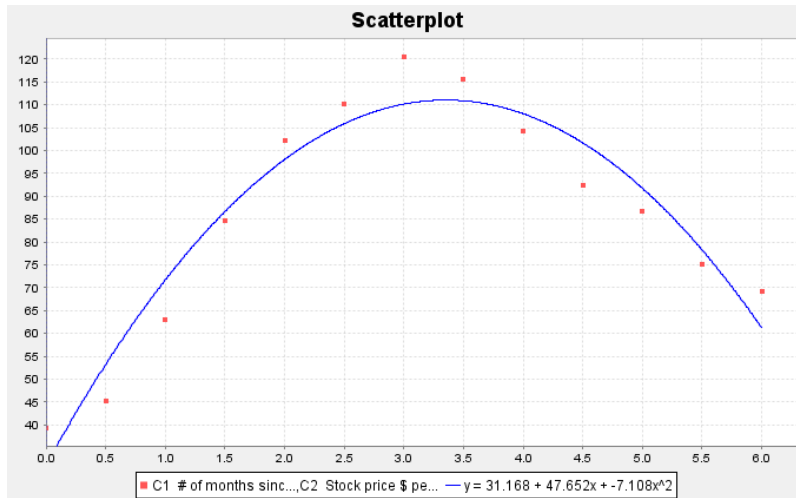


We will now find the quadratic curve that best fits the data.

### Scatterplot with Quadratic Curve

Coefficient of determination  $r^2 = 0.9284$

Standard Deviation of the Residual Errors = 7.5858



The quadratic curve seems to fit the data very well. The R-squared is very high at 0.9284 (92.84%) and the standard deviation of the residual errors is much lower indicating only a \$7.59 prediction error. This means the quadratic is a much better fit than the exponential curve or the regression line.

Let us look at the quadratic equation that the software found.

$$\hat{y} = 31.168 + 47.652x - 7.108x^2$$

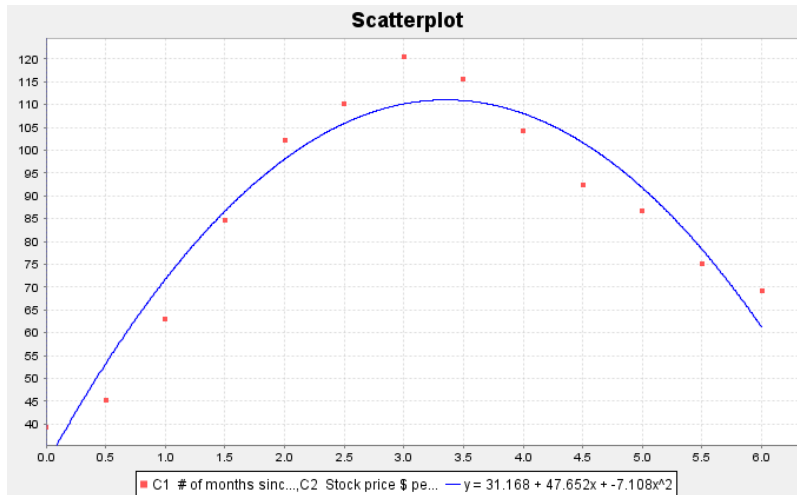
Notice that the number in front of the  $x^2$  term is negative ( $-7.108$ ). This is called the leading coefficient. When the leading coefficient is positive, the quadratic will have an opening up "U" shape, but as in this function, when the leading coefficient is negative, the quadratic will have an upside down "U" shape.

#### Finding the Vertex

Parabolas that open up have a minimum Y value and parabolas that open down have a maximum Y value. This can be important information to businesses trying to find an approximate minimum cost or maximum profits.

In the last example, we looked at some stock data that seemed to take on a parabolic shape. We found the equation of the quadratic curve that fit the data and confirmed that the curve does fit the data by looking at R-squared and the standard deviation of the residual errors.





$$\hat{y} = 31.168 + 47.652x - 7.108x^2$$

One might ask during what month the stock reached a maximum price and what was that maximum price. First notice that since the parabola opens down, the maximum occurred at the top of the parabola. We call this point the vertex. It is important to keep in mind that any point has an x coordinate and a y coordinate. In this problem x represented the number of months since January 1<sup>st</sup> and the y represented the stock price in dollars per share.

So to find the point in time (months) when the stock reached a maximum, we will need to find the x coordinate of the vertex. To find the maximum predicted price (dollars per share), we will need to find the y coordinate of the vertex.

Fortunately, algebra can help us. There is formula for finding the x coordinate of the vertex.

$$X \text{ coordinate of the vertex} = \frac{-1b}{2a}$$

The “b” is the number in front of x and the “a” is the number in front of x-squared. Let us use the formula to calculate the x coordinate of the vertex for the stock price data.

$$X \text{ coordinate of the vertex} = -1b / 2a = -1(47.652) \div 2(-7.108) = (-47.652) \div (-14.216) \approx 3.352$$

What does this tell us? Remember the units. The explanatory (X) variable in this problem was the number of months since January. Therefore, the model predicts that the maximum stock price occurred about 3.352 months after January 1<sup>st</sup>.

What is the predicted maximum stock price? For this, we will need the Y coordinate of the vertex.

Y coordinate of the vertex: Plug in the x coordinate of the vertex into the equation of the quadratic curve and compute the Y value. Make sure to follow the order of operations.

$$\hat{y} = 31.168 + 47.652x - 7.108x^2$$

$$\begin{aligned} Y &= 31.168 + 47.652(3.352) - 7.108(3.352)^2 \\ &= 31.168 + 47.652(3.352) - 7.108(11.235904) \\ &\approx 31.168 + 159.7295 - 79.8648 \\ &\approx 111.03 \end{aligned}$$

Therefore, the predicted maximum stock price is about \$111.03 per share.



### Making Predictions with the Quadratic Curve

Since the quadratic curve fits the stock price data pretty well, let us use the curve to make a prediction.

Let us predict the stock price in mid-February (month 2.5). We would need to plug in 2.5 for  $x$  in the equation of the quadratic curve and compute. Remember to follow the order of operations.

$$\hat{y} = 31.168 + 47.652x - 7.108x^2$$

$$Y = 31.168 + 47.652(2.5) - 7.108(2.5)^2$$

$$= 31.168 + 47.652(2.5) - 7.108(6.25)$$

$$\approx 31.168 + 119.13 - 44.425$$

$$\approx \$105.87$$

Therefore, the predicted stock price in mid-February is about \$105.87 per share.

Remember to be careful with extrapolation.

The scope of the  $x$  values for this data are between 0 and 6. So making predictions out of the scope is called extrapolation and may lead to large prediction errors.

Extrapolation: Let us predict the stock price in mid-September (month 9.5). Notice this is not in the scope of the  $x$  values. Let us plug in 9.5 for  $x$  in the equation of the quadratic curve and see what happens. Remember to follow the order of operations.

$$\hat{y} = 31.168 + 47.652x - 7.108x^2$$

$$Y = 31.168 + 47.652(9.5) - 7.108(9.5)^2$$

$$= 31.168 + 47.652(9.5) - 7.108(90.25)$$

$$\approx 31.168 + 452.694 - 641.497$$

$$\approx -\$157.64 \text{ (Negative!)}$$

So the predicted stock price is about  $-\$157.64$  per share. Notice this does not make much sense and seems to have a huge error in the prediction. That is what can happen when you extrapolate.

### Example 2

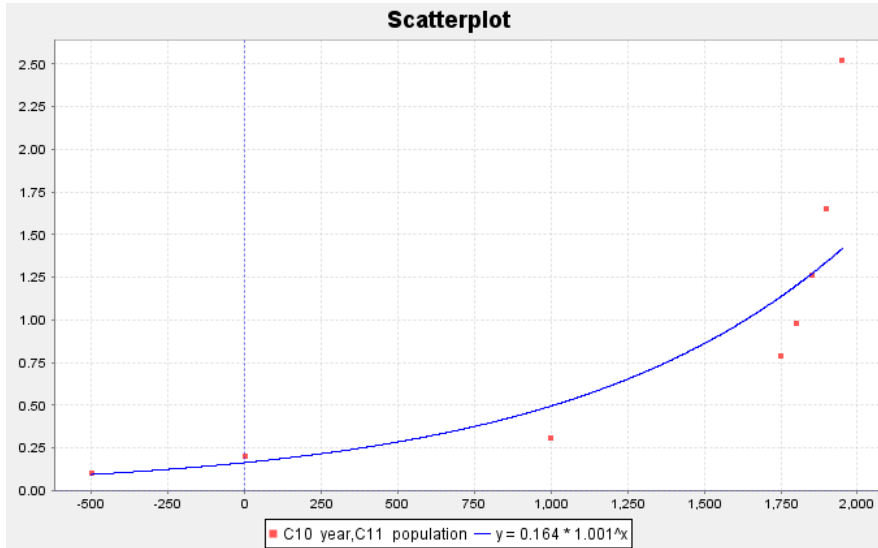
As I said earlier, do not make the mistake of thinking that quadratic functions are only useful when your scatterplot has a U shape. On the contrary, quadratic functions can be a good model for many curves.

For an example of this, let us look again at our world population data. Recall that this data gives the year and the world population from the year 500 BC to the year 1950 AD. The statistics software found that the equation for that exponential curve that fits the data was  $\hat{y} = 0.164(1.001)^x$ . Here is the data and a scatterplot of the data with the exponential function.

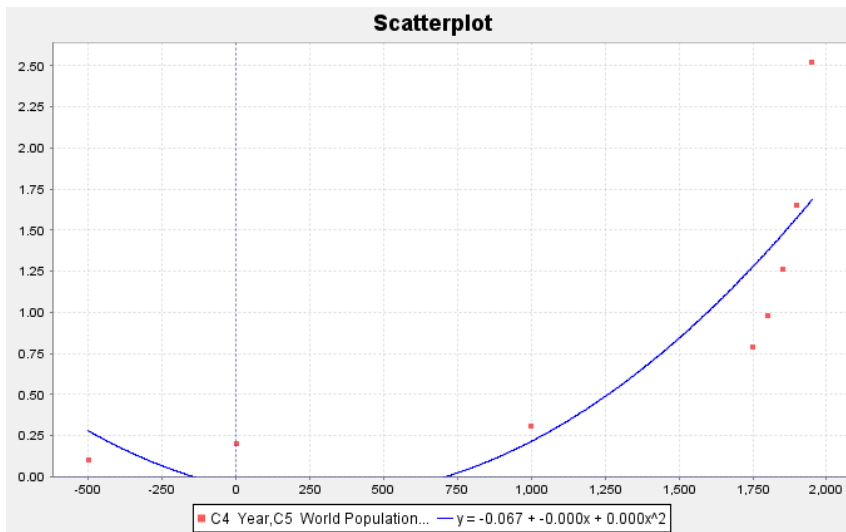
Year	World Population in Billions
-500	0.1
1	0.2



1000	0.31
1750	0.791
1800	0.978
1850	1.262
1900	1.65
1950	2.519



We said that the exponential function fits the data set better than the regression line, but still not perfectly. We may think about whether another function might fit the data better than the exponential. Plugging in this data into the statistics software, we have the program find the quadratic curve that best fits the data. Here is the scatterplot.



Notice that the quadratic function fits the data reasonably well. Looking at the graph, you may see that the leading coefficient says "0.000". This does not mean that the leading coefficient is zero. (If that were the case, this function would not be quadratic.) It just means that the number when rounded to three decimal places is zero. We can get

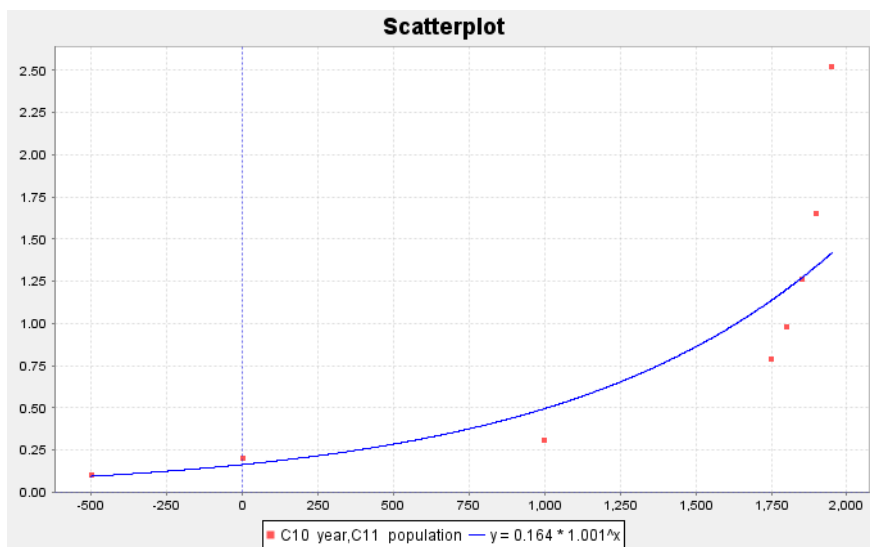
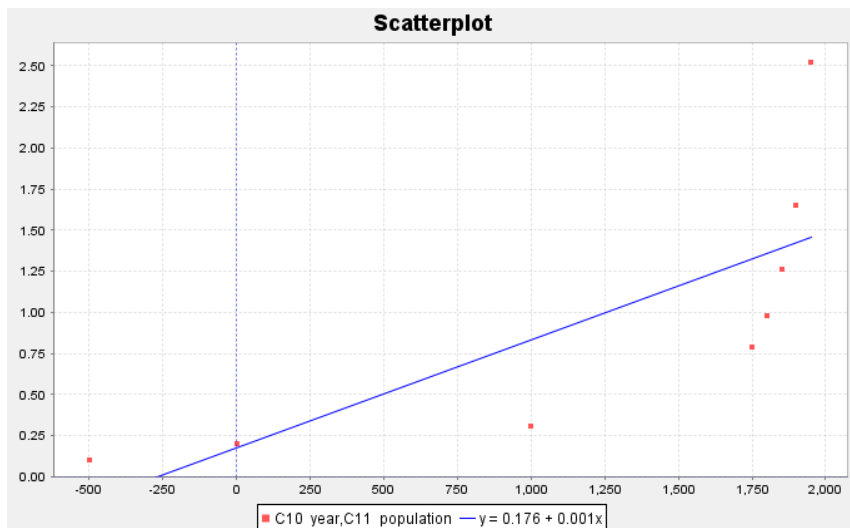


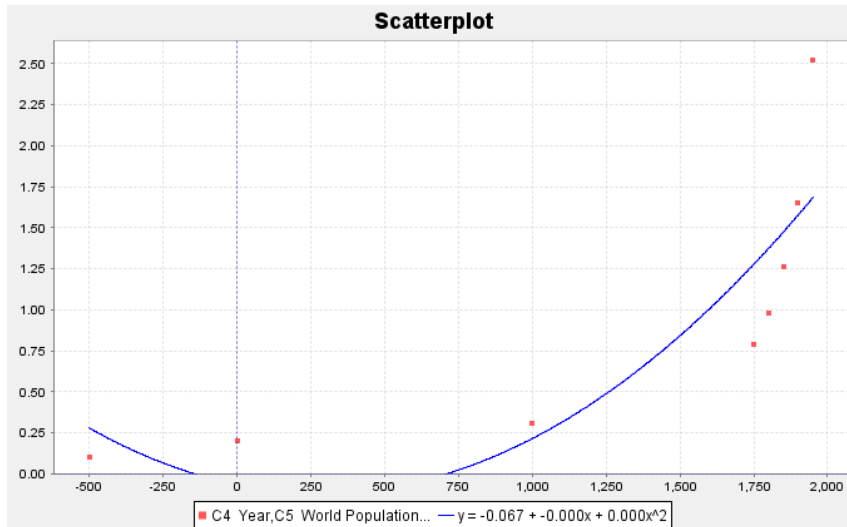
these numbers with better accuracy. We found the curve to be  $\hat{y} = -0.6674 - 0.00037x + 0.0000006479x^2$ . Notice the leading coefficient is positive, which corresponds to the graphs U shape.

Now we have a quandary. We found the linear and exponential functions that fit this data. Now we also have the quadratic function to think about. So which is the best fit for the data?

### Assessing the fit of a quadratic function

Let us start by looking again at the scatterplots. Which curve or line looks like it fits the data the best, the line, the exponential curve, or the quadratic curve?





We definitely can see that the curves fit the data better than the line, but it is hard to tell which of the curves fit the data better. Recall that one way to answer the question of best fit is to look at the R-squared values for the line and each of the two curves.

R-Squared (line) = 0.5878

R-Squared (Exponential curve) = 0.9078

R-Squared (Quadratic curve) = 0.7369

Remember that we do not like to use a more complicated model unless there is a significant improvement. We see from the R-squared values that the quadratic (73.7%) is significantly better than the line (58.8%) but not nearly as good as the exponential (90.8%). For this data set, it seems the exponential is the best model.

We also like to look at the standard deviation of the residual errors ( $S_e$ ). Recall that the curve with the smallest standard deviation is also an indication of best fit. We can use the statistics software to calculate the standard deviations for the line and the two curves.

$S_e$  (line) = 0.5721 Billion people

$S_e$  (exponential) = 0.3673 Billion people

$S_e$  (quadratic) = 0.5007 Billion people.

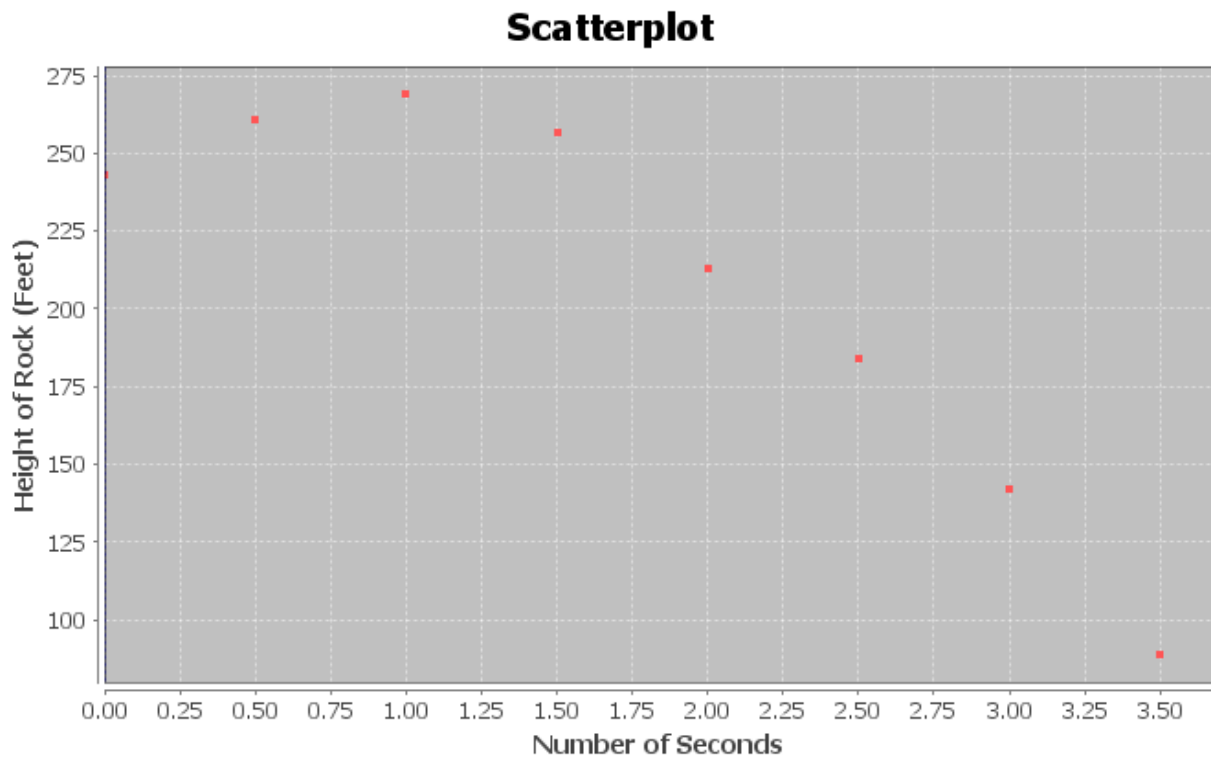
As with the R-squared, the exponential model seems to be the best fit for this data. It has not only the highest R-squared value, but also the lowest standard deviation of the residual errors.



## Problem Set Section 7C

NOTE: You will not need Statcato or StatKey to do this assignment. You only need to analyze the graphs and statistics provided.

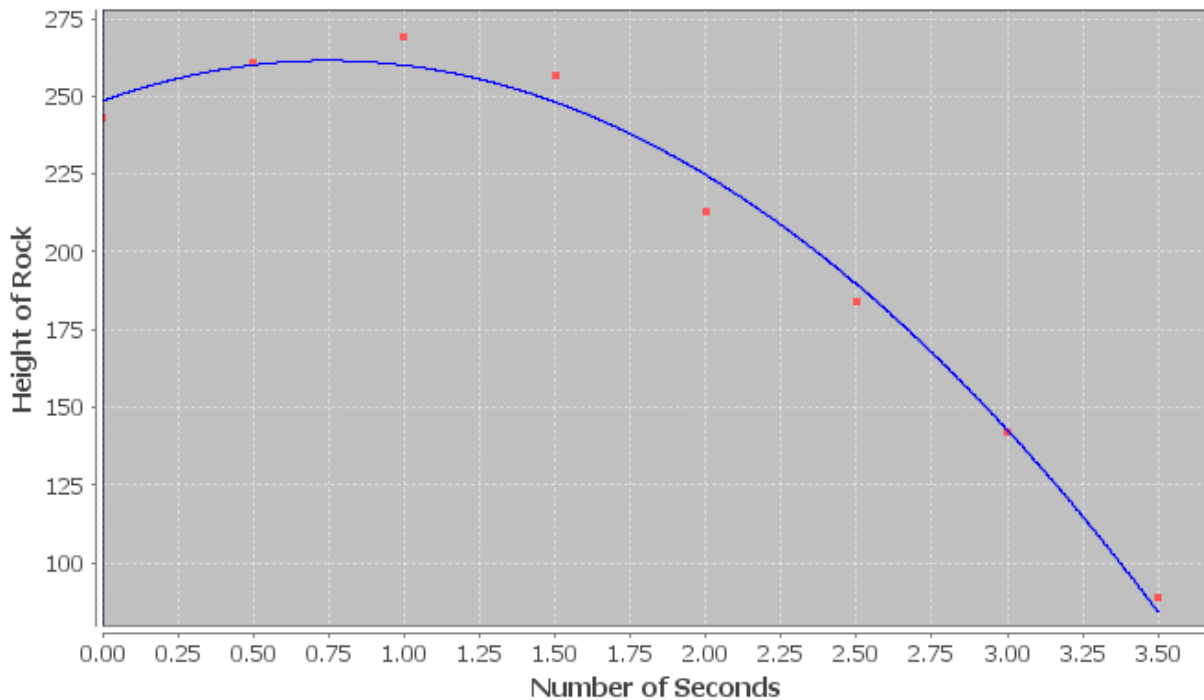
1. A rock was thrown off a 273-foot cliff and this data set gives the number of seconds and the corresponding height in feet of the rock. Let the number of seconds be the explanatory variable and the height be the response variable.
  - a) Look at the following scatterplot. Does the scatterplot have a parabolic shape? If so, would the parabola open up or down?



- b) The following scatterplot and statistics were created with Statcato and describe the quadratic relationship. Do you think that the quadratic function fits the data well? Are the points close to the curve?



## Scatterplot



### Non-Linear Modeling:

x (explanatory variable): Number of seconds

y (response variable): Height of rock (feet)

**Quadratic Model:**  $y = b_0 + b_1x + b_2x^2$

$b_0 = 248.50000$

$b_1 = 34.88095$

$b_2 = -23.38095$

Sample size = 8

Coefficient of determination  $r^2 = 0.9870$

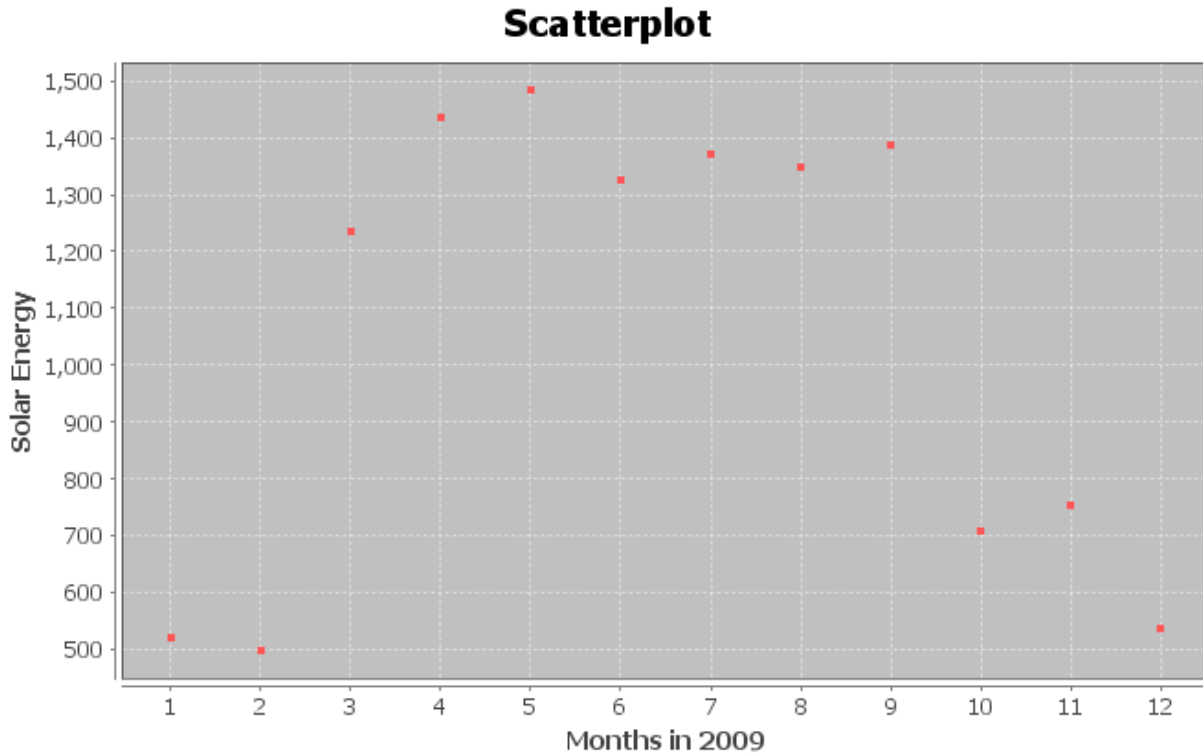
Standard Deviation of the Residual Errors = 8.7402

- What is the equation for the quadratic curve? Was the number in front of  $x^2$  positive or negative? What does this tell us about the shape of the curve.
- What is  $r^2$ ? Write a sentence explaining the meaning of  $r^2$  in this context.
- What was the standard deviation of the residual errors  $s_e$ ? Write two sentences explaining the two meanings of the standard deviation in this context.
- Use the formula  $\frac{-b}{2a}$  to predict the number of seconds will elapse before the rock reaches a maximum height. Plug in your answer into the equation of the quadratic curve for X and calculate the predicted maximum height?





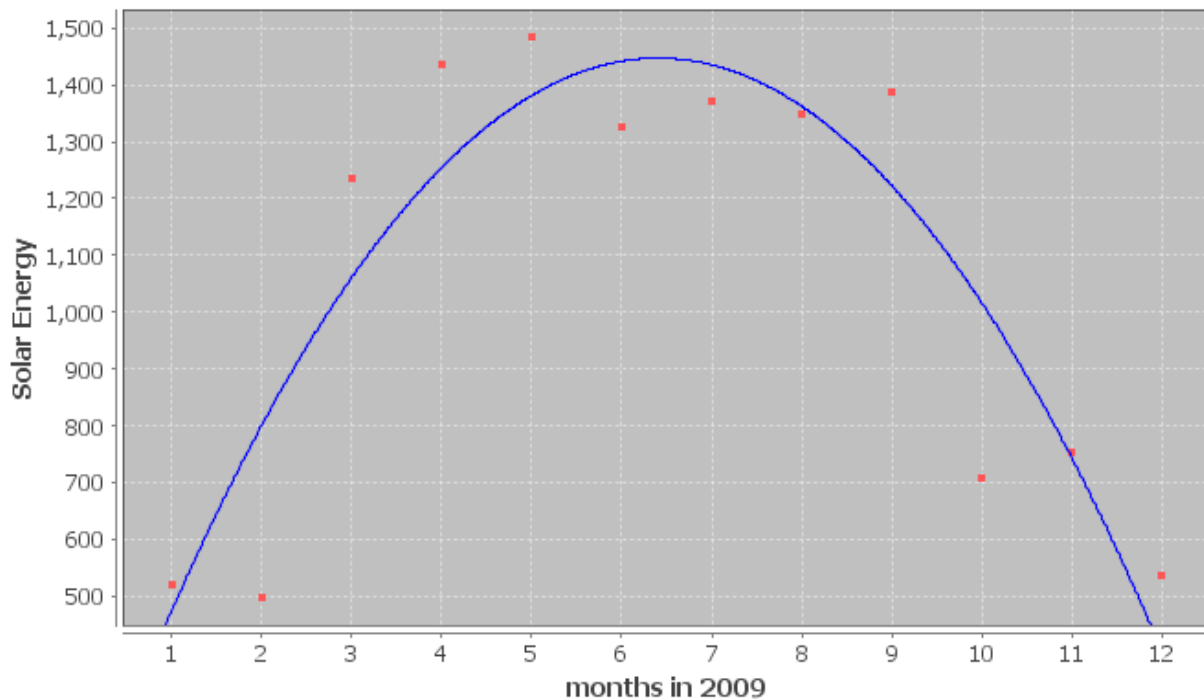
- g) What is the scope of the data (x values)?
- h) Use the quadratic function to predict the height of the rock after 1.8 seconds. How far off could this prediction be on average?
- i) Use the quadratic function to predict the height of the rock after 3.2 seconds. How far off could this prediction be on average?
- j) Do you think it would be all right to extrapolate use this model to predict the height of the rock after 20 seconds? Why or why not? If a person did make this prediction, would the prediction even make sense?
2. Delta college kept track of how much solar energy was made by their solar panels in kilowatt-hours (kWh) for every month in 2009. Let the explanatory variable be the month and the solar energy be the response variable.
- a) Look at the following scatterplot. Does the scatterplot have a parabolic shape? If so, would the parabola open up or down?



- b) The following scatterplot and statistics were created with Statcato and describe the quadratic relationship. Do you think that the quadratic function fits the data well? Are the points close to the curve?



## Plot



### Non-Linear Modeling:

x (explanatory variable): Month in 2009

y (response variable): Solar Energy (kWh)

**Quadratic Model:**  $y = b_0 + b_1x + b_2x^2$

$b_0 = 84.17045$

$b_1 = 425.60047$

$b_2 = -33.22890$

Sample size = 12

Coefficient of determination  $r^2 = 0.8202$

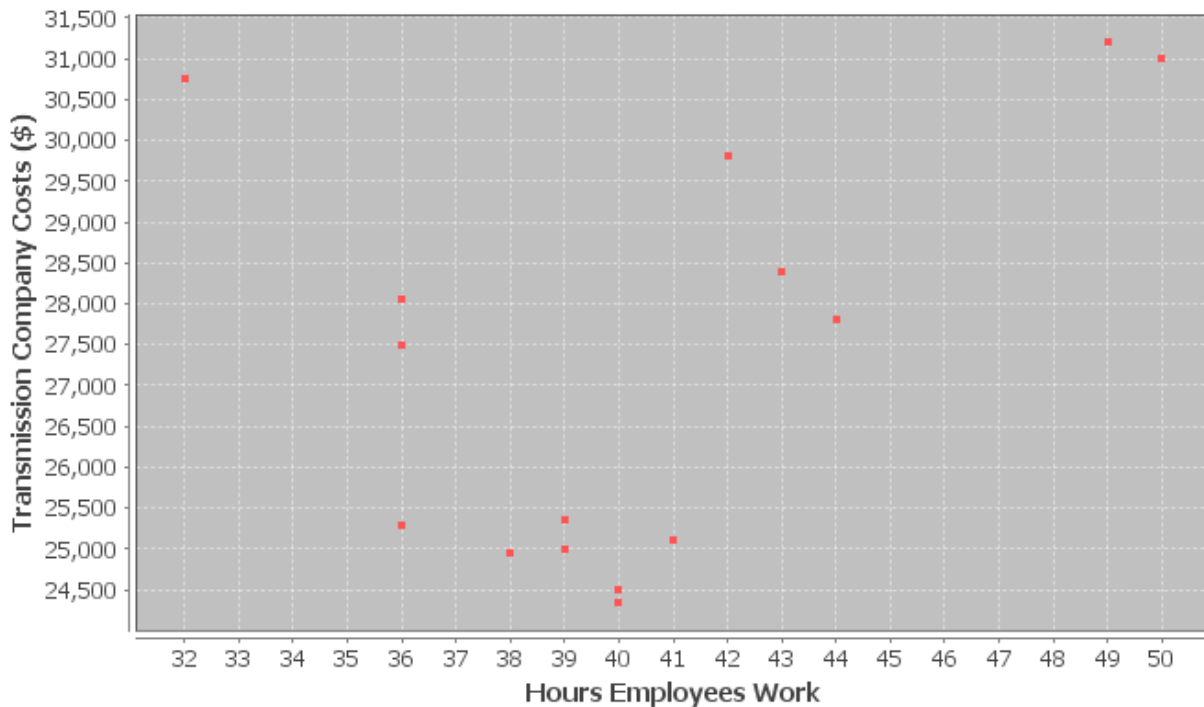
Standard Deviation of the Residual Errors = 189.8436

- What is the equation for the quadratic curve? Was the number in front of  $x^2$  positive or negative? What does this tell us about the shape of the curve.
- What is  $r^2$ ? Write a sentence explaining the meaning of  $r^2$  in this context.
- What was the standard deviation of the residual errors  $s_e$ ? Write two sentences explaining the two meanings of the standard deviation in this context.
- Use the formula  $\frac{-b}{2a}$  to predict what month will have the maximum solar energy. Plug in your answer into the equation of the quadratic curve for X and calculate the predicted maximum amount of energy.
- What is the scope of the data (x values)?



- h) Use the quadratic function to predict the amount of solar energy in mid-March (month 3.5). How far off could this prediction be on average?
- i) Use the quadratic function to predict the amount of solar energy in mid-October (month 10.5). How far off could this prediction be on average?
- j) Do you think it would be all right to extrapolate and use this model to predict the solar energy in January of 2029 (month 240)? Why or why not? If a person did make this prediction, would the prediction even make sense?
3. A company that manufactures transmissions wants to minimize their costs. They think there may be a relationship between their monthly costs and the average number of hours their employees work per week. Let the hours worked be the explanatory variable and the costs be the response variable.
- a) Look at the following scatterplot. Does the scatterplot have a parabolic shape? If so, would the parabola open up or down?

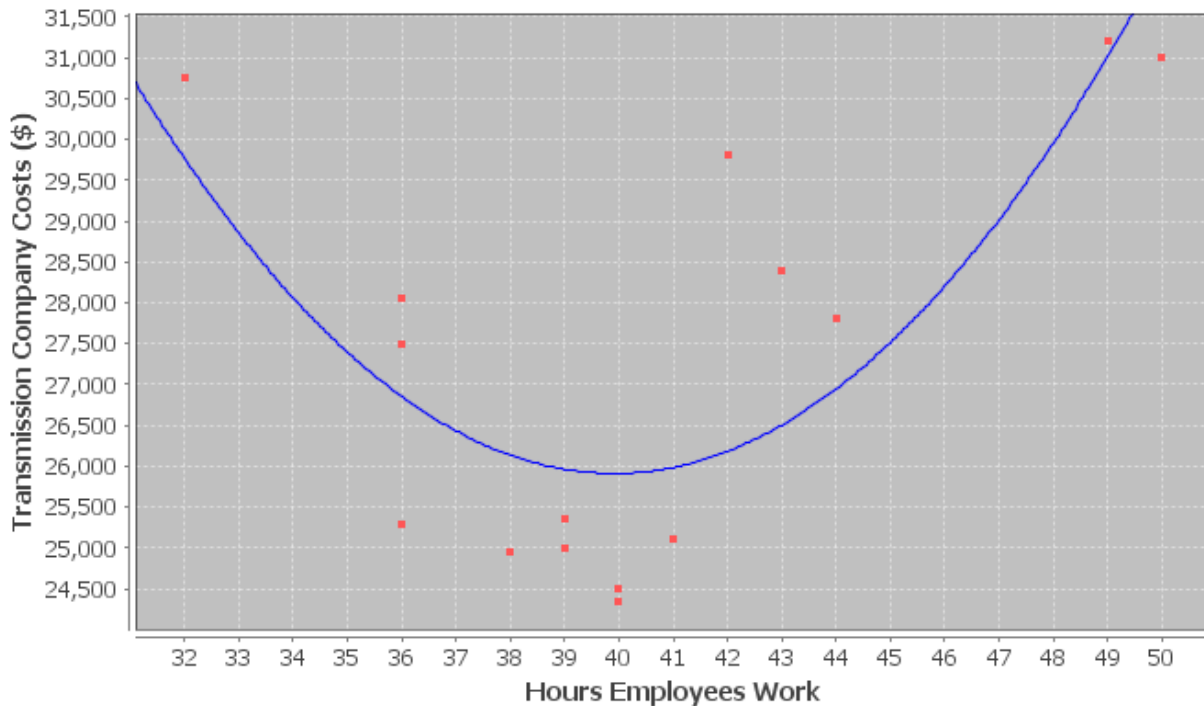
### Scatterplot



- b) The following scatterplot and statistics were created with Statcato and describe the quadratic relationship. Do you think that the quadratic function fits the data well? Are the points close to the curve?



## Scatterplot



### Non-Linear Modeling:

x (explanatory variable): Average Number Hours Employee Works

y (response variable): Transmission Company Costs (\$)..

**Quadratic Model:**  $y = b_0 + b_1x + b_2x^2$

$b_0 = 124202.34526$

$b_1 = -4926.25365$

$b_2 = 61.72503$

Sample size = 15

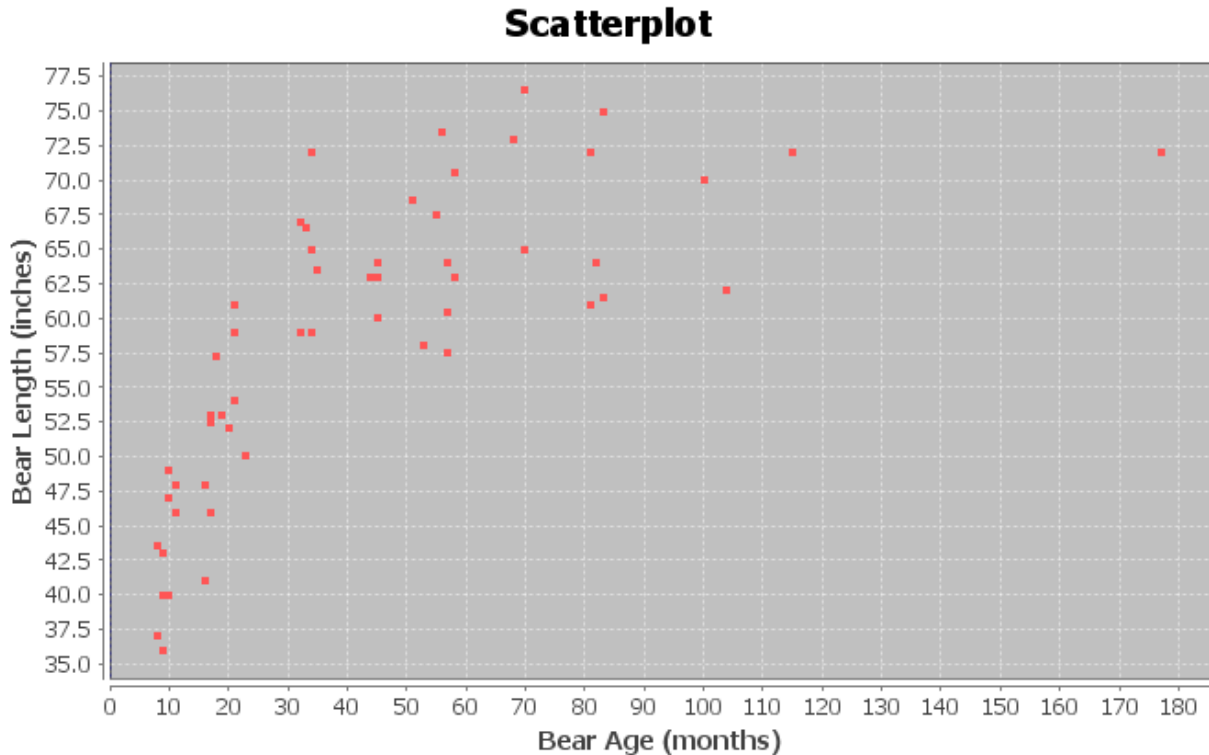
Coefficient of determination  $r^2 = 0.6404$

Standard Deviation of the Residual Errors = 1634.0807

- c) What is the equation for the quadratic curve? Was the number in front of  $x^2$  positive or negative? What does this tell us about the shape of the curve.
- d) What is  $r^2$ ? Write a sentence explaining the meaning of  $r^2$  in this context.
- e) What was the standard deviation of the residual errors  $s_e$ ? Write two sentences explaining the two meanings of the standard deviation in this context.
- f) Use the formula  $\frac{-b}{2a}$  to determine how much their employees should work in order to minimize costs. Plug in your answer into the equation of the quadratic curve for X and calculate the predicted minimum cost.



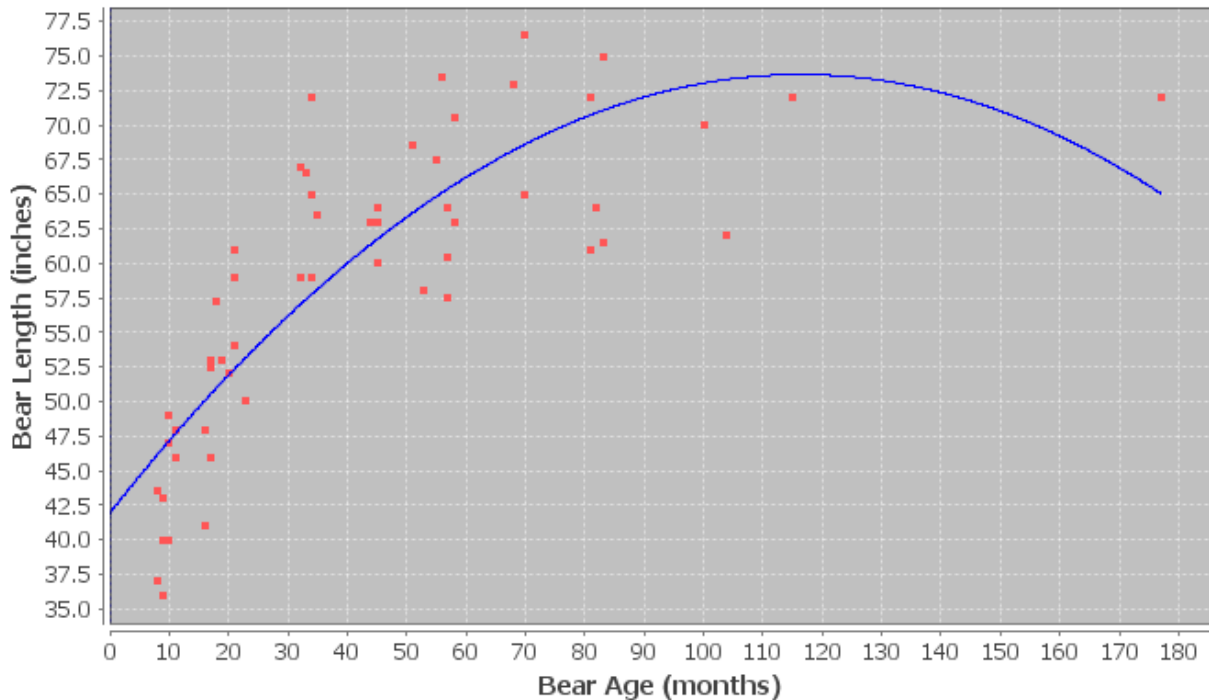
- g) What is the scope of the data (x values)?
- h) Use the quadratic function to predict the monthly costs when the employees work an average of 44 hours per week. How far off could this prediction be on average?
- i) Use the quadratic function to predict the monthly costs when the employees work an average of 35 hours per week. How far off could this prediction be on average?
- j) Do you think it would be all right to extrapolate a lot and use this model to predict monthly costs when employees work an average of 120 hours per week? Why or why not? If a person did make this prediction, would it have the same prediction error as parts (i) and (j)?
4. Let the bear age in months be the explanatory variable and the bear length in inches be the response variable.
- a) Look at the following scatterplot. Does the scatterplot have a parabolic shape? If so, would the parabola open up or down?



- b) The following scatterplot and statistics were created with Statcato and describe the quadratic relationship. Do you think that the quadratic function fits the data well? Are the points close to the curve?



## Scatterplot



### Non-Linear Modeling:

x (explanatory variable): Bear AGE (months)

y (response variable): Bear LENGTH (inches)

**Quadratic Model:**  $y = b_0 + b_1x + b_2x^2$

$b_0 = 41.93861$

$b_1 = 0.54530$

$b_2 = -0.00234$

Sample size = 54

Coefficient of determination  $r^2 = 0.6884$

Standard Deviation of the Residual Errors = 6.0897

- What is the equation for the quadratic curve? Was the number in front of  $x^2$  positive or negative? What does this tell us about the shape of the curve.
- What is  $r^2$ ? Write a sentence explaining the meaning of  $r^2$  in this context.
- What was the standard deviation of the residual errors  $s_e$ ? Write two sentences explaining the two meanings of the standard deviation in this context.
- Use the formula  $\frac{-b}{2a}$  to determine the age of a bear when it reaches its maximum length. Plug in your answer into the equation of the quadratic curve for X and calculate the predicted maximum length.
- What is the scope of the data (x values)?



- h) Use the quadratic function to predict the length of a bear that is four years (48 months) old. How far off could this prediction be on average?
  - i) Use the quadratic function to predict the length of a bear that is ten years (150 months) old. How far off could this prediction be on average?
  - j) Do you think it would be all right to extrapolate a lot and use this model to predict the length of a bear that is 30 years (360 months) old? Why or why not? If a person did make this prediction, would it have the same prediction error as parts (i) and (j)?
5. How can we identify a quadratic function if we only see the equation? How can we know from just the equation of the quadratic function whether it opens up or down?
6. How do we know if the quadratic function has a maximum or minimum point? Where does the maximum or minimum value occur? What are some applications where knowing the maximum or minimum will be important to know?
- 



## Chapter 7 Review Sheet

Here are the important topics to remember from this chapter.

- An exponential growth pattern looks like a backward L shape and increases very quickly from left to right.
- A logarithmic growth pattern looks like an upside down L shape and increases very slowly as the graph goes from left to right.
- Exponential and logarithmic decay patterns both look L shaped and decrease from left to right. The main difference is that a logarithmic decay curve can cross the x-axis but not the y-axis, while the exponential curve can cross the y-axis but not the x-axis.
- Exponential Curves have equation where the x is an exponent. The equation looks like  $y = a \cdot b^x$  where "a" is the y-intercept and "b" is the base. If the base is greater than 1, you will have an exponential growth curve. If the base is less than 1 you will have an exponential decay curve.
- You cannot use the exponential curve if the response variable (Y) has zero or negative numbers in the data set. (There may be ways to adjust the data though.)
- Logarithmic Curves have an "LN (X) in the equation. The equation looks like  $y = a + b \cdot \text{LN}(x)$ . If the number in front of the LN(x) is positive, you will have a logarithmic growth curve. If the number in front of the LN(x) is negative, you will have a logarithmic decay curve.
- You cannot use the logarithmic curve if the explanatory variable (X) has zero or negative numbers in the data set. (There may be ways to adjust the data though.)
- A traditional quadratic pattern has a parabolic "U" shape. The "U" may be facing up or down. A quadratic curve may still be a good model even if the shape is not "U" shaped since we can use a piece of the curve.
- The quadratic curve  $y = c + bx + ax^2$ . The quadratic curve opens up and has a minimum Y value if the leading coefficient "a" is positive. The quadratic curve opens down and has a maximum Y value if the leading coefficient "a" is negative. The quadratic curve works well with positive numbers, negative numbers and zero.
- The quadratic curve  $y = c + bx + ax^2$  has a maximum or minimum point at the vertex. The x coordinate of the vertex can be calculated with  $-b/2a$ . The y coordinate of the vertex can be calculated by plugging in  $-b/2a$  in for x in the formula. The vertex may not always make sense, especially if the vertex is out of the scope of the x values.
- R-squared is the percent of variability in the response variable (Y) that can be explained by the (exponential, logarithmic, or quadratic) relationship with the explanatory variable (X). R-squared is a very useful number to judge how well the curve is fitting the data. The higher the r-squared percentage the better the fit.
- The standard deviation of the residual errors measures how far the points in the scatterplot are from the (exponential, logarithmic, or quadratic) curve on average. It also tells us the average prediction error if we use the equation to make a prediction in the scope of the x values.
- The curve with the highest r-squared and lowest standard deviation is generally the best-fit curve. However, statisticians also look at things like outliers, residual plots, and histograms of the residuals when judging the fit of a curve.

---

### Problem Set Chapter 7 Review

(For #1-4) Multiple Choice: Match each of the following scatterplots with one of the following patterns:

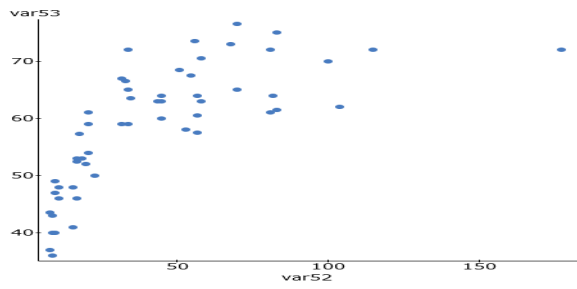
- Exponential Growth
- Logarithmic Growth
- Exponential/Log Decay
- Open Up Quadratic
- Open Down Quadratic



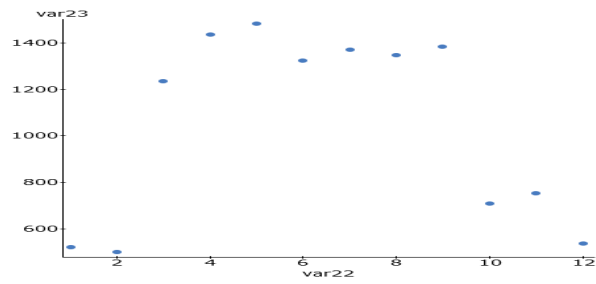
This material is from *Introduction to Data Analysis*, 2<sup>nd</sup> edition by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a "CC-BY" [Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by/4.0/) – 3/17/2021



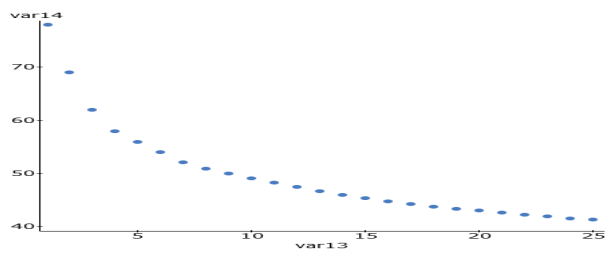
1.



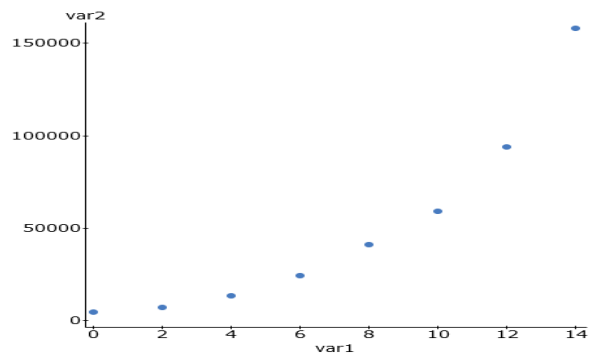
2.



3.



4.



5. A local business, decided to do an experiment. They wanted to see if there is a relationship between the number of lunch and snack breaks they gave their employees and how efficient their employees worked. Each week, a computer randomly selected how many breaks each employee would get, and then measured how efficient the employees were. The explanatory variable X was the number of breaks and the response variable Y was the efficiency rating percentage. After analyzing the data, we found that a quadratic curve fit the data pretty well and the following formula was found with statistics software.

$$Y = c + b x + a x^2$$

$$Y = 41.800 + 5.868 x + ^{-}0.163 x^2$$

a) Use the formula  $\frac{-1b}{2a}$  to find the number of breaks X that the company should give its employees per week in order to maximize their efficiency rating. (Follow the order of operations and show your work.)

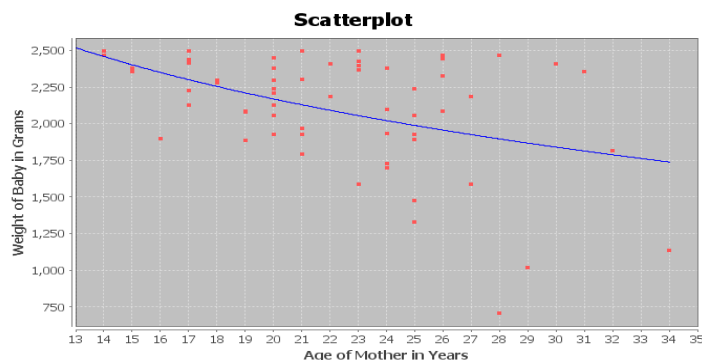
b) What is the predicted maximum efficiency rating Y if the company gives the employees the recommended number of breaks from part (a)? (Hint: Plug in your answer in part (a) into the formula for x (and x-squared) and work it out with your calculator. Follow the order of operations and show your work. Round your rating to the tenths place.)

(For #6-12) The following data describes the relationship between the age of a mother in years and the weight of underweight babies in grams. The age of the mother was the explanatory variable X and the weight of the underweight baby was the response variable Y. We used statistics software to find a Natural Logarithmic function that may fit the data.

Regression Equation:  $y = 4596.59332 + -810.36250 \text{ LN} ( x )$

$$r^2 = 0.1808$$

Standard Deviation of the Residual Errors = 356.8787 grams



6. What percent of the variability in baby weight can be explained by the logarithmic relationship to the age of the mother?

7. How far are the points from the log curve on average?

8. If we use the log curve and the age of the mother to predict the weight of the baby, how far off might that prediction be?



9. Use the formula  $y = 4596.59332 + -810.36250 \text{LN} ( x )$  to predict the babies weight if the mother was 33 years old. (Hint: Plug in 33 for x and work it out with your calculator. Follow the order of operations and round your answer to the ones place. Don't round during the calculation)

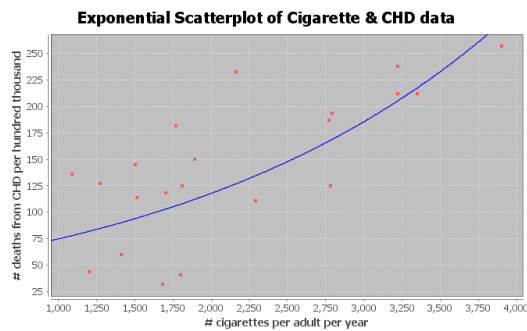
10. How well do you think the Log curve fits the data? Explain your answer with the scatterplot, r-squared, and standard deviation of the residual errors.

11. Does this study prove that a mother's age causes a baby to more underweight? Explain why or why not.

12. List some possible confounding variables that might influence a baby's weight other than just the age of the mother.

(For #13-22) The following data describes the relationship between smoking (# cigarettes per adult per year) and congestive heart disease (CHD) (# deaths per hundred thousand). The number of cigarettes was the explanatory variable X and the deaths by CHD was the response variable Y. Plugging the data into a statistics software, we tried both an exponential curve and a quadratic curve.

*Exponential Scatterplot, Regression Equation, R-squared, Standard Deviation of Residuals*

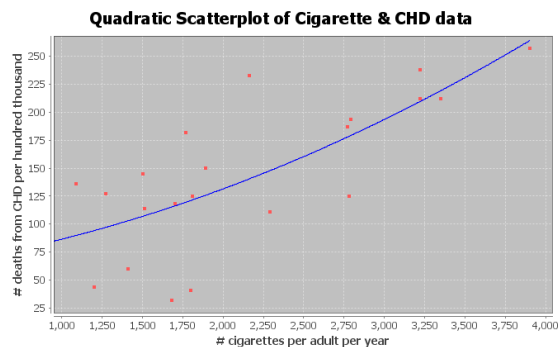


Exponential Equation:  $Y = 47.49274 ( 1.00045^X )$

(Exponential) R-squared = 0.3746

(Exponential) Standard Deviation of the Residual Errors = 48.297 CHD deaths (The computer said 0.4865 deaths but this is a mistake.)

*Quadratic Scatterplot, Regression Equation, R-squared, Standard Deviation of Residuals*



Quadratic Equation:  $Y = 58.59448 + 0.01939 X + 0.00000852869 X^2$

(Quadratic) R-squared = 0.5397

(Quadratic) Standard Deviation of the Residual Errors = 47.5847 CHD deaths



13. What percent of the variability in CHD deaths can be explained by the exponential relationship to the number of cigarettes?
  14. What percent of the variability in CHD deaths can be explained by the quadratic relationship to the number of cigarettes?
  15. Which curve (quadratic or exponential) had the strongest relationship? Explain your answer using R-squared.
  16. How far are the points from the exponential curve on average?
  17. How far are the points from the quadratic curve on average?
  18. Which curve (quadratic or exponential) were the points in the scatterplot closer to?
  19. If we use the exponential curve and the number of cigarettes per adult per year to predict the number of deaths by CHD, how far off might we be in that prediction.
  20. If we use the quadratic curve and the number of cigarettes per adult per year to predict the number of deaths by CHD, how far off might we be in that prediction.
  21. Which curve (quadratic or exponential) has less prediction error?
  22. Which curve (quadratic or exponential) was the better fit for the cigarette and CHD data? Explain why using the R-squared and the standard deviation of the residuals.
- 



# Appendix A: Answer Keys

## Introduction to Data Analysis (2<sup>nd</sup> Edition)

### Chapter 1 Answer Keys

#### Section 1A Answers

##### 1. Bear Data

Age: Quantitative (units: months)  
Month Bear Measured: Categorical  
Gender: Categorical  
Head Length: Quantitative (units: inches)  
Head Width: Quantitative (units: inches)  
Neck Circumference: Quantitative (units: inches)  
Length: Quantitative (units: inches)  
Chest: Quantitative (units: inches)  
Weight: Quantitative (units: pounds)

##### 2. Cereal Data

Name of Cereal: Categorical  
Manufacturer: Categorical  
Target: Categorical  
Shelf Displayed: Categorical  
Calories: Quantitative (units: number of calories per serving)  
Carbs: Quantitative (units: grams per serving)  
Fat: Quantitative (units: grams per serving)  
Fiber: Quantitative (units: grams per serving)  
Potassium: Quantitative (units: milligrams per serving)  
Protein: Quantitative (units: grams per serving)  
Sodium: Quantitative (units: milligrams per serving)  
Sugar: Quantitative (units: grams per serving)  
Vitamins: Quantitative (units: % of daily need per serving)  
Consumer Report Magazine: Quantitative (units: Consumer Report Rating Points)  
Serving Size: Quantitative (units: cups per serving)  
Weight: Quantitative (units: ounces per serving)

##### 3.

- a) Milligrams of Aspirin: Quantitative
- b) Types of Cars: Categorical
- c) Smoke Marijuana or not: Categorical
- d) Number of Bicycles: Quantitative
- e) Types of Birds: Categorical
- f) Grams of Gold: Quantitative
- g) Types of Cardio Classes: Categorical
- h) Number of Cardio Classes: Quantitative
- i) City: Categorical
- j) Money in Bank Accounts: Quantitative
- k) Zip Codes: Categorical
- l) Driver's License Numbers: Categorical
- m) Number of Taxis: Quantitative



## Section 1B Answers

1.

- a) Population of Interest: All students at the college.
- b) Method: Voluntary Response
- c) Will not represent the population very well. There is sampling Bias, since the individuals were not chosen randomly.

2.

- a) Population of Interest: All students at the high school.
- b) Method: Convenience
- c) Will not represent the population very well. There is sampling bias, since the individuals were not chosen randomly.

3.

- a) Population of Interest: All voters in Jamie's city.
- b) Method: Simple Random Sample
- c) Will represent the population well as long as there is no other types of bias present. No sampling bias.

4.

- a) Population of Interest: All employees at the company.
- b) Method: Census
- c) Census is better than a random sample. Will represent the population very well as long as there is no other types of bias present. No sampling bias.

5.

- a) Population of Interest: All people in Portland, Oregon.
- b) Method: Convenience
- c) Will not represent the population very well. There is sampling bias, since the individuals were not chosen randomly.

6.

- a) Population of Interest: All people in Toronto.
- b) Method: Simple Random Sample
- c) Will represent the population well as long as there is no other types of bias present. No sampling bias.

7.

- a) Population of Interest: All people that come to Hugo's library.
- b) Method: Census
- c) Census is better than a random sample. Will represent the population very well as long as there is no other types of bias present. No sampling bias.

8.

- a) Population of Interest: All people that use smart phones.
- b) Method: Voluntary Response
- c) Will not represent the population very well. Sampling bias, since the individuals were not chosen randomly.

9.

- a) Population of Interest: All students at that college.
- b) Method: Simple Random Sample
- c) Will represent the population well as long as there is no other types of bias present. No sampling bias.



## Section 1C Answers

1.

- a) Population: All people or objects to be studied. For example, all students at College of the Canyons.
- b) Census: Collecting data from everyone in your population. For example, collecting data from all of the students at college of the canyons.
- c) Sample: Collecting data from a subgroup of the population. For example, collecting data from fifty students at College of the Canyons.
- d) Bias: When data does not reflect the population. For example, friends and family will not represent the population of all people in Los Angeles, CA.
- e) Question Bias: Phrasing a question in order to force people to answer the way you want. For example, we want to collect data on smoking cigarettes, but give the person a lecture on how unhealthy cigarettes are before asking them.
- f) Response Bias: When someone is likely to lie about the answer to a question. For example, asking people how much they weigh in pounds. They may not give you a truthful answer.
- g) Sampling Bias: Not using randomization when collecting sample data. For example, collecting data from only your friends and family. This is not a random sample.
- h) Deliberate Bias: Falsifying or changing your data or leaving out groups from your population of interest. For example, a person might remove all of the data from people that disagreed with their opinion.
- i) Non-response Bias: When people are likely to not answer when asked to provide data. Randomly calling phone numbers to get data, but the person refuses to answer the phone.

2.

Population of interest: All people in the U.S.

Question Bias: The question was phrased to make people feel bad about answering no.

Response Bias: Vaccinations are a controversial issue and many people may feel scared to admit that they don't agree with vaccinations. There will be many people that lie.

Non-response Bias: There will be many people that randomly selected, but refuse to answer the question.

3.

Population of interest: All Americans.

Response Bias: Cocaine users would not feel comfortable answering the question honestly.

Non-response: Many people may be randomly selected, but will chose not to answer the question.

4.

Population of interest: All college students in Canada

Sampling Bias: The data was not collected randomly.

Deliberate Bias: Most of the colleges in Canada were left out since they only got data from a college near house.

Response Bias: Many people lie about their ages.

Non-response Bias: Many people will refuse to answer.

5.

Population of interest: All adults in Palmdale, CA.

Sampling Bias: The individuals were not selected randomly.

Deliberate Bias: Julie skipped streets that looked poor. These people are not being represented in the data.

Response Bias: People often lie about their income.

Non-response: Many people may not be home or refuse to answer the door.



6.

Population of interest: All students at the college

Response Bias: Many people may lie about their mental health status.

Non-response Bias: Many people may refuse to participate.

Question Bias: The question seems to make people feel bad about answering no.

7.

Population of interest: All pills made by the company.

Deliberate Bias: They deleted data that poorly reflected the pharmaceutical company.

8.

Population of interest: All cars made by this manufacturer

Deliberate Bias: They are leaving out all cars brought to private mechanics or other dealerships.

Non-response Bias: Many people may refuse to bring in the car for minor repairs.

Response Bias: Some people may lie about or forget to list problems with the car.

9.

Population of Interest: All people accused of a crime in the U.S.

Deliberate Bias: There is a conflict of interest. Northpointe should not be doing their own validation study. The validation study should be done by independent statisticians.

---

## Section 1D Answers

1. Observational Study: Collecting data without trying to control confounding variables. Data collected by an observational study can show relationships but cannot prove cause and effect.

2. Experiment: A scientific method for controlling confounding variables and proving cause and effect.

3. Explanatory Variable: The independent or treatment variable. In an experiment, this is the variable that causes the effect.

4. Response Variable: The dependent variable. In an experiment this the variable that measures the effect.

5. Confounding Variables (or lurking variables): Other variables that might influence the response variable other than the explanatory variable being studied.

6. Random assignment: A process for creating similar groups where you take a group of people or objects and randomly split them into two or more groups.

7. Placebo: A fake medicine or fake treatment used to control the placebo effect.

8. Placebo Effect: The capacity of the human brain to manifest physical responses based on the person believing something is true.

9. Single Blind: When only the person receiving the treatment does not know if it is real or a placebo.

10. Double Blind: When both the person receiving the treatment and the person giving the treatment does not know if it is real or a placebo.

11.

a) They did use random assignment. The problem says they were randomly put into three groups.



This material is from *Introduction to Data Analysis*, 2<sup>nd</sup> edition by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a “CC-BY” [Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by/4.0/) – 3/17/2021



b) Answers will vary. Confounding Variables: volume of the music, genetics, age, education level, etc.

c) Explanatory Variable: Type of Music

Response Variable: The amount of information they were able to memorize.

d) Control Group: The group that memorized information without music.

Treatment Groups: The groups that memorize information with their favorite music or with hated music.

e) Since the three groups were randomly assigned, they are likely to have similar characteristics (similar variety of ages, education levels, and genetics.) They must make the volume of the music the same for all participants. These steps will control confounding variables and allow the possibility of proving cause and effect.

f) Yes. This experiment controlled confounding variables and since the no music group did significantly better, it proves that listening to music does not cause a person to memorize information better. It shows that memorizing information in silence is better.

12.

a) They did use random assignment. The problem says that participants were randomly put into the control and treatment groups.

b) Answers will vary. Confounding Variables: amount of motion, genetics, age, diet, pregnancy, etc.

c) Explanatory Variable: Taking Dramamine or not.

Response Variable: The amount of motion sickness.

d) Control Group: The group that took a placebo.

Treatment Group: The groups that took Dramamine.

e) Since the two groups were randomly assigned, they are likely to have similar characteristics (similar variety of ages, education levels, and genetics.) They must make the amount of motion the same for all participants. These steps will control confounding variables and allow the possibility of proving cause and effect.

f) Yes. This experiment controlled confounding variables and since the treatment (Dramamine) group had significantly less motion sickness, it proves that Dramamine does decrease the amount of motion sickness.

13.

a) They did use random assignment. The problem says they were randomly put into two groups.

b) Answers will vary. Confounding Variables: age, education level, experience, type of job, where the applicant lives, poverty level, etc.

c) Explanatory Variable: Whether the applicant on the fake resume had a white or African American sounding name.

Response Variable: Whether or not the applicant received a call back for a job interview.

d) Control Group: The group that had a white sounding name.

Treatment Group: The group that had an African American sounding name.

e) Since the two groups were randomly assigned, they are likely to have similar characteristics (similar variety of ages, education levels, experience, type of job, where the applicant lives, poverty level, etc.) These steps will control confounding variables and allow the possibility of proving cause and effect.

f) Yes. This experiment controlled confounding variables and since the applicants with African American names received significantly less call backs for job interviews than for applicants with white sounding names, it shows that whether the applicant had a white or African American sounding name did influence the chances of getting a call back for a job interview. Shows there is racial discrimination in the job market in Boston and Chicago.



## Chapter 1 Review Sheet Answers

1.

- a) Categorical since the data would consist of words.
- b) Quantitative since it is numerical measurement data.
- c) Categorical since the data would consist of words.
- d) Categorical since the data would consist of words.
- e) Quantitative since it is numerical measurement data.
- f) Quantitative since it is numerical measurement data.

2.

- a) Jim can ask every 5<sup>th</sup> student that walks into the COC cafeteria about their salary. This would have a significant amount of sampling bias.
- b) Jim can put a survey on Facebook asking how money COC students make. This would have a significant amount of sampling bias.
- c) Jim can have a computer randomly select student ID numbers and then track down those students whose ID numbers were selected and ask them their salary. This would have no sampling bias.
- d) Jim can ask other students in his COC classes about their salary. This would have a significant amount of sampling bias since it is not a random sample.
- e) Jim can randomly select 10 section numbers at COC, and then go to those classes and get data from everyone in the class. Since he chose the groups randomly, this would not have much sampling bias.
- f) Jim could walk around the COC campus asking female students about their salary. Later he could walk around asking male students about their salary. Later he could compare the female and male student salaries. Since this method was not randomly selected, there would be a lot of sampling bias.

3.

Population: The collection of all people or objects to be studied. For example, a marine biologist could study all dolphins in the world.

Census: Collecting data from everyone in a population. This is the best way to collect data and minimizes sampling bias. For example, suppose our population of interest was the students at Valencia high school. We could collect data from every student at Valencia high school.

Sample: Collecting data from a small subgroup of the population. For example, if our population was all people in Palmdale, CA, we might collect data from fifty people in Palmdale.

Random: When everyone in the population has a chance to be included in the sample. Suppose our population is all COC students. We could have a computer randomly select student ID numbers and then collect data from those students.

Bias: When data does not represent the population. Asking your friends and family will not represent the population of all people in the world.

Statistic: A number calculated from sample data in order to understand the characteristics of the data. Sample mean averages, sample standard deviations, or sample percentages would all be examples of statistics.



4.

**Sampling Bias:** A type of bias that results from collecting sample data that is not random or representative of the population. For example, if our population was all adults in California, and our sample consists of asking our friends and family. To limit this bias, we could take a random sample instead.

**Question Bias:** A type of bias that results when someone phrases the question or gives extra information with the goal of swaying the person to answer a certain way. Instead of asking a person's opinion about raising taxes, the person first gives a speech about how they think raising taxes is terrible. To limit this bias we could simply ask if the person is for raising or lowering taxes and not give any extra information.

**Response Bias:** A type of bias that results when people do not answer truthfully or accurately. Asking people how much they weigh in pounds will result in many people lying about the answer. Instead of asking people, we could weigh them on a scale and assure them the data will not be released.

**Deliberate Bias:** A type of bias that results when the people collecting the data falsify the reports, delete data, or decide to not collect data from certain groups in the population. A common deliberate bias is to delete all of the data that makes your company look bad. We could avoid this bias by not deleting data or falsifying reports. Use the data to improve the company.

**Non-response Bias:** A type of bias that results when people refuse to participate or give data. When calling random phone numbers to collect data, many people will refuse to answer. To limit this bias, we may leave a message asking them to call us back and offering a gift card if they do.

5. Rachael will need a group of volunteers who want to participate in the experiment. She will need to randomly assign the volunteers into two groups. One group will be the treatment group and receive actual nicotine patches. The other group will be the control group and receive a fake patch (placebo). The placebo patch and the real patch should look identical. Patches should be given to patients using a double blind approach. No volunteer in the experiment will know if they are getting the real patch or a placebo. Also those directly giving the patch will not know either. This will control the placebo effect. Randomly assigning the groups will make them alike in many confounding variables. Rachael may also exercise direct control and manipulate the groups so that they are even more alike. There are many confounding variables including the level of addiction, the number of cigarettes smoked previously, genetics, age, gender, stress, job, etc. Answers may vary. Random assignment should control these confounding variables. If the experiment shows that those with the patch have a significantly higher percentage of quitting smoking, then it will prove that using the patch causes a person to quit smoking.

6.

An experiment creates two or more similar groups with either random assignment or using the same people twice. The similar groups control confounding variables and prove cause and effect. An observational study does not create similar groups and does not control confounding variables. An observational study just collects data and analyzes it, so it cannot prove cause and effect.

Experiment Example: Suppose we want to prove that drinking alcohol causes car accidents. We can have a group of volunteers that wish to participate. We create a driving course with cones. All of the volunteers drive the course sober and we keep track of the number of cones struck. All volunteers drive the same car, with no other distractions (no phones or radio). Then we allow the volunteers to drink alcohol until they all have similar blood alcohol content. Then they can re-drive the course and we keep track of the number of cones struck. If the number of cones is significantly more in the drunk drivers, we have proven that drinking alcohol causes car accidents.

Observational Study Example: Suppose we collect data on car accidents and how many of them involved drunk driving. There are many things that influence having a car accidents other than alcohol, so this data would not prove cause and effect.

---



## Introduction to Data Analysis (2<sup>nd</sup> Edition) Chapter 2 Answer Keys

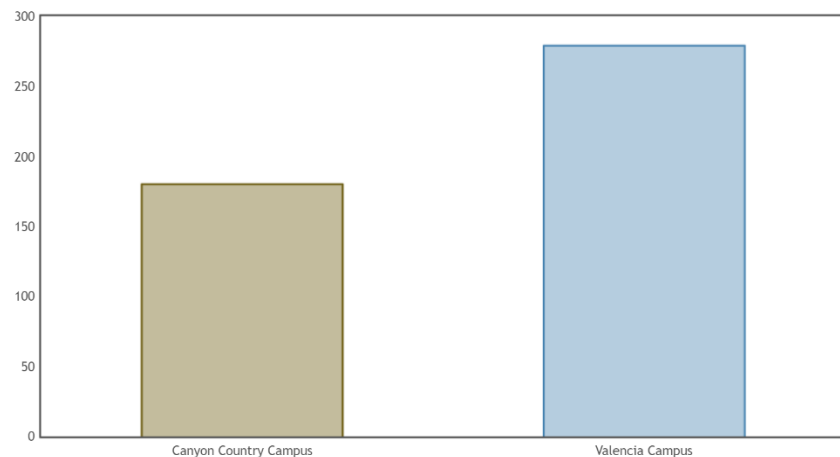
### Section 2A Answers

1. 3.9%
  2. 88.3%
  3. 0.61%
  4. 9.2%
  5. 21.7%
  6. 0.38%
  7. 65.1%
  8. 7.05%
  9. 0.014%
  10. 70.05%
  11. 0.58
  12. 0.926
  13. 0.08104
  14. 0.00772
  15. 0.0319
  16. 0.08
  17. 0.625
  18. 0.0352
  19. 0.00044
  20. 0.03
  21. 0.354
  22. 0.026
  23. 0.004
  24. 0.026
  25. 0.200
  26. 5.7%
  27. 12.3%
  28. 74.0%
  29. 2.7%
  30. 0.3%
  
  31.  $6064 \div 10528 \approx 0.576 = 57.6\%$  of the LGBTQ students feel unsafe at school because of their sexual orientation.
  32.  $8970 \div 10528 \approx 0.852 = 85.2\%$  of the LGBTQ students have been verbally harassed.
  33.  $5117 \div 10528 \approx 0.486 = 48.6\%$  of the LGBTQ students experienced cyberbullying.
  34.  $1369 \div 10528 \approx 0.130 = 13.0\%$  of the LGBTQ students were physically assaulted.
  35.  $57.6\% = 0.576$  of the LGBTQ students who were harassed or assaulted in school did not report the incident.
  36.  $63.5\% = 0.635$  of the LGBTQ students who reported an incident said that the school staff did not respond and told them to ignore it.
  37.  $45\% = 0.45$  of African American defendants were misclassified as high risk by the COMPASS program.
- 



## Section 2B Answers

1.

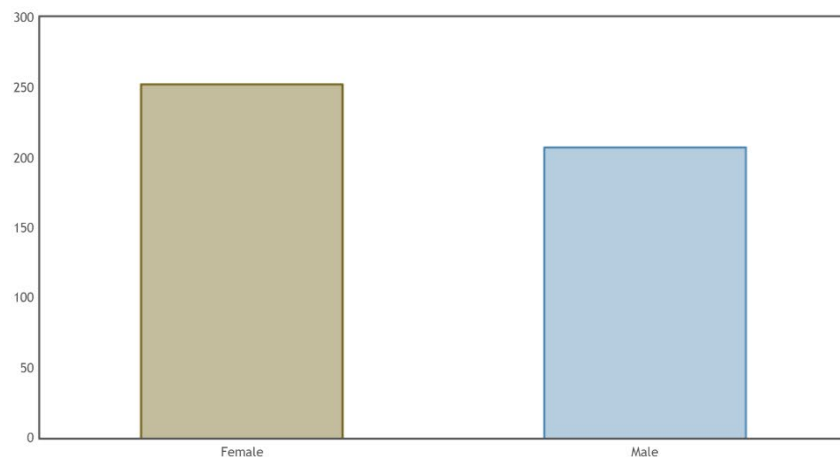


Summary Statistics

	Count	Proportion
Canyon Country Campus	180	0.392
Valencia Campus	279	0.608
Total	459	1.000

- There are more students at Valencia.
- There were 279 Math 075 students at the Valencia campus.
- There were 180 Math 075 students at the Canyon Country campus.
- 0.608 of the Math 075 students attend the Valencia campus.
- 0.392 of the Math 075 students attend the Canyon Country campus.
- 60.8% of the Math 075 students attend the Valencia campus.
- 39.2% of the Math 075 students attend the Canyon Country campus.

2.



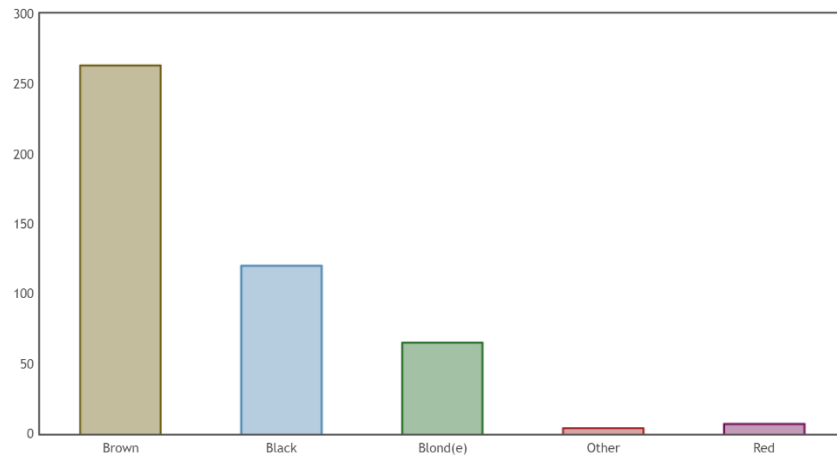
Summary Statistics

	Count	Proportion
Female	252	0.549
Male	207	0.451
Total	459	1.000

- There were more Math 075 students that identified as female than male.
- There were 252 Math 075 students that identified as female.
- There were 207 Math 075 students that identified as male.
- 0.549 of the Math 075 students identified as female.
- 0.451 of the Math 075 students identified as male.
- 54.9% of the Math 075 students identified as female.
- 45.1% of the Math 075 students identified as male.



3.



Summary Statistics

	Count	Proportion
Brown	263	0.573
Black	120	0.261
Blond(e)	65	0.142
Other	4	0.0087
Red	7	0.015
Total	459	1.000

- a) The hair color with the most Math 075 students was brown.
- b) The hair color with the least Math 075 students was “other”.
- c) 263 of the Math 075 students had brown hair.
- d) 65 of the Math 075 students had blonde hair.
- e) 0.015 of the Math 075 students had red hair.
- f) 0.261 of the Math 075 students had black hair.
- g) 1.5% of the Math 075 students had red hair.
- h) 26.1% of the Math 075 students had black hair.

4.

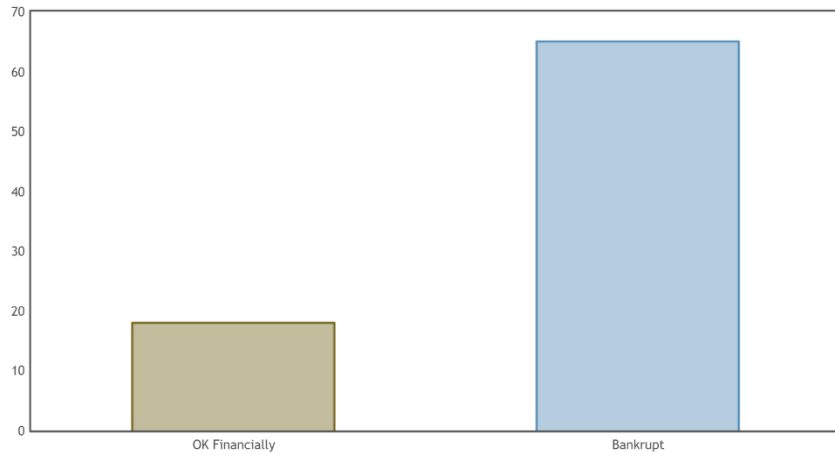
- a) Democratic party was most popular with Math 075 students.
- b) Independent political party was least popular with Math 075 students.
- c) 97 of the Math 075 students were republican.
- d) 185 of the Math 075 students were democrat.
- e) 19% of the Math 075 students identified as independent political party.
- f) 20% of the Math 075 students identified as other political party.
- g) 0.4 of the Math 075 students identified as democratic.
- h) 0.21 of the Math 075 students identified as republican.

5.

- a) In 2015, the most popular social media with Math 075 students was Instagram.
- b) In 2015, the least popular social media with Math 075 students was “other” social media.
- c) 94 of the Math 075 students prefer Snapchat.
- d) 137 of the Math 075 students prefer Instagram.
- e) 20% of the Math 075 students prefer Twitter.
- f) 8% of the Math 075 students prefer “other” social media.
- g) 0.3 of the Math 075 students prefer Instagram.
- h) 0.2 of the Math 075 students prefer Snapchat.



6.

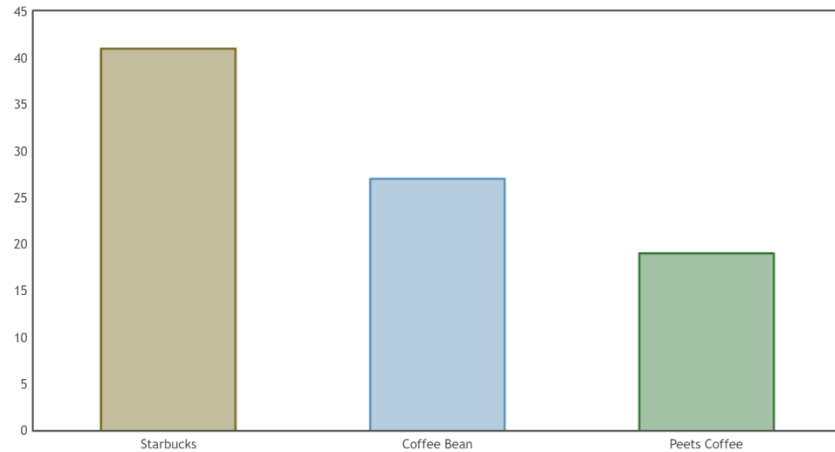


Summary Statistics

	Count	Proportion
OK Financially	18	0.217
Bankrupt	65	0.783
Total	83	1.000

- a) 0.783 of the retired NFL players had gone bankrupt.
- b) 78.3% of the retired NFL players had gone bankrupt.
- c) 0.217 of the retired NFL players were doing ok financially.
- d) 21.7% of the retired NFL players were doing ok financially.

7.



Summary Statistics

	Count	Proportion
Starbucks	41	0.471
Coffee Bean	27	0.31
Peets Coffee	19	0.218
Total	87	1.000

- a) 0.471 of the Math 075 students prefer Starbucks.
- b) 47.1% of the Math 075 students prefer Starbucks.
- c) 0.31 of the Math 075 students prefer Coffee Bean.
- d) 31% of the Math 075 students prefer Coffee Bean.
- e) 0.218 of the Math 075 students prefer Peet's Coffee.
- f) 21.8% of the Math 075 students prefer Peet's Coffee.

8.

- a) 9.5% (0.095) of the LGBTQ students were not planning to continue their education due to high victimization against their sexual orientation.
- b) 10.0% (0.100) of the LGBTQ students were not planning to continue their education due to high victimization against their gender expression.
- c) 5.4% (0.054) of the LGBTQ students were not planning to continue their education due to lower level victimization against their sexual orientation.



d) 5.2% (0.052) of the LGBTQ students were not planning to continue their education due to lower level victimization against their gender expression.

9.

a) 66.3% (0.663) of the LGBTQ students attending a school without a Gay-Straight Alliance program feel unsafe because of sexual orientation.

b) 50.2% (0.502) of the LGBTQ students attending a school with a Gay-Straight Alliance program feel unsafe because of sexual orientation.

c) 48.2% (0.482) of the LGBTQ students attending a school without a Gay-Straight Alliance program feel unsafe because of gender expression.

d) 39.1% (0.391) of the LGBTQ students attending a school with a Gay-Straight Alliance program feel unsafe because of gender expression.

---

## Section 2C Answers

1.

a) Ratio:  $0.653 \div 0.347 \approx 1.88$

The percentage of the women that preferred athletic wear is 1.88 times larger than the percentage of women that preferred traditional jeans.

b) Percent of Increase:  $\frac{(0.653-0.347)}{0.347} \times 100\% \approx 88.2\%$  (Large percent of increase)

c) The ratio and percent of increase were significantly high. Since the sample size was large enough, this data indicates that the percentage of the women that prefer athletic wear is significantly higher than the percent of the women that prefer jeans.

d) We would advise the company to increase their supply of women's athletic wear and decrease the supply of women's jeans.

2.

a) Ratio:  $0.163 \div 0.14 \approx 1.16$

The percentage of the patients on the medical/surgical ward is 1.16 times larger than the percentage of patients on the telemetry ward.

b) Percent of Increase:  $\frac{(0.163-0.14)}{0.14} \times 100\% \approx 16.3\%$  (Small percent of increase)

c) The sample size was large enough, but the ratio and percent of increase were low. This data indicates that the percentage of patients on the medical/surgical floor and the telemetry floor are about the same.

d) We would advise the hospital to set aside similar amount resources for both floors.

3.

a) Ratio:  $0.724 \div 0.276 \approx 2.62$

The percentage of employees without health insurance is 2.62 times larger than the percentage of employees with health insurance.

b) Percent of Increase:  $\frac{(0.724-0.276)}{0.276} \times 100\% \approx 162.3\%$  (Large percent of increase)





c) The ratio and percent of increase were significantly high. Since the sample size was large enough, this data indicates that the percentage of employees without health insurance is significantly larger than the percentage of employees with health insurance.

d) We would advise the company to increase access to their health insurance benefits.

4.

a) Ratio:  $0.228 \div 0.18 \approx 1.27$

The percentage of the people that took the medicine and improved was only 1.27 times larger than the percentage in the placebo group.

b) Percent of Increase:  $\frac{(0.228-0.18)}{0.18} \times 100\% \approx 26.7\%$  (Moderate percent of increase)

c) The sample size was large enough (barely), and the ratio and percent of increase were moderately high. This data indicates that the percentage of people that took the medicine and improved was not significantly larger than for the placebo group.

d) The experiment indicates that the depression medicine may not work. We recommend further testing.

5.

a) Ratio:  $0.45 \div 0.23 \approx 1.96$

The percentage of African American defendants misclassified as high risk is 1.96 times larger than the percentage white defendants misclassified as high risk.

b) Percent of Increase:  $\frac{(0.45-0.23)}{0.23} \times 100\% \approx 95.6\%$  (Large percent of increase)

c) The ratio and percent of increase were significantly high. Since the sample size was large enough, this data indicates that the percentage of African American defendants misclassified as high risk is significantly higher than the percentage white defendants misclassified as high risk.

d) We would advise the court system to stop using the COMPAS program to determine if a defendant will repeat their crime. The system seems to be racially biased.

6.

a) Ratio:  $0.18 \div 0.13 \approx 1.38$

The percentage of cars from Japan was 1.38 times larger than the percentage of cars from Germany.

b) Percent of Increase:  $\frac{(0.18-0.13)}{0.13} \times 100\% \approx 38.5\%$  (Moderate percent of increase.)

c) The ratio and percent of increase were not significantly high which would usually indicate that the percentage of cars made in Japan was slightly higher than for Germany. However, the sample size was really too small to make a determination with this data.

7.

a) Ratio:  $0.33 \div 0.29 \approx 1.14$

The percentage of cereals made by Kelloggs was only 1.14 times larger than the percentage of cereals made by General.

b) Percent of Increase:  $\frac{(0.33-0.29)}{0.29} \times 100\% \approx 13.8\%$  (Low percent of increase)

c) The ratio and percent of increase were low which would indicate that the percentages were close. However, the sample size was really too small to make a determination with this data.



8.

a) Ratio:  $0.67 \div 0.33 \approx 2.03$

The percentage of cereals made for adults was 2.03 times larger than the percentage of cereals made for children.

b) Percent of Increase:  $\frac{(0.67-0.33)}{0.33} \times 100\% \approx 103.0\%$  (High percent of increase)

c) The ratio and percent of increase were high which would usually indicate that the percentage of cereals for adults is significantly higher than the percentage for children. However, the sample size was really too small to make a determination with this data.

9.

a) Ratio:  $0.095 \div 0.054 \approx 1.76$

The percentage of highly victimized sexual orientation LGBTQ students not planning to continue school is 1.76 times larger than the percentage of lower level sexual orientation victimized LGBTQ students not planning to continue school.

b) Percent of Increase:  $\frac{(0.095-0.054)}{0.054} \times 100\% \approx 75.9\%$  (Large percent of increase)

c) The ratio and percent of increase were significantly high. Since the sample size was large enough, this data indicates that the percentage of highly victimized (sexual orientation) LGBTQ students not planning to continue school is significantly higher than the percentage of lower level victimized (sexual orientation) LGBTQ students not planning to continue school.

d) Ratio:  $0.1 \div 0.052 \approx 1.92$

The percentage of highly victimized gender expression LGBTQ students not planning to continue school is 1.92 times larger than the percentage of lower level gender expression victimized LGBTQ students not planning to continue school.

e) Percent of Increase:  $\frac{(0.1-0.052)}{0.052} \times 100\% \approx 92.3\%$  (Large percent of increase)

f) The ratio and percent of increase were significantly high. The sample size was large enough, so this data indicates that the percentage of highly victimized gender expression LGBTQ students not planning to continue school is significantly higher than the percentage of lower level victimized gender expression LGBTQ students not planning to continue school.

10.

a) Ratio:  $0.663 \div 0.502 \approx 1.32$

The percentage of LGBTQ students from schools without a Gay-Straight Alliance that feel unsafe due to sexual orientation is 1.32 times larger than the percentage of LGBTQ students from schools with a Gay-Straight Alliance that feel unsafe due to sexual orientation.

b) Percent of Increase:  $\frac{(0.663-0.502)}{0.502} \times 100\% \approx 32.1\%$  (Moderate percent of increase)

c) The ratio and percent of increase were moderately high. Since the sample size was large enough, this data indicates that the percentage of LGBTQ students from schools without a Gay-Straight Alliance (GSA) that feel unsafe due to sexual orientation is moderately higher than the percentage of LGBTQ students from schools with a Gay-Straight Alliance that feel unsafe due to sexual orientation. Schools without a GSA should consider instituting one.

d) Ratio:  $0.482 \div 0.391 \approx 1.23$

The percentage of LGBTQ students from schools without a Gay-Straight Alliance that feel unsafe due to gender expression is 1.23 times larger than the percentage of LGBTQ students from schools with a Gay-Straight Alliance that feel unsafe due to gender expression.



e) Percent of Increase:  $\frac{(0.482-0.391)}{0.391} \times 100\% \approx 23.3\%$  (Moderate percent of increase)

f) The ratio and percent of increase were moderately high. Since the sample size was large enough, this data indicates that the percentage of LGBTQ students from schools without a Gay-Straight Alliance (GSA) that feel unsafe due to gender expression is moderately higher than the percentage of LGBTQ students from schools with a Gay-Straight Alliance that feel unsafe due to gender expression. Schools without a GSA should consider instituting one.

11.

a) The percentage of call-backs for applicants with white sounding names was 1.5 times higher than for applicants with African American sounding names.

b) Percent of Increase:  $\frac{(0.1006-0.0670)}{0.0670} \times 100\% \approx 50.1\%$  (High percent of increase)

c) The ratio and percent of increase were high. Since the sample size was large enough, this data indicates that the percentage of call-backs for applicants with white sounding names was significantly higher than for applicants with African American sounding names.

d) Since the experiment controlled confounding variables, this may indicate racial discrimination in the labor market in Boston and Chicago.

---

## Section 2D Answers

1. 658 cars
2. 1472 people
3. 260 dogs
4. 77 cats
5. 315 bears
6. 20,246 car accidents
7. 10,800 cases of flu

8.

a) 0.15

b)  $0.15 \times 78300 \approx 11,745$

We estimate that there are 11,745 people in Chino Hills without insurance.

9.

a) 0.3

b)  $0.3 \times 305700 \approx 91,710$

We estimate that there are 91,710 people in Stockton which own guns.

10.

a) 0.093

b)  $0.093 \times 18400 \approx 1,711$

We estimate that there are 1,711 students at COC with diabetes.

11.

a) 0.159

b)  $0.159 \times 161000 \approx 25,599$

We estimate that there are 25,599 people in Lancaster struggling with hunger.



12.

- a) 0.0147
- b)  $0.0147 \times 136400 \approx 2,005$

We estimate that there are 2,005 people in Van Nuys with autism.

13.

- a) 0.0051
- b)  $0.0051 \times 1769000 \approx 9,022$

We estimate that there are 9,022 cars in San Francisco with defective air bags.

14.

- a) 0.148
- b)  $0.148 \times 305700 \approx 45,244$

We estimate that there are 45,244 people in Stockton, CA living in poverty.

15.

- a) 0.33
- b)  $0.33 \times 147 \approx 49$

We estimate that there are 49 doctors that have been sued for malpractice at that hospital.

16.

- a) 0.78
- b)  $0.78 \times 26682 \approx 20,812$

We estimate that there are 20,812 retired NFL players bankrupt or in financial stress.

17.

- a) 0.6
- b)  $0.6 \times 4374 \approx 2,624$

We estimate that there are 2,624 retired NBA players that have gone broke.

18.

- a) 0.576
- b)  $0.576 \times 244,000 \approx 140,544$

We estimate that there are 140,544 LGBTQ students between 13 and 17 years old in California that feel unsafe at their school due to sexual orientation.

19.

- a) 0.852
- b)  $0.852 \times 1994000 \approx 1,698,888$

We estimate that there are 1,698,888 LGBTQ students in the U.S. between 13 and 17 years old that have been verbally harassed at school.

20.

- a) 0.486
- b)  $0.486 \times 114000 \approx 55,404$

We estimate that there are 55,404 LGBTQ students between the ages of 13 and 17 in Florida that have experienced cyberbullying.



21.

- a) 0.13
- b)  $0.13 \times 113,000 \approx 14,690$

We estimate there are 14,690 LGBTQ students between in the ages of 13 and 17 in New York that have been physically assaulted.

---

## Chapter 2 Review Sheet Answers

- 1. Quantitative
- 2. Categorical
- 3. Categorical
- 4. Quantitative

- 5. 0.0385
- 6. 0.926
- 7. 0.0051

- 8. 55.8%
- 9. 0.32%
- 10. 9.3%

11.  $17/47 \approx 0.362$

12.  $0.362 \times 100\% = 36.2\%$

13. 0.333

14.  $0.333 \times 58 \approx 19$  HIV deaths by Tuberculosis

15. 1478 Intel processors

16. 850 AMD processors

17. About 4% of processors are made by Mobile.

18. About 35% of processors are made by AMD.

19. About 61% of processors are made by Intel.

20. Intel had the most processors.

21. Mobile made the least processors.

22. Percent Ratio for Intel and AMD =  $61/35 \approx 1.7$

The percentage of Intel is significantly higher as it is a larger sample size and the ratio is not close to 1.

---



**Introduction to Data Analysis (2<sup>nd</sup> Edition)**  
**Chapter 3 Answer Keys**

**Section 3A**

1.

2 x 4 table

	A	AB	B	O	All
Rh+	3	1	2	9	15
Rh-	2	1	0	2	5
All	5	2	2	11	20

2.

2 x 4 table

	ER	ICU	Med/Surg	SDS	All
F	2	1	4	2	9
M	3	2	2	4	11
All	5	3	6	6	20

3.

2 x 4 table

	ER	ICU	Med/Surg	SDS	All
Rh+	4	2	5	4	15
Rh-	1	1	1	2	5
All	5	3	6	6	20

4.

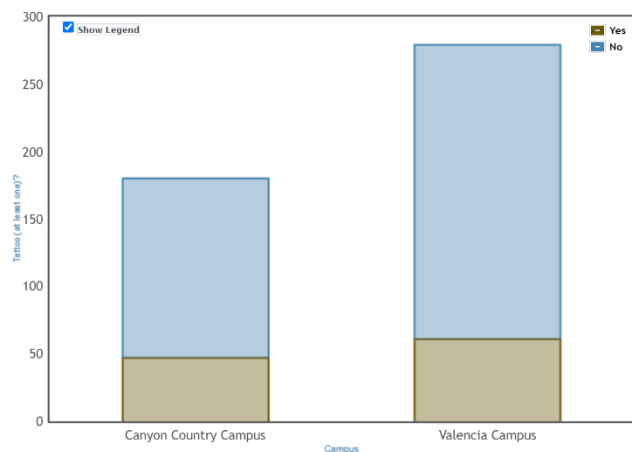
4 x 4 table

	ER	ICU	Med/Surg	SDS	All
A	1	0	2	2	5
AB	1	0	1	0	2
B	0	1	0	1	2
O	3	2	3	3	11
All	5	3	6	6	20



5.

a-b)



Counts Table [Switch Variables](#)

Tattoo (at least one)? \ Campus	Canyon Country Campus	Valencia Campus	Total
Yes	47	61	108
No	133	218	351
Total	180	279	459

Proportions [Row](#) [Column](#) [Overall](#)

c) Grand Total = 459 students

d) Total Valencia Campus = 279 students

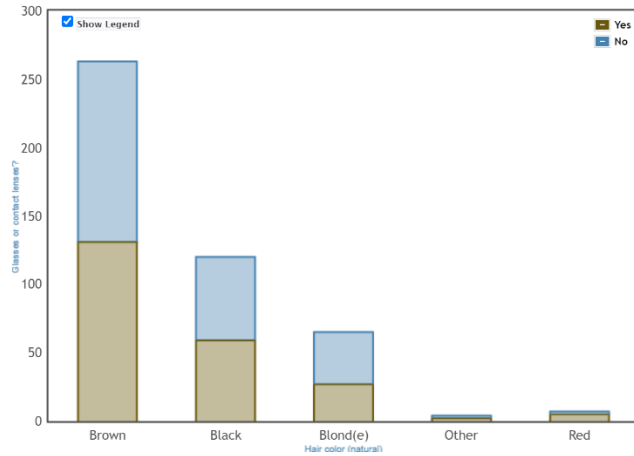
e) Total with Tattoo = 108 students

f) 133 students both went to Canyon Country campus and did not have a tattoo.

6.

a-b)





Counts Table [Switch Variables](#)

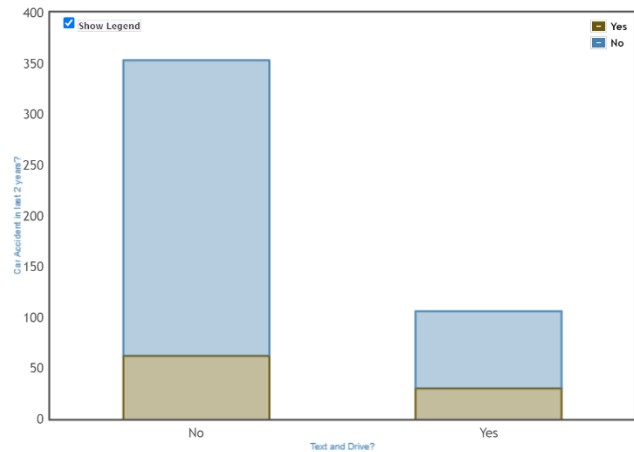
Glasses or contact lenses? \ Hair color (natural)	Brown	Black	Blond(e)	Other	Red	Total
Yes	131	59	27	2	5	224
No	132	61	38	2	2	235
Total	263	120	65	4	7	459

Proportions [Row](#) [Column](#) [Overall](#)

- c) Grand Total = 459 students
- d) 224 students need contacts or glasses.
- e) 263 students have brown hair.
- f) 61 students both do not need contacts or glasses and have black hair.

7.

a-b)



Counts Table [Switch Variables](#)

Car Accident in last 2 years? \ Text and Drive?	No	Yes	Total
Yes	62	30	92
No	291	76	367
Total	353	106	459

Proportions [Row](#) [Column](#) [Overall](#)

- c) Grand Total = 459 students
- d) 353 students said they do not text and drive. I do not believe that all of these students told the truth. The actual count may be lower because of response bias.
- e) 92 students have been in a car accident over the two year period.
- f) 30 students both said they have been in a car accident and admitted to texting and driving.

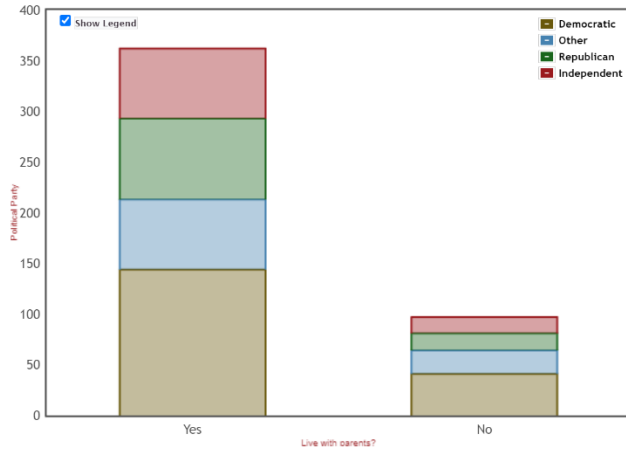
8.

a-b)



This material is from *Introduction to Data Analysis*, 2<sup>nd</sup> edition by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a “CC-BY” [Creative Commons Attribution 4.0 International License](#) – 3/17/2021





Counts Table [Switch Variables](#)

Political Party \ Live with parents?	Yes	No	Total
Democratic	144	41	185
Other	69	23	92
Republican	80	17	97
Independent	69	16	85
Total	362	97	459

Proportions [Row](#) [Column](#) [Overall](#)

- c) Grand Total = 459 students
- d) 97 students said they do not live with their parents.
- e) 85 students identify as “independent” political party.
- f) 144 students both live with parents and identify as democrat.

### Section 3B Answers

1.
  - a) 108 students have at least one tattoo.
  - b)  $108 \div 459 \approx 0.235$
  - c)  $0.235 \times 100\% = 23.5\%$
2.
  - a) 99 students prefer Facebook.
  - b)  $99 \div 459 \approx 0.216$
  - c)  $0.216 \times 100\% = 21.6\%$
3.
  - a) 351 students do not have a tattoo.
  - b)  $351 \div 459 \approx 0.765$
  - c)  $0.765 \times 100\% = 76.5\%$
4.
  - a) 137 students prefer Instagram.
  - b)  $137 \div 459 \approx 0.298$
  - c)  $0.298 \times 100\% = 29.8\%$
5.
  - a) 33 students both have a tattoo and prefer Facebook.
  - b)  $33 \div 459 \approx 0.072$
  - c)  $0.072 \times 100\% = 7.2\%$
- 6.



- a) 99 students both do not have a tattoo and prefer Instagram.
- b)  $99 \div 459 \approx 0.216$
- c)  $0.216 \times 100\% = 21.6\%$

7.

- a) 79 students both do not have a tattoo and prefer Snapchat.
- b)  $79 \div 459 \approx 0.172$
- c)  $0.172 \times 100\% = 17.2\%$

8.

- a) 9 students both have a tattoo and prefer "Other" social media.
- b)  $9 \div 459 \approx 0.020$
- c)  $0.020 \times 100\% = 2.0\%$

9.

- a)  $108 + 99 - 33 = 174$  students either have a tattoo or prefer Facebook.
- b)  $174 \div 459 \approx 0.379$
- c)  $0.379 \times 100\% = 37.9\%$

10.

- a)  $351 + 137 - 99 = 389$  students either do not have a tattoo or prefer Instagram.
- b)  $389 \div 459 \approx 0.847$
- c)  $0.847 \times 100\% = 84.7\%$

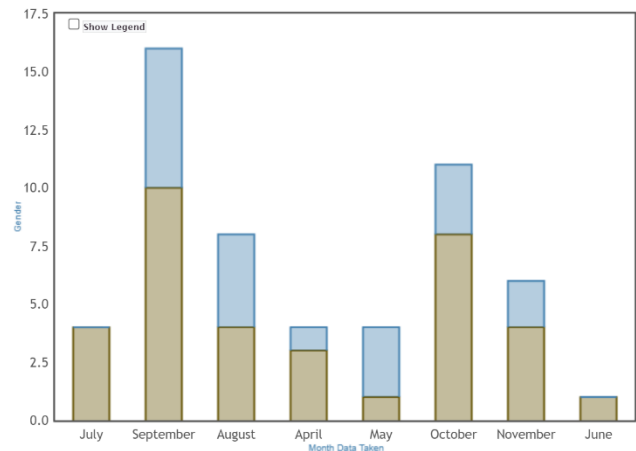
11.

- a)  $91 + 94 = 185$  students prefer either Twitter or Snapchat. (Variables do not intersect.)
- b)  $185 \div 459 \approx 0.403$
- c)  $0.403 \times 100\% = 40.3\%$

12.

- a)  $108 + 38 - 9 = 137$  students either have a tattoo or prefer "Other" social media.
- b)  $137 \div 459 \approx 0.298$
- c)  $0.298 \times 100\% = 29.8\%$

13.



Counts Table [Switch Variables](#)

Gender \ Month Data Taken	July	September	August	April	May	October	November	June	Total
male	4	10	4	3	1	8	4	1	35
female	0	6	4	1	3	3	2	0	19
Total	4	16	8	4	4	11	6	1	54

Proportions [Row](#) [Column](#) [Overall](#)

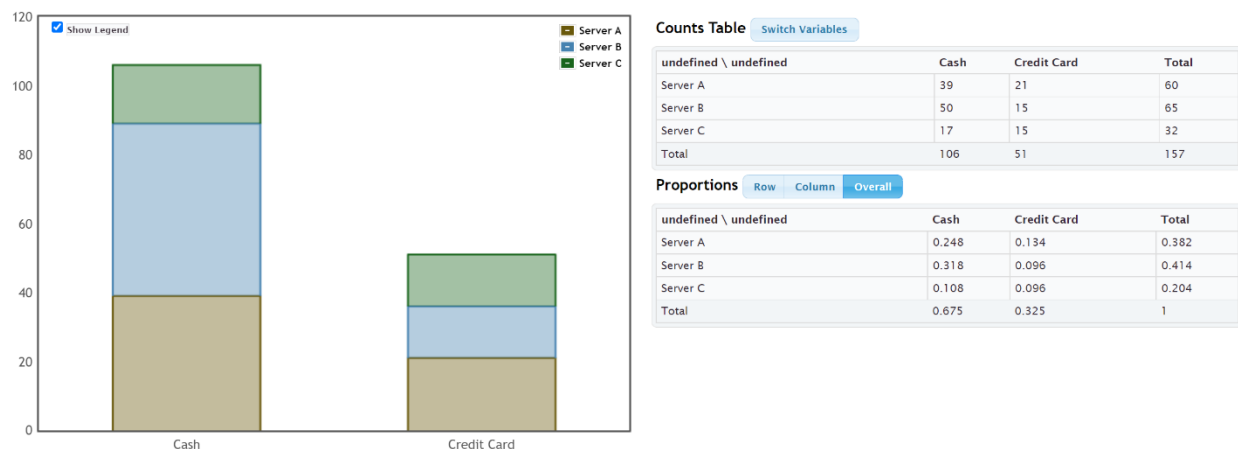
Gender \ Month Data Taken	July	September	August	April	May	October	November	June	Total
male	0.074	0.185	0.074	0.056	0.019	0.148	0.074	0.019	0.648
female	0	0.111	0.074	0.019	0.056	0.056	0.037	0	0.352
Total	0.074	0.296	0.148	0.074	0.074	0.204	0.111	0.019	1

- a) 0.296 (29.6%) of the bears were measured in September.
- b) 0.352 (35.2%) of the bears were female.



- c) 0.111 (11.1%) of the bears were both female and had data taken in September.  
 d)  $0.296 + 0.352 - 0.111 = 0.537$  (53.7%) of the bears were either female or had data taken in September.

14.



- a) 0.675 (67.5%) of the bills were paid with cash.  
 b) 0.414 (41.4%) of the bills had server B as the server.  
 c) 0.318 (31.8%) of the bills were both served by server B and paid in cash.  
 d)  $0.675 + 0.414 - 0.318 = 0.771$  (77.1%) of the bills were either served by server B or paid in cash.

### Section 3C Answers

1.

- a) 108 students have at least one tattoo.  
 b) 38 students both have a tattoo and prefer Instagram.  
 c)  $38 \div 108 \approx 0.352$  of the tattoo students prefer Instagram.  
 d)  $0.352 \times 100\% = 35.2\%$  of the tattoo students prefer Instagram.

2.

- a) 91 students prefer Twitter?  
 b) 78 students both do not have a tattoo and prefer Twitter.  
 c)  $78 \div 91 \approx 0.857$  of the Twitter students do not have a tattoo.  
 d)  $0.857 \times 100\% = 85.7\%$  of the Twitter students do not have a tattoo.

3.

- a) 351 students do not have a tattoo.  
 b) 66 students both do not have a tattoo and prefer Facebook.  
 c)  $66 \div 351 \approx 0.188$  of the no tattoo students prefer Facebook.  
 d)  $0.188 \times 100\% = 18.8\%$  of the no tattoo students prefer Facebook.

4.

- a) 94 students prefer Snapchat.  
 b) 15 students both have a tattoo and prefer Snapchat.  
 c)  $15 \div 94 \approx 0.160$  of the Snapchat students have a tattoo.  
 d)  $0.160 \times 100\% = 16.0\%$  of the Snapchat students have a tattoo.



5.

- a) 180 students went to the Canyon Country campus.
- b) 138 students both drive alone and went to the Canyon Country campus.
- c)  $138 \div 180 \approx 0.767$  of the Canyon Country campus students drove alone to school.
- d)  $0.767 \times 100\% = 76.7\%$  of the Canyon Country campus students drove alone to school.

6.

- a) 46 students were dropped off by someone.
- b) 14 students were both dropped off and went to the Canyon Country campus.
- c)  $14 \div 46 \approx 0.304$  of the dropped off students went to the Canyon Country campus.
- d)  $0.304 \times 100\% = 30.4\%$  of the dropped off students went to the Canyon Country campus.

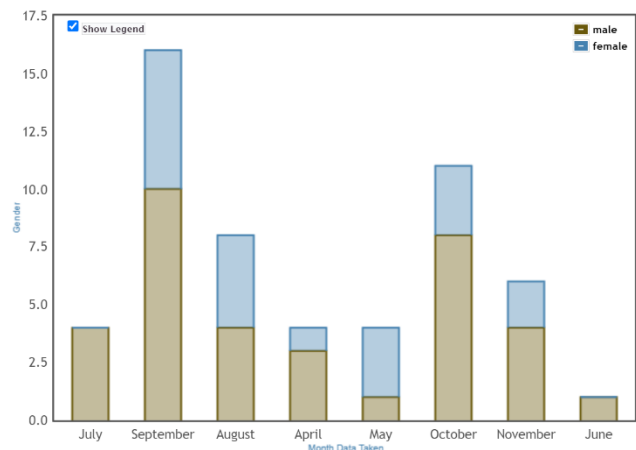
7.

- a) 279 students went to the Valencia campus.
- b) 22 students both carpool and went to the Valencia campus.
- c)  $22 \div 279 \approx 0.079$  of the Valencia campus students carpool to school.
- d)  $0.079 \times 100\% = 7.9\%$  of the Valencia campus students carpool to school.

8.

- a) 24 students used public transportation to school.
- b) 17 students both used public transportation and went to the Valencia campus.
- c)  $17 \div 24 \approx 0.708$  of the public transportation students went to the Valencia campus.
- d)  $0.708 \times 100\% = 70.8\%$  of the public transportation students went to the Valencia campus.

9.



Counts Table [Switch Variables](#)

Gender \ Month Data Taken	July	September	August	April	May	October	November	June	Total
male	4	10	4	3	1	8	4	1	35
female	0	6	4	1	3	3	2	0	19
Total	4	16	8	4	4	11	6	1	54

Proportions [Row](#) [Column](#) [Overall](#)

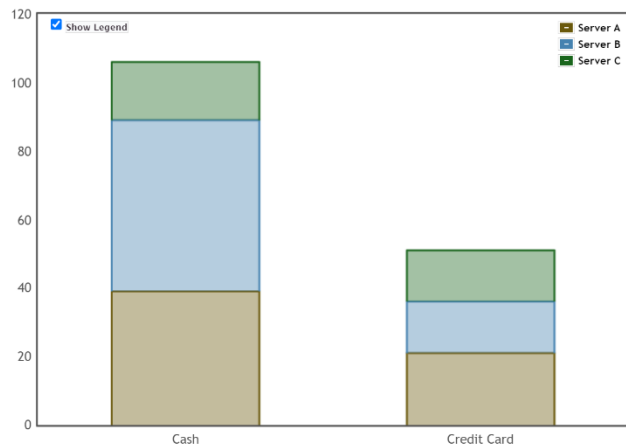
Gender \ Month Data Taken	July	September	August	April	May	October	November	June	Total
male	0.114	0.286	0.114	0.086	0.029	0.229	0.114	0.029	1
female	0	0.316	0.211	0.053	0.158	0.158	0.105	0	1
Total	0.074	0.296	0.148	0.074	0.074	0.204	0.111	0.019	1

- a) 0.211 (21.1%) of the female bears were measured in August.
- b) 0.114 (11.4%) of the male bears were measured in August.
- c) The proportions in part (a) and (b) look significantly different. (85.1% increase)



- d) 0.229 (22.9%) of the female bears were measured in October.
- e) 0.158 (15.8%) of the male bears were measured in October
- f) The proportions in part (d) and (e) have a 44.9% increase. They are different but may not be significant based on the small sample size.
- g) Overall these two percentages indicate a difference between female and male bears in when they were measured. This difference in conditional probabilities may indicate that gender is related to the month the bear was measured.
- h) No. Just because two variables are related, does not prove causation. This data was an observational study. It needed to use experimental design to prove cause and effect.

10.



Counts Table [Switch Variables](#)

undefined \ undefined	Cash	Credit Card	Total
Server A	39	21	60
Server B	50	15	65
Server C	17	15	32
Total	106	51	157

Proportions [Row](#) [Column](#) [Overall](#)

undefined \ undefined	Cash	Credit Card	Total
Server A	0.368	0.412	0.382
Server B	0.472	0.294	0.414
Server C	0.16	0.294	0.204
Total	1	1	1

- a) 0.412 (41.2%) of the credit card customers were served by server A.
- b) 0.294 (29.4%) of the credit card customers were served by server B.
- c) 0.294 (29.4%) of the credit card customers were served by server C.
- d) The conditional proportions in part (a), (b) and (c) do not look significantly different. Server A had a higher percentage, but servers B and C were the same.
- e) Since the conditional proportions were not significantly different, this probably indicates that using a credit card is not related to the server.

### Ch 3 Review Problem Answers



This material is from *Introduction to Data Analysis*, 2<sup>nd</sup> edition by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a “CC-BY” [Creative Commons Attribution 4.0 International License](#) – 3/17/2021

1.

	Everyone	Kids	Parents	All
Female Pet	2	0	2	4
Male Pet	4	1	4	9
All	6	1	6	13

2.  $119/280 = 0.425 = 42.5\%$
3.  $49/280 = 0.175 = 17.5\%$
4.  $35/280 = 0.125 = 12.5\%$
5.  $21/280 = 0.075 = 7.5\%$
6.  $91/280 = 0.325 = 32.5\%$
7.  $119/280 = 0.425 = 42.5\%$
8.  $28/105 \approx 0.267 = 26.7\%$
9.  $21/49 \approx 0.429 = 42.9\%$
10. Significantly different
11. Data suggests that grade level is related to being democrat.

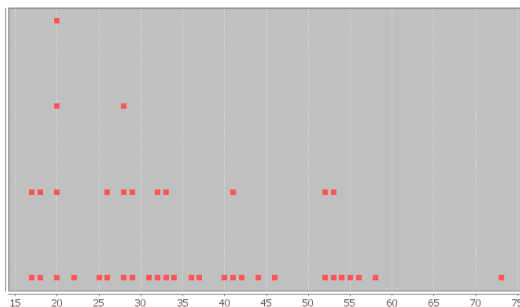
---

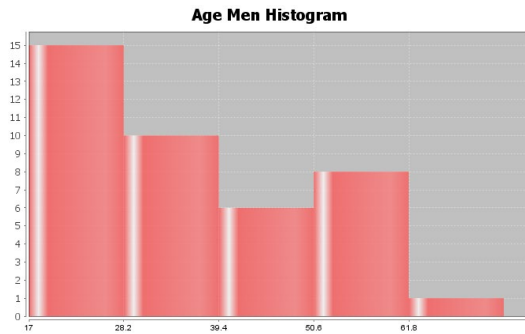
## Introduction to Data Analysis Chapter 4 Answer Key

### Section 4A Answers

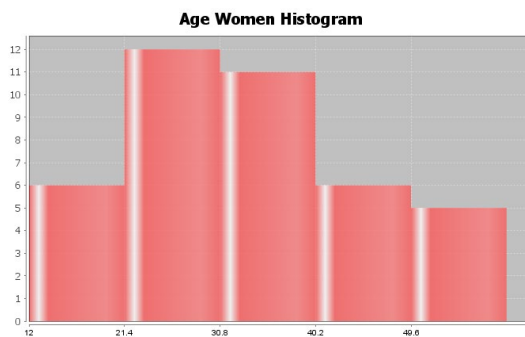
1. Men's Age (years): Skewed Right

**Age Men Dot Plot**

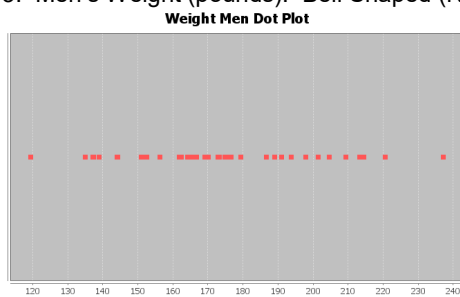


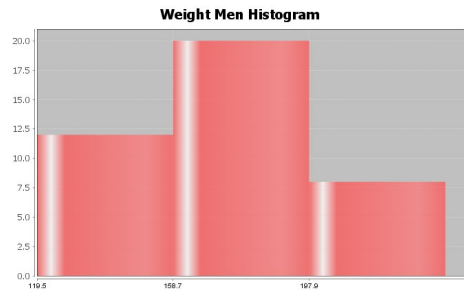


2. Women's Age (years): Almost Bell Shaped (Slightly Skewed Right)

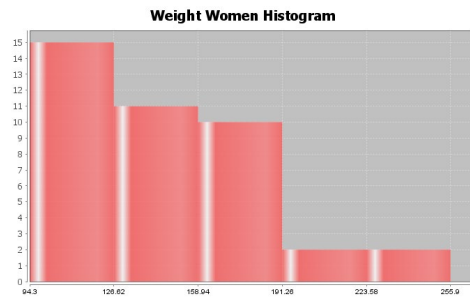
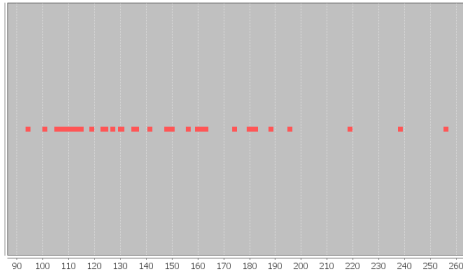


3. Men's Weight (pounds): Bell Shaped (Normal)

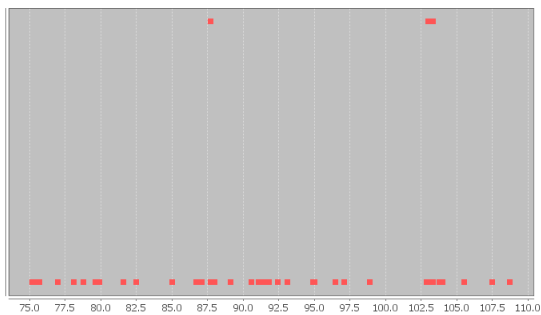




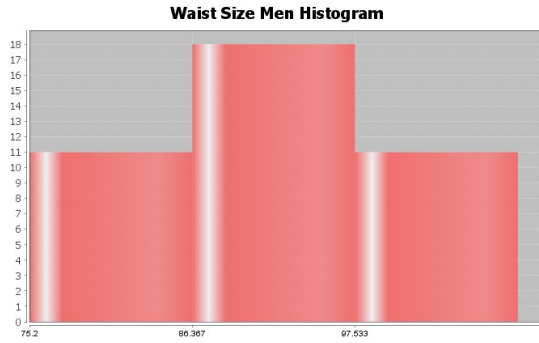
4. Women's Weight (pounds): Skewed Right  
**Weight Women Dot Plot**



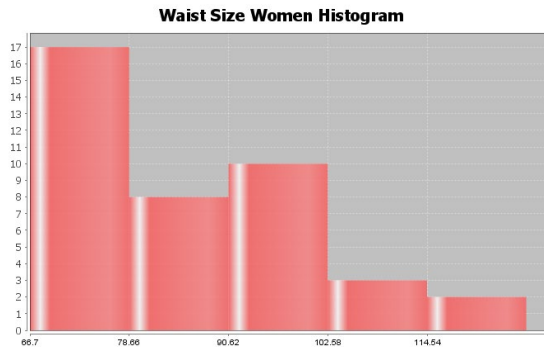
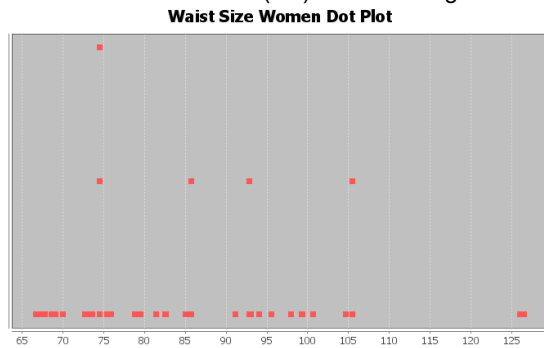
5. Men's Waist Size (cm): Bell Shaped (Normal)  
**Waist Size Men Dot Plot**



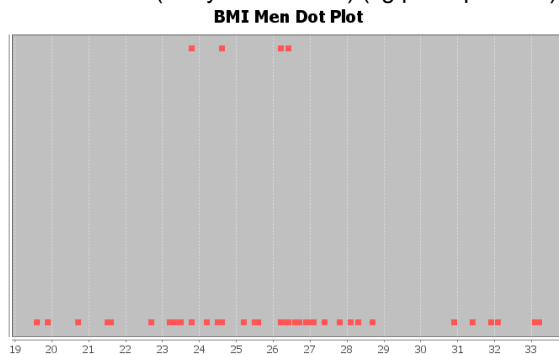


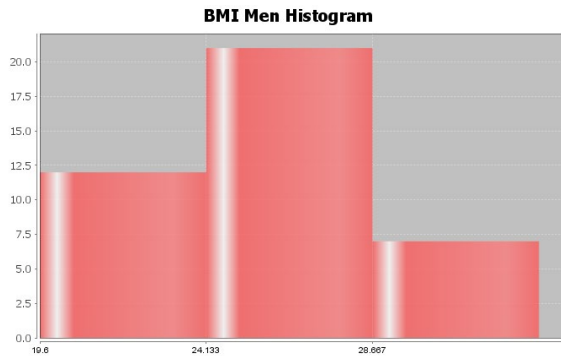


6. Women's Waist Size (cm): Skewed Right

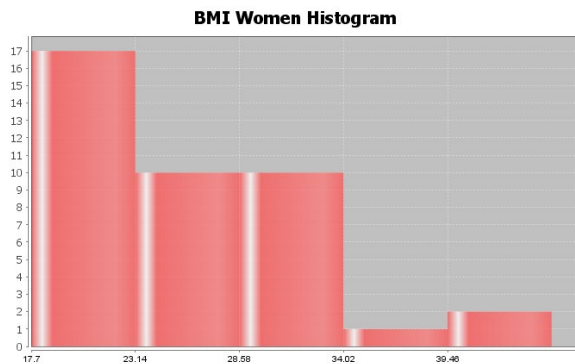
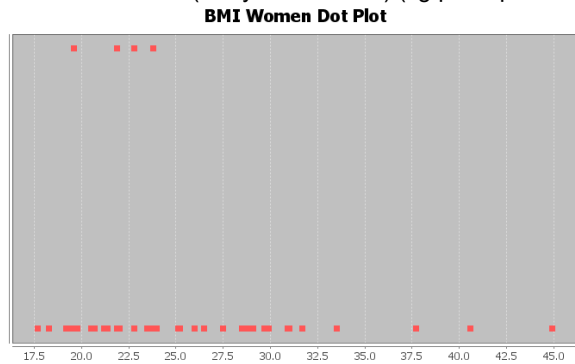


7. Men's BMI (Body Mass Index) (kg per sq meters): Bell Shaped (Normal)





8. Women's BMI (Body Mass Index) (kg per sq meters): Skewed Right



### Section 4B Answers

1. Men's Age (years): Skewed Right

Best Measure of Center: Median

### Descriptive Statistics

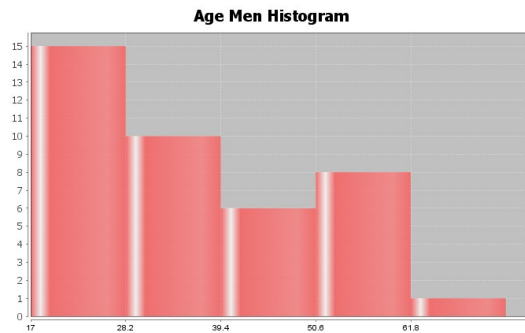
Variable	Mean
C15 Men Age (years)	35.475



Variable	Median	Mode	N for mode
C15 Men Age (years)	32.5	20.0	4

Variable	Min	Max
C15 Men Age (years)	17.0	73.0

$$\text{Midrange} = (17 + 73) / 2 = 45$$



2. Women's Age (years): Almost Bell Shaped (Slightly Skewed Right)

Best Measure of Center: Median (If said skewed)

Best Measure of Center: Mean (If said almost Bell Shaped)

### Descriptive Statistics

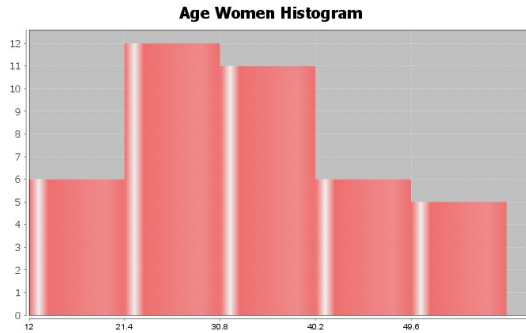
Variable	Mean
C1 Women Age (years)	33.225

Variable	Median	Mode	N for mode
C1 Women Age (years)	31.5	23.0	4

Variable	Min	Max
C1 Women Age (years)	12.0	59.0

$$\text{Midrange} = (12 + 59) / 2 = 35.5$$





3. Men's Weight (pounds): Bell Shaped (Normal)

Best Measure of Center: Mean

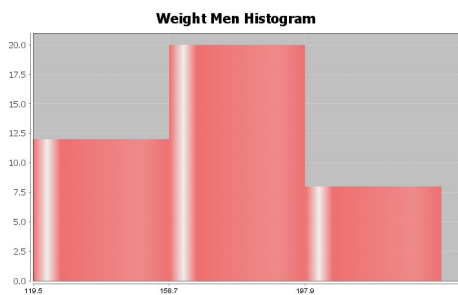
**Descriptive Statistics**

Variable	Mean
C17 Men Wt (Lbs)	172.55

Variable	Median	Mode	N for mode
C17 Men Wt (Lbs)	169.95	*	0

Variable	Min	Max
C17 Men Wt (Lbs)	119.5	237.1

Midrange =  $(119.5 + 237.1) / 2 = 178.3$



4. Women's Weight (pounds): Skewed Right

Best Measure of Center: Median

**Descriptive Statistics**



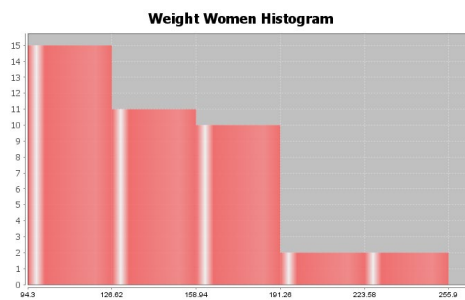
This material is from *Introduction to Data Analysis*, 2<sup>nd</sup> edition by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a “CC-BY” [Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by/4.0/) – 3/17/2021

Variable	Mean
C3 Women Wt (Lbs)	146.220

Variable	Median	Mode	N for mode
C3 Women Wt (Lbs)	135.8	*	0

Variable	Min	Max
C3 Women Wt (Lbs)	94.3	255.9

$$\text{Midrange} = (94.3 + 255.9) / 2 = 175.1$$



5. Men's Waist Size (cm): Bell Shaped (Normal)

Best Measure of Center: Mean

**Descriptive Statistics**

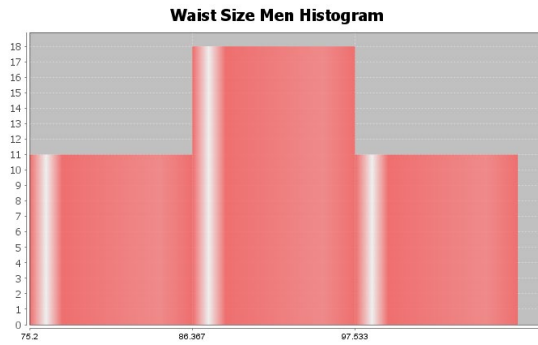
Variable	Mean
C18 Men Waist (cm)	91.285

Variable	Median	Mode	N for mode
C18 Men Waist (cm)	91.200	87.7, 103.0, 103.3	2

Variable	Min	Max
C18 Men Waist (cm)	75.2	108.7

$$\text{Midrange} = (75.2 + 108.7) / 2 = 91.95$$





6. Women's Waist Size (cm): Skewed Right

Best Measure of Center: Median

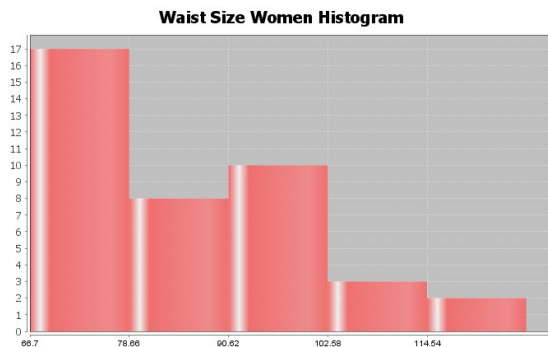
**Descriptive Statistics**

Variable	Mean
C4 Women Waist (cm)	85.033

Variable	Median	Mode	N for mode
C4 Women Waist (cm)	81.95	74.5	3

Variable	Min	Max
C4 Women Waist (cm)	66.7	126.5

Midrange =  $(66.7 + 126.5) / 2 = 96.6$



7. Men's BMI (Body Mass Index) (kg per sq meters): Bell Shaped (Normal)

Best Measure of Center: Mean

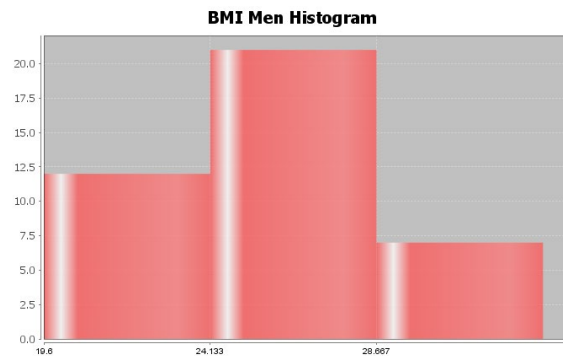
**Descriptive Statistics**

Variable	Mean
C23 Men BMI	25.998

Variable	Median	Mode	N for mode
C23 Men BMI	26.2	26.4, 24.6, 23.8, 26.2	2

Variable	Min	Max
C23 Men BMI	19.6	33.2

Midrange =  $(19.6 + 33.2) / 2 = 26.4$



8. Women's BMI (Body Mass Index) (kg per sq meters): Skewed Right

Best Measure of Center: Median

**Descriptive Statistics**

Variable	Mean
C9 Women BMI	25.74

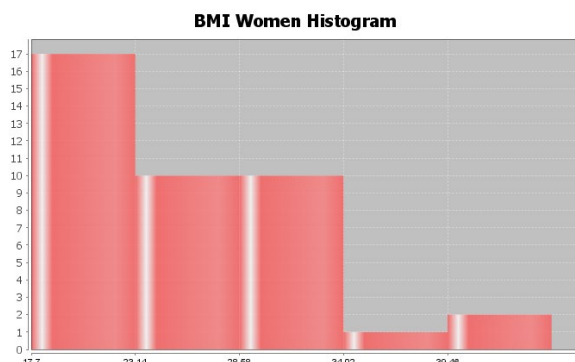
Variable	Median	Mode	N for mode
C9 Women BMI	23.9	22.8, 19.6, 23.8, 21.9	2

Variable	Min	Max



C9 Women BMI 17.7 44.9

$$\text{Midrange} = (17.7 + 44.9) / 2 = 31.3$$



### Section 4C Answers

1. Mean =  $84/18 \approx 4.7$
2. Mean =  $326/12 \approx 27.2$
3. Mean =  $68/7 \approx 9.71$
4. Mean =  $53.8/12 \approx 4.48$
5. Mean =  $33.21/11 \approx 3.019$
6. Answers may vary (10,11,12,14,15,16)
7. Answers may vary (9,10,11,12,14,15,16,17)
8. Answers may vary (18,19,20,21,21.5,22,23,24,25)
9. Answers may vary (17,18,19,20,21,21.5,22,23,24,25,26)

10. The numbers are balanced around 10. 5 and 15 are 5 places from 10. 6 and 14 are four places from 10. 7 and 13 are both three places from 10. 8 and 12 are two places from 10. 9 and 11 are both one place from 10. The total distance from 10 for numbers above = total distance from 10 for numbers below.

### Section 4D Answers

1.  
Mean = 7  
Sum of Squares = 154  
Sample Size (total frequency) = 6  
Degrees of Freedom =  $n-1 = 6-1 = 5$   
Standard Deviation = square root (154/5) = square root (30.8)  $\approx 5.5$
2.  
Mean = 8  
Sum of Squares = 100  
Sample Size (total frequency) = 8  
Degrees of Freedom =  $n-1 = 8-1 = 7$   
Standard Deviation = square root (100/7) = square root (14.28571429)  $\approx 3.8$
3. Bear Ages (Months): Shape Skewed Right  
Mean and Standard Deviation are NOT accurate (not bell shaped)





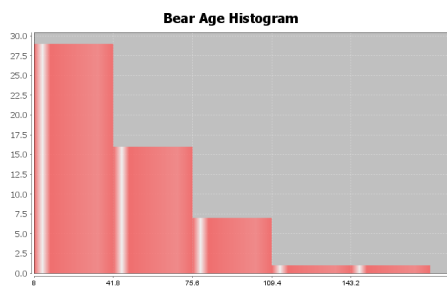
Standard Deviation Sentence: Typical bear ages were 33.7 months from the mean of 43.5 months.

### Descriptive Statistics

Variable	Mean	Standard Deviation	Variance
C1 AGE (months)	43.519	33.721	1137.085

Variable	IQR
C1 AGE (months)	41.0

Variable	Range
C1 AGE (months)	169.0



4. Bear Neck Circumference (Inches): Bell Shaped

Mean and Standard Deviation are accurate (bell shaped)

Standard Deviation Sentence: Typical bear neck sizes were 5.64 inches from the mean of 20.56 inches.

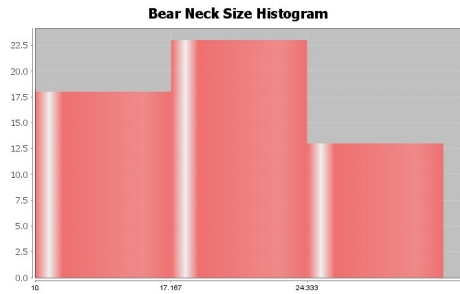
### Descriptive Statistics

Variable	Mean	Standard Deviation	Variance
C6 Neck Circum (in)	20.556	5.641	31.818

Variable	IQR
C6 Neck Circum (in)	8.125

Variable	Range
C6 Neck Circum (in)	21.5





5. Bear Length (Inches): Skewed Left

Mean and Standard Deviation are NOT accurate (not bell shaped)

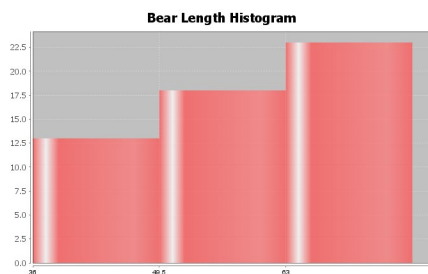
Standard Deviation Sentence: Typical bear lengths were 10.70 inches from the mean of 58.62 inches.

**Descriptive Statistics**

Variable	Mean	Standard Deviation	Variance
C7 Length (in)	58.617	10.701	114.509

Variable	IQR
C7 Length (in)	16.875

Variable	Range
C7 Length (in)	40.5



6. Bear Chest Size (Inches): Bell Shaped

Mean and Standard Deviation are accurate (bell shaped)

Standard Deviation Sentence: Typical bear chest sizes were 9.4 inches from the mean of 35.7 inches.

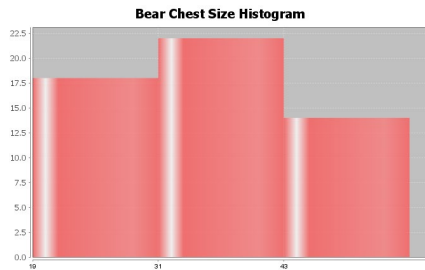
**Descriptive Statistics**

Variable	Mean	Standard Deviation	Variance
C8 Chest (in)	35.663	9.352	87.455



Variable	IQR
C8 Chest (in)	15.25

Variable	Range
C8 Chest (in)	36.0



7. Bear Weight (pounds): Skewed Right

Mean and Standard Deviation are NOT accurate (not bell shaped)

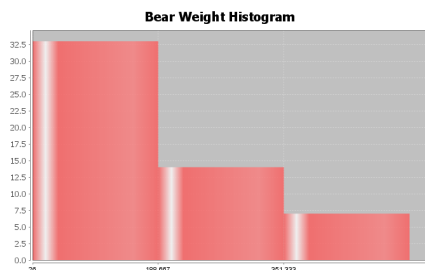
Standard Deviation Sentence: Typical bear lengths were 121.8 pounds from the mean of 182.9 pounds.

### Descriptive Statistics

Variable	Mean	Standard Deviation	Variance
C9 Weight (Lbs)	182.889	121.801	14835.535

Variable	IQR
C9 Weight (Lbs)	158.0

Variable	Range
C9 Weight (Lbs)	488.0



8. Bear Head Length (inches): Bell Shaped

Mean and Standard Deviation are accurate (bell shaped)

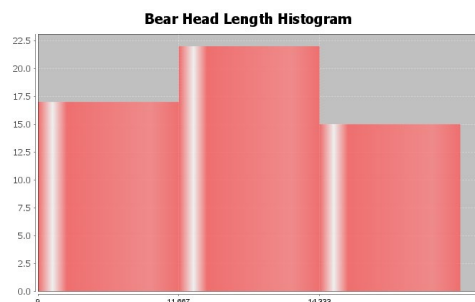
Standard Deviation Sentence: Typical bear head lengths were 2.14 inches from the mean of 12.95 inches.

### Descriptive Statistics

Variable	Mean	Standard Deviation	Variance
C4 Head Length (In)	12.954	2.144	4.597

Variable	IQR
C4 Head Length (In)	3.0

Variable	Range
C4 Head Length (In)	8.0



9. Answers may vary (16,17,18,19,21,22,23,24) (This example checked with statistics software, see below)

### Descriptive Statistics

Variable	Mean	Standard Deviation
3D#9	20.0	2.928

Variable	N total
3D#9	8



10. Answers may vary (9,10,11,12,13,27,28,29,30,31) (This example checked with statistics software, see below)

### Descriptive Statistics

Variable	Mean	Standard Deviation
3D#10	20.0	9.603

Variable	N total
3D#10	10

### Section 4E Answers

- 1a. Mean Average
- 1b. Mean Average
- 1c. Standard Deviation
- 1d. One Standard Deviation is Typical
- 1e. 68%
- 1f. Two Standard Deviations (or more) is considered unusual
- 1g. 2.5%
- 1h. 2.5%
- 1i. First calculate the unusual high cutoff by adding two standard deviations to the mean. Then look on the dotplot and see if any dots are higher than the cutoff.
- 1j. First calculate the unusual low cutoff by subtracting two standard deviations from the mean. Then look on the dotplot and see if any dots are lower than the cutoff.

2.

This data measured the lengths of the head of 54 bears in inches.

The data was bell shaped (normal).

The best measure of center was the mean of 12.95 inches. So the average length of the bear heads was 12.95 inches.

The best measure of spread was the standard deviation of 2.14 inches. This implies that typical bear head lengths were 2.14 inches from the mean. In fact, typical bear heads were between 10.81 inches and 15.10 inches.

There were no unusual values in the data set. The smallest bear head was 9 inches and the largest was 17 inches. Neither was unusual.

Unusual high cutoff = 17.24 (no values above 17.24)

Unusual low cutoff = 8.67 (no values below 8.67)

3.

This data measured the neck circumference of 54 bears in inches.

The data was bell shaped (normal).

The best measure of center was the mean of 20.56 inches. So the average bear neck circumference was 20.56 inches.

The best measure of spread was the standard deviation of 5.64 inches. This implies that typical bear neck sizes were 5.64 inches from the mean. In fact, typical bear neck circumferences were between 14.92 inches and 26.2 inches.

There were no unusual values in the data set. The smallest bear neck circumference was 10 inches and the largest was 31.5 inches. Neither was unusual.



Unusual high cutoff = 31.84 (no values above 31.84)

Unusual low cutoff = 9.28 (no values below 9.28)

4.

This data measured the chest size of 54 bears in inches.

The data was bell shaped (normal).

The best measure of center was the mean of 35.66 inches. So the average chest size of the bears was 35.66 inches.

The best measure of spread was the standard deviation of 9.35 inches. This implies that typical bear chest sizes were 9.35 inches from the mean. In fact, typical bear chest sizes were between 26.31 inches and 45.01 inches.

The smallest bear chest size was 19 inches. This was not unusual. The largest bear chest size was 55 inches. This was unusually high. This was the only unusual value in the data set.

Unusual high cutoff = 54.36 (there was one value above 54.36)

Unusual low cutoff = 16.96 (no values below 16.96)

5.

This data measured the diastolic blood pressure of 40 women in (mm of mercury).

The data was bell shaped (normal).

The best measure of center was the mean of 67.4 mm of mercury. So the average diastolic blood pressure of these women was 67.4 mm of mercury.

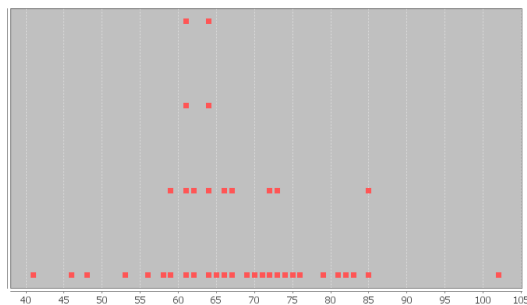
The best measure of spread was the standard deviation of 11.6 mm of mercury. This implies that typical diastolic blood pressures were 11.6 mm of mercury from the mean. In fact, typical diastolic blood pressures for these women were between 55.8 and 79.0 mm of mercury.

The lowest diastolic blood pressure for these women was 41 mm of mercury. This was unusually low. The highest diastolic blood pressure was 102 mm of mercury. This was unusually high. There were no other unusual values in the data set.

Unusual high cutoff = 90.6 (there was one value (102) that was above 90.6)

Unusual low cutoff = 44.2 (there was one value (41) that was below 44.2)

**Diastolic Blood Pressure Dot Plot**



6.

This data measured the wrist circumference of 40 women in inches.

The data was bell shaped (normal).

The best measure of center was the mean of 5.07 inches. So the average wrist size of these women was 5.07 inches.

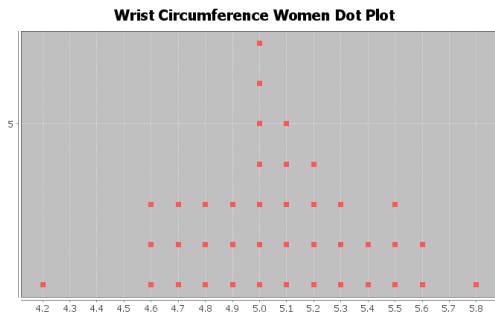
The best measure of spread was the standard deviation of 0.33 inches. This implies that typical wrist sizes for these women were 0.33 inches from the mean. In fact, typical wrist sizes for these women were between 4.74 and 5.40 inches.

The smallest wrist circumference for these women was 4.2 inches. This was unusually low. The largest wrist circumference was 5.8 inches. This was unusually high. There were no other unusual values in the data set.

Unusual high cutoff = 5.73 (there was one value (5.8) that was above 5.73)

Unusual low cutoff = 4.41 (there was one value (4.2) that was below 4.41)





7.

This data measured the height of 40 men in inches.

The data was bell shaped (normal).

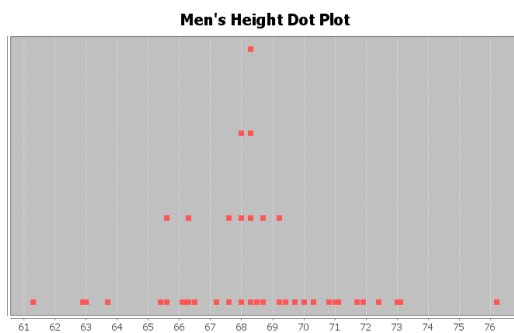
The best measure of center was the mean of 68.34 inches. So the average height of these men was 68.34 inches.

The best measure of spread was the standard deviation of 3.02 inches. This implies that typical heights of these men were 3.02 inches from the mean. In fact, typical heights for these men were between 65.32 inches and 71.36 inches.

The shortest man in the data was 61.3 inches. This height was unusually low. The tallest man in the data was 76.2 inches. This height was unusually high. There were no other unusual values in the data set.

Unusual high cutoff = 74.38 (there was only one value (76.2) that was above 74.38)

Unusual low cutoff = 62.30 (there was only one value (61.3) that was below 62.30)



8.

This data measured the weight of 40 men in pounds.

The data was bell shaped (normal).

The best measure of center was the mean of 172.55 pounds. So the average weight of these men was 172.55 pounds.

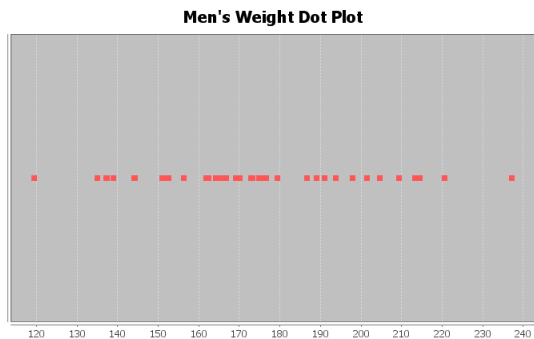
The best measure of spread was the standard deviation of 26.33 pounds. This implies that typical weights of these men were 26.33 pounds from the mean. In fact, typical weights for these men were between 146.22 pounds and 198.88 pounds.

The lightest man in the data was 119.5 pounds. This weight was unusually low. The heaviest man in the data was 237.1 pounds. This weight was unusually high. There were no other unusual values in the data set.

Unusual high cutoff = 225.21 (there was only one value (237.1) that was above 225.21)

Unusual low cutoff = 119.89 (there was only one value (119.5) that was below 119.89)





### Answers for Chapter 4 Review Problems

1. Men's Diastolic BP

Shape: Skewed Left

Best Measure of Center (Best Average): Median

2. Men's Heights (inches)

Shape: Bell Shaped (Normal)

Best Measure of Center (Best Average): Mean

3. Men's Pulse Rates

Shape: Skewed Right

Best Measure of Center (Best Average): Median

4.

Mean =  $216.1 / 13 \approx 16.62$

5.

Standard Deviation: How far typical values are from the mean in a bell shaped (normal) data set.

6.

The mean and standard deviation are only accurate if the data is bell shaped (normal).

7. Middle 68%

8. Top 2.5%

9. Bottom 2.5%

10. Bell Shaped (Normal)

11. 478 total students

12. Yes. The mean and standard deviation are accurate representations of center and spread because the data set is bell shaped (normal).

13. Average Math Intimidation score = 6.159 (mean)

14. Average Distance from the mean = 2.418 (standard deviation)

15.  $3.741 \leq$  typical math intimidation scores  $\leq 8.577$

16. Unusual High Cutoff = 10.995





17. Unusual Low Cutoff = 1.323

18. No. There are no unusually high values. (No values above the unusual high cutoff of 10.995)

19. None

20. Yes. There was one unusually low math intimidation score.

21. There are many people that answered 1. This was an unusually low value since it was below the unusually low cutoff of 1.323.

---



This material is from *Introduction to Data Analysis*, 2<sup>nd</sup> edition by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a “CC-BY” [Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by/4.0/) – 3/17/2021

# Introduction to Data Analysis (2<sup>nd</sup> edition)

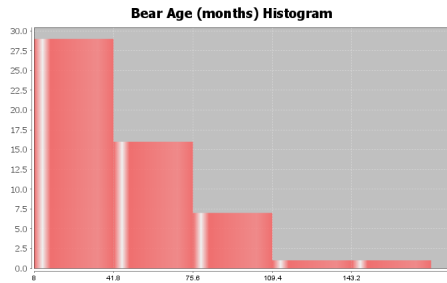
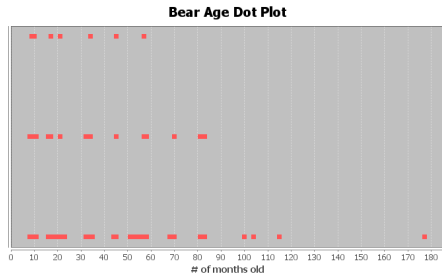
## Chapter 5 Answer Key

### Section 5A Answers

1. Bear Ages (Months)

Shape: Skewed Right

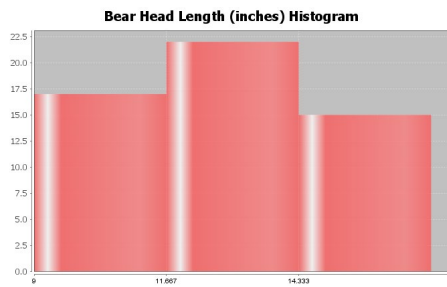
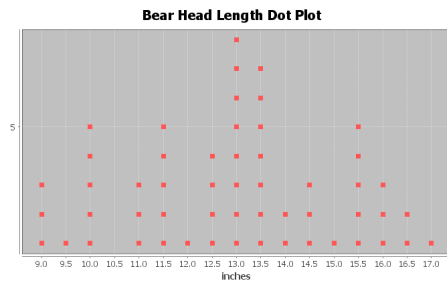
Best Measure of Center: Median



2. Head Length (inches)

Shape: Bell Shaped

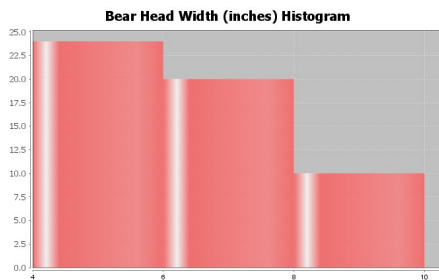
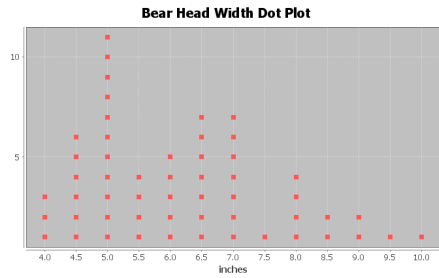
Best Measure of Center: Mean



### 3. Bear Head Width (Inches)

Shape: Skewed Right

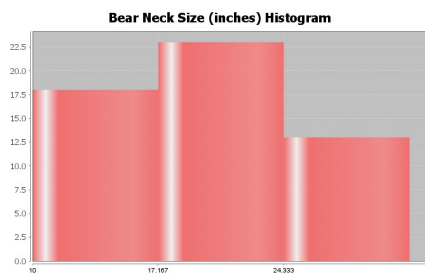
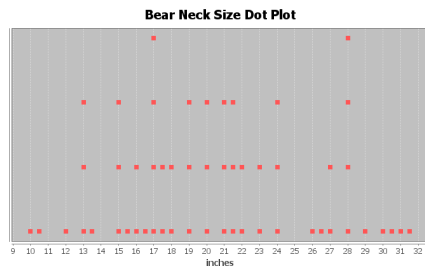
Best Measure of Center: Median



### 4. Bear Neck Size (inches)

Shape: Bell Shaped

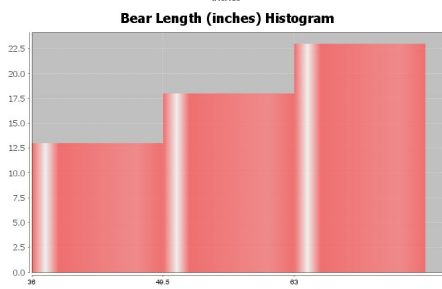
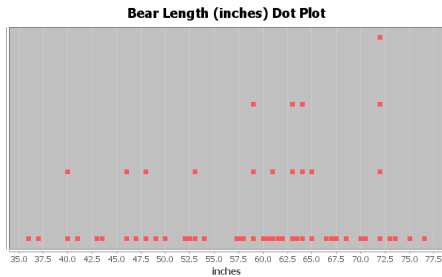
Best Measure of Center: Mean



5. Bear Length (inches)

Shape: Skewed Left

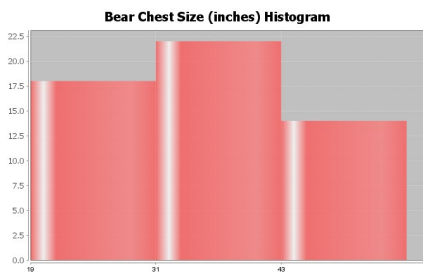
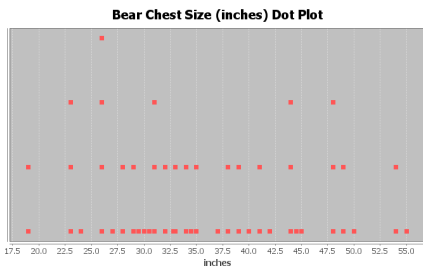
Best Measure of Center: Median



6. Bear Chest Size (inches)

Shape: Bell Shaped

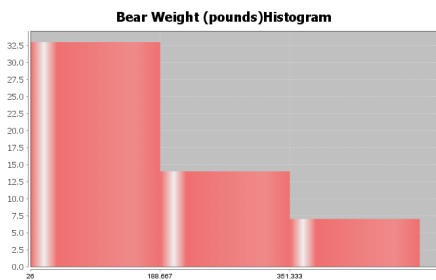
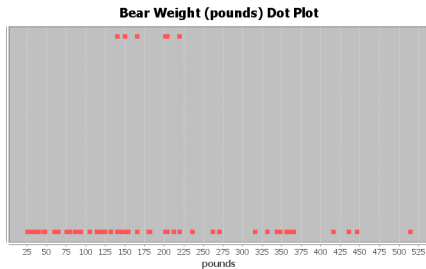
Best Measure of Center: Mean



7. Bear Weight (pounds)

Shape: Skewed Right

Best Measure of Center: Median



**Section 5B Answers**

- 1a. Median = 17 (one number in middle)
- 1b. Median =  $(7.2 + 10.4)/2 = 8.8$  (two numbers in middle)
- 1c. Median = 71 (one number in middle)
- 1d. Median = 157 (one number in middle)
- 1e. In order first: 1.9, 2.3, 2.8, 4.6, 6.1, 7.5, 8.3, 9.4  
Median =  $(4.6 + 6.1)/2 = 5.35$  (two numbers in middle)
- 1f. In order first: 18, 19, 20, 21, 23, 25, 26, 28, 29, 31, 32  
Median = 25 (one number in middle)

2. Bear Data Medians

**Descriptive Statistics**

Variable	Median
C1 AGE (months)	34.0
C4 Head Length (In)	13.0
C5 Head Width (In)	6.0
C6 Neck Circum (in)	20.0
C7 Length (in)	60.75



<b>C8 Chest (in)</b>	34.0
<b>C9 Weight (Lbs)</b>	150.0

### Section 5C Answers

1a. (answers may vary, these do not include the median in Q1 and Q3 calculation)

Median = 17

Q1 =  $(8+9)/2 = 8.5$

Q3 =  $(26+29)/2 = 27.5$

IQR =  $27.5-8.5 = 19$

Five Number Summary: 5, 8.5, 17, 27.5, 36

1b. (answers may vary, these do not include the median in Q1 and Q3 calculation)

Median =  $(7.2 + 10.4)/2 = 8.8$

Q1 = 5.1

Q3 = 14.7

IQR =  $14.7 - 5.1 = 9.6$

Five Number Summary = 2.1 , 5.1 , 8.8 , 14.7 , 16.0

1c. (answers may vary, these do not include the median in Q1 and Q3 calculation)

Median = 71

Q1 = 41

Q3 = 88

IQR =  $88 - 41 = 47$

Five Number Summary = 31 , 41 , 71 , 88 , 103

1d. (answers may vary, these do not include the median in Q1 and Q3 calculation)

Median = 157

Q1 =  $(152+154)/2 = 153$

Q3 =  $(163+164)/2 = 163.5$

IQR =  $163.5 - 153 = 10.5$

Five Number Summary = 150 , 153 , 157 , 163.5 , 165

1e. (answers may vary, these do not include the median in Q1 and Q3 calculation)

In order first: 1.9, 2.3, 2.8, 4.6, 6.1, 7.5, 8.3, 9.4

Median =  $(4.6 + 6.1)/2 = 5.35$

Q1 =  $(2.3 + 2.8)/2 = 2.55$

Q3 =  $(7.5 + 8.3)/2 = 7.9$

IQR =  $7.9 - 2.55 = 5.35$

Five Number Summary = 1.9 , 2.55 , 5.35 , 7.9 , 9.4

1f. (answers may vary, these do not include the median in Q1 and Q3 calculation)

In order first: 18, 19, 20, 21, 23, 25, 26, 28, 29, 31, 32

Median = 25

Q1 = 20

Q3 = 29

IQR =  $29 - 20 = 9$

Five Number Summary = 18 , 20 , 25 , 29 , 32

2. (Answers may vary depending on computer program used. These are from Statcato)

### Descriptive Statistics

Variable	Q1	Median	Q3	IQR
<b>C1 AGE (months)</b>	17.0	34.0	58.0	41.0
<b>C4 Head Length (In)</b>	11.5	13.0	14.5	3.0



<b>C5 Head Width (In)</b>	5.0	6.0	7.0	2.0
<b>C6 Neck Circum (in)</b>	16.375	20.0	24.5	8.125
<b>C7 Length (in)</b>	49.75	60.75	66.625	16.875
<b>C8 Chest (in)</b>	28.75	34.0	44.0	15.25
<b>C9 Weight (Lbs)</b>	84.5	150.0	242.5	158.0

<b>Variable</b>	<b>Min</b>	<b>Max</b>
<b>C1 AGE (months)</b>	8.0	177.0
<b>C4 Head Length (In)</b>	9.0	17.0
<b>C5 Head Width (In)</b>	4.0	10.0
<b>C6 Neck Circum (in)</b>	10.0	31.5
<b>C7 Length (in)</b>	36.0	76.5
<b>C8 Chest (in)</b>	19.0	55.0
<b>C9 Weight (Lbs)</b>	26.0	514.0

- 2a. Bear Age (months) Five Number Summary: 8 , 17 , 34 , 58 , 177
- 2b. Bear Head Length (inches) Five Number Summary: 9 , 11.5 , 13 , 14.5 , 17
- 2c. Bear Head Width (inches) Five Number Summary: 4 , 5 , 6 , 7 , 10
- 2d. Bear Neck Size (inches) Five Number Summary: 10 , 16.375 , 20 , 24.5 , 31.5
- 2e. Bear Length (inches) Five Number Summary: 36 , 49.75 , 60.75 , 66.625 , 76.5
- 2f. Bear Chest Size (inches) Five Number Summary: 19 , 28.75 , 34 , 44 , 55
- 2g. Bear Weight (pounds) Five Number Summary: 26 , 84.5 , 150 , 242.5 , 514

### Section 5D Answers

1a. (answers may vary, these do not include the median in Q1 and Q3 calculation)

Median = 25

Q1 = 20.5

Q3 = 29.5

IQR = 29.5 – 20.5 = 9

Unusual High Cutoff (for Skewed Data) =  $Q3 + (1.5 \times IQR) = 29.5 + (1.5 \times 9) = 43$

Unusual Low Cutoff (for Skewed Data) =  $Q1 - (1.5 \times IQR) = 20.5 - (1.5 \times 9) = 7$

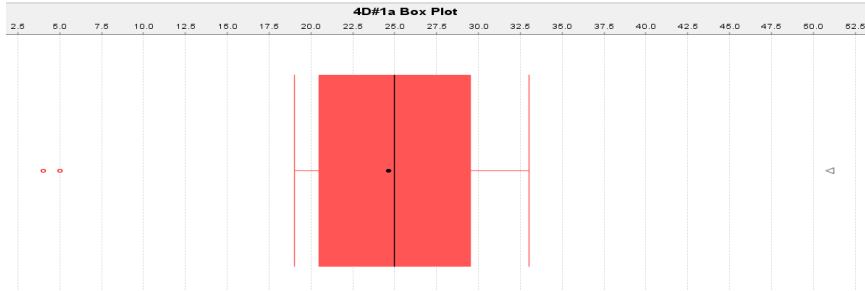
Unusually High values: 51

Unusually Low values: 4 and 5

High Whisker (largest # that is not unusual): 33

Low Whisker (smallest # that is not unusual): 19





1b. (answers may vary, these do not include the median in Q1 and Q3 calculation)

Median = 34.5

Q1 = 32.5

Q3 = 36.5

IQR = 4

Unusual High Cutoff (for Skewed Data) =  $Q3 + (1.5 \times IQR) = 36.5 + (1.5 \times 4) = 42.5$

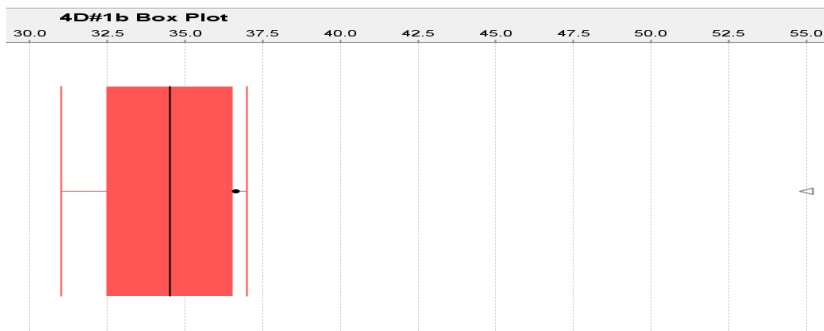
Unusual Low Cutoff (for Skewed Data) =  $Q1 - (1.5 \times IQR) = 32.5 - (1.5 \times 4) = 26.5$

Unusually High values: 55

Unusually Low values: none

High Whisker (largest # that is not unusual): 37

Low Whisker (smallest # that is not unusual): 31



1c. (answers may vary, these do not include the median in Q1 and Q3 calculation)

Median = 11.25

Q1 = 10.85

Q3 = 11.65

IQR = 0.8

Unusual High Cutoff (for Skewed Data) =  $Q3 + (1.5 \times IQR) = 11.65 + (1.5 \times 0.8) = 12.85$

Unusual Low Cutoff (for Skewed Data) =  $Q1 - (1.5 \times IQR) = 10.85 - (1.5 \times 0.8) = 9.65$

Unusually High values: 15.1

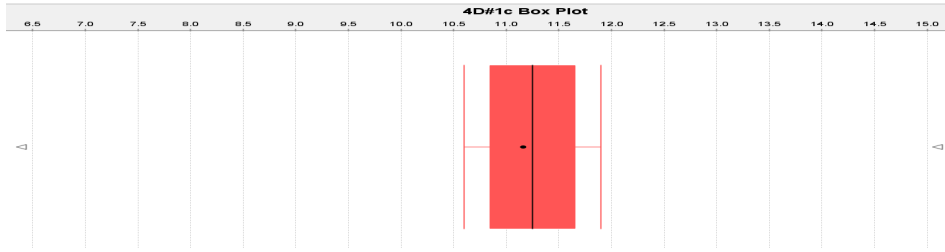
Unusually Low values: 6.4

High Whisker (largest # that is not unusual): 11.9

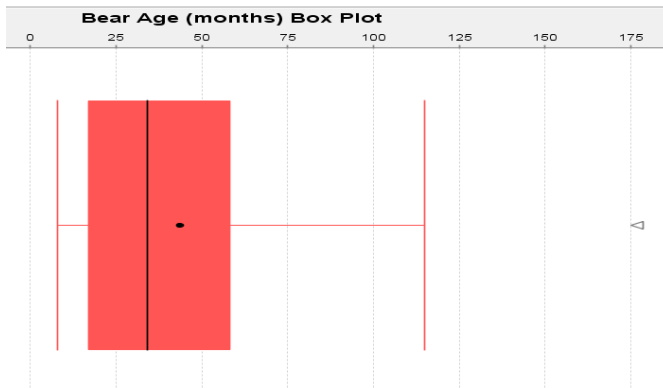
Low Whisker (smallest # that is not unusual): 10.6







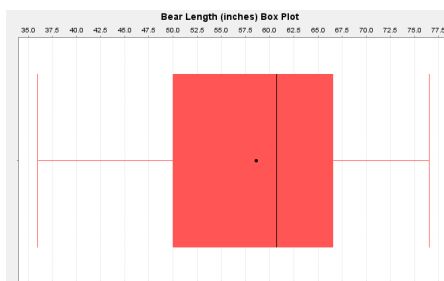
2a.



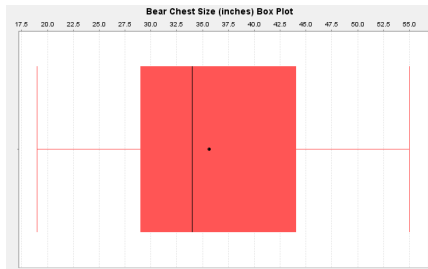
2b.



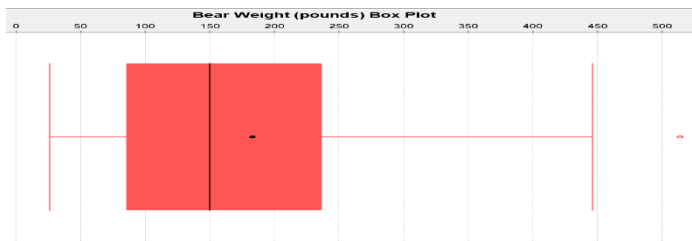
2c.



2d.



2e.



3. Spring Bears (These are approximate values from the graph.)  
Average Weight = Median  $\approx$  153 pounds

4. Spring Bears (These are approximate values from the graph.)  
Typical Spread = IQR  $\approx$  28 pounds  
137 pounds (Q1)  $\leq$  Typical Spring Bear Weights  $\leq$  165 pounds (Q3)

5. Yes. There was one unusual value in the spring bear data.  
Unusual value  $\approx$  197 pounds (This is an approximate value from the graph.)

### Section 5E Answers

- 1a. Q1 is a measure of position.
- 1b. Mean is a measure of center.
- 1c. Variance is a measure of spread.
- 1d. Midrange is a measure of center.
- 1e. Standard Deviation is a measure of spread.
- 1f. Minimum value is a measure of position.
- 1g. Q3 is a measure of position.
- 1h. Mode is a measure of center.
- 1i. IQR is a measure of spread.
- 1j. Median is a measure of center.
- 1k. Range is a measure of spread.
- 1l. Maximum value is a measure of position.

2.

Mean: The mean is the center or average for bell shaped data sets that balances the distances. If this data was bell shaped we would use the mean of \$1149.05 as the average.

Standard Deviation: The standard deviation measures how far typical values are from the mean in a bell shaped data set. If this data was bell shaped, then the typical values would be \$516 from the mean.



Variance: Variance is a measure of spread used in ANOVA testing that is equal to the standard deviation squared.

Q1: The first quartile tells us that approximately 25% of the values in the data set are lower than \$703.45.

Median: The median is a center or average when the data is in order. If this data set was skewed, we would use the median as our average. So if the data was skewed the average salary would be \$1015.74.

Q3: The third quartile tells us that approximately 75% of the values in the data set are lower than \$1496.11.

IQR: The interquartile range is the most accurate measure of spread for skewed data sets. It measures the spread for the middle 50% of the data. If this data was skewed, we would say that typical salaries are \$792.66 from each other.

Mode: The mode is a measure of center that gives the number or numbers in the data that appear most often. There was no mode in the salary data since all the salaries appeared only once. "N for mode" tells us how many times the mode appears.

Min: The minimum is the smallest value in the data set. The lowest salary in the data was \$371.57 and all other salaries in the data set are greater than \$371.57.

Max: The maximum is the largest value in the data set. The highest salary in the data was \$2396.28 and all other salaries in the data set are lower than \$2396.28.

Range: The overall range of a data set is a quick measure of spread that is not very accurate because it does not measure typical values and may be influenced by unusual values. It is calculated by the subtracting the max and the min. The overall range of this data set was \$2024.71. So all values in the data were within \$2024.71 from each other.

N Total: The sample size or total frequency tells you how many numbers are in the data set. In this case there were 35 salaries in this data set.

### Answers to Chapter 5 Review Sheet Problems

1. Shape = Skewed Right

Use the Median & IQR for center and spread.

2. Shape = Skewed Left

Use the Median & IQR for center and spread.

3. Shape = Bell Shaped (Normal)

Use the mean and standard deviation for center and spread.

4. (Answers may vary)

$$\text{Median} = (37 + 41)/2 = 78/2 = 39$$

$$Q1 = (26+28)/2 = 27$$

$$Q3 = (48+51)/2 = 49.5$$

$$\text{IQR} = Q3 - Q1 = 49.5 - 27 = 22.5$$

5. Interquartile Range (IQR): The interquartile range or IQR measures how far typical values are from each other in skewed data sets. IQR measure the spread for the middle 50% of the data values. To calculate IQR you subtract  $Q3 - Q1$ .

6. We should use the median as our center (average) and the IQR as our spread when the data is skewed (or not bell shaped).

7. Women's Cholesterol

8. Milligrams per deciliter (mg per dL)

9. Skewed Right



10. 38 numbers in data set
  11. Yes. The median and IQR are accurate measures of center and spread because the data is skewed.
  12. Average = 215 mg per dL (median)
  13. Typical Distance from each other = 186.75 mg per dL (IQR)
  14. 124.5 mg per dL (Q1)  $\leq$  typical values  $\leq$  311.25 mg per dL (Q3)
  15. No. No unusual low values on the boxplot.
  16. Yes. The boxplot shows three unusually high values.
  17. Unusual Values in the Data: 596 mg per dL, 600 mg per dL, and 920 mg per dL
  18. About 75%
  19. About 25%
  20. About 50%
  21. 531 mg per dL
  22. False. There were the same amount of numbers.
- 



**Introduction to Data Analysis  
Chapter 6 Answer Key**

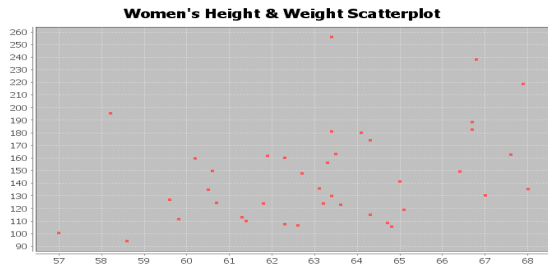
**Section 6A Answers**

1.

Explanatory Variable (X): Height of woman

Response Variable (Y): Weight of woman

Though height and weight may respond to each other, we chose weight to be the response variable. The thinking was that as a woman grows taller, her weight may increase. Weight changes may not indicate that a woman's height is changing.



The scatterplot seems to show a positive linear trend. The dots tend to increase from left to right and could be close to a line.

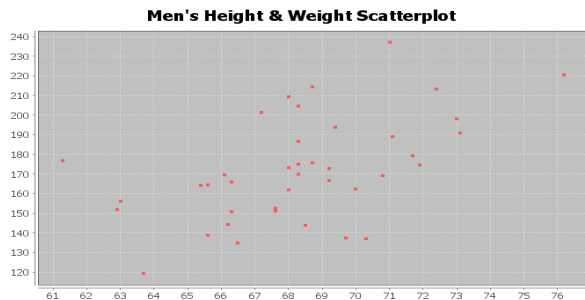
There are a couple points at (58.2 in, 196 Lb) and (63.4 in, 256 Lb) that don't seem to fit the pattern and could be unusual points (outliers).

2.

Explanatory Variable (X): Height of man

Response Variable (Y): Weight of man

Though height and weight may respond to each other, we chose weight to be the response variable. The thinking was that as a man grows taller, his weight may increase. Weight changes may not indicate that a man's height is changing.



The scatterplot seems to show a positive linear trend. The dots tend to increase from left to right and could be close to a line.

There does not seem to be any points that are not following the linear pattern. No outliers.

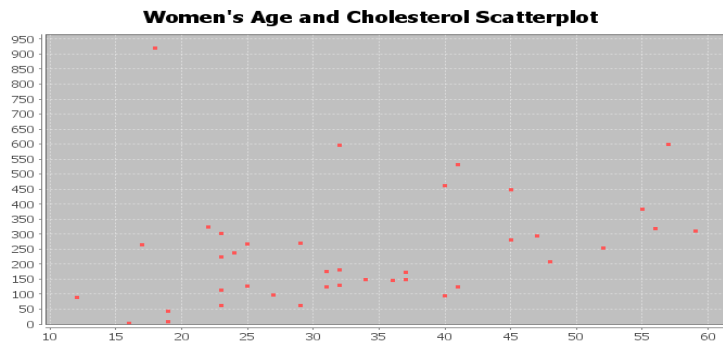


3.

Explanatory Variable (X): Age of woman

Response Variable (Y): Cholesterol of woman

A woman's cholesterol may change as a response to getting older, but age probably does not change in response to cholesterol.



The scatterplot seems to show a positive linear trend. The dots tend to increase from left to right and could be close to a line.

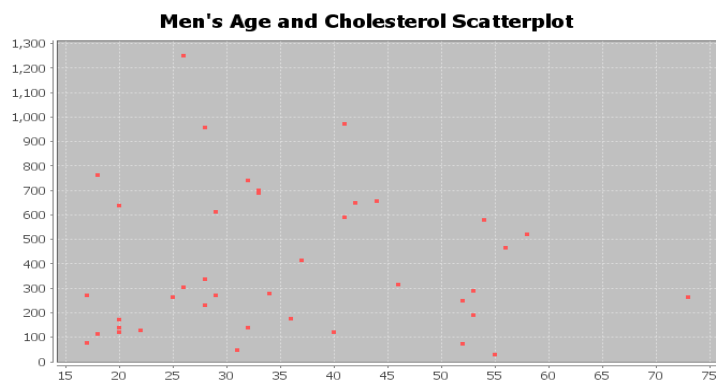
There is one point at (18 years old, 920 mg per dL) that doesn't seem to fit the pattern and is unusual (outlier).

4.

Explanatory Variable (X): Age of man

Response Variable (Y): Cholesterol of man

A man's cholesterol may change as a response to getting older, but age probably does not change in response to cholesterol.



The scatterplot does not show any linear or curved trend. The dots seem to be all over without a distinguishable pattern. This indicates there is probably no relationship between men's age and cholesterol.

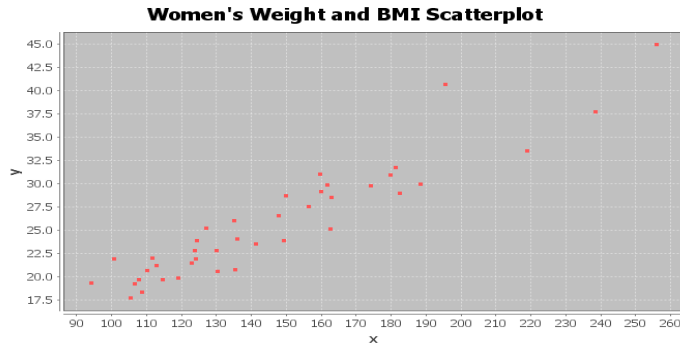
Since there is not linear or curved trend, all of the points are scattered making it difficult to judge what is unusual or not. They all look unusual.



5.

Explanatory Variable (X): Weight of woman  
Response Variable (Y): Body Mass Index (BMI) of woman

Weight and Body Mass Index respond to each other, so we can choose either to be the response variable. It comes down to what variable are we more interested in predicting. I chose body mass index as the response variable (y) because I was interested in predicting BMI from weight.



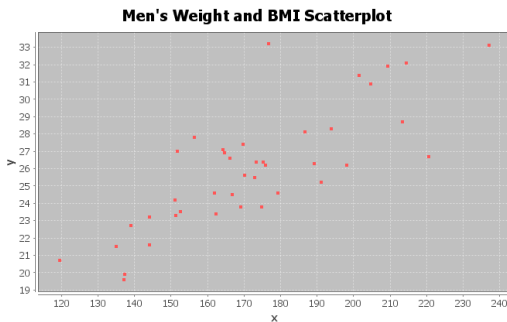
The scatterplot seems to show a positive linear trend. The dots tend to increase from left to right and could be close to a line.

There do not appear to be any outliers. All the points appear close to a line.

6.

Explanatory Variable (X): Weight of man  
Response Variable (Y): Body Mass Index (BMI) of man

Weight and Body Mass Index respond to each other, so we can choose either to be the response variable. It comes down to what variable are we more interested in predicting. I chose body mass index as the response variable (y) because I was interested in predicting BMI from weight.



The scatterplot seems to show a positive linear trend. The dots tend to increase from left to right and could be close to a line.

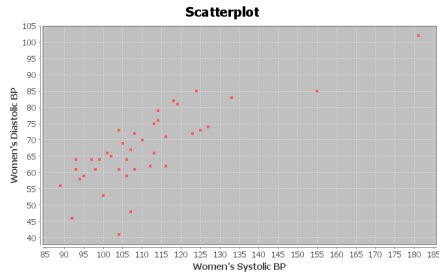
There do not appear to be any outliers. All the points appear close to a line. The most unusual point was (178 Lbs, 33.2 kg/m<sup>2</sup>), but this does not seem to be very far from the linear pattern.



7.

Explanatory Variable (X): Systolic Blood Pressure woman  
Response Variable (Y): Diastolic Blood Pressure woman

Systolic blood pressure and diastolic blood pressure respond to each other, so we can chose either to be the response variable. It comes down to what variable we are more interested in predicting. I chose diastolic blood pressure as the response variable (y) because I was interested in predicting diastolic blood pressure.



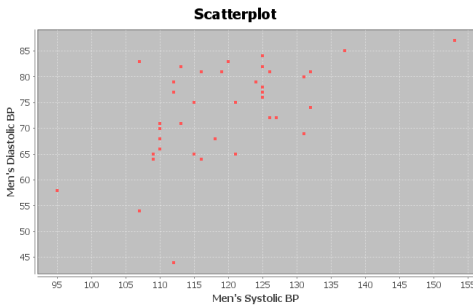
The scatterplot seems to show a positive linear trend. The dots tend to increase from left to right and are close to a line.

There do not appear to be any outliers. All the points appear close to a line.

8.

Explanatory Variable (X): Systolic Blood Pressure man  
Response Variable (Y): Diastolic Blood Pressure man

Systolic blood pressure and diastolic blood pressure respond to each other, so we can chose either to be the response variable. It comes down to what variable we are more interested in predicting. I chose diastolic blood pressure as the response variable (y) because I was interested in predicting diastolic blood pressure.



The scatterplot seems to show some positive linear trend. The dots tend to increase from left to right and could be close to a line.

There seems to be one unusual point at (112 mg/dL , 44 mg/dL). This may be an outlier. Another point to consider is (107 mg/dL , 84 mg/dL) but this does not seem to be very far from the linear pattern.

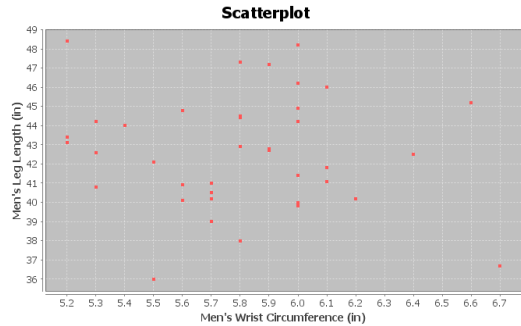




9.

Explanatory Variable (X): Wrist Circumference man  
Response Variable (Y): Leg Length man

The variables may respond to each other, so we can chose either to be the response variable. It comes down to what variable we are more interested in predicting. I chose leg length as the response variable (y) because I was interested in seeing if we can predict leg length from the wrist size.



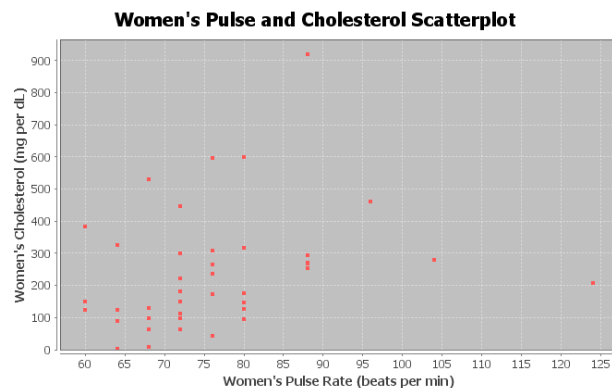
The scatterplot does not show any linear or curved trend. The dots seem to be all over without a distinguishable pattern. This indicates there is probably no relationship between men's wrist circumference and leg length.

Since there is not linear or curved trend, all of the points are scattered making it difficult to judge what is unusual or not. They all look unusual.

10.

Explanatory Variable (X): Pulse Rate Woman (beats per min)  
Response Variable (Y): Cholesterol of Woman (mg per dL)

The variables may respond to each other, so we can chose either to be the response variable. It comes down to what variable we are more interested in predicting. Checking cholesterol is more difficult as it requires a blood test, while pulse is relatively easy to check. I chose cholesterol as the response variable (y) because that is more difficult to measure and I was interested in seeing if we can predict cholesterol from pulse.



There does seem to be some positive linear trend, though the points are not as close to a line as I would like. The points show a slight upward trend from left to right.

There seems to be one unusual point (outlier) at (88 BPM , 920 mg/dL). Another point to consider is (124 BPM, 201 mg/dL). This may also be unusual.



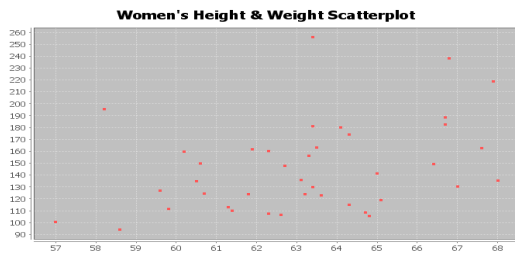
## Section 6B Answers

1.

Explanatory Variable (X): Height of woman

Response Variable (Y): Weight of woman

Though height and weight may respond to each other, we chose weight to be the response variable. The thinking was that as a woman grows taller, her weight may increase. Weight changes may not indicate that a woman's height is changing.



Correlation Coefficient  $r = +0.3644$

This means that there is a weak positive linear correlation between the height and weight of the women in the data set.

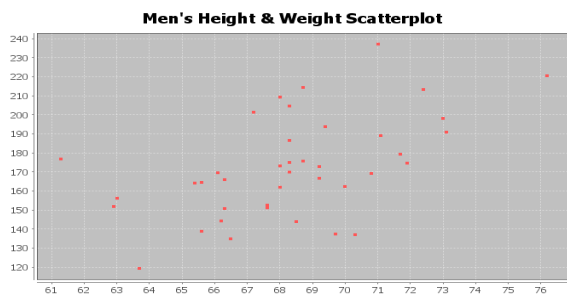
There are a couple points at (58.2 in, 196 Lb) and (63.4 in, 256 Lb) that don't seem to fit the pattern and could be unusual points (outliers). The correlation coefficient  $r$  is not very strong, indicating that these outliers are influential.

2.

Explanatory Variable (X): Height of man

Response Variable (Y): Weight of man

Though height and weight may respond to each other, we chose weight to be the response variable. The thinking was that as a man grows taller, his weight may increase. Weight changes may not indicate that a man's height is changing.



Correlation Coefficient  $r = +0.5222$

This tells us that there is a moderate positive correlation between the height and weight of the men in the data set.

There does not seem to be any points that are not following the linear pattern. No outliers. The correlation coefficient  $r$  being moderately high confirms that there are probably no influential outliers.

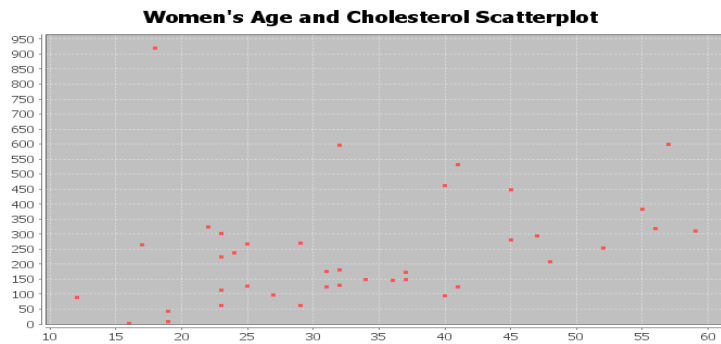


3.

Explanatory Variable (X): Age of woman

Response Variable (Y): Cholesterol of woman

A woman's cholesterol may change as a response to getting older, but age probably does not change in response to cholesterol.



Correlation Coefficient  $r = +0.3022$

This means that there is a weak positive linear correlation between the age and cholesterol of the women in the data set.

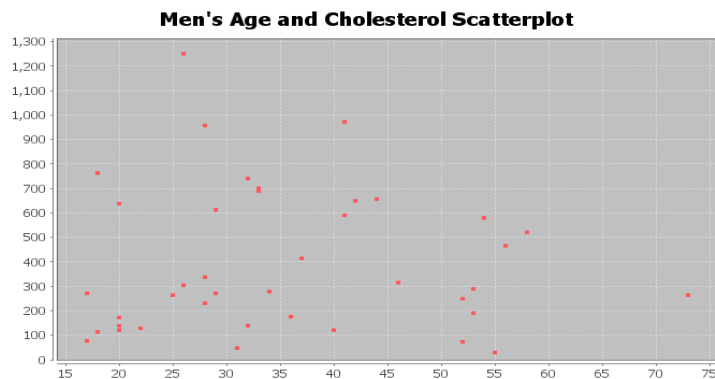
There is one point at (18 years old, 920 mg per dL) that doesn't seem to fit the pattern and is unusual (outlier). The correlation coefficient  $r$  is not very strong, indicating that this outlier is influential.

4.

Explanatory Variable (X): Age of man

Response Variable (Y): Cholesterol of man

A man's cholesterol may change as a response to getting older, but age probably does not change in response to cholesterol.



Correlation Coefficient  $r = -0.0154$

The correlation coefficient is close to zero, so there is no correlation between men's age and cholesterol.

Since there is not linear or curved trend, all of the points are scattered making it difficult to judge what is unusual or not. They all look unusual. The correlation coefficient confirms this. There is no correlation.

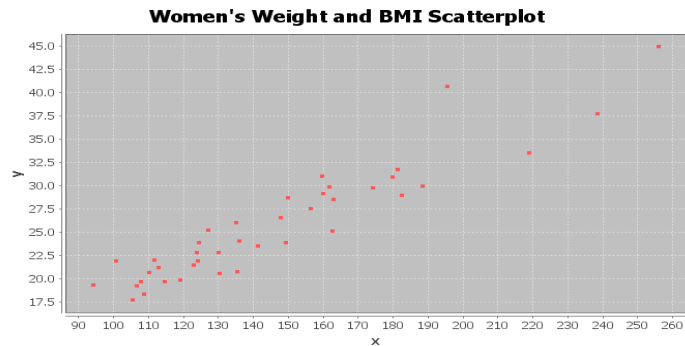


5.

Explanatory Variable (X): Weight of woman

Response Variable (Y): Body Mass Index (BMI) of woman

Weight and Body Mass Index respond to each other, so we can choose either to be the response variable. It comes down to what variable are we more interested in predicting. I chose body mass index as the response variable (y) because I was interested in predicting BMI from weight.



Correlation Coefficient  $r = +0.9361$

There is a very strong positive correlation between the weight and BMI of the women in the data set.

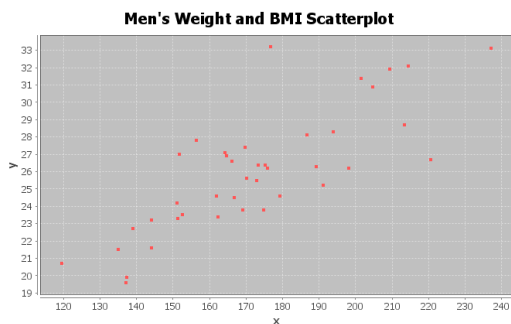
There do not appear to be any outliers. All the points appear close to a line. The correlation coefficient being so strong indicates there are no outliers.

6.

Explanatory Variable (X): Weight of man

Response Variable (Y): Body Mass Index (BMI) of man

Weight and Body Mass Index respond to each other, so we can choose either to be the response variable. It comes down to what variable are we more interested in predicting. I chose body mass index as the response variable (y) because I was interested in predicting BMI from weight.



Correlation Coefficient  $r = +0.7997$

There is a strong positive correlation between the weight and BMI of the men in the data set.

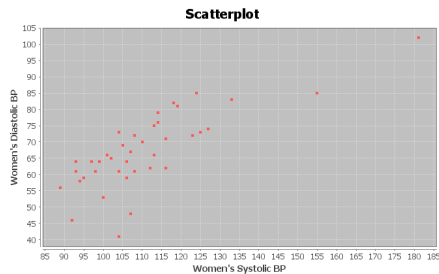
There do not appear to be any outliers. All the points appear close to a line. The most unusual point was (178 Lbs, 33.2 kg/m<sup>2</sup>), but this does not seem to be very far from the linear pattern. The correlation coefficient confirms this since it is strong. So this possible outlier is not influential.



7.

Explanatory Variable (X): Systolic Blood Pressure woman  
Response Variable (Y): Diastolic Blood Pressure woman

Systolic blood pressure and diastolic blood pressure respond to each other, so we can chose either to be the response variable. It comes down to what variable we are more interested in predicting. I chose diastolic blood pressure as the response variable (y) because I was interested in predicting diastolic blood pressure.



Correlation Coefficient  $r = +0.7854$

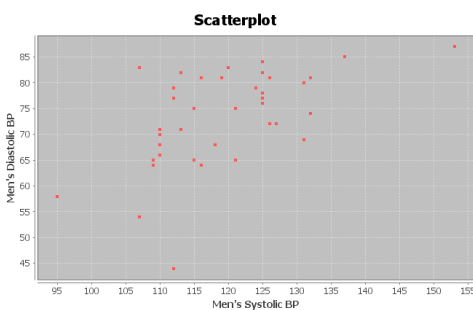
This indicates that there is a strong positive correlation between the systolic and diastolic blood pressure of the women in the data set.

There do not appear to be any outliers. All the points appear close to a line. The correlations coefficient  $r$  confirms this since it is strong. There are no influential outliers.

8.

Explanatory Variable (X): Systolic Blood Pressure man  
Response Variable (Y): Diastolic Blood Pressure man

Systolic blood pressure and diastolic blood pressure respond to each other, so we can chose either to be the response variable. It comes down to what variable we are more interested in predicting. I chose diastolic blood pressure as the response variable (y) because I was interested in predicting diastolic blood pressure.



Correlation Coefficient  $r = +0.5517$

This tells us there is a moderate positive correlation between the systolic and diastolic blood pressure of the men in this data set.

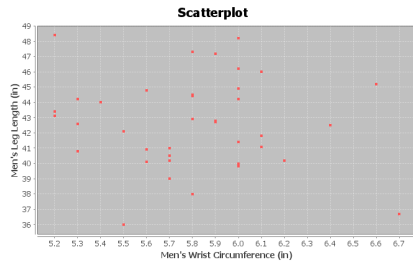
There seems to be one unusual point at (112 mg/dL , 44 mg/dL). This may be an outlier. Another point to consider is (107 mg/dL , 84 mg/dL) but this does not seem to be very far from the linear pattern. The correlation coefficient is only moderate and not strong. The (112 , 44) may be having a small influence on the correlation.



9.

Explanatory Variable (X): Wrist Circumference man  
Response Variable (Y): Leg Length man

The variables may respond to each other, so we can chose either to be the response variable. It comes down to what variable we are more interested in predicting. I chose leg length as the response variable (y) because I was interested in seeing if we can predict leg length from the wrist size.



Correlation Coefficient  $r = -0.0789$

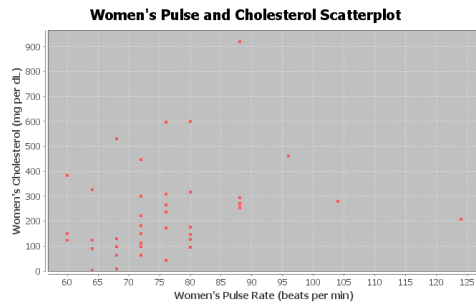
The correlation coefficient is close to zero, so there is no correlation between the wrist circumference and leg length of the men in the data set.

Since there is not linear or curved trend, all of the points are scattered making it difficult to judge what is unusual or not. They all look unusual. The correlation coefficient confirms this. There is no correlation.

10.

Explanatory Variable (X): Pulse Rate Woman (beats per min)  
Response Variable (Y): Cholesterol of Woman (mg per dL)

The variables may respond to each other, so we can chose either to be the response variable. It comes down to what variable we are more interested in predicting. Checking cholesterol is more difficult as it requires a blood test, while pulse is relatively easy to check. I chose cholesterol as the response variable (y) because that is more difficult to measure and I was interested in seeing if we can predict cholesterol from pulse.



Correlation Coefficient  $r = +0.2659$

This tells us that there is weak positive correlation between the pulse and cholesterol of the women in the data set.

There seems to be one unusual point (outlier) at (88 BPM , 920 mg/dL). Another point to consider is (124 BPM, 201 mg/dL). This may also be unusual. The correlation coefficient is very weak indicating these are both influential outliers.



## Section 6C Answers

1a. 79.0% of the variability in the men's weight can be explained by the linear relationship with the waist size. This tells us there is a very strong relationship between the variables.

1b. 0.31% of the variability in the men's weight can be explained by the linear relationship with the pulse rate. This tells us there is no relationship between these variables.

1c. 12.4% of the variability in the men's weight can be explained by the linear relationship with the systolic blood pressure. This tells us there is a weak relationship between these variables.

1d. 15.0% of the variability in the men's weight can be explained by the linear relationship with the diastolic blood pressure. This tells us there is a weak relationship between these variables.

1e. 0.07% of the variability in the men's weight can be explained by the linear relationship with the cholesterol. This tells us there is no relationship between these variables.

1f. 64% of the variability in the men's weight can be explained by the linear relationship with the body mass index. This tells us there is a very strong relationship between the variables.

1g. 13.8% of the variability in the men's weight can be explained by the linear relationship with the leg length. This tells us there is a weak relationship between these variables.

1h. 40.3% of the variability in the men's weight can be explained by the linear relationship with the elbow circumference. This tells us there is a moderate relationship between these variables.

1i. 27.0% of the variability in the men's weight can be explained by the linear relationship with the wrist circumference. This tells us there is a moderate relationship between these variables.

1j. 67.5% of the variability in the men's weight can be explained by the linear relationship with the arm length. This tells us there is a very strong relationship between the variables.

2.

$$r^2 = (0.7287)^2 = 0.531 = 53.1\%$$

53.1% of the variability in total trash can be explained by the linear relationship with paper trash.

Confounding Variables (answers may vary): metal trash, food trash, plastic trash, amount of recycling, number of trash trucks running

No. Correlation is not causation. We can say that there is a relationship or correlation but that does not imply that one variable causes another. There are many factors involved.

3.

$$r^2 = (0.5862)^2 = 0.344 = 34.4\%$$

34.4% of the variability in metal trash can be explained by the linear relationship with plastic trash.

Confounding Variables (answers may vary): paper trash, food trash, total trash, amount of recycling, number of trash trucks running

No. Correlation is not causation. We can say that there is a relationship or correlation but that does not imply that one variable causes another. There are many factors involved.



4.

$$r^2 = (0.5833)^2 = 0.340 = 34.0\%$$

34.0% of the variability in total trash can be explained by the linear relationship with food trash.

Confounding Variables (answers may vary): metal trash, paper trash, plastic trash, amount of recycling, number of trash trucks running

No. Correlation is not causation. We can say that there is a relationship or correlation but that does not imply that one variable causes another. There are many factors involved.

5.

$$r^2 = (-0.8713)^2 = 0.759 = 75.9\%$$

75.9% of the variability in miles per gallon can be explained by the linear relationship with horsepower.

Confounding Variables (answers may vary): weight of car, type of gas, type of engine, type of carburetor, freeway or road driving, wind resistance

No. Correlation is not causation. We can say that there is a relationship or correlation but that does not imply that one variable causes another. There are many factors involved.

6.

$$r^2 = (0.9404)^2 = 0.884 = 88.4\%$$

88.4% of the variability in profit can be explained by the linear relationship with the number of cars sold.

Confounding Variables (answers may vary): costs of company, number of cars available, type of cars, area, talent of the sales employees, number of employees

No. Correlation is not causation. We can say that there is a relationship or correlation but that does not imply that one variable causes another. There are many factors involved.

7.

$$r^2 = (0.6727)^2 = 0.453 = 45.3\%$$

45.3% of the variability in number of flowers can be explained by the linear relationship with amount of fertilizer.

Confounding Variables (answers may vary): type of flowers, weather, climate, temperature, area, amount of carbon dioxide, quality of soil before fertilizer was added

No. Correlation is not causation. We can say that there is a relationship or correlation but that does not imply that one variable causes another. There are many factors involved.

8.

$$r^2 = (-0.9429)^2 = 0.889 = 88.9\%$$

88.9% of the variability in this stock price can be explained by the linear relationship with the number of weeks (time).

Confounding Variables (answers may vary): type of stock, overall stock market trends, national debt, unemployment rates

No. Correlation is not causation. We can say that there is a relationship or correlation but that does not imply that one variable causes another. There are many factors involved.





9.

Men's Body Mass Index multivariable study

Age/BMI:  $r^2 = 0.071 = 7.1\%$

Height/BMI:  $r^2 = 0.008 = 0.8\%$

Weight/BMI:  $r^2 = 0.640 = 64.0\%$

Waist/BMI:  $r^2 = 0.731 = 73.1\%$

Cholesterol/BMI:  $r^2 = 0.012 = 1.2\%$

Follow up: Waist size had the strongest relationship with body mass index. Weight also had a strong relationship. A study of body mass index should focus on waist size and weight. Age had a weak relationship. Height and Cholesterol had virtually no relationship with BMI.

Surprises (answers may vary): Height is used in the calculation of body mass index, yet the correlation study indicated no relationship. This was surprising.

10:

Bear Weight multivariable study

Bear Age / Bear Weight:  $r^2 = 0.561 = 56.1\%$

Bear Head Length / Bear Weight:  $r^2 = 0.696 = 69.6\%$

Bear Head Width / Bear Weight:  $r^2 = 0.614 = 61.4\%$

Bear Neck Size / Bear Weight:  $r^2 = 0.873 = 87.3\%$

Bear Length / Bear Weight:  $r^2 = 0.747 = 74.7\%$

Bear Chest Size / Bear Weight:  $r^2 = 0.928 = 92.8\%$

Follow up: Chest Size of the bear had the strongest relationship with weight. Neck size and overall length had very strong relationships with weight also. Age, Head width and head length also had strong relationships with weight.

Surprises (answers may vary): All of the variables were related to weight. There were not any variables that were not related to the weight and they were all pretty strong relationships.



## Section 6D Answers

1.

$$\text{Slope} = r \text{ times } S_y / S_x = 0.7287 \times 12.46 / 4.268 = 2.1784$$

Slope Sentence: For every one ton increase in paper trash, the total trash increases about 2.178 tons.

$$Y \text{ int} = y \text{ mean} - (\text{slope})(x \text{ mean}) = 27.44 - (2.1784 \times 9.428) = 27.44 - 20.0538 = 6.902$$

Y int Sentence: If there was zero tons of paper trash, there would still be about 6.902 tons of total trash.

$$\text{Equation of Regression Line: } Y = 6.902 + 2.178 X$$

2.

$$\text{Slope} = r \text{ times } S_y / S_x = 0.5862 \times 1.091 / 1.065 = 0.6005 = 0.601$$

Slope Sentence: For every one ton increase in plastic trash, the metal trash increases about 0.6 tons.

$$Y \text{ int} = y \text{ mean} - (\text{slope})(x \text{ mean}) = 2.218 - (0.6005 \times 1.911) = 2.218 - 1.14755 = 1.07$$

Y int Sentence: If there was zero tons of plastic trash, there would still be about 1.07 tons of metal trash.

$$\text{Equation of Regression Line: } Y = 1.07 + 0.601 X$$

3.

$$\text{Slope} = r \text{ times } S_y / S_x = 0.5833 \times 12.46 / 3.297 = 2.204$$

Slope Sentence: For every one ton increase in food trash, the total trash increases about 2.204 tons.

$$Y \text{ int} = y \text{ mean} - (\text{slope})(x \text{ mean}) = 27.44 - (2.204 \times 4.816) = 27.44 - 10.614 = 16.826$$

Y int Sentence: If there was zero tons of food trash, there would still be about 16.826 tons of total trash.

$$\text{Equation of Regression Line: } Y = 16.826 + 2.204 X$$

4.

$$\text{Slope} = r \text{ times } S_y / S_x = 0.9404 \times 175.615 / 7.512 = 21.985$$

Slope Sentence: For every one car sold, the profits increases about 21.985 thousand dollars (\$21985).

$$Y \text{ int} = y \text{ mean} - (\text{slope})(x \text{ mean}) = 420.25 - (21.985 \times 21.667) = 420.25 - 476.349 = -56.099$$

Y int Sentence: If there was zero cars sold, the profits would be about -56.099 thousand dollars (loss of \$56099).

$$\text{Equation of Regression Line: } Y = -56.099 + 21.985 X$$

5.

$$\text{Slope} = r \text{ times } S_y / S_x = 0.6727 \times 1.356 / 1.165 = 0.783$$

Slope Sentence: For every one pound of fertilizer added, the number of flowers per square foot increase about 0.783.

$$Y \text{ int} = y \text{ mean} - (\text{slope})(x \text{ mean}) = 13.867 - (0.783 \times 3.387) = 13.867 - 2.652 = 11.215$$

Y int Sentence: If there was zero fertilizer, there would still be about 11.215 flowers per square foot.

$$\text{Equation of Regression Line: } Y = 11.215 + 0.783 X$$



6.

$$\text{Slope} = r \text{ times } S_y / S_x = -0.9429 \times 17.031 / 5.916 = -2.714$$

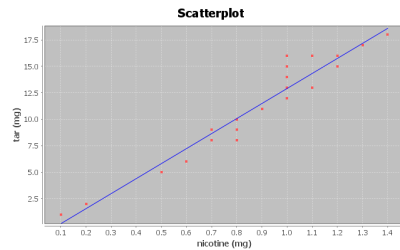
Slope Sentence: For every one week that goes by, the stock price decreases about \$2.71 on average.

$$Y \text{ int} = y \text{ mean} - (\text{slope})(x \text{ mean}) = 270.6 - (-2.714 \times 10.5) = 270.6 - (-28.497) = 270.6 + 28.497 = 299.097$$

Y int Sentence: At week zero, the stock price was \$299.10 per share.

$$\text{Equation of Regression Line: } Y = 299.097 - 2.714 X$$

7.



Correlation Coefficient  $r = 0.9614$

There is a very strong positive correlation between the amount of nicotine and tar. The regression line fits the data very well with no outliers. The regression line will be very accurate for predicting tar.

Regression:

Regression equation  $Y = b_0 + b_1X$

$$b_0 = -1.2713$$

$$b_1 = 14.2076$$

Y-intercept = -1.2713

If there was zero mg of nicotine, then the amount of tar would be about -1.2713 mg.

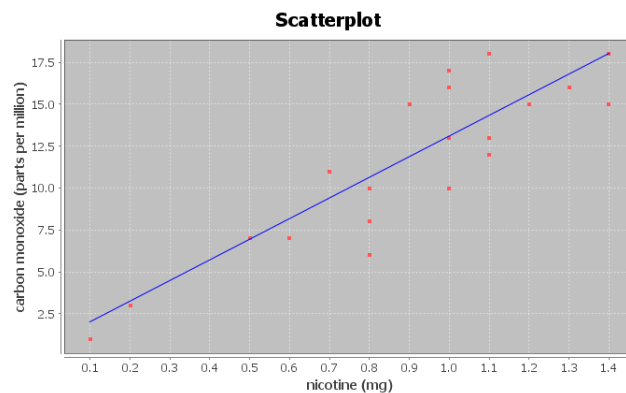
*Note: The Y-intercept interpretation doesn't make sense in context because an x value of zero is not in the scope of the x values on the scatterplot. The formula is not designed to predict tar when nicotine is zero.*

Slope = 14.2076

For every 1 mg of nicotine added to a cigarette, they add 14.2076 mg of tar.

$$\text{Regression Line Equation: } Y = -1.2713 + 14.2076 X$$

8.



Correlation Coefficient  $r = 0.8633$

There is a strong positive correlation between the amount of nicotine and carbon monoxide. The regression line fits the data very well with no outliers. The regression line will be very accurate for predicting carbon monoxide.

Regression:

Regression equation  $Y = b_0 + b_1X$

$b_0 = 0.7950$

$b_1 = 12.3057$

Y-intercept = 0.7950

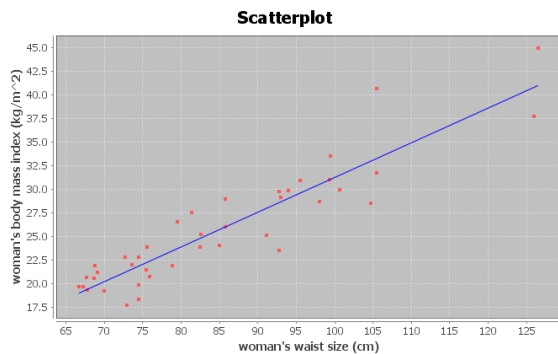
If a cigarette with zero mg of nicotine is lighted, it will still release 0.7950 ppm of carbon monoxide.

Slope = 12.3057

For every 1 mg of nicotine added to a cigarette, the amount of carbon monoxide increases 12.3057 ppm.

Regression Line Equation:  $Y = 0.7950 + 12.3057 X$

9.



Correlation Coefficient  $r = 0.9181$

There is a very strong positive correlation between the women's waist size and body mass index in the data set. The regression line fits the data very well with no outliers. The regression line will be very accurate for predicting body mass index from waist size.

Regression:

Regression equation  $Y = b_0 + b_1X$

$b_0 = -5.5117$

$b_1 = 0.3675$

Y-intercept = -5.5117

If a woman had a waist size of zero centimeters, she would have a predicted body mass index of -5.5117.

*Note: The Y-intercept interpretation doesn't make sense in context because an x value of zero is not in the scope of the x values on the scatterplot. The formula is not designed to predict BMI for a waist size of zero. A waist size of zero and a body mass index of negative 5.5 are both impossible.*

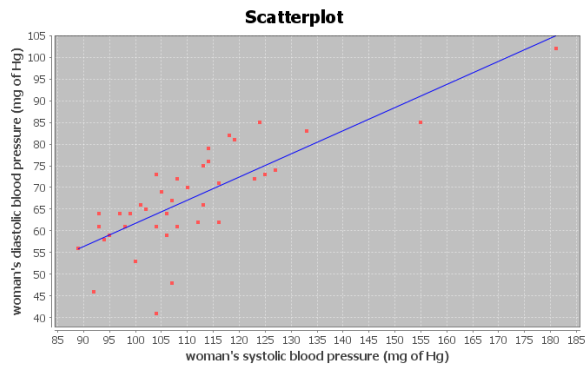
Slope = 0.3675

For every 1 cm increase in waist size, the women's body mass index increases 0.3675.

Regression Line Equation:  $Y = -5.5117 + 0.3675 X$



10.



Correlation Coefficient  $r = 0.7854$

There is a strong positive correlation between the women's systolic and diastolic blood pressure in the data set. The regression line fits the data very well with no influential outliers. The regression line will be very accurate for predicting diastolic BP from systolic BP.

Regression:

Regression equation  $Y = b_0 + b_1X$

$b_0 = 8.3079$

$b_1 = 0.5335$

Y-intercept = 8.3079

If a woman had a systolic blood pressure of zero mm of Hg, she would have a predicted diastolic blood pressure of 8.3079 mm of Hg.

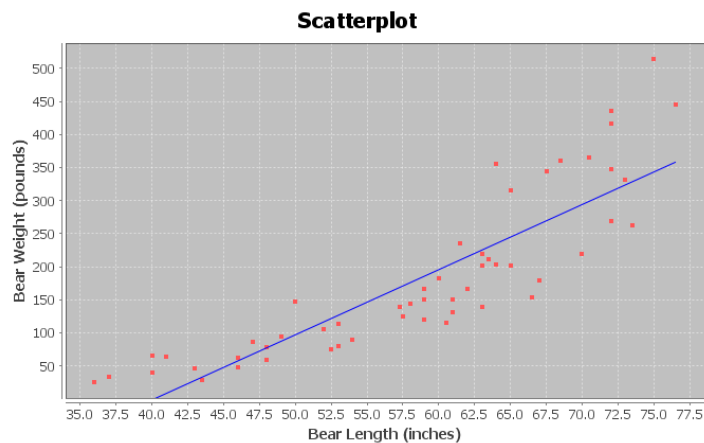
*Note: The Y-intercept interpretation doesn't make sense in context because an x value of zero is not in the scope of the x values on the scatterplot. The formula is not designed to predict diastolic blood pressure for a systolic blood pressure of zero. A living person cannot have a blood pressure of zero.*

Slope = 0.5335

For every 1 mm of Hg increase in systolic blood pressure, the women's diastolic blood pressure increases 0.5335 mm of Hg.

Regression Line Equation:  $Y = 8.3079 + 0.5335 X$

11.



Correlation Coefficient  $r = 0.8644$

There is a strong positive correlation between the length and weight of the bears. The regression line fits the data very well with no influential outliers. The regression line will be very accurate for predicting bear weights.

Regression:

Regression equation  $Y = b_0 + b_1X$

$b_0 = -393.8391$

$b_1 = 9.8390$

Y-intercept = -393.8391

If a bear has a length of zero inches, then the bear would have a predicted weight of -393.8391 pounds.

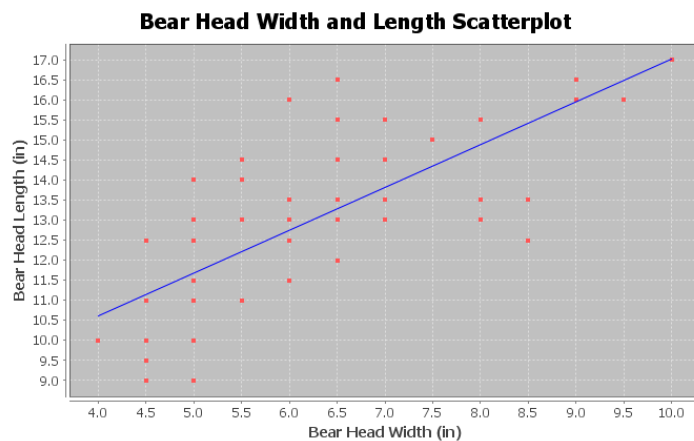
*Note: The Y-intercept interpretation doesn't make sense in context because an x value of zero is not in the scope of the x values on the scatterplot. The formula is not designed to predict a bear weight from a length of zero. A bear length of zero and a weight of -393.8391 pounds are both impossible.*

Slope = 9.8390

For every 1 inch longer a bear gets, the weight of the bear increases about 9.8390 pounds.

Regression Line Equation:  $Y = -393.8391 + 9.8390 X$

12.



Correlation Coefficient  $r = 0.7535$

There is a strong positive correlation between the head width and head length of the bears. The regression line fits the data very well with no influential outliers. The regression line will be accurate for predicting bear head lengths from the bear head width.

Regression:

Regression equation  $Y = b_0 + b_1X$

$b_0 = 6.3362$

$b_1 = 1.0683$

Y-intercept = 6.3362

If a bear has a head width of zero inches, then the bear would have a predicted head length of 6.3362 inches.

*Note: The Y-intercept interpretation doesn't make sense in context because an x value of zero is not in the scope of the x values on the scatterplot. The formula is not designed to predict a bear head length from a head width of zero. A bear head width of zero is impossible.*

Slope = 1.0683

For every 1 inch wider a bear's head gets, the head length of the bear increases about 1.0683 inches.

Regression Line Equation:  $Y = 6.3362 + 1.0683 X$



## Section 6E Answers

1.

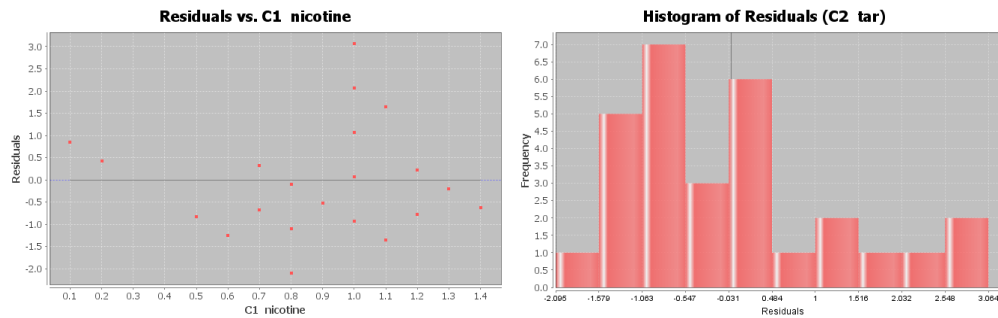
Se = 1.2984 mg of tar

The points in the scatterplot are 1.2984 mg of tar from the regression line on average.

If we use a nicotine value in the scope of the x values and regression line to predict the amount of tar a cigarette has, our prediction could have an average error of 1.2984 mg of tar.

The residual plot does not look evenly spread out. It looks “V” shaped (fan shaped). It is more spread out on the right side of the graph and less spread out on the left side of the graph.

The histogram is not bell shaped (skewed right). It is also not centered at zero.



2.

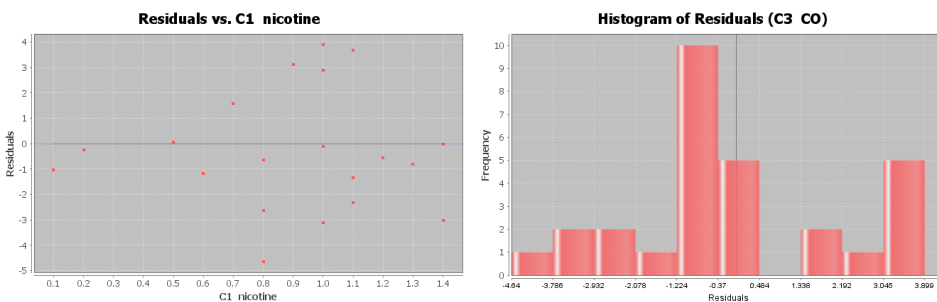
Se = 2.2961 PPM

The points in the scatterplot are 2.2961 parts per million (ppm) from the regression line on average.

If we use a nicotine value in the scope of the x values and regression line to predict the amount of carbon monoxide a cigarette releases when burned, our prediction could have an average error of 2.2961 ppm.

The residual plot does not look evenly spread out. It looks “V” shaped (fan shaped). It is more spread out on the right side of the graph and less spread out on the left side of the graph.

The histogram does look relatively bell shaped. However it is not centered at zero.



3.

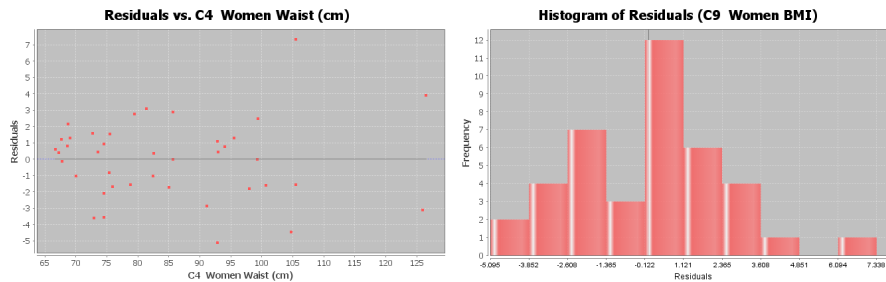
$$Se = 2.4761 \text{ kg/m}^2$$

The points in the scatterplot are 2.4761 kg/m<sup>2</sup> from the regression line on average.

If we use a woman's waist size in the scope of the x values and regression line to predict the woman's body mass index, our prediction could have an average error of 2.4761 kg/m<sup>2</sup>.

Overall the residual plot is pretty evenly spread out. There is one point that is far away from the regression line in the top right section of the residual plot that does make the graph look slightly "V" shaped (fan shaped).

The histogram does look relatively bell shaped. It is also centered at zero.



4.

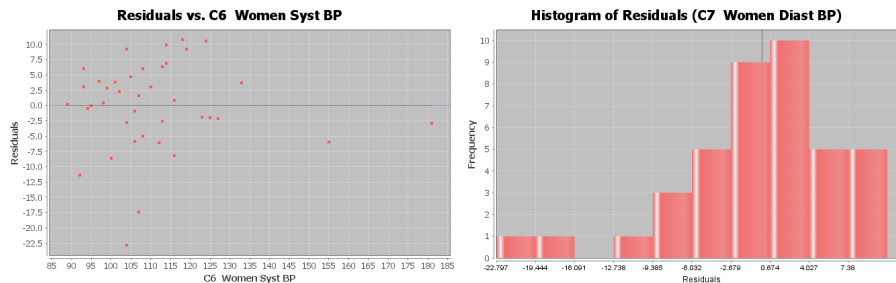
$$Se = 7.2912 \text{ mm of Hg}$$

The points in the scatterplot are 7.2912 mm of Hg from the regression line on average.

If we use a woman's systolic blood pressure in the scope of the x values and regression line to predict the woman's diastolic blood pressure, our prediction could have an average error of 7.2912 mm of Hg.

The residual plot does not look evenly spread out. It looks "V" shaped (fan shaped). It is more spread out on the left side of the graph and less spread out on the right side of the graph.

The histogram does not look bell shaped (skewed left). The center is a little off from zero.





5.

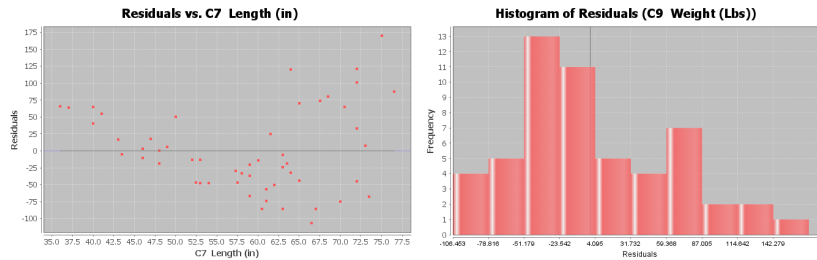
Se = 61.8272 pounds

The points in the scatterplot are 61.8272 pounds from the regression line on average.

If we use a bears length in the scope of the x values and regression line to predict the bears weight, our prediction could have an average error of 61.8272 pounds.

The residual plot does not look evenly spread out. It looks “V” shaped (fan shaped). It is more spread out on the right side of the graph and less spread out on the left side of the graph.

The histogram does not look bell shaped (skewed right). The center is a little off from zero.



6.

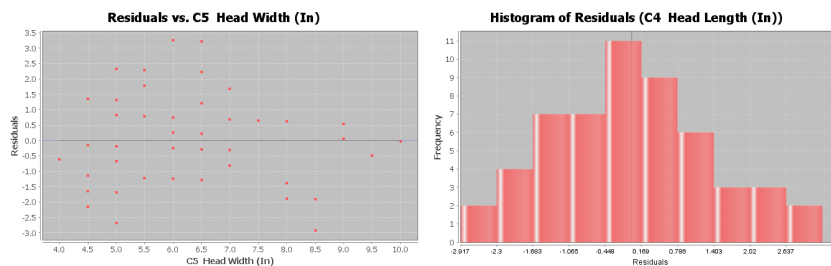
Se = 1.4231 inches

The points in the scatterplot are 1.4231 inches from the regression line on average.

If we use a bears head width in the scope of the x values and regression line to predict the bears head length, our prediction could have an average error of 1.4231 inches.

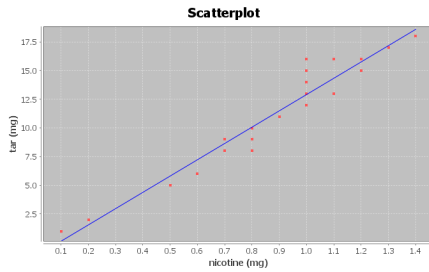
The residual plot does not look evenly spread out. It looks “V” shaped (fan shaped). It is more spread out on the left side of the graph and less spread out on the right side of the graph.

The histogram does look bell shaped. The histogram also looks centered at zero.



## Section 6F Answers

1a.



Correlation Coefficient  $r = 0.9614$

There is a very strong positive correlation between the amount of nicotine and tar. The regression line fits the data very well with no outliers. The regression line will be very accurate for predicting tar.

1b.

Regression Line Equation:  $Y = -1.2713 + 14.2076 X$

1c.

0.1 mg nicotine  $\leq$  scope of X values  $\leq$  1.4 mg of nicotine

Zero is an extrapolation since it does not fall in the scope of the X values. This is why the Y intercept -1.2713 does not make sense. You cannot have a negative amount of tar. The formula is not designed to plug in zero for x. Not surprising the predicted Y value from an extrapolation can be dramatically wrong.

1d.

Prediction for X = 0.8 mg nicotine:

$$Y = -1.2713 + 14.2076 X$$

$$Y = -1.2713 + 14.2076 (0.8)$$

$$Y = -1.2713 + 11.36608$$

$$Y = 10.09478 \approx 10.1 \text{ mg of tar}$$

A cigarette with 0.8 mg of nicotine would be predicted to have 10.1 mg of tar. This prediction could have an average error of 1.2984 mg of tar (Se). So our prediction of 10.1 mg of tar could be about 1.3 mg too high or too low.

1e.

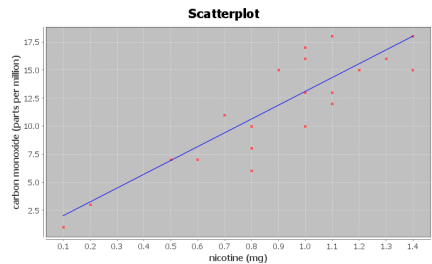
Prediction for X = 4.75 mg nicotine:

We should not use this formula to predict the amount of tar for a cigarette that has 4.75 mg of nicotine. The formula was not designed to plug in 4.75 for X. 4.75 is way out of the scope of the x values and would be an extrapolation. It could result in dramatic errors.

1f. Answers may vary. Tar is a dangerous substance to put into the body. Cigarette smoking has been linked to lung cancer and other diseases.



2a.



Correlation Coefficient  $r = 0.8633$

There is a strong positive correlation between the amount of nicotine and carbon monoxide. The regression line fits the data very well with no outliers. The regression line will be very accurate for predicting carbon monoxide.

2b.

Regression Line Equation:  $Y = 0.7950 + 12.3057 X$

2c.

0.1 mg nicotine  $\leq$  scope of X values  $\leq$  1.4 mg of nicotine

Zero is an extrapolation since it does not fall in the scope of the X values. The Y intercept may not make sense. Though the formula is not designed to plug in zero for x, the number may have some meaning here. If a cigarette with zero mg of nicotine is lighted, it will still release 0.7950 ppm of carbon monoxide.

2d.

Prediction for X = 1.2 mg nicotine:

$$Y = 0.7950 + 12.3057 X$$

$$Y = 0.7950 + 12.3057 (1.2)$$

$$Y = 0.7950 + 14.76684$$

$$Y = 15.56184 \approx 15.6 \text{ PPM of carbon monoxide}$$

A cigarette that has 1.2 mg of nicotine would be predicted to release 15.6 ppm of carbon monoxide when burned. This prediction could have an average error of 2.2961 parts per million (ppm) (Se). So our prediction of 15.6 ppm of carbon monoxide could be about 2.3 ppm too high or too low.

2e.

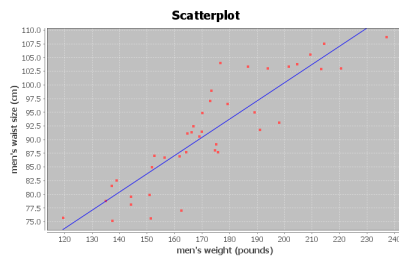
Prediction for X = 4.75 mg nicotine:

We should not use this formula to predict the amount of tar for a cigarette that has 4.75 mg of nicotine. The formula was not designed to plug in 4.75 for X. 4.75 is way out of the scope of the x values and would be an extrapolation. It could result in dramatic errors.

2f. Answers may vary. Carbon Monoxide is a very dangerous gas to take into the lungs. Cigarette smoke has been linked to lung cancer and other diseases.



3a.



Correlation Coefficient  $r = 0.8889$

There is a strong positive correlation between the weight and waist size of these men. The regression line fits the data very well with no outliers. The regression line will be very accurate for predicting waist size.

3b.

Regression Line Equation:  $Y = 33.8291 + 0.3330 X$

3c.

120 pounds  $\leq$  scope of X values  $\leq$  237 pounds

Zero is an extrapolation since it does not fall in the scope of the X values. The Y intercept does not make sense. The formula is not designed to plug in zero for x. It is impossible for a man to have a weight of zero pounds.

3d.

Prediction for X = 200 pounds:

$$Y = 33.8291 + 0.3330 X$$

$$Y = 33.8291 + 0.3330 (200)$$

$$Y = 33.8291 + 66.6$$

$$Y = 100.4291 \approx 100.4 \text{ cm}$$

A man that weighs 200 pounds would be predicted to have a waist size of about 100.4 cm. This prediction could have an average error of 4.5763 cm (Se). So our prediction of 100.4 cm could be about 4.6 cm too high or too low.

3e.

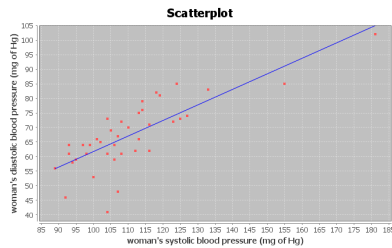
Prediction for X = 400 pounds:

We should not use this formula to predict the waist size of a man that is 400 pounds. The formula was not designed to plug in 400 for X. 400 pounds is way out of the scope of the x values and would be an extrapolation. It could result in dramatic errors.

3f. Answers may vary. This could have applications in the medical field and clothing industry.



4a.



Correlation Coefficient  $r = 0.7854$

There is a strong positive correlation between the women's systolic and diastolic blood pressure in the data set. The regression line fits the data very well with no influential outliers. The regression line will be very accurate for predicting diastolic BP from systolic BP.

4b.

Regression Line Equation:  $Y = 8.3079 + 0.5335 X$

4c.

89 pounds  $\leq$  scope of X values  $\leq$  181 pounds

Zero is an extrapolation since it does not fall in the scope of the X values. The Y intercept does not make sense. The formula is not designed to plug in zero for x. It is impossible for a living woman to have a systolic blood pressure of zero.

4d.

Prediction for  $X = 135$  mm of Hg:

$$Y = 8.3079 + 0.5335 X$$

$$Y = 8.3079 + 0.5335 (135)$$

$$Y = 8.3079 + 72.0225$$

$$Y = 80.3304 \approx 80.3 \text{ mm of Hg}$$

A woman with a systolic blood pressure of 135 mm of Hg would be predicted to have a diastolic blood pressure of about 80.3 mm of Hg. This prediction could have an average error of 7.2912 mm of Hg (Se). So our prediction of 80.3 mm of Hg could be about 7.3 mm of Hg too high or too low.

4e.

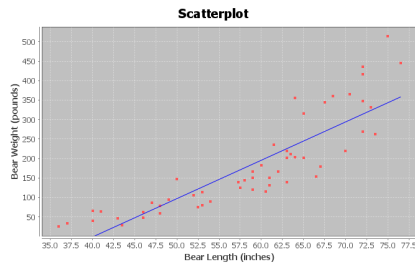
Prediction for  $X = 240$  mm of Hg:

We should not use this formula to predict the diastolic blood pressure for a woman with a systolic blood pressure of 240. The formula was not designed to plug in 240 for X. 240 mm of Hg is way out of the scope of the x values and would be an extrapolation. It could result in dramatic errors.

4f. Answers may vary. High blood pressure is a dangerous condition and needs to be studied to better understand how to help people.



5a.



Correlation Coefficient  $r = 0.8644$

There is a strong positive correlation between the length and weight of bears in the data set. The regression line fits the data very well with no influential outliers. The regression line will be very accurate for predicting the weight of bears from the length.

5b.

Regression Line Equation:  $Y = -393.8391 + 9.8390 X$

5c.

36 inches  $\leq$  scope of X values  $\leq$  76.5 inches

Zero is an extrapolation since it does not fall in the scope of the X values. The Y intercept does not make sense. The formula is not designed to plug in zero for x. It is impossible for a bear to be zero inches long and it is impossible for a bear to weigh -393.8 pounds.

5d.

Prediction for  $X = 72$  inches:

$$Y = -393.8391 + 9.8390 X$$

$$Y = -393.8391 + 9.8390 (72)$$

$$Y = -393.8391 + 708.408$$

$$Y = 314.5689 \approx 314.6 \text{ pounds}$$

A bear that is 72 inches long would have a predicted weight of 314.6 pounds.

This prediction could have an average error of 61.8272 pounds (Se). So our predicted bear weight of 314.6 pounds could be about 61.8 pounds too high or too low.

5e.

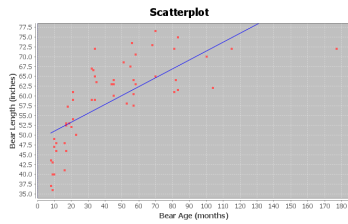
Prediction for  $X = 18$  inches:

We should not use this formula to predict the weight of a young bear that is 18 inches long. The formula was not designed to plug in 18 for X. 18 inches is way out of the scope of the x values and would be an extrapolation. It could result in dramatic errors.

5f. Answers may vary. This analysis may be important to anyone studying bears.



6a.



Correlation Coefficient  $r = 0.7188$

There is a strong positive correlation between the age and length of bears in the data set. The regression line fits the data very well with no influential outliers. The regression line will be very accurate for predicting the length of bears from the age.

6b.

Regression Line Equation:  $Y = 48.6903 + 0.2281 X$

6c.

8 months  $\leq$  scope of X values  $\leq$  177 months

Zero is an extrapolation since it does not fall in the scope of the X values. The Y intercept does not make sense. The formula is not designed to plug in zero for x. A bear zero months old (newborn) is not 48.7 inches long.

6d.

Prediction for X = 120 months:

$$Y = 48.6903 + 0.2281 X$$

$$Y = 48.6903 + 0.2281 (120)$$

$$Y = 48.6903 + 27.372$$

$$Y = 76.0623$$

A bear that is 120 months (10 years) old would have a predicted length of 76.1 inches.

This prediction could have an average error of 7.5109 inches (Se). So our predicted bear length of 76.1 inches could be about 7.5 inches too high or too low.

6e.

Prediction for X = 0 months:

We should not use this formula to predict the weight of a newborn bear at 0 months. The formula was not designed to plug in 0 for X. Zero months is way out of the scope of the x values and would be an extrapolation. It would result in dramatic error.

6f. Answers may vary. This analysis may be important to anyone studying bears.



## Answers to Problems from Chapter 6 Review Sheet

1.

Explanatory Variable: the “X” variable

Response Variable: This is the “Y” variable and should respond to the explanatory variable. This is the focus of the correlation study and the variable you want to make predictions about.

Correlation Coefficient “r”: A statistic between -1 and +1 that measures the strength and direction of the correlation.

r-squared: The square of the correlation coefficient tells us the percent of variability in the Y variable that can be explained by the linear relationship with the X variable.

Slope: A rate of change that measures the amount of increase or decrease in the y variable per unit of x.

Y-intercept: The predicted Y value when X is zero.

Residual: The vertical distance that each point in the scatterplot is above or below the regression line.

Standard Deviation of the Residual Errors (Se): A statistic that measure the average distance that the points in the scatterplot are from the regression line. It also measures the average amount of error if the regression line is used to make a prediction.

2. The response variable Y should respond to the explanatory variable x, but the key to choosing the variables is to know which variable you plan to make predictions about. The focus of your correlation study and the variable you want to make predictions about should be the response variable (Y). The variable you are not making predictions about is the explanatory variable (x).

3a. Both variables may respond to each other. Sanvi should make the hours of sleep the explanatory variable (x) since she wants to predict the number of migraines.

3b. Both variables may respond to each other. Sanvi should make the number of migraines the response variable (y) since the focus of her study is migraine headaches and she wants to predict the number of migraines.

3c. She will not be able to prove the lack of sleep causes migraines because correlation is not causation. There are many confounding variables that may influence migraines besides sleep.

4.

Scatterplot C: No Correlation ( $r = 0.023$ )

Scatterplot D: Strong Negative Correlation ( $r = -0.993$ )

Scatterplot B: Moderate Positive Correlation ( $r = 0.592$ )

5a. Slope = 2.489075718 (about 2.49)

5b. Y intercept = 34.80952773 (about 34.81)

6. The scatterplot and the correlation coefficient r indicate that there is a moderate positive correlation between the wrist size and the BMI of these women.

7. 4.2 in < scope of x values (wrist circumference) < 5.8 in

8. Slope = 10.9407

For every 1 inch increase in wrist circumference, the Body mass index is increasing 10.9407 kg/m<sup>2</sup>.





9.  $r^2 = 0.3446 \times 100\% = 34.46\%$

10.

34.46% of the variability in the women's body mass index can be explained by the linear relationship with wrist size.

11. (Answers may vary) Weight, Height, Muscle Mass, Diet, Exercise, Genetics

12. No. Correlation does not prove causation.

13.  $Se = 5.0568 \text{ kg/m}^2$  (BMI units)

14.

The average distance the points are from the line is  $5.0568 \text{ kg/m}^2$ .

If we predict a woman's body mass index from her wrist size, we could have an average error of  $5.0568 \text{ kg/m}^2$ .

15.

The residual plot is V shaped (fan shaped).

16.

There does not appear to be any curve pattern in the residual plot.

17.

No. The histogram does not look very bell shaped.

18.

No. The graph does not appear to be centered at zero.

19.

$Y = 48.802 - 8.367(4.5) = 48.802 - 37.6515 = 11.1505 \text{ kg/m}^2$

20.

Prediction could have an average error of  $5.0568 \text{ kg/m}^2$ . (Standard Deviation of the Residuals)

21.

No. The scope of the X values is 4.2 to 5.8 inches. A child's wrist of 3.1 inches is out of the scope and would be an extrapolation if we used it to predict BMI. It would have a large error in the prediction.

---



## Introduction to Data Analysis (2<sup>nd</sup> Edition) Chapter 7 Answer Keys

### Section 7A Answers

1.

a)

The scatterplot shows an exponential growth pattern. The scope of X values is from 0 years to 14 years since 1995. This would represent years 1995 – 2009.

b)

The exponential growth curve fits the data very well. The points are very close to the curve. The data appears to have a strong exponential relationship.

c)

There were 8 ordered pairs in the data.

d)

Exponential Curve Equation:  $\hat{y} = 4945.11427 (1.28424)^x$

Y-intercept (Predicted Y value when  $x = 0$ ): 4945.11427 MW

Base: 1.28424

Base is greater than 1, confirming that this is an exponential growth curve. If the base were less than 1, it would be a decay curve.

e)

$r^2 = 0.9965$

r-squared sentence: 99.65% of the variability in wind power can be explained by the exponential relationship with the years since 1995.

The r-squared value also confirms that there is an extremely strong exponential relationship between the variables.

f)

The points in the scatterplot are about 4039.2 megawatts (MW) from the exponential curve on average.

If we use the exponential curve equation and a year in the scope of the x values to make a prediction, our prediction could have an average error of 4039.2 MegaWatts (MW).

g)

$y = 4945.11427 (1.28424)^x$

$y = 4945.11427 (1.28424)^{(7)}$

$y = 4945.11427 (5.761337922)$

$Y = 28490.47437 \approx 28,490.5$  MW

The exponential curve predicts that by 2002 (year 7) there should be about 28,490.5 MW of wind power generated worldwide.

This prediction could have an average error of about 4039.2 MW (Se) too high or too low.



h)

$$y = 4945.11427 (1.28424)^x$$

$$y = 4945.11427 (1.28424)^{(13)}$$

$$y = 4945.11427 (25.84642641)$$

$$Y = 127813.5321 \approx 127,813.5 \text{ MW}$$

The exponential curve predicts that by 2008 (year 13) there should be about 127,813.5 MW of wind power generated worldwide.

This prediction could have an average error of about 4039.2 MW (Se) too high or too low.

i)

No. We should not plug in 70 into the equation. Year 70 would be an extremely bad extrapolation and could result in a huge error. The standard deviation of the residuals would not apply as the prediction error since 70 is way out of the scope of the x values.

2.

a)

The scatterplot shows an exponential decay pattern. The scope of the X-value are from 1 – 25 months since January 2010. These represent February 2010 to March 2012.

b)

The exponential decay curve fits the data moderately well. The points are somewhat close to the curve. It appears to have a moderate exponential relationship.

c)

There were 25 ordered pairs in the data.

d)

$$\text{Exponential Curve: } y = 63.85340 (0.97985)^x$$

Y-intercept (predicted y value when  $X = 0$ ) is 63.85340 thousand dollars (\$65,853.40)

The base is 0.97985. Notice the base is less than 1. This indicates that this is an exponential decay curve.

e)

$$r^2 = 0.8337$$

r-squared sentence: 83.37% of the variability in the retirement account balance can be explained by the exponential relationship with the months since January 2010.

We thought by the graph that there was only a moderate relationship, but the r-squared value indicates that there is a strong exponential relationship between the variables.



f)

The points in the scatterplot are about 4.182 thousand dollars from the exponential curve on average.

If we use the exponential curve equation and a month in the scope of the x values to make a prediction, our prediction could have an average error of 4.182 thousand dollars.

g)

$$y = 63.85340 (0.97985)^x$$

$$y = 63.85340 (0.97985)^{(11.5)}$$

$$y = 63.85340 (0.791289431)$$

$$Y = 50.52652055 \approx 50.5 \text{ thousand dollars}$$

The exponential curve predicts that by month 11.5 there should be about 50.5 thousand dollars left in the retirement account.

This prediction could have an average error of about 4.182 thousand dollars too high or too low.

h)

$$y = 63.85340 (0.97985)^x$$

$$y = 63.85340 (0.97985)^{(24.5)}$$

$$y = 63.85340 (0.607309571)$$

$$Y = 38.77878096 \approx 38.8 \text{ thousand dollars}$$

The exponential curve predicts that by month 24.5 there should be about 38.8 thousand dollars left in the retirement account.

This prediction could have an average error of about 4.182 thousand dollars too high or too low.

i)

No. We should not plug in 480 into the equation. Month 480 would be an extremely bad extrapolation and could result in a huge error. Without adding to the account, the retirement account balance would have run out long before then. The standard deviation of the residuals would not apply as the prediction error.

3.

a) The scatterplot shows an exponential growth pattern. The scope of x values is from 0 years to 22 years since 1990. These represent the years 1990 to 2012.

b)

The exponential growth curve fits the data very well. The points are very close to the curve. It appears to have a strong exponential relationship.

c)

There were 8 ordered pairs in the data.



d)

Exponential Curve:  $y = 497.44019 (1.07211)^x$

Y-intercept (predicted Y value when X = 0): \$497.44019

Base = 1.07211

The base is greater than 1. This indicates that this equation describes an exponential growth curve.

e)

$r^2 = 0.9806$

r-squared sentence: 98.06% of the variability in the savings account balance can be explained by the exponential relationship with the years since 1990.

The r-squared value indicates that there is a very strong exponential relationship between the variables.

f)

The points in the scatterplot are about \$110.94 from the exponential curve on average.

If we use the exponential curve equation and a year in the scope of the x values to make a prediction, our prediction could have an average error of \$110.94.

g)

$y = 497.44019 (1.07211)^x$

$y = 497.44019 (1.07211)^{16}$

$y = 497.44019 (3.046698999)$

$Y = 1515.550529 \approx \$1515.55$

The exponential curve predicts that by 2006 (year 16) there should be about \$1515.55 in the savings account.

This prediction could have an average error of about \$110.94 too high or too low.

h)

$y = 497.44019 (1.07211)^x$

$y = 497.44019 (1.07211)^{21}$

$y = 497.44019 (4.315451939)$

$Y = 2146.679233 \approx \$2146.68$

The exponential curve predicts that by 2011 (year 21) there should be about \$2146.68 in the savings account.

This prediction could have an average error of about \$110.94 too high or too low.

i)

No. We should not plug in 50 into the equation. Year 50 would be an extremely bad extrapolation and could result in a huge error. If money in this account is invested, the investment may go bad at some point or there may be recession. Also, the standard deviation of the residuals would not apply as the prediction error.



4.

4a. The scatterplot shows an exponential decay pattern. The scope of x values (metal distance) are from 0.5 mm to 6 mm.

b)

The exponential decay curve fits the data very well. The points are very close to the curve. It appears to have a strong exponential relationship.

c)

There were 214 ordered pairs in the data.

d)

Exponential Curve:  $y = 74.91083 (0.61685)^x$

Y-intercept (predicted Y value when X is zero) = 74.91083 Watts per square cm

Base = 0.61685

Notice the base is less than 1. This indicates that the exponential equation corresponds to an exponential decay curve.

e)

$r^2 = 0.9126$

r-squared sentence: 91.26% of the variability in ultrasonic response (Watts per square cm) can be explained by the exponential relationship with the metal distance in mm.

The r-squared value indicates that there is a very strong exponential relationship between the variables.

f)

The points in the scatterplot are about 8.239 Watts per  $\text{cm}^2$  from the exponential curve on average.

If we use the exponential curve equation and a metal distance in the scope of the x values to make a prediction, our ultrasound response prediction could have an average error of 8.239 Watts per  $\text{cm}^2$ .

g)

$y = 74.91083 (0.61685)^x$

$y = 74.91083 (0.61685)^{(2.83)}$

$y = 74.91083 (0.254805137)$

$Y = 19.08766435 \approx 19.1$  Watts per  $\text{cm}^2$

The exponential curve predicts that if the metal is 2.83 mm away, the ultrasound response will be about 19.1 Watts per  $\text{cm}^2$ .

This prediction could have an average error of about 8.239 Watts per  $\text{cm}^2$  too high or too low.



h)

$$y = 74.91083 (0.61685)^x$$

$$y = 74.91083 (0.61685)^{4.51}$$

$$y = 74.91083 (0.113164408)$$

$$Y = 8.477239743 \approx 8.5 \text{ Watts per cm}^2$$

The exponential curve predicts that if the metal is 4.51 mm away, the ultrasound response will be about 8.5 Watts per  $\text{cm}^2$ .

This prediction could have an average error of about 8.239 Watts per  $\text{cm}^2$  too high or too low.

i)

No. We should not plug in 12.75 mm into the equation for  $x$ . 12.75 mm would be an extremely bad extrapolation and could result in a huge error. The standard deviation of the residuals would not apply as the prediction error.

5.

Exponential curves are only defined for bases that are positive and not equal to 1. Raising a positive number to an exponent will always give you a positive result. Hence you can plug in zero or negative numbers for  $x$ , but it is impossible for the exponential curve to give negative  $Y$  values or a  $Y$  value of zero.

---

## Section 7B Answers

1.

a)

The scatterplot shows a decay pattern. A logarithmic decay curve may work well. The scope of the  $X$  values are between 2 years and 25 years since 1980. This represents years 1982 – 2005.

b)

The logarithmic decay curve fits the data very well. The points look very close to the curve. There seems to be a strong logarithmic relationship between the variables.

c)

There were 24 ordered pairs in the data.

d)

$$r^2 = 0.8688$$

86.88% of the variability in the number of drunk driving fatal accidents can be explained by the logarithmic relationship with the year since 1980.

e)

Standard Deviation of the Residuals ( $Se$ ) = 1036.1736

The points in the scatterplot are 1036.2 drunk driving fatal accidents from the logarithmic curve on average.

If we use the logarithmic curve and a year since 1980 in the scope of the  $x$  values in order to predict the number of drunk driving fatal accidents, our prediction could have an average error of 1036.2 accidents too few or too many.



f)

Logarithmic Curve Equation:  $y = 23389.29331 + (-3753.61219) \ln(x)$

g)

The number in front of  $\ln(x)$  is negative. This indicates that the equation describes a logarithmic decay curve. This does agree with part (a).

h)

$$y = 23389.29331 + (-3753.61219) \ln(x)$$

$$y = 23389.29331 + (-3753.61219) \ln(12.5)$$

$$y = 23389.29331 + (-3753.61219) 2.525728644$$

$$y = 23389.29331 + (-9480.605828)$$

$$Y = 13908.687 \approx 13909$$

The logarithmic curve predicts that in year 12.5 we would have about 13,909 fatal car crashes due to drunk driving.

This prediction could have an average error of 1036.1736 (Se) too low or too high.

i)

$$y = 23389.29331 + (-3753.61219) \ln(x)$$

$$y = 23389.29331 + (-3753.61219) \ln(23.75)$$

$$y = 23389.29331 + (-3753.61219) 3.16758253$$

$$y = 23389.29331 + (-11889.8764)$$

$$Y = 11499.41691 \approx 11499$$

The logarithmic curve predicts that in year 23.75 we would have about 11,499 fatal car crashes due to drunk driving.

This prediction could have an average error of 1036.1736 (Se) too low or too high.

j)

I think extrapolation in this circumstance would be very bad. The logarithmic decay will quickly go below zero and start predicting negative number of drunk driving fatal crashes, which is impossible. We should not plug in 70 into the formula. We will not be able to use the Se for the prediction error in 70 is not in the scope of the x-values. It will likely have much greater error.

2.

a)

The scatterplot shows a logarithmic growth pattern. A logarithmic growth curve may work well. The scope of the x-values (bear ages) is between 8 months and 177 months.

b)

The logarithmic growth curve fits the data very well. The points look very close to the curve. There seems to be a strong logarithmic relationship between the variables.

c)

There are 54 ordered pairs in the data.





d)

$$r^2 = 0.7539$$

75.39% of the variability in bear lengths can be explained by the logarithmic relationship with the bears age.

e)

Standard Deviation of the Residuals (Se) = 5.3594 inches (bear length)

The points in the scatterplot are about 5.4 inches from the logarithmic curve on average.

If we use the logarithmic curve and a bears age in the scope of the x-values to predict the length of the bear, our bear length prediction could have an average error of 5.4 inches too low or too high.

f)

Logarithmic Curve Equation:  $y = 19.12622 + 11.37504 \ln(x)$

g)

The number in front of the  $\ln(x)$  is positive. This indicates that the equation is describing a logarithmic growth curve. This agrees with the graph in part (a).

h)

$$y = 19.12622 + 11.37504 \ln(x)$$

$$y = 19.12622 + 11.37504 \ln(48)$$

$$y = 19.12622 + 11.37504 (3.871201011)$$

$$y = 19.12622 + 44.03506635$$

$$y = 63.16128635 \approx 63.2 \text{ inches}$$

The logarithmic curve predicts that a 4 year old bear (48 months) would have a length of about 63.2 inches.

This prediction could have an average error of 5.4 inches (Se) too low or too high.

i)

$$y = 19.12622 + 11.37504 \ln(x)$$

$$y = 19.12622 + 11.37504 \ln(120)$$

$$y = 19.12622 + 11.37504 (4.787491743)$$

$$y = 19.12622 + 54.45791007$$

$$y = 73.58413007 \approx 73.6 \text{ inches}$$

The logarithmic curve predicts that a 10 year old bear (120 months) would have a length of about 73.6 inches.

This prediction could have an average error of 5.4 inches (Se) too low or too high.

j)

This logarithmic growth pattern seems to fit the bears well even for older bears. It is still not good to extrapolate. 50 years (600 months) is way out of the scope of the x values. We will not be able to use the Se for the prediction error. It may have more error.



3.

a)

The scatterplot shows a logarithmic growth pattern. A logarithmic growth curve may work well. The scope of the x-value temperatures is between about 14 degrees Kelvin and 852 degrees Kelvin.

b)

The logarithmic growth curve fits the data very well. The points look very close to the curve. There seems to be a strong logarithmic relationship between the variables.

c)

There are 236 ordered pairs in the data.

d)

$$r^2 = 0.9628$$

96.28% of the variability in copper expansion (cubic cm) can be explained by the logarithmic relationship with temperature (Kelvin).

r-squared also indicates a very strong relationship between the variables.

e)

Standard Deviation of the Residuals (Se) = 1.1149 cubic cm

The points in the scatterplot are about 1.1149 cubic cm from the logarithmic curve on average.

If we use the logarithmic curve and the temperature to predict the copper expansion, our prediction could have an average error of 1.1149 cubic cm too low or too high.

f)

Logarithmic Curve Equation:  $y = -16.99737 + 5.77086 \ln(x)$

g)

The number in front of the  $\ln(x)$  is positive. This indicates that the equation is describing a logarithmic growth curve. This agrees with the graph in part (a).

h)

$$y = -16.99737 + 5.77086 \ln(x)$$

$$y = -16.99737 + 5.77086 \ln(400)$$

$$y = -16.99737 + 5.77086 (5.991464547)$$

$$y = -16.99737 + 34.5759031$$

$$y = 17.5785331 \approx 17.58 \text{ cm}^3$$

The logarithmic curve predicts that at a temperature of 400 degrees Kelvin, copper would expand about 17.58 cubic centimeters.

This prediction could have an average error of about 1.11  $\text{cm}^3$  (Se) too low or too high.



3h

$$y = -16.99737 + 5.77086 \ln(x)$$

$$y = -16.99737 + 5.77086 \ln(600)$$

$$y = -16.99737 + 5.77086 (6.396929655)$$

$$y = -16.99737 + 36.91578547$$

$$y = 19.91841547 \approx 19.92 \text{ cm}^3$$

The logarithmic curve predicts that at a temperature of 600 degrees Kelvin, copper would expand about 19.92 cubic centimeters.

This prediction could have an average error of about 1.11 cm<sup>3</sup> (Se) too low or too high.

i)

The logarithmic curve fits the data very well for higher temperatures, yet without knowing about copper expansion, it is hard to determine if it will follow this pattern beyond the scope of the x values. I would not plug in 1000 degrees into this formula. It may have more error. The standard deviation of the residuals would not apply.

4.

a)

The scatterplot shows a logarithmic growth pattern. A logarithmic growth curve may work well. The scope of the x-values is between about 0.25 u and 0.63 u.

b)

The logarithmic growth curve fits the data moderately well. The points look somewhat close to the curve. There seems to be a moderate logarithmic relationship between the variables.

c)

There are 25 ordered pairs in the data set.

d)

$$r^2 = 0.9410$$

94.10% of the variability in the atomic energy released can be explained by the logarithmic relationship with the atomic defect.

The graph showed only a moderate relationship, but the r-squared indicates there is a very strong logarithmic relationship between the variables.

e)

Standard Deviation of the Residuals (Se) = 341.8581 MeV

The points in the scatterplot are 341.8581 MeV from the logarithmic curve on average.

If we use the logarithmic curve and the atomic defect in the scope to predict the atomic energy released, our prediction could have an average error of 341.8581 MeV too low or too high.



f)

Logarithmic Curve Equation:  $y = 2241.60751 + 4720.83078 \ln(x)$

g)

The number in front of the  $\ln(x)$  is positive. This indicates that the equation is describing a logarithmic growth curve. This agrees with the graph in part (a).

h)

$$y = 2241.60751 + 4720.83078 \ln(x)$$

$$y = 2241.60751 + 4720.83078 \ln(0.37)$$

$$y = 2241.60751 + 4720.83078 (-0.994252273)$$

$$y = 2241.60751 + (-4693.696735)$$

$$y = -2452.089225 \approx -2452.1 \text{ Mega Electron Volts (MeV)}$$

The logarithmic curve predicts that if the atomic defect was 0.37 atomic defect units (u), we would have about -2452.1 Mega Electron Volts (MeV) of energy released.

This prediction could have an average error of 341.8581 MeV too low or too high.

i)

$$y = 2241.60751 + 4720.83078 \ln(x)$$

$$y = 2241.60751 + 4720.83078 \ln(0.56)$$

$$y = 2241.60751 + 4720.83078 (-0.579818495)$$

$$y = 2241.60751 + (-2737.22499)$$

$$y = -495.6174892 \approx -495.6 \text{ Mega Electron Volts (MeV)}$$

The logarithmic curve predicts that if the atomic defect was 0.56 atomic defect units (u), we would have about -495.6 Mega Electron Volts (MeV) of energy released.

This prediction could have an average error of 341.8581 MeV too low or too high.

j)

0.9 is not in the scope of the x-values. We should not extrapolate. It may have more error in the prediction. The standard deviation of the residuals would not apply.

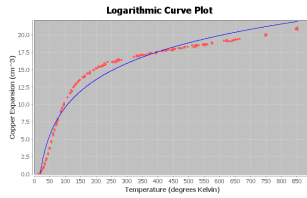
5.

To use a logarithmic curve, the explanatory variable (x) cannot be zero or negative. You can only plug in positive values into a logarithm. However the once you plug in a positive number for x, the  $\ln(x)$  value can be negative. So in logarithmic equations, the x must be positive, but the y can be zero or negative.

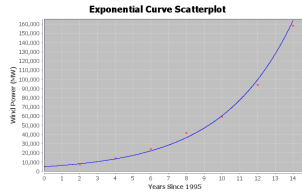


6.

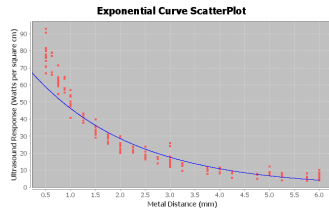
### Log Growth



### Exponential Growth



### Logarithmic and Exponential Decay (Same basic shape)



Logarithmic curves and exponential curves are inverses of each other. If you switch the x and y variables in an exponential curve and then solve for y, you would get a logarithmic curve. If you switch the x and y variables in a logarithmic curve and then solve for y, you would get an exponential curve.

---

### Section 7C Answers

1.

a)

The data shows an opening down parabolic shape indicating there is a maximum height of the rock. A quadratic curve may fit the data well.

b)

The quadratic curve fits the data very well. The points are close to the curve. There appears to be a strong quadratic relationship.

c)

$$y = 248.50000 + 34.88095 x + (-23.38095) x^2$$

The leading coefficient (number in front of  $x^2$ ) is negative. This tells us that the equation corresponds to a quadratic curve that opens down and has a maximum y value at the vertex.



d)

$$r^2 = 0.9870$$

98.70% of the variability in the rock height can be explained by the quadratic relationship with the time in seconds.

The r-squared also tells us there is a very strong quadratic relationship between the variables.

e)

Standard Deviation of Residuals = 8.7402 ft.

The average distance that the points are from the quadratic curve is 8.7402 feet.

If we use the quadratic curve and the time in seconds to predict the rock height, we could have an average error of 8.7402 feet too high or too low.

f)

$$y = c + b x + a x^2$$

$$y = 248.50000 + 34.88095 x + (-23.38095) x^2$$

Number of Seconds for max =

x coordinate of vertex =

$$-1b/2a = -1(34.88095) / 2(-23.38095)$$

$$= -34.88095 / -46.7619$$

$$= 0.745926705 \text{ seconds.}$$

The maximum height will happen in about 0.746 seconds.

Max Height = Y value when x = 0.745926705

$$y = 248.50000 + 34.88095 x + (-23.38095) x^2$$

$$y = 248.50000 + 34.88095 (0.745926705) + (-23.38095) (0.745926705)^2$$

$$y = 248.50000 + 34.88095 (0.745926705) + (-23.38095) (0.556406649)$$

$$y = 248.50000 + 26.0186321 + (-13.00931604)$$

$$y = 261.5093161 \approx 261.5 \text{ feet}$$

The maximum predicted height of the rock is 261.5 feet.

g)

0 seconds  $\leq$  Scope of X values  $\leq$  3.5 seconds

h)

$$y = 248.50000 + 34.88095 x + (-23.38095) x^2$$

$$y = 248.50000 + 34.88095 (1.8) + (-23.38095) (1.8)^2$$

$$y = 248.50000 + 34.88095 (1.8) + (-23.38095) (3.24)$$

$$y = 248.50000 + 62.78571 + (-75.754278)$$

$$y = 235.531432 \approx 235.5 \text{ feet}$$



At 1.8 seconds the rock will be predicted to have a height of 235.5 feet

This prediction could have an average error of 8.7 feet too high or too low.

i)

$$y = 248.50000 + 34.88095 x + (-23.38095) x^2$$

$$y = 248.50000 + 34.88095 (3.2) + (-23.38095) (3.2)^2$$

$$y = 248.50000 + 34.88095 (3.2) + (-23.38095) (10.24)$$

$$y = 248.50000 + 111.61904 + (-239.420928)$$

$$y = 120.698112 \approx 120.7 \text{ feet}$$

At 3.2 seconds the rock will be predicted to have a height of 120.7 feet

This prediction could have an average error of 8.7 feet too high or too low.

j)

The rock will not follow this pattern very long. It will hit the ground. We should not plug in 20 into the equation. The equation will start predicting a negative height, which is impossible.

2.

a)

The data shows an opening down parabolic shape indicating there is a predicted maximum solar energy. A quadratic curve may fit the data well.

b)

The quadratic curve fits the data very well. The points are close to the curve. There appears to be a strong quadratic relationship.

c)

$$y = 84.17045 + 425.60047 x + -33.22890 x^2$$

The leading coefficient (number in front of  $x^2$ ) is negative. This tells us that the equation corresponds to a quadratic curve that opens down and has a maximum  $y$  value at the vertex.

d)

$$r^2 = 0.8202$$

82.02% of the variability in solar energy can be explained by the quadratic relationship with the month.

The r-squared also tells us there is a very strong quadratic relationship between the variables.

e)

Standard Deviation of Residuals = 189.8436 kWh.

The average distance that the points are from the quadratic curve is about 189.8 kWh.

If we use the quadratic curve and the month to predict the amount of solar energy, we could have an average error of 189.8 kWh too high or too low.



f)

$$y = c + b x + a x^2$$

$$y = 84.17045 + 425.60047 x + -33.22890 x^2$$

Number of Months for max =

x coordinate of vertex =

$$-1b/2a = -1(425.60047) / 2(-33.22890)$$

$$= -425.60047 / -66.4578$$

$$= 6.404071004 \text{ month} \approx 6.4 \text{ months.}$$

The predicted maximum solar energy will happen at about month 6.4 (mid june).

Max Energy = Y value when x = 6.404071004

$$y = 84.17045 + 425.60047 x + -33.22890 x^2$$

$$y = 84.17045 + 425.60047 (6.404071004) + -33.22890 (6.404071004)^2$$

$$y = 84.17045 + 425.60047 (6.404071004) + -33.22890 (41.01212543)$$

$$y = 84.17045 + (2725.575629) + (-1362.787815)$$

$$y = 1446.958264 \approx 1447.0 \text{ kWh}$$

The maximum predicted solar energy is 1447.0 kWh.

g)

1 month  $\leq$  Scope of X values  $\leq$  12 months

h)

$$y = 84.17045 + 425.60047 x + -33.22890 x^2$$

$$y = 84.17045 + 425.60047 (3.5) + -33.22890 (3.5)^2$$

$$y = 84.17045 + 425.60047 (3.5) + -33.22890 (12.25)$$

$$y = 84.17045 + (1489.601645) + (-407.054025)$$

$$y = 1166.71807 \approx 1166.7 \text{ kWh}$$

The predicted solar energy in mid-march (month 3.5) is 1166.7 kWh.

This prediction could have an average error of 189.8 kWh too high or too low.

i)

$$y = 84.17045 + 425.60047 x + -33.22890 x^2$$

$$y = 84.17045 + 425.60047 (10.5) + -33.22890 (10.5)^2$$

$$y = 84.17045 + 425.60047 (10.5) + -33.22890 (110.25)$$

$$y = 84.17045 + (4468.804935) + (-3663.486225)$$

$$y = 889.48916 \approx 889.5 \text{ kWh}$$



This material is from *Introduction to Data Analysis*, 2<sup>nd</sup> edition by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a “CC-BY” [Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by/4.0/) – 3/17/2021



The predicted solar energy in mid-october (month 10.5) is 889.5 kWh.

This prediction could have an average error of 189.8 kWh too high or too low.

j)

The solar energy will not follow this pattern long. We should not plug in month 240. Not only is it an extrapolation, but the equation will also predict a negative solar energy. This would be impossible.

3.

a)

The data shows an opening up parabolic shape indicating there may be a minimum cost possible. A quadratic curve may fit the data.

b)

The quadratic curve fits the data moderately. The points are somewhat close to the curve. There appears to be a moderate quadratic relationship.

c)

$$y = 124202.34526 + (-4926.25365)x + 61.72503x^2$$

The leading coefficient (number in front of  $x^2$ ) is positive. This tells us that the equation corresponds to a quadratic curve that opens up and has a minimum y value at the vertex.

d)

$$r^2 = 0.6404$$

64.04% of the variability in transmission company monthly costs can be explained by the quadratic relationship with the number of hours worked.

The r-squared also tells us there is a strong quadratic relationship between the variables.

e)

Standard Deviation of Residuals = \$1634.0807

The average distance that the points are from the quadratic curve is about \$1634.08 .

If we use the quadratic curve and the average number of hours worked to predict the transmission company costs, we could have an average error of \$1634.08 too high or too low.

f)

$$y = c + b x + a x^2$$

$$y = 124202.34526 + (-4926.25365)x + 61.72503x^2$$

Number of Hours work for min cost =

x coordinate of vertex =

$$-1b/2a = -1(-4926.25365) / 2(61.72503)$$

$$= 4926.25365 / 123.45006$$



= 39.90482994 ≈ 39.9 hours.

The transmission company should have its employees work 39.9 hours per week to minimize costs.

Min Cost = Y value when  $x = 39.90482994$

$$y = 124202.34526 + (-4926.25365)x + 61.72503x^2$$

$$y = 124202.34526 + (-4926.25365)(39.90482994) + 61.72503(39.90482994)^2$$

$$y = 124202.34526 + (-4926.25365)(39.90482994) + 61.72503(1592.395452)$$

$$y = 124202.34526 + (-196581.3141) + 98290.65706$$

$$y = \$25911.68818 \approx \$25911.69$$

The predicted minimum costs of the company is \$25,911.69 if the employees work 39.9 hours per week.

g)

32 hours ≤ Scope of X values ≤ 50 hours

h)

$$y = 124202.34526 + (-4926.25365)x + 61.72503x^2$$

$$y = 124202.34526 + (-4926.25365)(44) + 61.72503(44)^2$$

$$y = 124202.34526 + (-4926.25365)(44) + 61.72503(1936)$$

$$y = 124202.34526 + (-216755.1606) + 119499.6581$$

$$y = \$26946.84276 \approx \$26,947$$

If the employees work 44 hours a week, the predicted monthly costs would be about \$26,947.

This prediction could have an average error of \$1634.08 too high or too low.

i)

$$y = 124202.34526 + (-4926.25365)x + 61.72503x^2$$

$$y = 124202.34526 + (-4926.25365)(35) + 61.72503(35)^2$$

$$y = 124202.34526 + (-4926.25365)(35) + 61.72503(1225)$$

$$y = 124202.34526 + (-172418.8778) + 75613.16175$$

$$y = \$27396.62926 \approx \$27,397$$

If the employees work 35 hours a week, the predicted monthly costs would be about \$27,397.

This prediction could have an average error of \$1634.08 too high or too low.

j)

It seems that if employees work too little or too much, the costs begin to rise. This trend will probably continue even out of the scope of the  $x$  values. I still would not plug in 120 into the equation. It is an excessive extrapolation and will probably result in a large error. The standard deviation of the residuals will not apply.



4.

a)

The data shows a opening down parabolic shape indicating there is a maximum length of the bears. A quadratic curve may fit the data well.

b)

The quadratic curve fits the data well. The points are close to the curve. There appears to be a strong quadratic relationship.

c)

$$y = 41.93861 + 0.54530x + (-0.00234)x^2$$

The leading coefficient (number in front of  $x^2$ ) is negative. This tells us that the equation corresponds to a quadratic curve that opens down and has a maximum y value at the vertex.

d)

$$r^2 = 0.6884$$

68.84% of the variability in bear length can be explained by the quadratic relationship with the bears age.

The r-squared also tells us there is a strong quadratic relationship between the variables.

e)

Standard Deviation of Residuals = 6.0897 inches.

The average distance that the points are from the quadratic curve is about 6.1 inches.

If we use the quadratic curve and the bears age to predict the bears length, we could have an average error of about 6.1 inches too high or too low.

f)

$$y = c + b x + a x^2$$

$$y = 41.93861 + 0.54530x + (-0.00234)x^2$$

Age of bear (months) for max length =

x coordinate of vertex =

$$-1b/2a = -1(0.54530) / 2(-0.00234)$$

$$= -0.54530 / -0.00468$$

$$= 116.517094 \approx 116.5 \text{ months old}$$

The maximum length will happen when a bear is about 116.5 months old.

Max Bear Length = Y value when  $x = 116.517094$

$$y = 41.93861 + 0.54530x + (-0.00234)x^2$$

$$y = 41.93861 + 0.54530 (116.517094) + (-0.00234) (116.517094)^2$$

$$y = 41.93861 + 0.54530 (116.517094) + (-0.00234) (13576.23319)$$



$$y = 41.93861 + 63.53677136 + (-31.76838567)$$

$$Y = 73.70699569 \approx 73.7 \text{ inches}$$

The maximum predicted length of the bears is 73.7 inches.

g)

8 months  $\leq$  Scope of X values  $\leq$  177 months

h)

$$y = 41.93861 + 0.54530x + (-0.00234)x^2$$

$$y = 41.93861 + 0.54530(48) + (-0.00234)(48)^2$$

$$y = 41.93861 + 0.54530(48) + (-0.00234)(2304)$$

$$y = 41.93861 + 26.1744 + (-5.39136)$$

$$Y = 62.72165 \approx 62.7 \text{ inches}$$

A bear 48 months old will be predicted length of 62.7 inches.

This prediction could have an average error of 6.0897 inches too high or too low.

i)

$$y = 41.93861 + 0.54530x + (-0.00234)x^2$$

$$y = 41.93861 + 0.54530(150) + (-0.00234)(150)^2$$

$$y = 41.93861 + 0.54530(150) + (-0.00234)(22500)$$

$$y = 41.93861 + 81.795 + (-52.65)$$

$$Y = 71.08361 \approx 71.1 \text{ inches}$$

We predict that the length of a bear 150 months old will be 71.1 inches.

This prediction could have an average error of 6.0897 inches too high or too low.

j)

We should not plug in 360 months. The equation will start predicting that the length of the bear will shrink, which is impossible. It will result in a huge error. The standard deviation of the residuals will not apply since 360 is out of the scope of the x values.

5.

The quadratic curves have a formula of the form  $y = c + b x + a x^2$ . The key is that the highest power of x is an x-squared in the formula. This tells us it is quadratic. If the number in front of the x-squared is positive, the curve will open up. If the number in front of the x-squared is negative, the curve will open down.



6.

If the number in front of the x-squared is positive, the curve will open up and will therefore have a minimum Y value at the vertex. If the number in front of the x-squared is negative, the curve will open down and will therefore have a maximum Y value at the vertex.

(Answers may vary)

There are many applications when maximum and minimum are useful. Maximizing profits and minimizing costs are vital in business. Minimizing cases of disease and maximizing the number of people vaccinated are vital in the medical field. Maximizing productivity and minimizing errors are vital in sports and business.

---

### Answers for Chapter 7 Review Sheet Problems

1. Logarithmic Growth (b)
2. Open Down Quadratic (e)
3. Decay (Exponential or Logarithmic) (c)
4. Exponential Growth (a)

5a.

$$-1b / 2a = -1(5.868) / 2(-0.163) = -5.868 / -0.326 = 18 \text{ breaks}$$

5b.

Max occurs when  $x = 18$

$$Y = 41.8 + 5.868(18) + (-0.163)(18)^2$$

$$Y = 41.8 + 5.868(18) + (-0.163)(324)$$

$$Y = 41.8 + 105.624 + (-52.812)$$

$$Y = 94.612$$

The maximum efficiency for the company is 94.6 if the employees are given 18 breaks each week.

6.

18.08% (r-squared converted to %)

7.

356.8787 grams (Se)

8.

356.8787 grams (Se)

9.

Predicted Baby Weight if mother is 33 years old: 1763.154711 (about 1763 grams)

10.

The log curve does not fit the data that well since the r-squared is low at 18.08% and the standard deviation is high at 356.8787 grams. I would classify it as a weak to moderate relationship.



11.

No. Just because there is a relationship does not imply causation.

12. (Answers may vary)

Confounding Variables: Health of the mother, Genetics of mother, genetics of the father, diet of mother during pregnancies, alcohol or drugs during pregnancy

13.

37.46% (r-squared for exponential curve)

14.

53.97% (r-squared for quadratic curve)

15.

The quadratic curve is the better fit and the stronger relationship since the r-squared percentage was higher.

16.

48.297 CHD deaths (standard deviation of residuals for exponential)

17.

47.5847 CHD deaths (standard deviation of residuals for quadratic)

18.

The points were closer to quadratic curve since the standard deviation of the residuals was smaller for the quadratic.

19.

48.297 CHD deaths (standard deviation of residuals for exponential)

20.

47.5847 CHD deaths (standard deviation of residuals for quadratic)

21.

The quadratic had less prediction error since the standard deviation of the residuals was smaller for the quadratic.

22.

The quadratic curve was the better fit since it had the higher r-square percentage and the lower standard deviation of the residuals.

---

