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.

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

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 type 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 as a category. 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.

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.

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, a list of peoples zip codes are numbers but an average zip code 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.

Practice Problems Section 1A

1. The following spreadsheet can be found on the website www.matt-teachout.org. Just click on the "statistics" tab and then "data sets". This data was taken from bears. Use the bear data to classify each column of data as categorical or quantitative. If the data is quantitative, what are the units? If the data 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 spreadsheet can be found on the website www.matt-teachout.org. Just click on the "statistics" tab and then "data sets". This data was taken from various cereals. Use the cereal data to classify each column of data as categorical or quantitative. If the data is quantitative, what are the units? If the data is categorical, indicate how many different options there are in that category.

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)	Vitamins (Percent of daily need per serving)	Consumer Report Magazine Rating	Serving Size (Cups per serving)	Weight (Ounces per serving)
Captain Crunch	Quaker	Child	Middle	120	12	2	0	35	1	220	12	25	18	0.75	1
Cocoa Puffs	General	Child	Middle	110	12	1	0	35	1	180	13	25	23	1	1
Trix	General	Child	Middle	110	13	1	0	25	1	140	12	25	28	1	1
Apple Jacks	Kelloggs	Child	Middle	110	11	0	1	30	2	125	14	25	33	1	1
Corn Chex	Ralston	Adult	Bottom	110	22	0	0	25	2	280	3	25	41	1	1
Corn Flakes	Kelloggs	Adult	Bottom	100	21	0	1	35	2	290	2	25	46	1	1
Nut & Honey	Kelloggs	Adult	Middle	120	15	1	0	40	2	190	9	25	30	0.67	1
Smacks	Kelloggs	Child	Middle	110	9	1	1	40	2	70	15	25	31	0.75	1
Multi-Grain	General	Adult	Bottom	100	15	1	2	90	2	220	6	25	40	1	1
Cracklin	Kelloggs	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
Nutri-Grain	Kelloggs	Adult	Top	140	21	2	3	130	3	220	7	25	41	0.67	1.33
Product-19	Kelloggs	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	28	1	1.5
Wheat Chex	Ralston	Adult	Bottom	100	17	1	3	115	3	230	3	25	50	0.67	1
Oatmeal	General	Adult	Top	130	13.5	2	1.5	120	3	170	10	25	30	0.5	1.25
Life	Quaker	Child	Middle	100	12	2	2	95	4	150	6	25	45	0.67	1
Maypo	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
Cheerios	General	Child	Bottom	110	17	2	2	105	6	290	1	25	51	1.25	1
Special K	Kelloggs	Adult	Bottom	110	16	0	1	55	6	230	3	25	53	1	1

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

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

Section 1B – Collecting Data

Vocabulary

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.

One of the most important goals in data science is to learn about the world around us (populations). 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.

A simple random sample is one where 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. 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 a 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.

I have had many students ask me if sample size is important. Isn't 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.

Method 5: Cluster Sample

A cluster sample is one where data is collected from groups of people in a population instead of one at a time. For example, a company that has 250 stores worldwide might have a computer randomly select ten stores and get data from those people that work at those ten stores. Notice this would be a random sample since every employee has a chance to be in the data. If their store was chosen, then they will be included in the sample. This is not a simple random sample however, since they are not choosing one at time. This example is sometimes called a "random cluster sample". While it is a good method for collecting data, it has less flexibility than a simple random sample. Think of it this way. In a simple random sample, any grouping is possible, but in this random cluster example, only groups of people that work at the same store can be chosen. It is still a random sample though, and would tend to be more representative of the population than non-random samples like convenience or voluntary response.

It is good to note that the goal of a cluster sample should be to choose the groups of people randomly. If we choose groups of people that are convenient to collect data from, our cluster sample will have more sampling bias and will not represent the population nearly as well.

Method 6: Stratified Sample

One of the most common studies done in statistics is to compare groups. We may compare data from 2016 to data from this year. We may compare people living in Canada to people living in Australia. To compare groups, you need to collect a stratified sample.

Some people in statistics explain a stratified sample as comparing two or more groups in one population. I like to think of it as comparing two or more populations. Whether you explain a stratified sample as comparing groups in one population or comparing populations, the key is that you are comparing.

For example, we may want to compare the percentage of adults in the U.S. with diabetes to the percentage of children in the U.S. with diabetes. Some statistics authors think of this as comparing adults and children in the one population of all people in the U.S. I like to think of it as comparing the population of U.S. adults to the population of U.S. children.

Another example may be to compare the mean average salary of people working in London, England to the mean average salary of people working in Toronto, Canada. Again, a stratified sample is needed because we are comparing.

To do a stratified sample, we often take a simple random sample from each group. I like to think of it as taking a simple random sample from each population you want to compare. In the previous example, we may collect a simple random sample of adults in the U.S. and another simple random sample of children in the U.S. We then can calculate the sample percentages that are diabetic from each sample and use statistical methods to compare them. For the salary example, we can collect a simple random sample of salaries for people working in London, and another simple random sample for people working in Toronto. The goal is then to use statistical methods to compare the mean average salaries.

It should be noted that when taking a stratified sample, we should use randomization. Again, if we just take a convenience sample from each group or voluntary response sample from each group, we will likely have a lot more bias and the data will not reflect the population (or populations) as well as we would like.

Many people confuse a cluster sample with a stratified sample because they both involve groups. The goal of a cluster is to get data on and analyze one population, not to compare. You are just collecting data from groups of people from that one population instead of one at a time. The goal of a stratified sample is to compare two or more populations so we need to collect data from each population.

Method 7: Systematic Sample

A systematic sample is one where we use a system to collect the sample. Usually it involves collecting data from every fifth person that comes in your store or every twentieth person on a list.

For example, let us suppose we want to collect a sample of students from our college. We could look at an alphabetical list of the names of all students that attend our college and then chose every 50th person on the list. Is this a random data set? Ask yourself this question. Does everyone on this list have a chance to be chosen? No. Only the 50th, 100th, 150th, 200th and so forth have a chance. People from 1-49 have no chance. People from 51-99 have no chance. Therefore, it is not a random sample. This may not be random, but we may make the argument that it is representative of the population. This method would have less bias than convenience or voluntary response samples. There is a way to incorporate randomization into the method. Many data scientists have a computer chose a random number between 1 and 50. Suppose it is 33. Then they collect data from the 33rd person on the alphabetical list. Now, from there use the system of choosing every 50th person. Therefore, they would choose the 33rd person, then the 83rd person, then the 133rd person and so on. Making the first choice random, makes the whole data set random, because everyone on the list now has a chance to be chosen.

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.
- If you cannot do a census, then use a random sample of some sort. It may be a simple random sample, random cluster, or a random systematic sample. The main thing is that if you are collecting a sample, randomization needs to be involved.
- Voluntary response samples and convenience samples tend to be very biased and should be avoided if possible.

Practice Problems Section 1B

Directions: For each of the following, identify the population of interest. Then identify the method used to collect the data (census, systematic, convenience, voluntary response, cluster, stratified, or simple random). Explain why you chose your answer and if the method will represent the population of interest or not?

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.

2. Rick works for a sports equipment manufacturing company. He wants to compare the opinion of his older employees to the new employees. To do this, he separates all the employees into two groups, employees that have been with company five or more years and those that have been with the company less than five years. He then chooses 12 of his most trusted older employees and 16 new employees that have proven themselves and ask what they think about changing the medical insurance coverage.

3. 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.

4. 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.

5. Rachael works at the Democrat recruiting center in her hometown. To determine what percent of people will vote for the Democratic candidate, she obtains a list of all residents in her town and decides to ask every 40th person on the list.

6. Laya is passionate about bringing an NFL football team to her city. She needs to take an opinion poll about how people in her city would feel about raising taxes in order to build a stadium for a professional football team. She randomly selects 75 streets in her city and asks every person living on those streets.

7. 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.

8. 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.

9. 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.

10. 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.

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

12. 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 60 classes and collect data from all of the students taking those classes.

13. A clothing store is designing a new line of athletic wear. They want to compare the percentage of teenagers that prefer the new line of athletic wear to the percentage of adults that prefer the new line of athletic wear. They take a random sample of teenagers and ask them about the new athletic wear. Then they take a random sample of adults and ask them about the new athletic wear.

14. Brian is collecting data for his statistics class project on the amount of time people spend on social media per day. He asks people in his college classes and at his church how many minutes they spend on social media per day.

15. A store that sells BBQ's in North Carolina wants to know what percentage of people own a "smoker BBQ". They ask every third person that enters the store if they own a smoker BBQ or not.

Section 1C – Bias

Vocabulary

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

Bias: When data does not reflect the population.

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 put 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 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.

   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 #2-10: 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 and if there is bias, what groups of people were not represented.

2. We are interested in calculating the percent of children in LA County that are up to date with their vaccines. To figure this out, a person put a survey up on the yahoo webpage asking the following question: "Is your child up to date with their vaccines?" The computer will keep track of the number of people that answer yes or no.

3. 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?"

4. We are interested in finding out how many people in the U.S. have had whooping cough this year. To figure this out, we called every major hospital in the United States and asked how many people at their hospital were diagnosed with whooping cough this year.

5. 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.

6. 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.

7. 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.

8. A college wants to collect data on their students to see how often they use the health office for mental health counseling. They randomly select 35 classes and collect data from all of the students taking those classes. They 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?"

9. 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.

10. 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.

Section 1D – Experimental Design

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: 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.

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

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

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

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 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 verses 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

Ruler Experiment Directions: *Divide class into groups of three or four. Each group will need a ruler and their cell phones. It is best to stand up during this activity. Procedure: Student A will hold the cell phone in their dominant hand and then hold their non-dominant hand straight out in front of them with their hand curved. The fingers should not be very close or very far away from the thumb. While student A is texting, student B holds the bottom of the ruler up inside of student A's non-dominate hand. Student B should hold the ruler from below student A's hand. The top of student A's hand should be about 5 inches on the ruler. Student B releases the ruler and student A tries to catch it. Student C records the number of inches on the top of the ruler before caught. Student C will take the catch length, subtract off the 5 inches, and then record the difference. If student A misses the ruler all together, then student C will just put "drop". Each student should attempt to catch the ruler while texting three times.*
Then repeat the process, but this time the students will attempt to catch the ruler with their non-dominant hand without a cell phone. Continue until all students have done the experiment three times without the cell phone and three times with the cell phones. Alternate the person releasing the ruler and the time before released. Collect the data for the "with phone" catches and drops in one column. In another column, collect the data for the "no phone" catches. When done, give the data to the instructor. Put the without cell phone/with cell phone data up on the board without names. The instructor or a student will collate the following results for the whole class: the mean average catch length with the cell, the mean average catch length without the cell, the total number of drops with the cell, the total number of drops without the cell.

Use your class data to answer the following questions as group. If you were absent on the day your class did the ruler experiment use the following data.

Ruler Experiment Data (Previous Class)

	With Phone	No Phone
Mean Average Catch (inches)	10.3 inches	8.2 inches
Number of Drops	41 drops	7 drops

1. What is the explanatory (treatment) variable? What was the response variable?

2. Why did we bother to have the person catch the yardstick without the phone?
   Wouldn't it of been quicker to just record the catching with the cell phone?

3. What were the two groups of people in the experiment? Were they alike?
   Why didn't we randomly assign the groups?

4. What are some of the confounding variables in this experiment?
   What are some steps that we took to control these variables?

5. Was this experiment blind, double blind, or neither? How do you know?

6. What did the class data show? Does texting cause slow reflexes?
   How do you think this experiment might apply to driving while texting?

(#7-11) Define the following terms and give an example of each.

7. Observational Study

8. Experiment

9. Explanatory Variable

10. Response Variable

11. Confounding Variables

12. Random Assignment

13. Placebo

14. Placebo Effect

15. Single Blind

16. Double Blind

(#17-21) Directions: Determine if each of the following studies are an observational study or an experiment. Explain why. Can the study prove cause and effect or just a relationship? Why? If the study is an experiment, list some confounding variables that need to be controlled.

17. 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 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.

18. Unemployment has become a very important topic in the United States and worldwide. We wish to understand how unemployment may be related to the tax rate. To shed light on this issue, we took a random sample of countries around the world and compared the average tax rate to the unemployment rate.

19. Tuberculosis (TB) is a disease that affects millions of people worldwide. TB is a contagious bacterial infection that affects the lungs. Doctors have long speculated that the percentage of people with Tuberculosis is higher in low income, crowded cities. A medical study was done to see if there is a relationship between low income, crowded cities and a high percentage of people with Tuberculosis. They took a random sample of cities and collected data about the size and the number of people. They then compared it to the number of cases of tuberculosis.

20. 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 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.

21. A study was done to determine if there is an association between obesity and diabetes. Obesity and diabetes data was taken from a random sample of adults.

Introduction to Categorical & Quantitative Data Analysis

Vocabulary

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.

Statistic: A number calculated from sample data in order to understand the characteristics of the data. For example, a sample mean average, a sample standard deviation, or a sample percentage.

Parameter: A population value, which is sometimes calculated from an unbiased census, but is often just a guess about what someone thinks the population value might be. For example, a population mean average or a population percentage.

Introduction

We learned that, in order to learn about the world around us, we need to collect and analyze data. Our goal is to understand populations. Sometimes we can collect data from everyone in the population (census) and sometimes we can only collect data from a small subgroup of the population (sample). Either way, once we have the data, we need to be able to analyze it. This chapter focuses on the basics of data analysis. If you remember, there are two types of data, quantitative (numerical measurements) and categorical (labels). We analyze quantitative data very differently than categorical data, so it is always vital to ask yourself a couple key questions.

- Was the data collected correctly, either an unbiased census or an unbiased large random sample?
- Is the data quantitative or categorical?
- Is their one data set or are we trying to analyze relationships between two data sets?

We will learn about rules for judging sample sizes in the next few chapters. This chapter focuses on being able to analyze the sample data or census data you have.

When analyzing data we rely on numbers calculated from the data that can help us understand the key features of the data set. If these numbers were calculated from a sample, they are called statistics. If these numbers are calculated from an unbiased census, they are called parameters. Most of the time, we only have sample data, so it is vital to understand and explain statistics.

Note on calculation: We live in the age of "big data". No one today calculates statistics by hand, especially for a data set of ten-thousand values. Even a sample of one-hundred can be overwhelming to calculate. Statisticians and data scientists rely on computers to calculate statistics. The focus should be on understanding the meaning and correct use of the statistic, not on calculating by hand with a calculator.

Section 1E – Categorical Data Analysis

Vocabulary

Percentage (%): An amount out of 100. For example if 72 out of every one-hundred employees opts to use a company's HMO insurance, we would say that 72% of the employees are using the HMO insurance.

Proportion: The decimal equivalent of a percentage. To calculate, divide the percentage by 100 and remove the percent symbol.

Proportion and Percentage Conversions

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".

To convert a decimal proportion into a percentage, we multiply the proportion by 100%. This moves the decimal point two places to the right. Do not forget to add the % symbol.

Example: Convert 0.047 into a percentage.

$0.047 \times 100\% = 4.7\%$

To convert a percentage into a decimal proportion, we divide by 100 and remove the percentage symbol. This moves the decimal two places to the left. Do not forget to remove the % symbol.

Example: Convert 52.9% into a decimal proportion.

$52.9\% = 52.9 \div 100 = 0.529$

Calculating Proportions and Percentages from Categorical Data

In order to calculate a decimal proportion from categorical data, you will need to find the amount (count, frequency) and divide by the total.

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

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.

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

Should we round the answer? Proportions and Percentages are usually rounded to the three significant figures. Proportions are usually rounded to the thousandths place (3rd 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 value you are rounding. 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.

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

How to 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.

Percentage = Decimal Proportion x 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.

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

Percentage = Decimal Proportion x 100% = 0.126 x 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.

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

Important Note

There are three ways to describe the proportion for 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.

Calculating a Frequency (Count) from a Percentage

How to calculate a count (frequency) from a percentage or proportion. Sometimes a percentage is given in a scientific report or in an article. For more advanced proportion analysis, the computer programs usually require the actual count (frequency). So it is important to be able to find the frequency from percentage information.

Start by converting the percentage into a proportion.

Proportion = Percentage ÷ 100 (and remove the percent symbol %).

Now multiply the proportion times the total to get the amount (frequency). This often called taking a "percentage of a total". It is important to round your answer to the ones place since is the number of people or objects that have a certain characteristic.

Count (Frequency) = Decimal Proportion × Total.

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.

Amount of people with hypertension = 0.32 x 603403 = 193088.96 ≈ 193,089

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

Bar Charts and Pie Charts

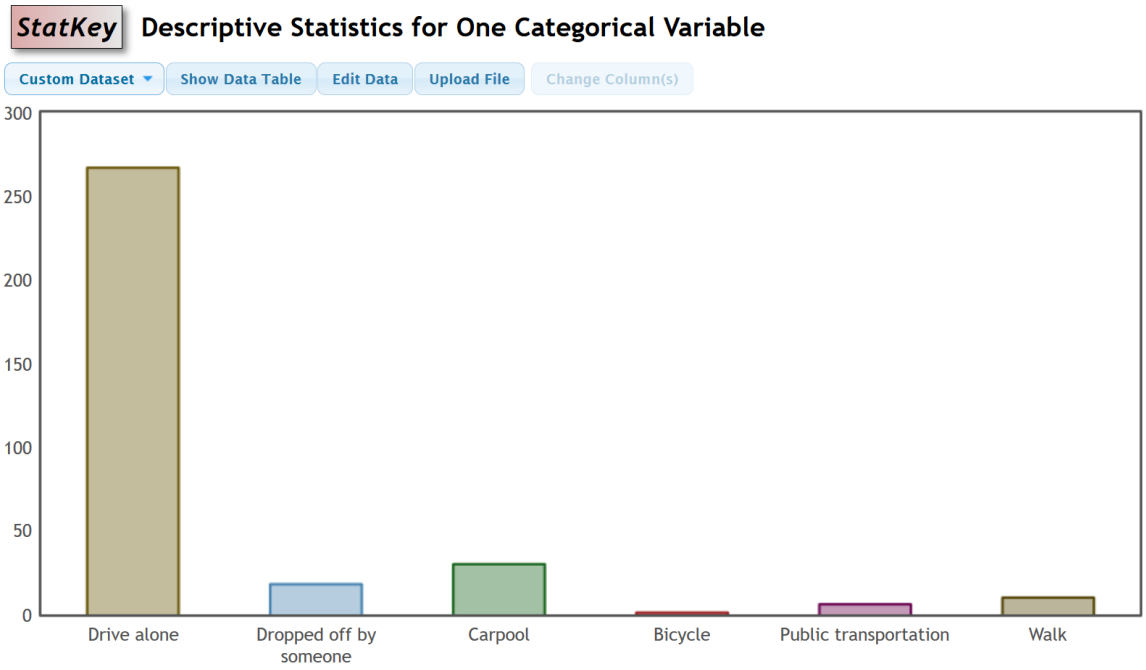
A quick way to count how many people or objects have a certain label is to create a Bar Chart or Pie Chart. There are many different statistics software that we could use to create these graphs. They are useful to show the characteristics of categorical data.

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 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". No click "OK".

For example, I copied and pasted the "transportation data" from the Math 140 Fall 2015 survey data at www.matt-teachout.org 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.



Summary Statistics

	Count	Proportion
Drive alone	267	0.804
Dropped off by someone	18	0.054
Carpool	30	0.09
Bicycle	1	0.003
Public transportation	6	0.018
Walk	10	0.03
Total	332	1.000

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, if my not be in raw form. Often it is just the counts (frequencies). In that case, when you go to the "edit data" button, 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, 267
Dropped off by someone, 18
Carpool, 30
Bicycle, 1

Public Transportation, 6
Walk, 10

Creating a Pie Chart with Raw Categorical Data and Statcato

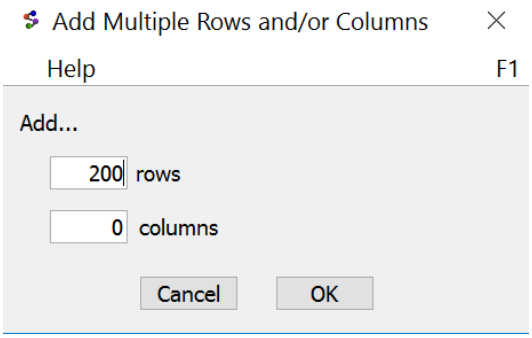
A pie chart is a very useful graph and can give the count (or frequency) for each variable and the percentages for each variable.

To create a pie chart with Statcato, open your excel spreadsheet. Copy and paste your column of categorical data from Excel into Statcato. Before pasting, be sure to click on the gray at the top of the column in Statcato, since titles must go in the gray. Now click on the graph menu at the top and then "pie chart". Click on "data values from a worksheet" and then under "data" put in the column. If your data is in the first column, you will click on "C1". If it is in the second column, you will click on "C2", and so on. Give the chart a title and click on "Show Legends" and "Show Values/Percentages for each Pie Sector". You can sort the graph by category or by frequency (counts). If you click on "sort by category", the pieces will be put in alphabetical order clockwise around the circle. If you click on "sort by frequency," then the chart will be organized from the smallest section to the largest section clockwise around the circle.

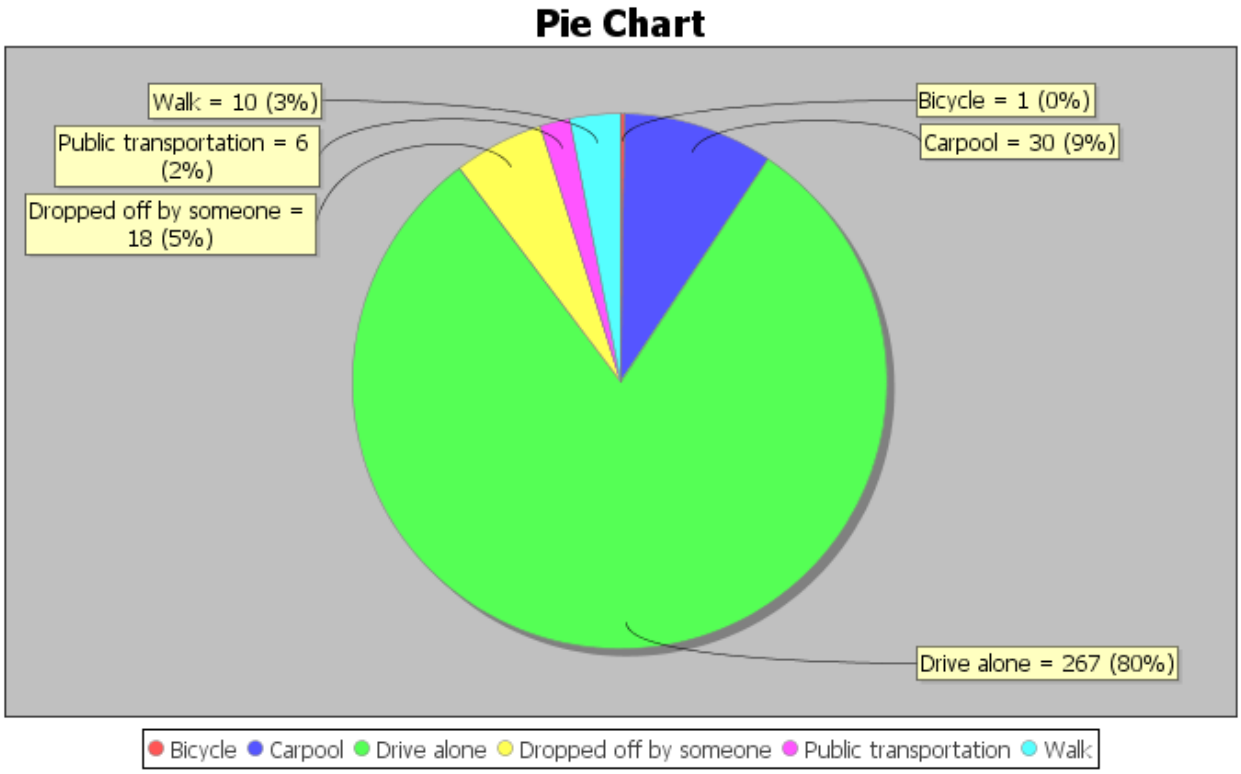
Graph Menu => Pie Chart => Data Values from a Worksheet => Sort by Categories or Frequencies, Show Legend, Show Values/Percentages

Let us use the same example, and open the transportation data from "Math 140 Survey Data" from fall 2015.

Important Reminder: If your data set is over 300 entries, you will need to add some rows to Statcato. The math 140-survey data had close to 350 students, so we will need to add some rows to the spreadsheet in Statcato before copy and pasting from Excel. (I added 200 more rows to Statcato before I tried to copy and paste.)



Once you have added enough rows in Statcato, copy and paste the column of data that says "Transportation" in Statcato. Do not forget to put the title in the gray cell at the top. Now go to the graph menu and make a pie chart. We will show two versions of the graph. One if you sort by categories and the other if you sort by frequencies. That way you can see the difference and which one you like better. The following graph was sorted by categories. Notice it gives the same counts as StatKey, though the proportions have been converted into percentages and rounded to two significant figures. You can copy and paste the graph into a Word or Pages document, by going to the "graph" button on the left side of the graph and click on "copy graph to clipboard".



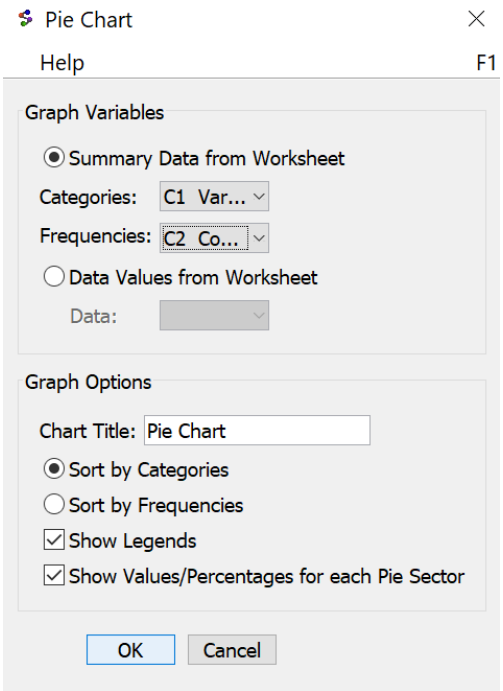
Notice at the touch of a button, the computer can tell us all of the counts (frequencies) and all of the percentages. We can answer all sorts of questions about how these students get to the college.

Creating a Pie Chart or Bar Chart with Summary Data and Statcato

Categorical data is often given in summarized form with the variables and the counts. Statcato cannot make bar charts from raw data, but it can make a bar chart from summary counts. Statcato can also make a pie chart from summarized data. Suppose we do not have access to the raw categorical transportation data. Suppose we only knew the variable labels and the counts. First type in the variables and counts (frequencies) into two columns of Statcato. We will use the transportation data again. Note that titles like "variable" or "count" must be typed in the gray where it says "Var".

	C1	C2
Var	Variable	Count
1	Drive Alone	267
2	Dropped off by someone	18
3	Carpool	30
4	Bicycle	1
5	Public Transportaion	6
6	Walk	10

Now go to the graph menu and then "pie chart". Click on "Summary Data from Worksheet". Give the columns for the categories and the columns for the frequencies.

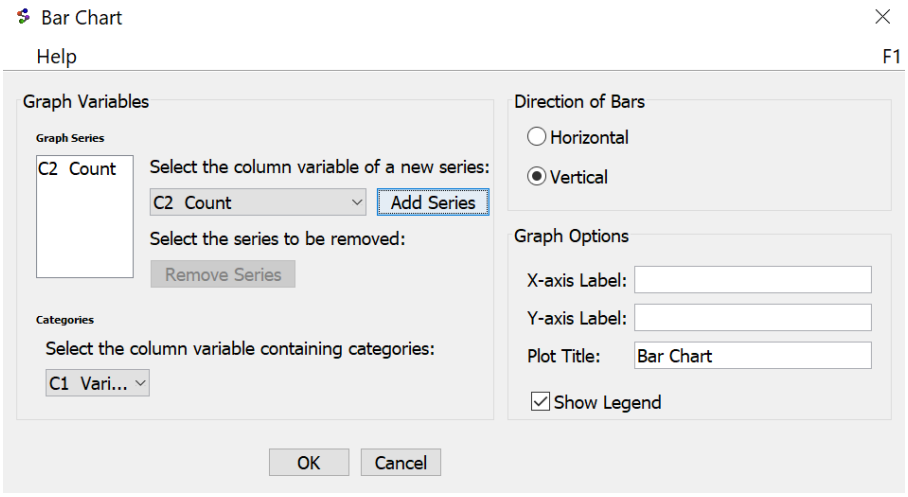


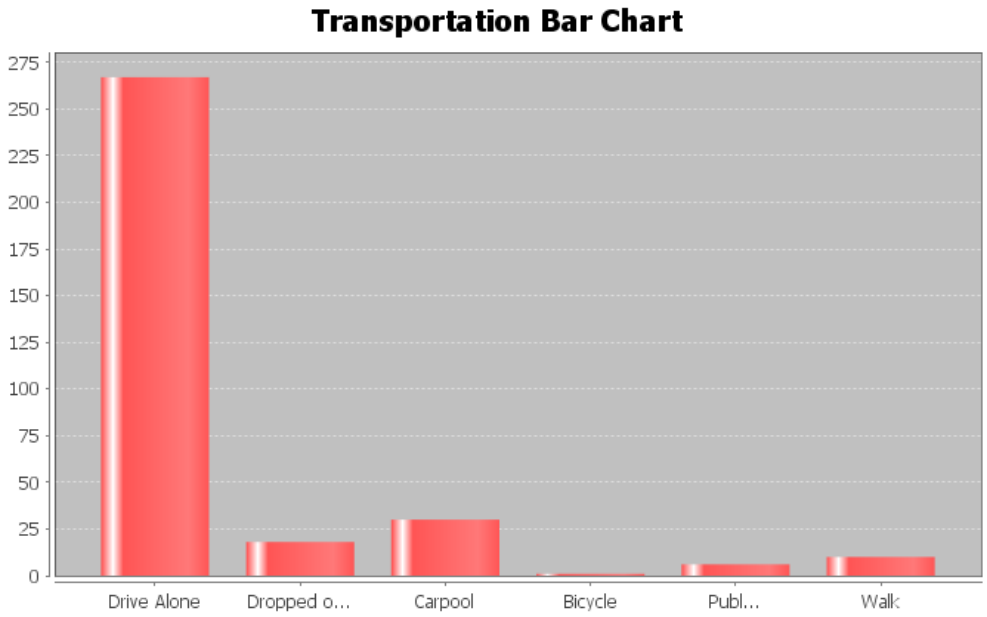
Notice the pie chart looks the same as the one we created with raw data.

We can also create a bar chart from summary categorical data. Again, type in the summary counts and variables into two columns of Statcato. Then go to the graph menu in Statcato and click on "Bar Chart". Statcato will want to know what column has your variable names and the column that has your counts.

Under "Select the column variable of a new series", pick the column with your counts (frequencies). Mine was in column 2. Now click "Add Series". Under "Select the column variable containing categories" select the column that has your variable names. Mine was in column 1. Type in a title and "show legend" and press OK. You can make the bars vertical or horizontal as well. I used vertical in this example.

	C1	C2
Var	Variable	Count
1	Drive Alone	267
2	Dropped off by someone	18
3	Carpool	30
4	Bicycle	1
5	Public Transportaion	6
6	Walk	10





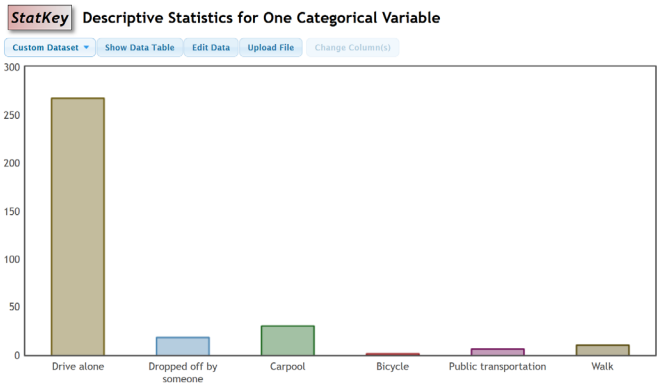
Comparing Percentages

Sometimes we want to 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". There are three different ways of calculating the percentage of increase. Any of these formulas give the same answer.

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

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

For example, let us look at the transportation bar chart found with StatKey. Suppose we want to compare the percentage of math 140 students that carpool verse the percentage that were dropped off. We can calculate the percent of increase from the counts, proportions or percentages. It is important to recognize which is the lower count (frequency) and which is the higher count. In this case, the number of students that carpool was higher than the number of students that were dropped off. The key question is was it significantly higher.



Summary Statistics

	Count	Proportion
Drive alone	267	0.804
Dropped off by someone	18	0.054
Carpool	30	0.09
Bicycle	1	0.003
Public transportation	6	0.018
Walk	10	0.03
Total	332	1.000

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

Percent of Increase = $\frac{(Higher\ Proportion - Lower\ Proportion)}{Lower\ Proportion} \times 100\% = \frac{(0.09-0.054)}{0.054} \times 100\% \approx 66.7\%$

Percent of Increase = $\frac{(Higher\ \% - Lower\ \%)}{Lower\ \%} \times 100\% = \frac{(9\%-5.4\%)}{5.4\%} \times 100\% \approx 66.7\%$

Notice this tells us that the proportion of students that carpool is 66.7% higher than the proportion that are dropped off. This difference seems statistically significant.

Note: In chapter 3 and chapter 4, we will learn how to use confidence intervals, test statistics, and P-values to determine significant differences. These are generally more accurate than the percent of increase calculation.

Statistical Significance verses Practical Significance

Sometimes when there is a statistically significant difference, it does not necessarily mean it is of practical use. In the last example, we saw that the number of students that carpool was a 66.7% higher than the number of students that are dropped off. Does this mean that college should make a special parking lot for all of the Math 140 students that carpool? Probably not. We are only talking about a difference of 12 total students a semester. College of the Canyons has thousands of students. So even though the percent of increase is significant, the data is not really of practical use in the sense that I would be careful of making huge decisions from the 66.7%.

Binomial Proportions with Statcato *(Optional Topic)*

Sometimes we want to know a percentage or proportion associated with a categorical event happening multiple times. One example of this is called a binomial proportion. A binomial proportion can be calculated from categorical data with only two outcomes (winning or losing, smoking or not, drinking alcohol or not). These are often referred to as "success" and "failure". The individuals must be independent of each other and the event (success) percentage *(p)* must be the same all the time. To calculate a binomial percentage, you will need a computer program and three bits of information, the number of events (number of successes), the event proportion *(p),* and the sample size *(n).*

Example

Categorical data often has a requirement of at least 10 success and at least 10 failures. Suppose we collect a random sample of 72 people and ask them whether they smoke cigarettes or not. Is 72 a large enough data set? Are we likely to get 10 or more people that smoke and 10 or more people that do not smoke? We can use Statcato to calculate this binomial percentage. According to the center for disease control, about 15.5% of adults in the U.S. smoke cigarettes.

Probability (percentage) of 10 or more people smoking =?

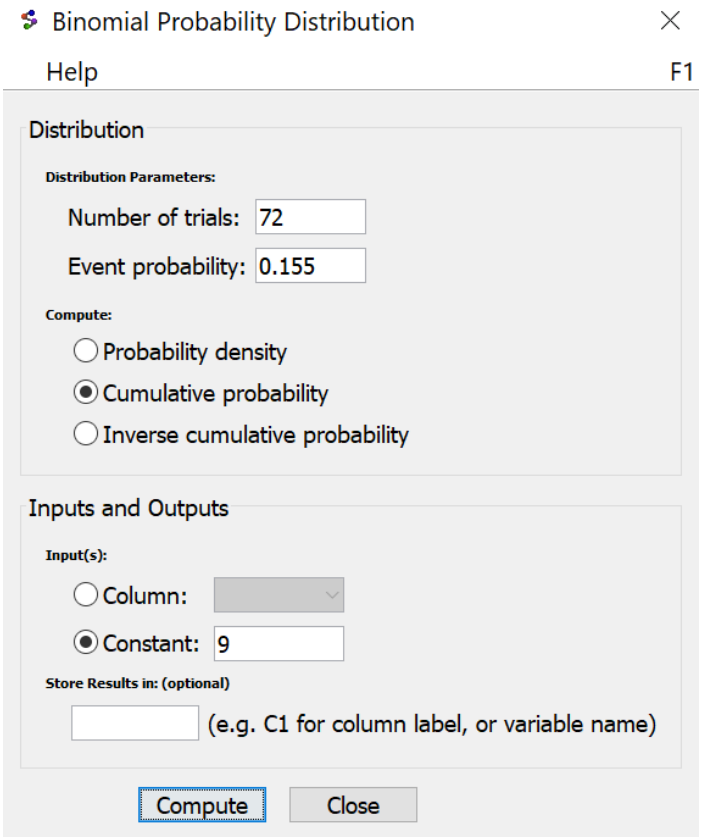
Number of Trials = Sample Size *(n)* = 72
Number of Events *(X)* = 10
Event Probability *(p)* = 0.155

Calculating binomial percentages can be challenging. Here is the formula that computer programs use.

Binomial Probability of X events: P(X) = $C(n,x)p^X(1-p)^{n-X}$

The problem with this formula is we have to calculate it for X = 10, X = 11, X = 12, … , X = 72 and then add all the proportions together. That is very difficult. It is best to let a computer program do the heavy lifting.

Open Statcato and click on "Calculate" menu. Then click on "probability distributions" and "binomial". Statcato is limited in the sense that it only calculates binomial percentage for either equal to (probability density) or less than or equal to (cumulative probability). So if we are calculating a greater than question, we must think about the opposite (less than or equal to). In this problem, we want to find 10 or more. The opposite of this would be 9 or less. Therefore, we will calculate the percentage for 9 or less, and then subtract the answer from 100%. This is sometimes called a "complement" proportion. In Statcato, put in the following. Under "Number of trials", put in the sample size 72. Under "constant" put in the number of events 9. Under "Event probability", put in 0.155. Now push the "Cumulative Probability" button and push "compute".



Binomial Distribution: n=72, p=0.155

Input: 9.0

Type: Cumulative probability

X	P(<=X)
9.0	0.304036

Notice the probability of getting 9 or less is 0.304 or 30.4%. This is the complement percentage to what we are looking for. So the probability of getting 10 or more people that smoke should be 100% − 30.4% = 69.6%. This may not be a high enough percentage to assure us that we will get at least 10 people that smoke. I would recommend collecting more data (increase the sample size).

Example

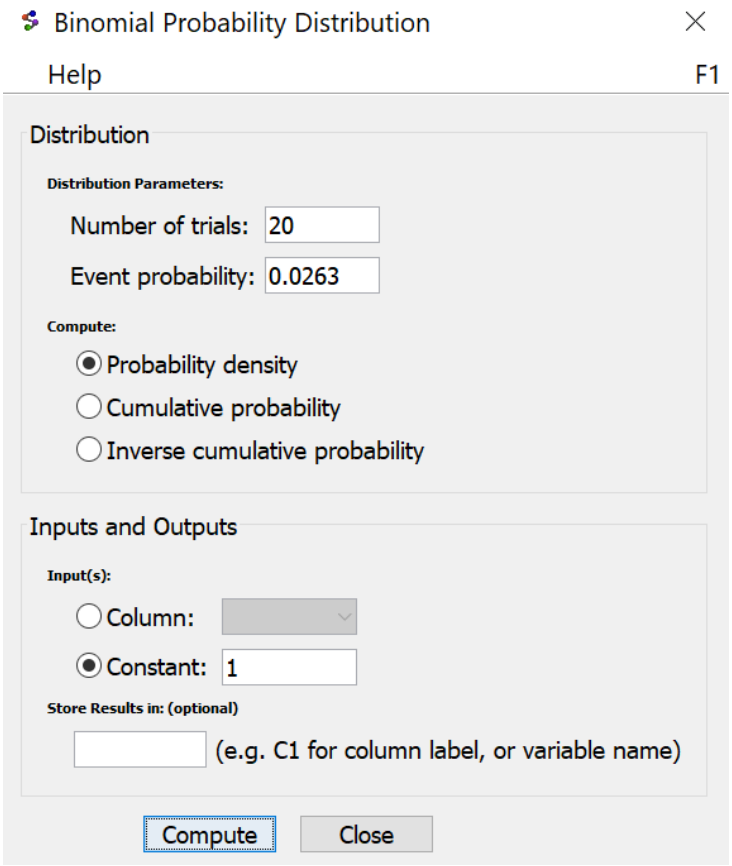
Suppose a person is playing a game of roulette that has a 1/38 or 2.63% chance of winning. The gambler plans to play the game 20 times. What is the probability that he or she wins just once?

Open Statcato and click on “Calculate” menu. Then click on “probability distributions” and “binomial”. Remember to calculate equal, you need to click on the “probability density button”.

Number of Trials = 20

Event Probability = 0.0263

Number of Events = 1 *(Put this in the “constant” box.)*



Binomial Distribution: n=20, p=0.0263

Input: 1.0

Type: Probability density

X	P(X)
1.0	0.317003

Notice the answer can be found under "P(X)". So the gambler has a 0.317 (31.7%) chance of winning the game once.

Problem Set Section 1E

1. Convert each of the following percentages into a proportion. Do not round the answers.

a) 75%
b) 2.75%
c) 0.664%
d) 0.082%
e) 39.7%
f) 8.6%
g) 0.189%
h) 0.0025%
i) 3.16%
j) 250%
k) 96.1%
l) 0.48%
m) 0.007%
n) 8.73%
o) 66.2%
p) 9%
q) 100%

2. Convert each of the following proportions into a percentage. Do not round the answers.

a) 0.057
b) 0.812
c) 0.0033
d) 0.0214
e) 0.0613
f) 0.451
g) 0.00045
h) 0.0779
i) 0.046
j) 0.3161
k) 0.0027
l) 0.051
m) 0.0058
n) 0.847
o) 1
p) 0.00022
q) 0.0204

(#3-10) Directions: Convert the given percentages into proportions. Then use the following formula to find the estimated amounts. Round your answers to the ones place.

$$Estimated\ Amount = Proportion \times Total$$

3. 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.

4. 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.

5. 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.

6. According to a news report by 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.

7. 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?

8. 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?

9. 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?

10. 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?

(#11-15) Directions: Use the following formulas to calculate the proportions, percentages and the percent of increase. Then answer the given questions.

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

Percentage = Decimal Proportion $\times$ *100%*

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

11. An article at 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 said they prefer athletic wear and 74 said they prefer traditional jeans. Calculate the decimal proportions and the percentages for both athletic wear and traditional jeans. Then calculate the percentage of increase. Does the percent of increase look statistically significant? Do you think it is practically significant? Explain.

12. The article at www.seattletimes.com also said that jean companies are creating more and more stretchy jeans to compete with the growing trend of women preferring athletic wear. Assume that a random sample of 197 total women were asked if they prefer stretchy jeans or athletic wear. Assume 103 said they prefer athletic wear and 94 said they prefer stretchy jeans. Calculate the decimal proportions and the percentages for both athletic wear and stretchy jeans. Then calculate the percentage of increase. Does the percent of increase look statistically significant? Do you think it is practically significant? Explain.

13. 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 were admitted to the medical/surgical ward and 49 were admitted to telemetry. Calculate the decimal proportions and the percentages for both medical/surgical and telemetry. Then calculate the percentage of increase. Does the percent of increase look statistically significant? Do you think it is practically significant? Explain.

14. A company found that of their 348 total employees, 96 employees have health insurance and 252 employees do not have health insurance. Calculate the decimal proportions and the percentages for both having health insurance and not having health insurance. Then calculate the percentage of increase. Does the percent of increase look statistically significant? Do you think it is practically significant? Explain.

15. 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 indicated significant improvement in their depression symptoms. Of the 61 people in the placebo group, 11 indicated significant improvement in their depression symptoms. Calculate the decimal proportions and the percentages for the medicine and placebo groups. Then calculate the percentage of increase. Does the percent of increase look statistically significant? Do you think it is practically significant? Explain.

(#16-20) Directions: Go to www.matt-teachout.org, click on the "statistics" tab and then "data sets". Open the indicated data set and copy the indicated column of categorical data. Go to www.lock5stat.com and click on StatKey. Under the "descriptive statistics and graphs" menu, click on "one categorical variable". Click on the "edit data" button and paste in the column. Check the box for "raw data" and "data has a header row" and push OK. Then answer the questions. Use the following formula for the percent of increase calculation.

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

16. Open the "Math 140 Survey Data Fall 2015" at www.matt-teachout.org. Look at the campus data. Use StatKey to make a bar chart, and a summary of the proportions and counts. What proportion of the students went to Valencia? What proportion of the students went to the Canyon Country campus? Calculate the percentage of increase. Does the percent of increase look statistically significant? Do you think it is practically significant? Explain.

17. Open the "Math 140 Survey Data Fall 2015" at www.matt-teachout.org. Look at the gender data. Use StatKey to make a bar chart, and a summary of the proportions and counts. What proportion of the students identified as female? What proportion of the students identified as male? Calculate the percentage of increase. Does the percent of increase look statistically significant? Do you think it is practically significant? Explain.

18. Open the "Math 140 Survey Data Fall 2015" at www.matt-teachout.org. Look at the hair color data. Use StatKey to make a bar chart, and a summary of the proportions and counts. Which hair color had the highest proportion? Which hair color had the lowest proportion?

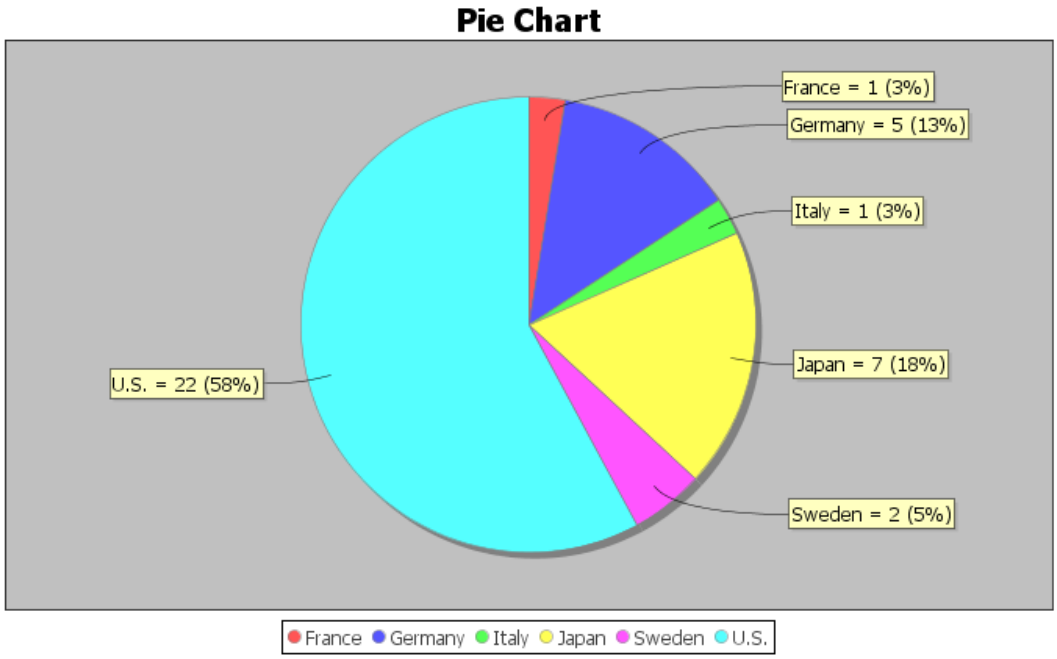
19. Open the "Math 140 Survey Data Fall 2015" at www.matt-teachout.org. Look at the political part data. Use StatKey to make a bar chart, and a summary of the proportions and counts. What proportion of the students identified as democratic? What proportion of the students identified as republican. Calculate the percentage of increase. Does the percent of increase look statistically significant? Do you think it is practically significant? Explain.

20. Open the "Math 140 Survey Data Fall 2015" at www.matt-teachout.org. Look at the "month of birthday" data. This data has numbers in it. Explain why this is categorical data and not quantitative. Use StatKey to make a bar chart, and a summary of the proportions and counts. Which month had the highest percentage? Which month had the lowest percentage?

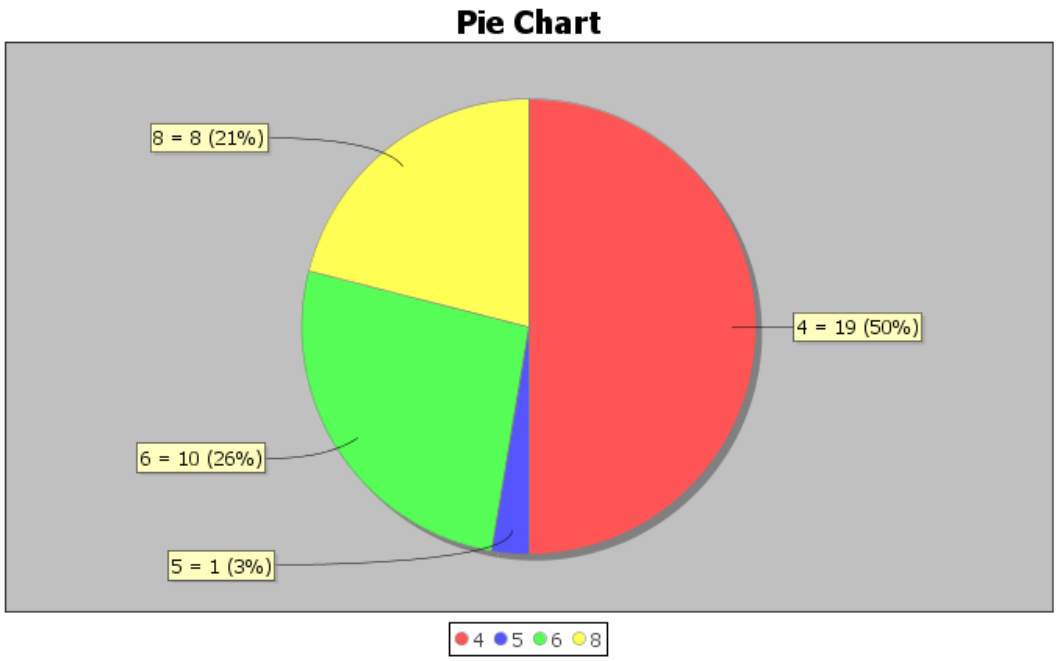
(#21-25) Use the following pie charts from Statcato to answer the following questions. Use the following formula for the percent of increase calculation.

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

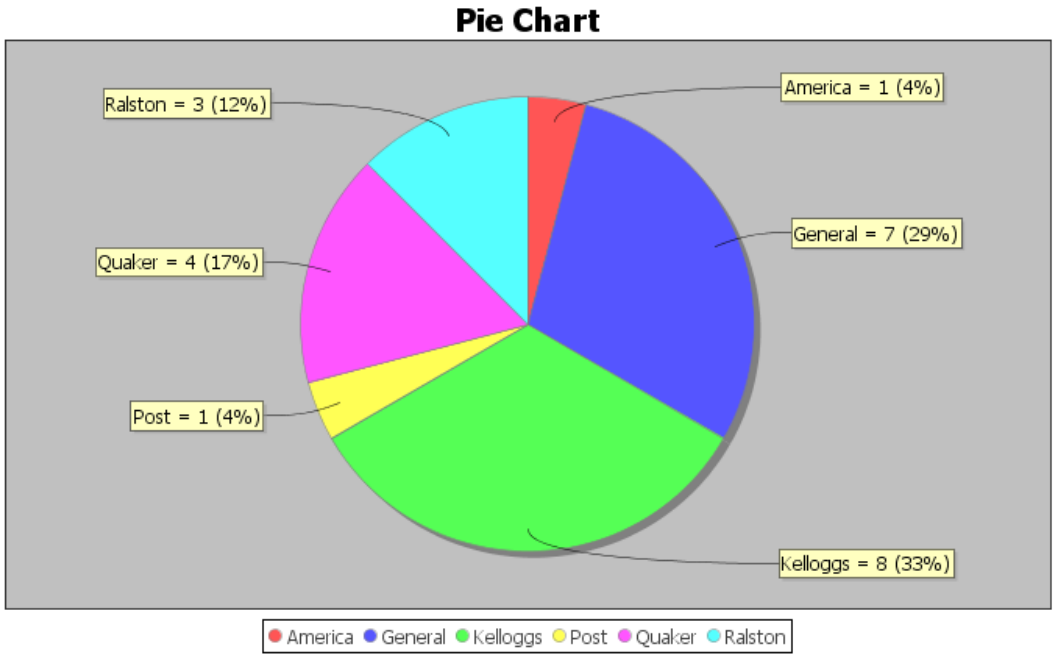
21. The following pie chart was created from the "car data" at www.matt-teachout.org. What percentage of the cars were made in France? How many of the cars were made in the U.S.? What proportion of the cars were made in Sweden? Calculate the percent of increase to compare Japan and Germany. Does the percent of increase look statistically significant? Do you think it is practically significant? Explain.



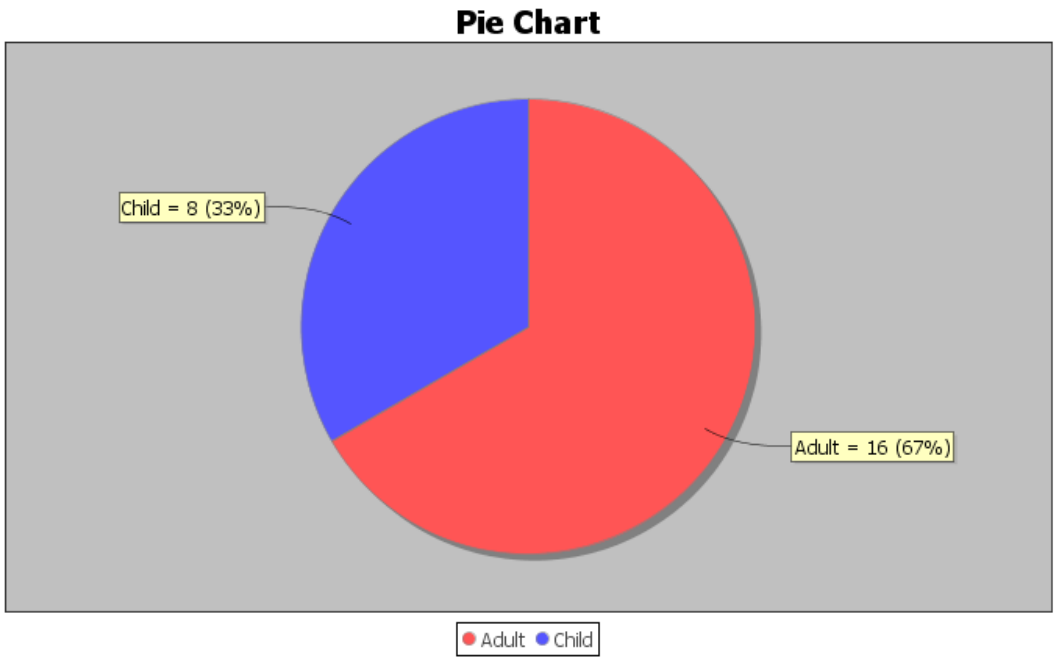
22. The following pie chart was created from the "car data" at www.matt-teachout.org. What percentage of the cars four cylinders? How many of the cars have eight cylinders? What proportion of the cars six cylinders? Calculate the percent of increase to compare four and eight cylinder cars. Does the percent of increase look statistically significant? Do you think it is practically significant? Explain.



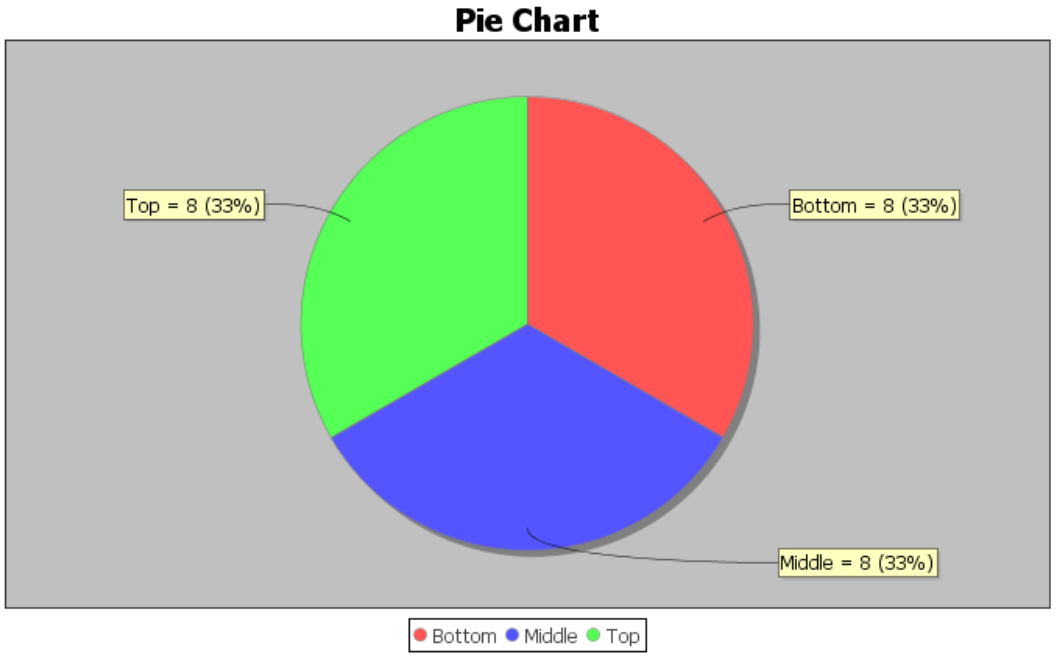
23. The following pie chart was created from the "cereal data" at www.matt-teachout.org. What percentage of the cereals did Quaker make? How many of the cereals did Ralston make? What proportion of the cereals did General make? Calculate the percent of increase to compare Kelloggs and Quaker. Does the percent of increase look statistically significant? Do you think it is practically significant? Explain.



24. The following pie chart was created from the "cereal data" at www.matt-teachout.org. What percentage of the cereals were targeted toward adults? What percentage of the cereals were targeted toward children? Calculate the percent of increase to compare adult cereals and children cereals. Does the percent of increase look statistically significant? Do you think it is practically significant? Explain.



25. The following pie chart was created from the "cereal data" at www.matt-teachout.org. What percentage of the cereals are displayed on the top shelf? How many of the cereals are displayed on the bottom shelf? What proportion of the cereals are displayed on the middle shelf? Calculate the percent of increase to compare the top and bottom shelf cereals. Does the percent of increase look statistically significant? Do you think it is practically significant? Explain.



Optional Binomial Probability Questions

(#26-30) Directions: Open Statcato on either one of the in-class or TLC computers. Go to the "calculate" menu, click on "probability distributions" and then "binomial". Enter the total under "number of trials" and the proportion under "event probability". Under "constant" put in the number of successes. Check "probability density" if you want to calculate equal to. Check "cumulative probability" to calculate less than or equal to. For greater than or equal to, subtract your less than or equal to (opposite) answer from one. Assume the questions meet the requirements for calculating a binomial probability.

26. To win at a dice game, the player must role two dice and get a 7 or 11 sum. This game has a 22.2% chance of winning. Suppose a player rolls the dice 18 times.

a) What is the probability that they win exactly once?
b) What is the probability that they win two times or less?
c) What is the probability that they do not win at all? (This means she wins zero times.)
d) What is the probability that they win three times or less?
e) What is the probability that they win four or more times? (Subtract your answer in (d) from one.)
f) What is the probability that they win four times or less?
g) What is the probability that they win five or more times? (Subtract your answer in (f) from one.)

27. A car company thinks that their minivan transmissions have a 12% defective rate. A total of 84 minivans were brought in to a service center this month.

a) What is the probability that exactly 11 of them need to have their transmission replaced?
b) What is the probability that exactly 8 of them need to have their transmission replaced?
c) What is the probability that 12 or less of the minivans will need their transmission replaced?
d) What is the probability that 13 or more of the minivans will need their transmission replaced? (Subtract your answer in (c) from one.)
e) What is the probability that 6 or less of the minivans will need their transmission replaced?
f) What is the probability that 7 or more of the minivans will need their transmission replaced? (Subtract your answer in (e) from one.)

28. Suppose we take a random sample of 57 total people and ask them if they smoke cigarettes or not. Assume that the population percentage for smoking in the U.S. is 15.5%.

a) What is the probability that we will get 9 or less people that smoke in the data set?

b) We need to have at least 10 people in the data set that smoke. What is the probability that we will get 10 or more people that smoke in the data set? (Subtract your answer in part (a) from 1.) Is this percentage high enough for us to be confident that 57 people is a large enough data set? Explain.

29. Suppose we take a random sample of 57 total people and ask them if they smoke cigarettes or not. Assume that the population percentage for non-smokers in the U.S. is 84.5%.

a) What is the probability that we will get 9 or less people that do not smoke in the data set?

b) We need to have at least 10 people in the data set that do not smoke. What is the probability that we will get 10 or more people that smoke in the data set? (Subtract your answer in part (a) from 1.)

30. Suppose a person is playing a game of roulette that has a 2.63% probability of winning. The person plays the game forty times.

a) What is the probability that they do not win at all? (The probability they win zero times.)
b) What is the probability that they win exactly one time?
c) What is the probability that they win two or less times?
d) What is the probability that they win three or more times? (Subtract your answer in (c) from one.)
e) What is the probability that they win one or less times?
f) What is the probability that they win two or more times? (Subtract your answer in (e) from one.)

Section 1F – Normal Quantitative Data Analysis

Vocabulary

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.

Normal Data: Data that is bell shaped, symmetric and unimodal. Also referred to as data that has a normal distribution.

Sample Size: Also called the total frequency.

Average: Also called the center of the data. A single number that represents a typical person or object in the data set.

Variability: Also called the spread. A measure of how spread out a data set is. A large spread tells us that the data is less consistent and the more difficult to predict. A small spread tells us that the data is more consistent and easier to predict.

Mean Average ($\bar{x}$): The balancing point for distances in a data set. The average for a data set that is normal.

Standard Deviation: The average or typical distance that points in a data set are from the mean. The measure of typical spread (typical variability) for a data set that is normal.

Maximum: The largest number in a data set.

Minimum: The smallest number in a data set.

Outliers: Unusual values in the data set.

Introduction

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?

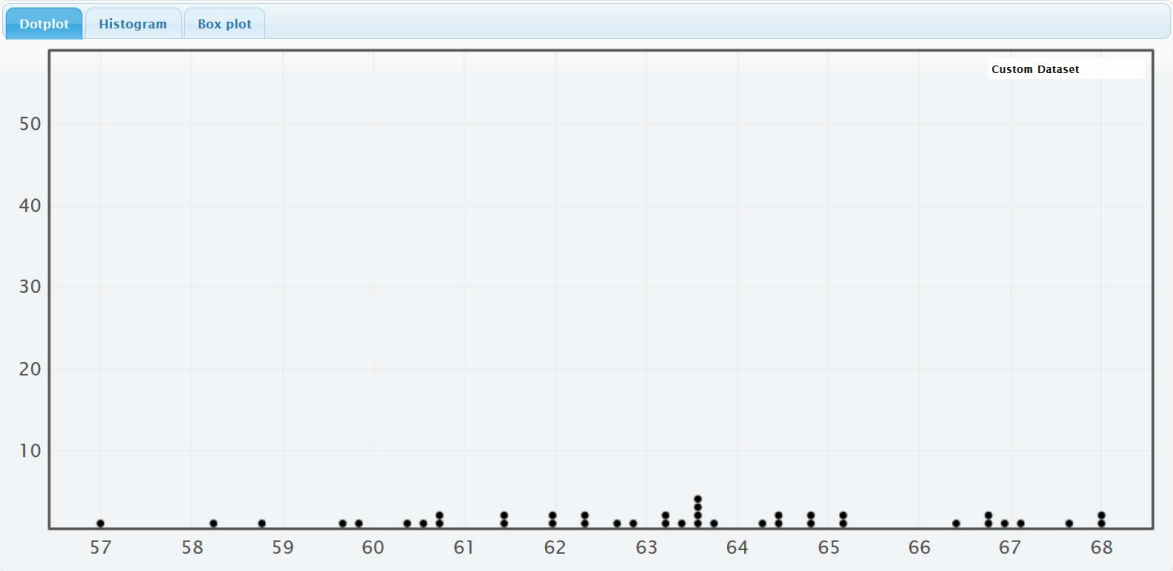
The goal in analyzing quantitative data is to find the average, spread and unusual values. In statistics, there are many types of averages, many types of spreads. Shape helps us determine which averages and spreads are most accurate for the data.

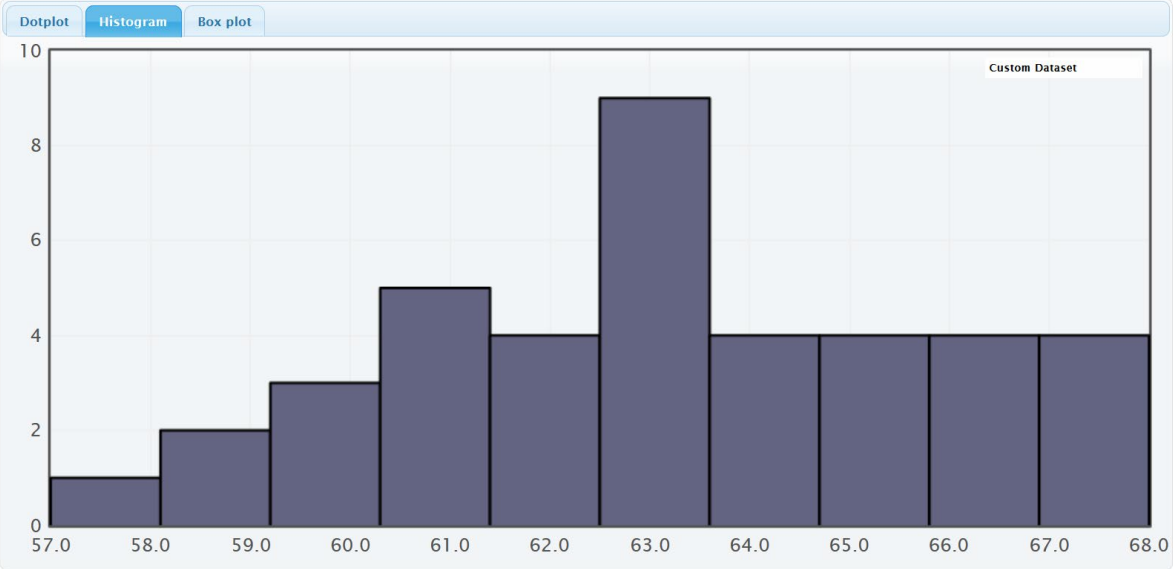
Quantitative Statistics and Graphs with StatKey

The most common quantitative statistics we like to look at are the mean, median, standard deviation, first quartile, third quartile, interquartile range, max, min, and range. 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 are also called "bins" or "buckets". Another graph we like to look at is the boxplot. A boxplot is a graph of the first quartile, median, and third quartile as well as potential outliers.

All of these graphs and statistics can be made with StatKey. Let us look at an example. Go to www.matt-teachout.org and click on the "statistics" tab and then the "data sets" tab. Look for the "Health Data" excel file. Open the data set and copy the women's heights data. Notice the data is quantitative. It measures the height in inches of the women and it seems reasonable to look for an average height of these women.

Go to 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. Copy and paste the women’s height data into StatKey. Uncheck 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. Check the box that says, “Data has a header row”. This means the data set has a title. Now push “OK”. Notice StatKey gives you the sample statistics, a dot plot, a histogram and a boxplot.

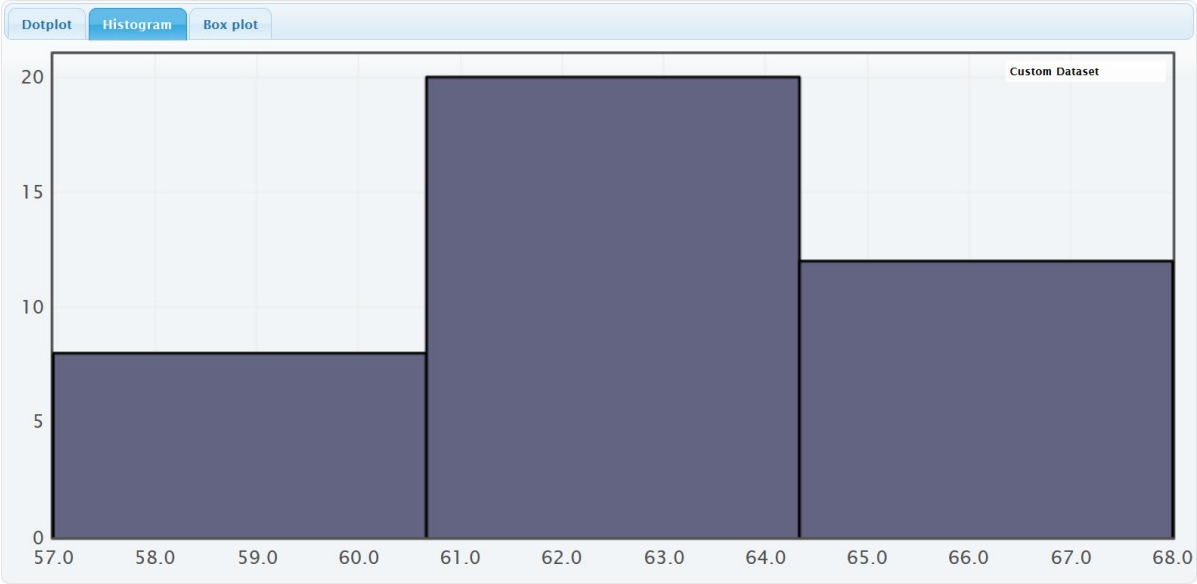




On the right of this histogram, you will see a slider that can adjust the number of “buckets” or bins. The smaller the data set the less bins you should have. This data set only has 40 numbers, so we want only a few bars. If we slide it to 3 buckets, we get the following.

Histogram Controls Set Limits

Number of buckets: 3



This data has a very special shape. It is called bell shaped or normal. Normally distributed data has the highest bar in the middle and about equal number of bars decreasing from the middle. It looks like bell. We see that this data set is relatively normal (bell shaped) or "normally distributed". StatKey has also given us summary statistics. Which statistics are most accurate for normal data?

Summary Statistics

Statistic	Value
Sample Size	40
Mean	63.195
Standard Deviation	2.741
Minimum	57
Q_1	61.350
Median	63.350
Q_3	64.900
Maximum	68

Mean and Standard Deviation

Important Note about Shape: The mean and standard deviation should only be used if the data set is normal. The mean and standard deviations are not accurate if the data does not have a normal shape.

Mean ($\bar{x}$): The mean is a type of average used for data that is normally distributed. The mean balances the distances between all the numbers in the data set and the mean. Think of it this way. If you tool all the numbers in the data set below the mean, measured their distances from the mean, then added up those distances. That total distance for numbers below the mean would be equal to the total distance for numbers above the mean. The mean is calculated by adding up all the numbers in a data set ($\sum x$) and then dividing by how many numbers are in the data set (sample size "n").

$$\bar{x} = \frac{\sum x}{n}$$

Standard Deviation (S): We said that the mean balances the distances in a data set. The standard deviation calculates the average distance numbers are from the mean. It is the most accurate measure of typical spread for data sets that are normally distributed. To calculate the standard deviation, computer programs take every single number in the data set and subtract the mean. Since those differences can be negative sometimes, they computer squares all the differences and then adds up the squares. This is a famous calculation called "sum of squares". Since we want the average distance, we divide by n – 1 (degrees of freedom) and take the square root at the end to undo all the squares. Never calculate this by hand. It is a long calculation that should be left to a computer program.

$$S = \sqrt{\frac{\sum(x - \bar{x})^2}{n - 1}}$$

Why do we study spread? Spread is a measure of how much variability is in the data set. Think of it this way. Suppose we were looking at exam scores in a history class that are normally distributed. If a data set were very spread out, then the standard deviation would be quite large. This would mean that the scores had a lot of variability. We had A's, B's, C's, D's, and F's. The exam scores are not consistent, and the history teacher will have a hard time predicting how her class will do. If the data set has a small spread, then the standard deviation would be quite small. The exam scores are very consistent. Maybe everyone in the class got an A or a high B. It is easier to predict how the class will do.

Statistics for Normal Data

Quantitative Variable and Units

Sample Size (n)

Maximum Value

Minimum Value

Average: Mean ($\bar{x}$)

Spread: Standard Deviation (s)

Typical Values: One standard deviation from the mean. Here is a formula that is sometimes used.

$$\bar{x} - s \leq \text{typical values} \leq \bar{x} + s$$

Outliers (unusual values): More than two standard deviations from the mean. Here are formulas that are sometimes used.

Unusually Low Values (Low outliers) $\leq \bar{x} - 2s$

Unusually High Values (High outliers) $\geq \bar{x} + 2s$

Women's Height Example

Quantitative Variable and Units: Women's heights in inches

Sample Size (n): There were 40 women in the data set.

Maximum Value: The tallest woman in the data set was 68 inches.

Minimum Value: The shortest woman in the data set was 57 inches.

Average: Mean $(\bar{x})$. The average height of the women in the data was 63.195 inches.

Spread: Standard Deviation (s). The typical spread for this data was 2.741 inches. Typical women in the data were 2.741 inches from the mean.

Typical Values: Add and subtract the mean and standard deviation. Typical women in the data set have a height between 60.454 inches and 65.936 inches. We will see later that these values are the cutoffs for the middle 68% for normal data.

$\bar{x} - s \leq$ typical values $\leq \bar{x} + s$

$63.195 - 2.741 \leq$ typical values $\leq 63.195 + 2.741$

$60.454 \leq$ typical values ≤ 65.936

Outliers (unusual values): Add and subtract the mean and two standard deviations. Unusually tall women are 68.677 inches or higher. There are no unusually tall women in this data set. Unusually short women are 57.713 inches or lower. This means that the minimum value of 57 inches was unusually low. We will see later that these values are the cutoffs for the top and bottom 2.5% for normal data.

Unusually Low Values (Low outliers) $\leq \bar{x} - 2s = 63.195 - (2 \times 2.741) = 57.713$ inches

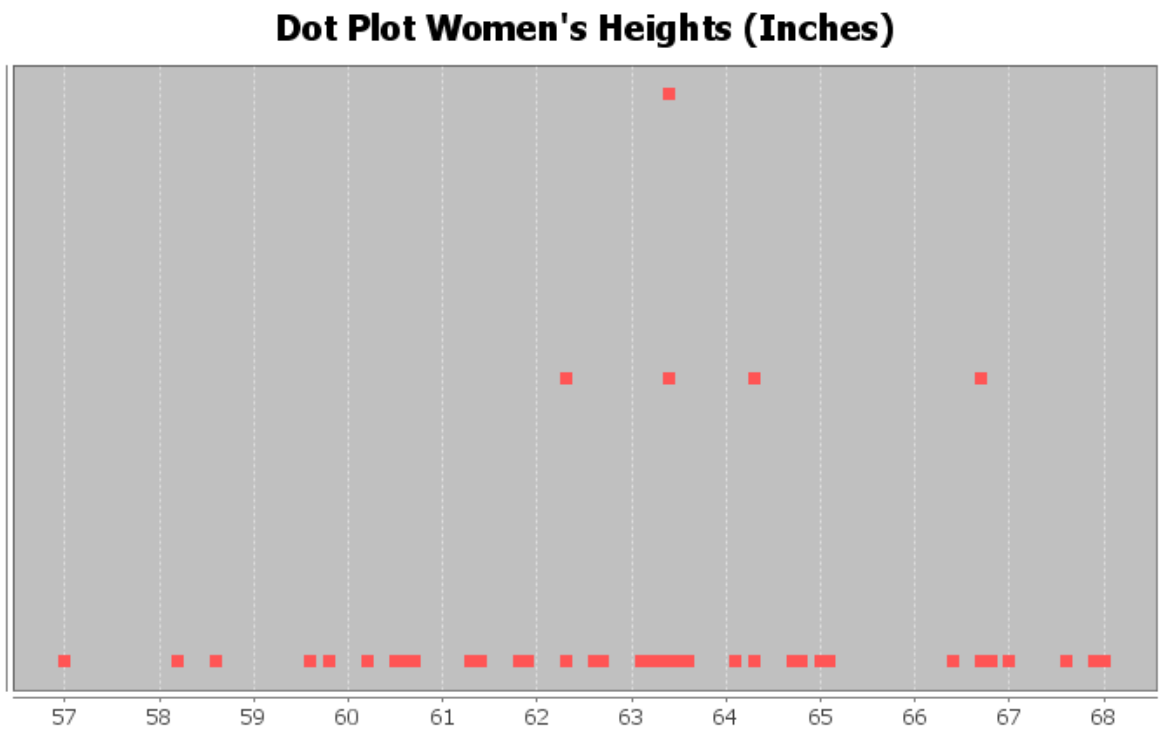
Unusually High Values (High outliers) $\geq \bar{x} + 2s = 63.195 + (2 \times 2.741) = 68.677$ inches

Quantitative Statistics and Graphs with Statcato

You can also make dot plots, histograms and sample statistics with Statcato. Copy and paste women's heights into a column of Statcato. The data set is only 40 values, so you will not need to add rows to Statcato. To make a dot plot, go to the graph menu and click on dot plot. Then click on the column of data you want to use. Then push ok.

Making a dot plot in Statcato: *Graph => Dot plot => Pick a column => OK*

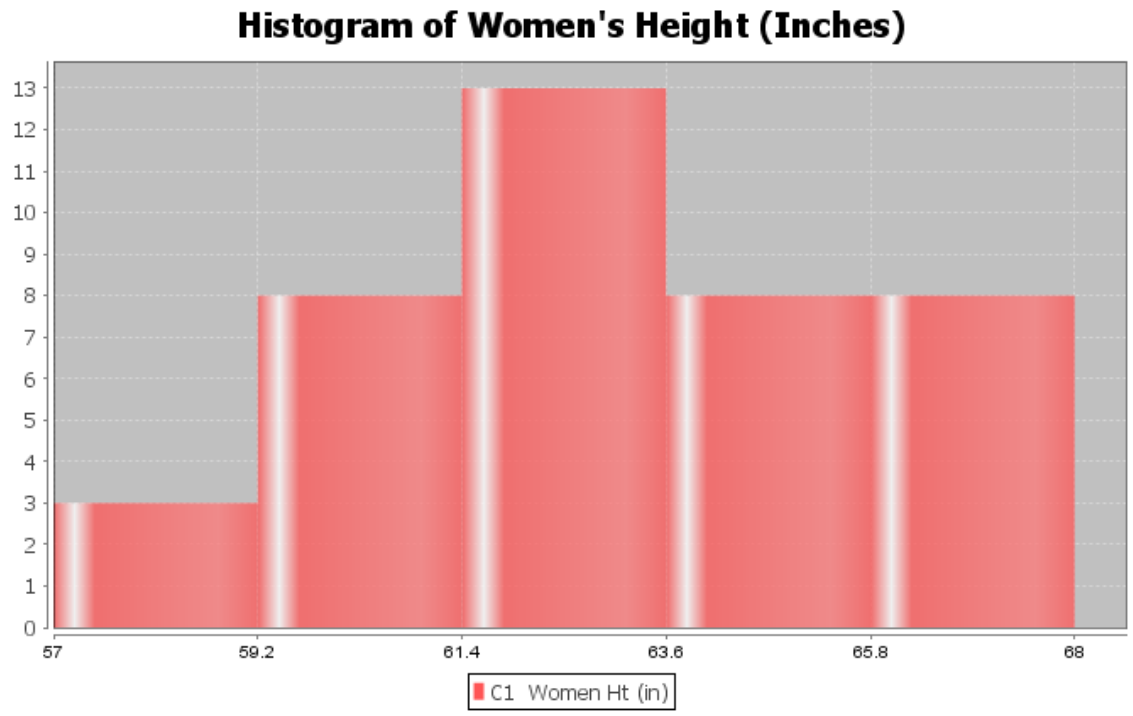
Here is the dot plot for the 40 women's heights.



To make a histogram in Statcato, go to the graph menu, and then click on histogram. Chose a column of data and how many bars (bins) you want. Then chose ok.

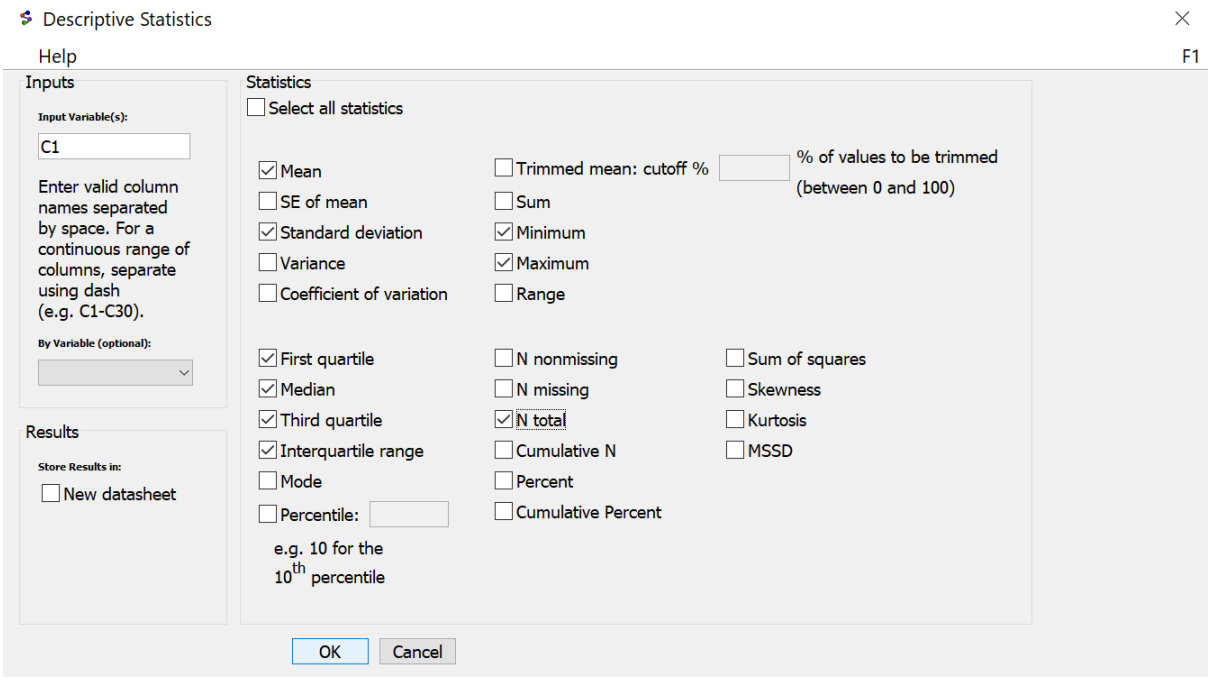
Making a histogram in Statcato: *Graph => Histogram => Pick a column => Chose number of bins => OK*

Note about bins: *If you chose too many bars then the histogram starts to look very crazy and you will have a hard time seeing the shape. Remember the goal is to break the dots up into groups. For example, in this health data there are only 40 women. I would not want 40 bins since that would give me about one bar per dot. If it were a small data set like the health data, I would do about three bins. Remember, the more bins you have, the more difficult it is to see the shape. This graph has five bins.*



Notice again that 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). I prefer to call it bell shaped or normal.

We can also calculate all of the sample statistics with Statcato. Go to the “Statistics” menu. Then click “Basic Statistics” and “Descriptive Statistics”. I had pasted the data into column 1, so type in “C1” under “input variable”. Check the boxes for statistics that you want and push “OK”.



Z-scores

In normal data, we often want to find out how many standard deviations a number (X-value) is from the mean. This is called a "Z-score". Here is a common formula. In later chapters, we will see that we can also use the Z-score as a test statistic to measure significance.

$$Z = \frac{(X\ value - Mean)}{Standard\ Deviation}$$

Example: In the last example, we saw that the women's height data was normally distributed with a mean of 63.195 inches and a standard deviation of 2.741 inches. Suppose a woman is 72 inches tall. What would be the Z-score for her height? Is she unusually tall?

It is important when calculating a Z-score that you subtract the X value and the mean first. Then divide by the standard deviation. Most people in statistics round Z-scores to the hundredths place (two numbers to the right of the decimal).

$$Z = \frac{(X\ value - Mean)}{Standard\ Deviation} = \frac{(72 - 63.195)}{2.741} = +3.21233 \approx +3.21$$

If the X-value is below the mean, the Z-score will be negative. If the X-value is above the mean, the Z-score will be positive. This Z-score was positive. So the woman that is 72 inches tall is 3.21 standard deviations above the mean. Is this unusual?

Remember the formula above for finding the cutoff for unusual values for normal data. Notice it is two standard deviations above and below the mean. Two standard deviations above the mean would be a Z-score of +2. Two standard deviations below the mean would be a Z-score of −2. So a common way to judge if a number is unusual (outlier) for normal data is to look at the Z-score.

Unusual High Values for Normal Data: $Z \geq +2$

Unusual Low Values for Normal Data: $Z \leq -2$

Hence, since the woman's Z-score was greater than or equal to +2, she is unusually tall compared to the women in the data set.

Example: The women's height data was normally distributed with a mean of 63.195 inches and a standard deviation of 2.741 inches. One woman in the data set was 57 inches tall and we said was unusually short. If you recall, her height was below the unusual low cutoff of 57.713 inches. What would be the Z-score for her height?

$$Z = \frac{(X\ value - Mean)}{Standard\ Deviation} = \frac{(57 - 63.195)}{2.741} \approx -2.26$$

Since the X-value is below the mean, the Z-score will be negative. So the woman that is 57 inches tall is 2.26 standard deviations below the mean. Remember if the Z-score is less than −2, it is unusually low. This confirms what we already knew.

Typical Z-scores: Remember that typical values are within one standard deviation from the mean. This would mean that typical Z-scores are between −1 and +1.

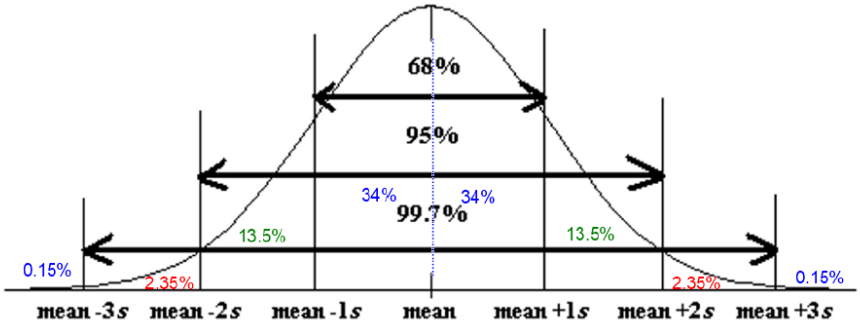
$-1 \leq$ Typical Z-scores $\leq +1$

A woman with a height of 61 inches would have a Z-score of −0.80. Notice that this Z-score is between −1 and +1 on the number line. Therefore, 61 inches is a typical height for women in this data set.

Note: Not all values are typical or unusual. A person that is 1.5 standard deviations from the mean would be neither typical (Z-score not between −1 and +1) nor unusual (Z-score not greater than +2 or less than −2).

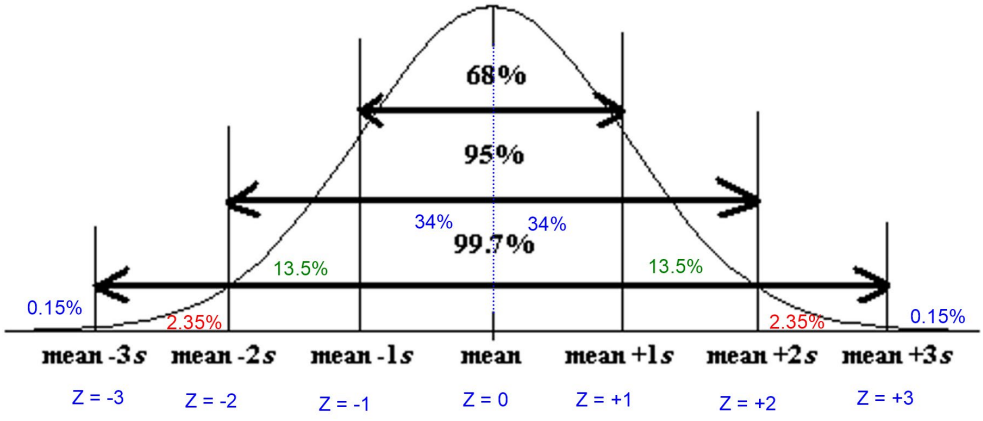
Empirical Rule

There is common percentages that go with normal (bell-shaped) data. Usually about 68% of normal data will be within one standard deviation of the mean (typical). About 95% of normal data will be within two standard deviations of the mean. About 99.7% of normal data will be within three standard deviations of the mean. These percentages are often referred to as the "Empirical Rule" or the "68-95-99.7 Rule".



Notice that we can use the 68%, 95% and 99.7% to figure out the sections. Since 68% makes up the middle two symmetric sections, we know each section is about 34%. Similarly, the middle four sections make up about 95%. Subtract out the middle two sections (68%) gives 27%. Divide that in half and you get two sections each making up 13.5% of the normal data. The middle six sections make up about 99.7%. Subtract out the middle four sections (95%) gives 4.7%. Divide that in half and you get two sections each making up 2.35% of the normal data. The end sections are calculated in a similar manner (100%-99.7% = 0.3%). Divide that into two symmetric tails and we get that each tail should be about 0.15%.

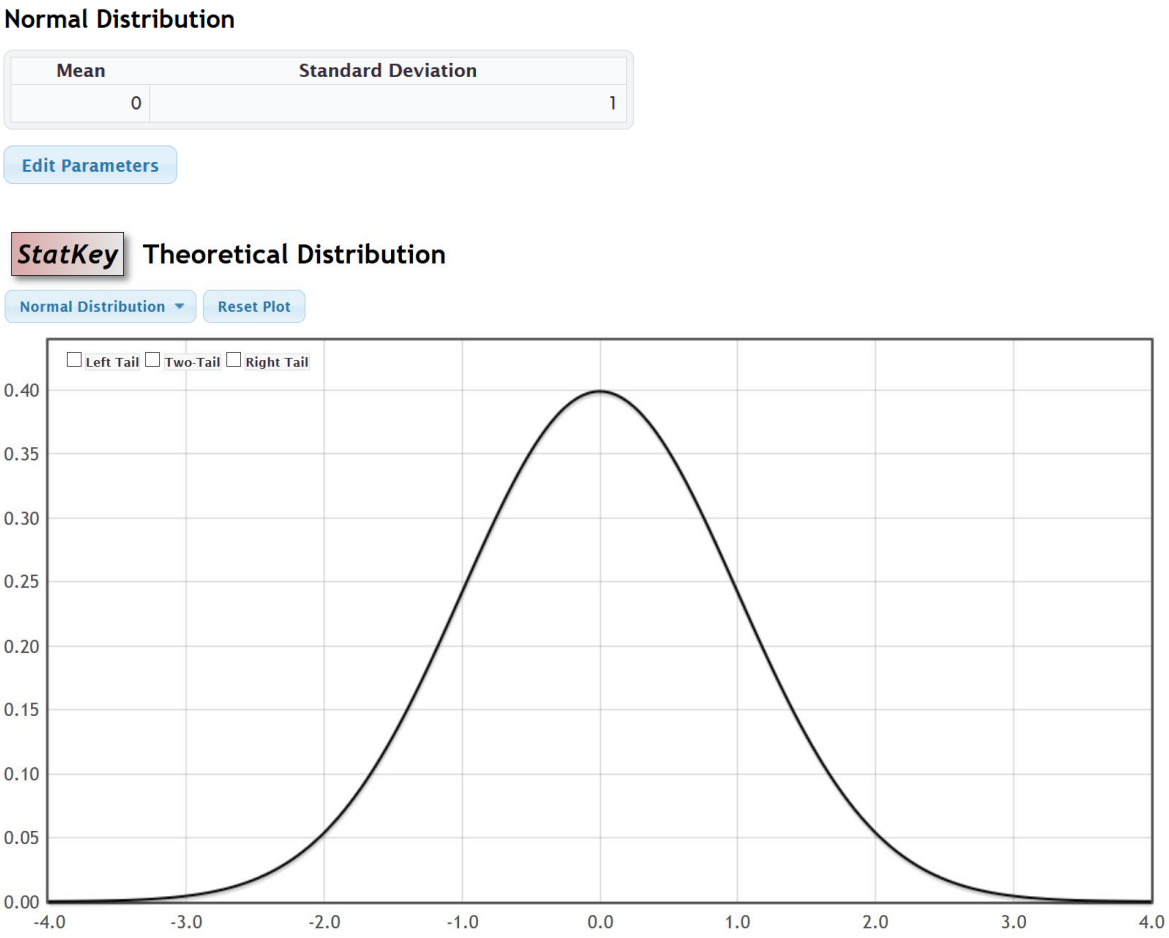
Remember the number of standard deviations from the mean is the Z-score. You can write the Z-scores for the bottom values in the Empirical rule. This is often called the "Standard Normal Curve". Notice the center of the curve is the mean (Z-score of zero) and the standard deviation of this curve is exactly one. When a computer program refers to a normal curve with a mean of zero and a standard deviation of one, they are talking about Z-scores and the Standard Normal Curve.



Many data sets are normal. We will see in the next chapter that many sampling distributions have a normal shape as well. It is therefore important to be able to calculate percentages associated with normal data and normal curves. Confidence Intervals and P-value are both extremely important topics that we will cover in chapter 3 and chapter 4 that involve the empirical rule and calculating percentages associated with normal curves.

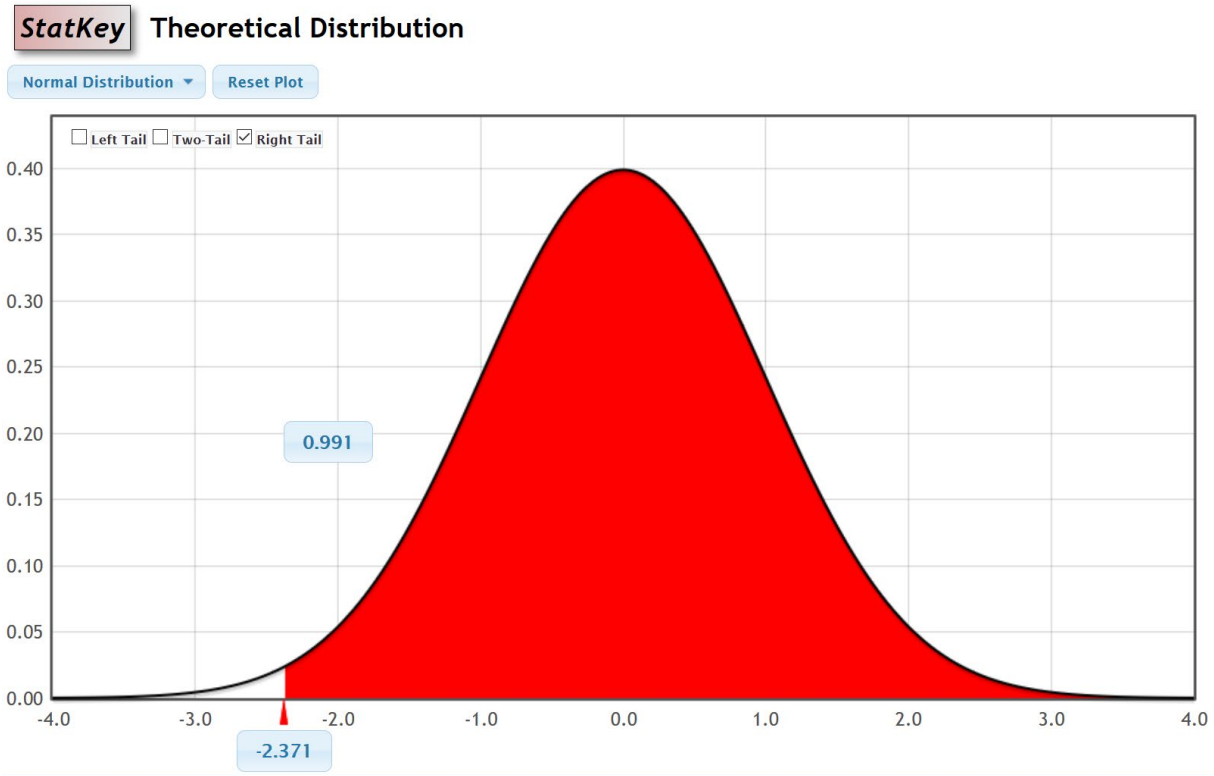
Calculating Percentages for Normal Curves with StatKey

Computer software programs can calculate percentages associated with normal quantitative data. Go to www.lock5stat.com and click on "StatKey". Under the "Theoretical Distributions" menu click on "Normal". Notice the parameters are set at a mean of zero and a standard deviation of one. Remember this means it is set up to find Z-scores or to find percentages associated with Z-scores. The curve is sometimes called a "density curve".
The idea is that the total area under the curve is 100%, so to find a percentage you find the area under the curve.



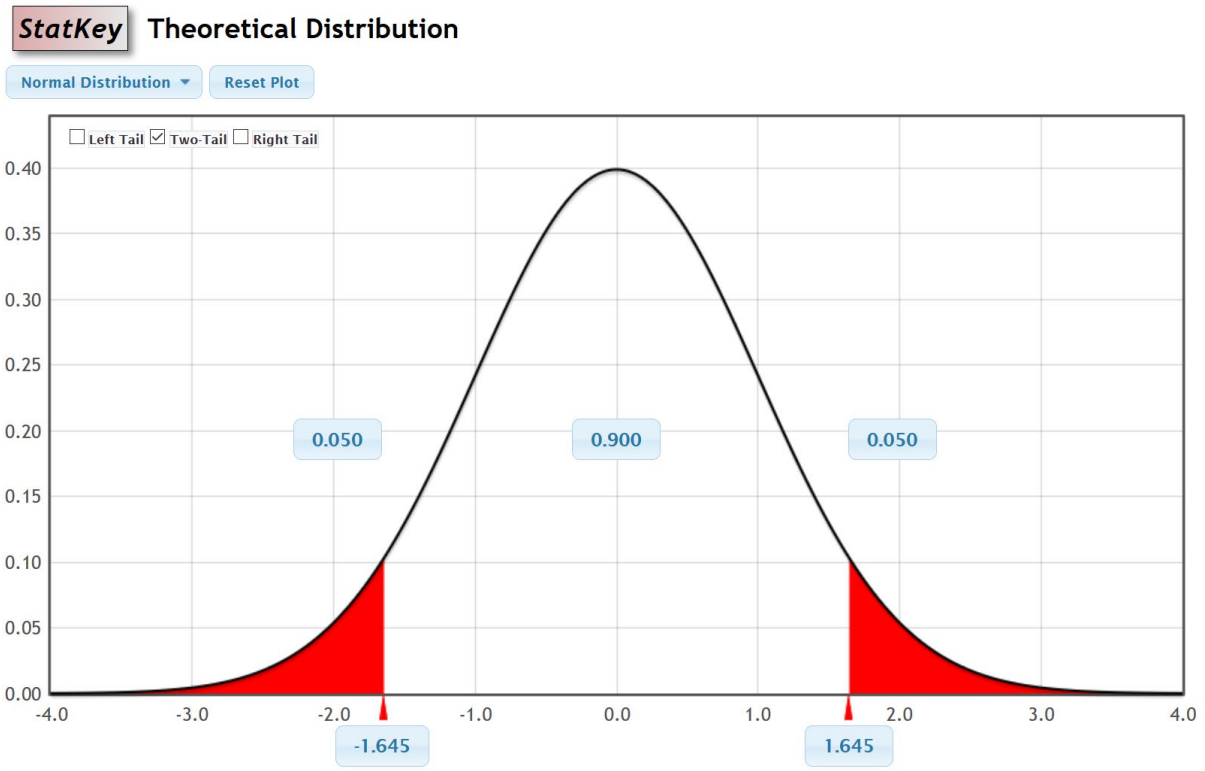
Notice that the curve has three buttons on the top left (Left Tail, Two-Tail, and Right Tail).

Example: Suppose we want to find what percent of normal data has a Z-score of −2.371 or above. Since we are looking for above, click the right tail button. The upper box is the percentage and the lower box is the Z-score. In this case, we know the Z-score and are looking for the percentage. So in the bottom box type in −2.371.



Notice the top box is the answer, 99.1% of normal data values will have a Z-score of −2.371 or higher.

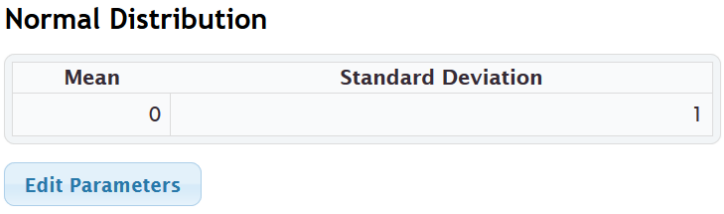
Example: Push the "reset plot" button. Suppose we want to find the two Z-scores that 90% (0.9) of normal data values are in between. Since we are looking for "in between", click the two-tail button. The upper boxes are the percentages and the lower boxes are the Z-scores. In this case, we know the percentage in between, but need to find the Z-scores. So in the upper middle box type in the decimal proportion equivalent of 90% (0.9).



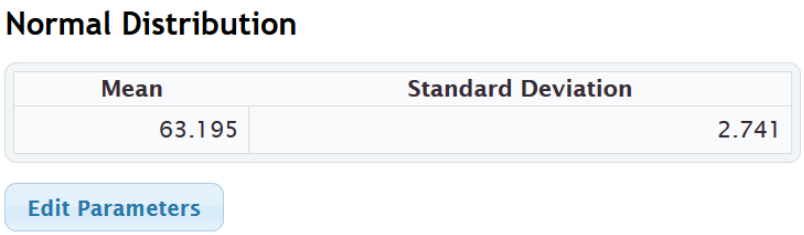
Notice the Z-score answers we are looking for are at the bottom. Therefore, the middle 90% of normal data values have a Z-score between −1.645 and +1.645. These are the famous Z-scores for 90% confidence intervals that we will study in later chapters.

Percentages for any normal data

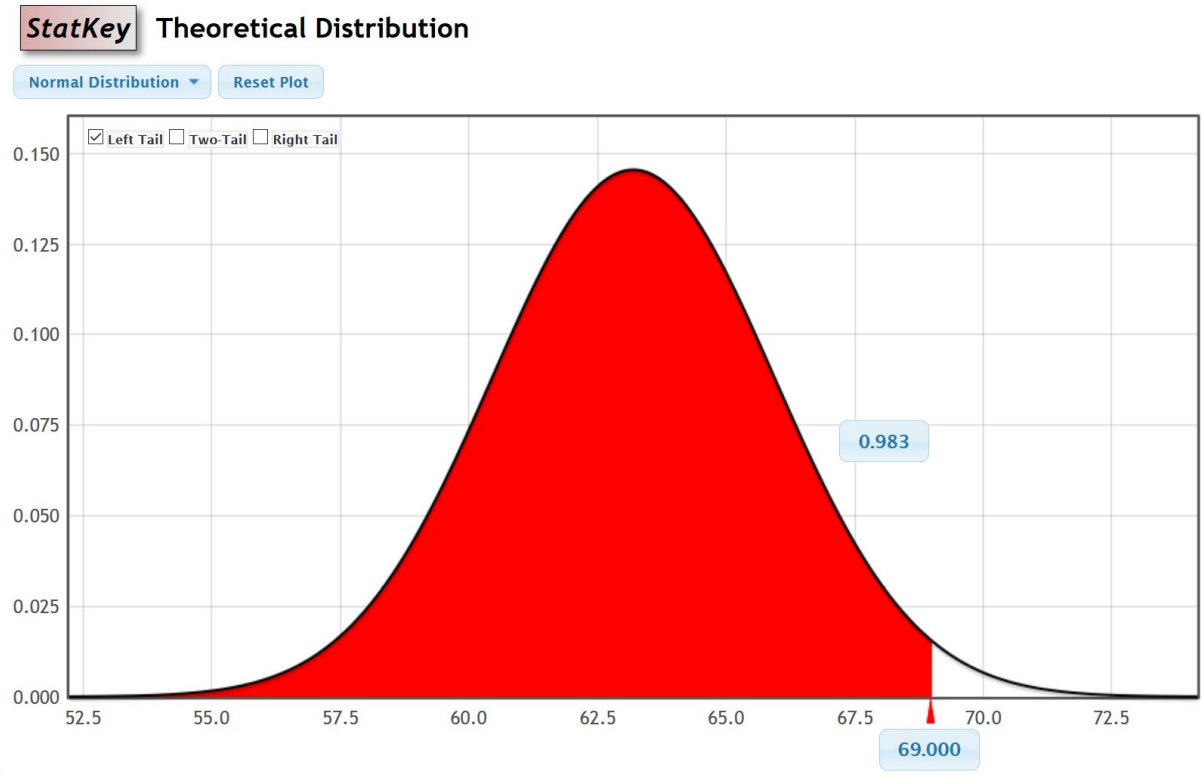
We often want to calculate percentages for normal quantitative data without calculating Z-scores first. StatKey can do that as well. Push the "reset plot" button. Right now, the mean is set at zero and the standard deviation is at one.



Example: Suppose we want to calculate percentages associated with the women's height data we studied earlier. We found that the women's heights were normally distributed with a mean of 63.195 inches and a standard deviation of 2.741 inches. Click on the button that says, "edit parameters" and put those numbers into StatKey.

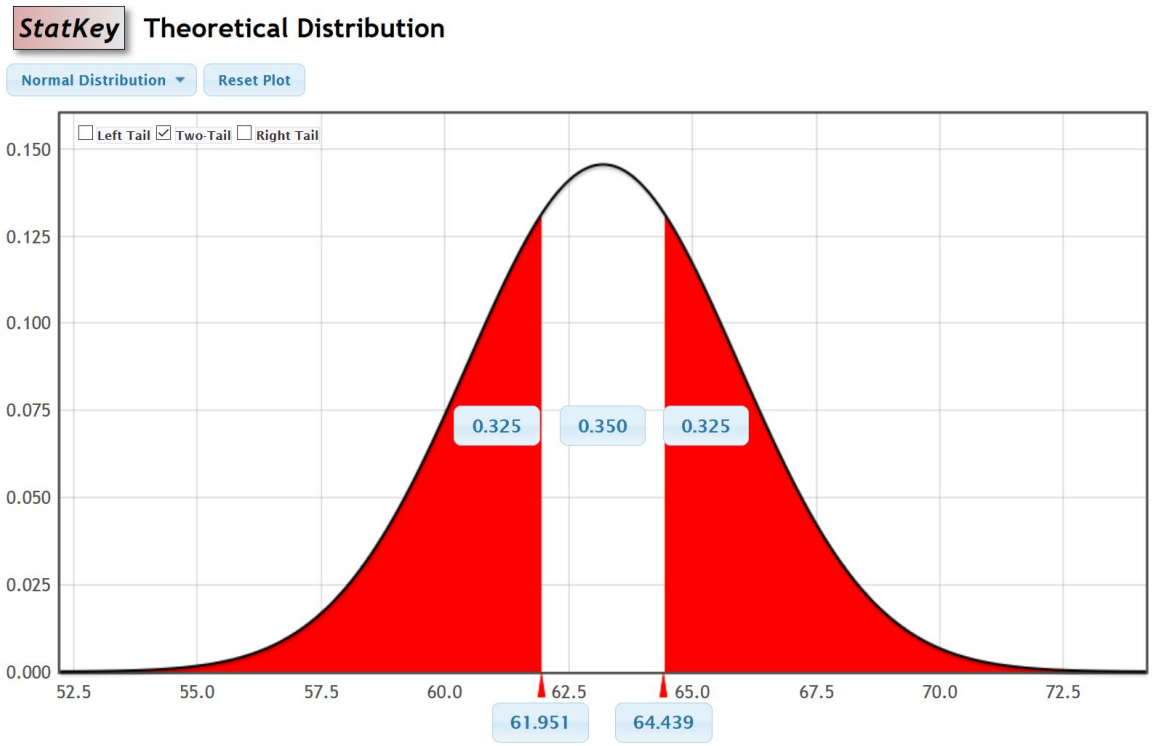


Suppose we want to know what percentage of women in the data have a height of 69 inches or less. Since we are looking for "less than", click on left tail. Remember the top box is the percentage (proportion). The bottom box is now the height. Since we know the height is 69, type in 69 into the bottom box. The proportion in the top box is our answer. So about 98.3% of the women in the sample data have a height below 69 inches.



Note: Be careful about generalizing results of sample data to the population. This does not mean that 98.3% of all women have a height of 69 inches or below. As we learned in chapter one, samples may have bias and not represent the population.

Example: Suppose we wish to find the two heights corresponding to the middle 35%. That is the two heights that 35% of women are in between. Just push the "two-tail" button and put 0.35 in the upper middle box. The answer will be in the two lower boxes.

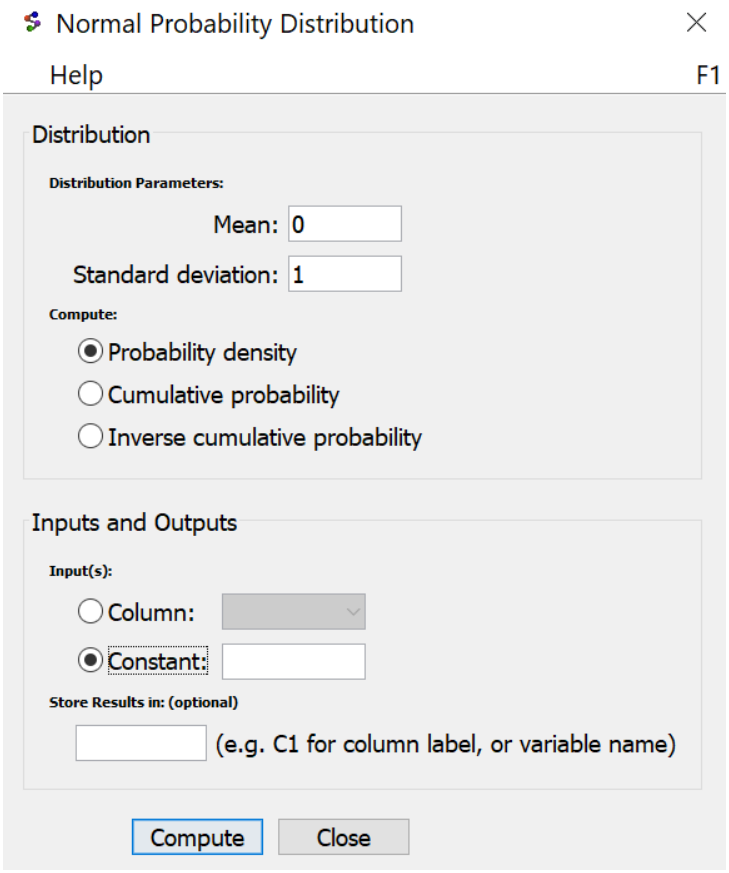


So about 35% of the women in the data have a height between 61.951 and 64.439 inches.

Note: These percentages are based on perfectly normal curves, yet real data is rarely perfectly normal. There are actually 15 women in the data had a height between 61.951 and 64.439. This was actually 37.5%. This is off from the theoretical percentage because the data was not perfectly normal. It is important to realize that theoretical distributions rarely match up exactly with real data.

Calculating Percentages for Normal Curves with Statcato

Z-scores, X-values and percentages for normal curves can also be calculated with Statcato. Go to the "Calculate" menu, click on "Probability Distributions" and then "Normal".



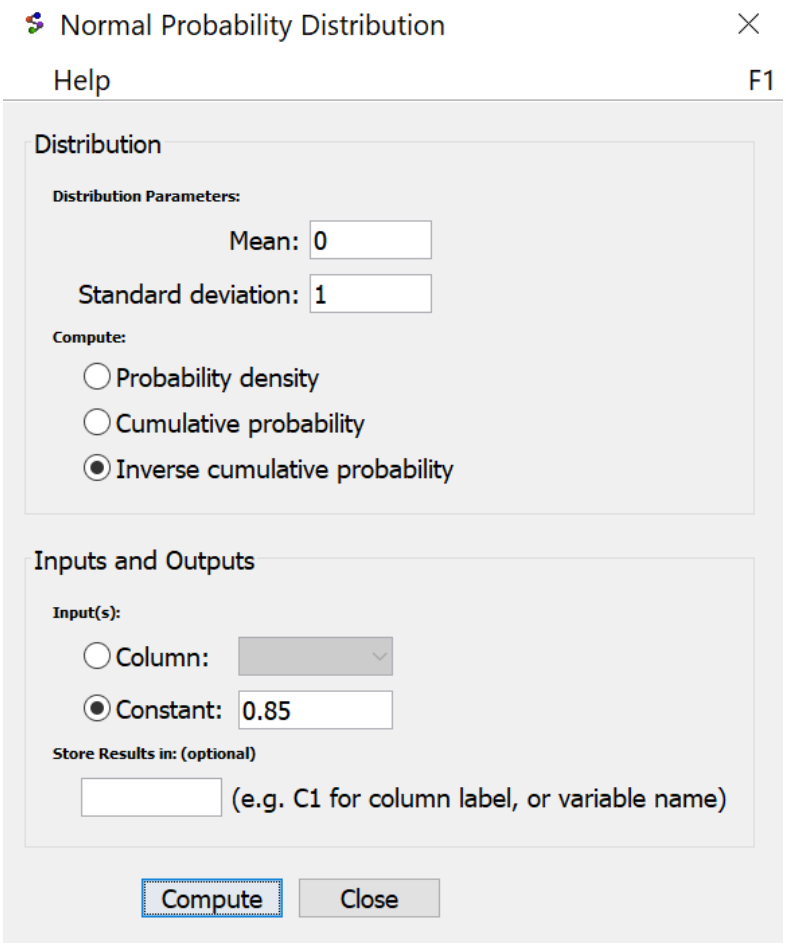
If you leave the mean at zero and the standard deviation at one, then Statcato is set up to calculate Z-scores or percentages from Z-scores. To calculate a Z-score from a percentage less than the Z-score, put in the proportion (decimal equivalent of the percentage) into the box that says "constant". Then click "inverse cumulative" and "compute". For example, what is the Z-score that 85% of values in a normal data set are less than? The answer is under "X". The Z-score is 1.0365.

Normal Distribution: mean = 0.0 stdev = 1.0

Input: 0.85

Type: Inverse cumulative probability

P(<=X)	X
0.85	1.0365



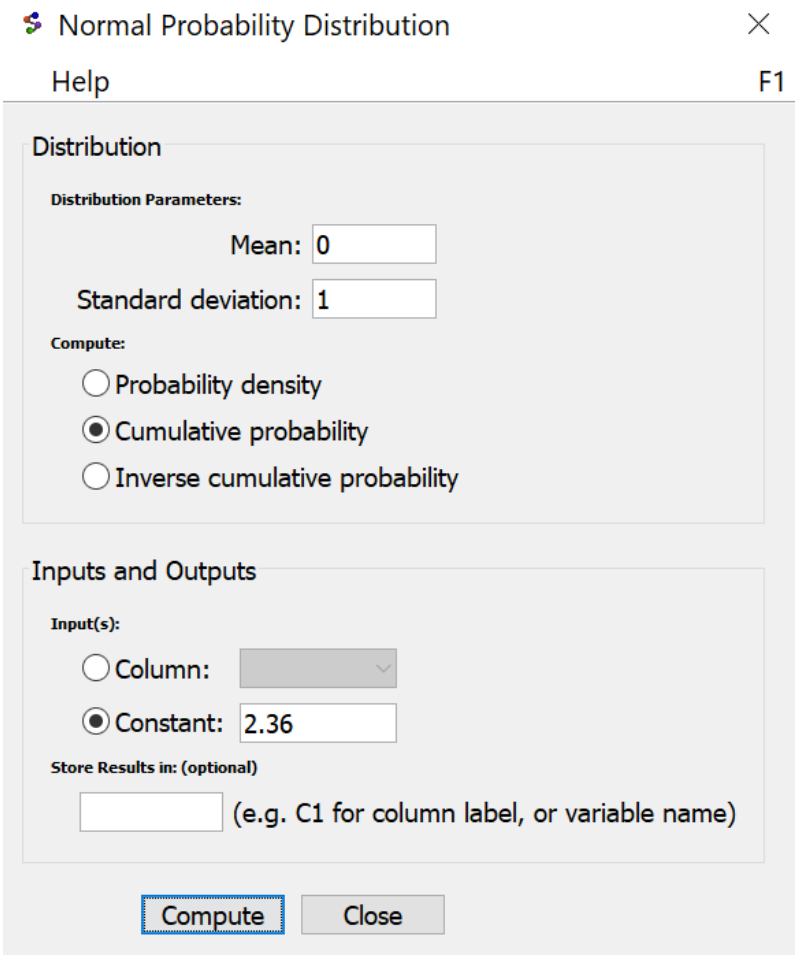
Suppose we want to find the percentage less than a Z-score of 2.36. Put 2.36 in the constant box and press "Cumulative Probability". The answer is under "P(< = X)". So the answer 0.990863 or about 99.1%.

Normal Distribution: mean = 0.0 stdev = 1.0

Input: 2.36

Type: Cumulative probability

X	P(<=X)
2.36	0.990863



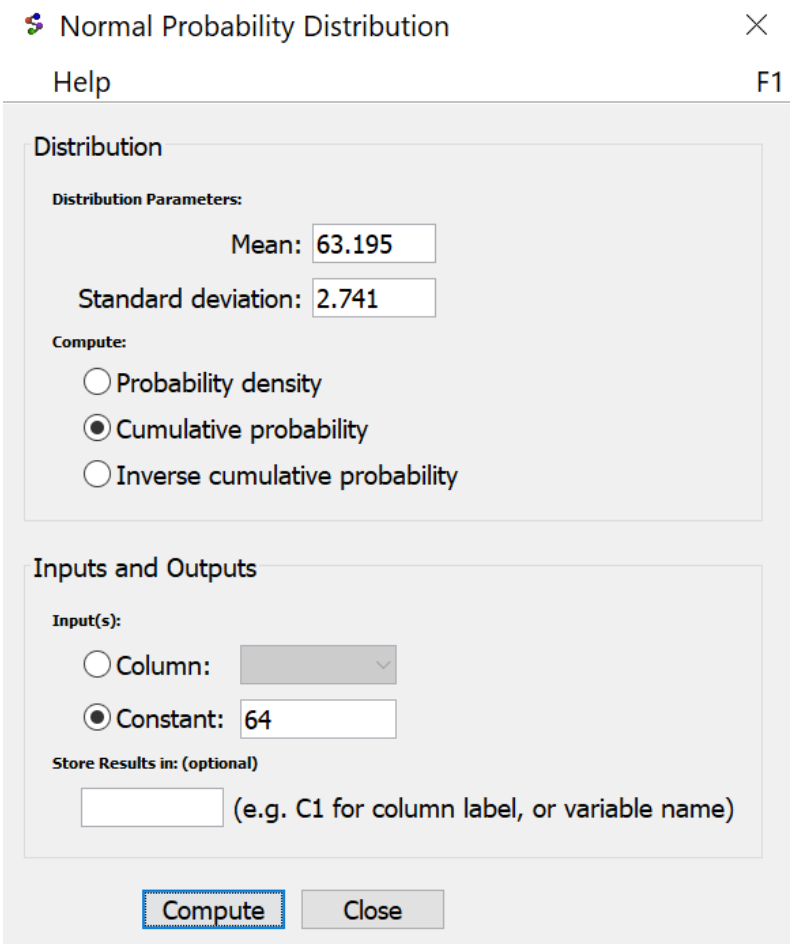
We can also calculate X-values and percentage for those X-values for normally distributed data. We need to input the mean and standard deviation into Statcato. For example, earlier we saw some random sample data for women's heights was normally distributed with a mean of 63.195 and a standard deviation of 2.741. Suppose we want to find the percentage of women in the data that have a height below 64 inches. We see that the answer is 0.615502 or about 61.6%. Note that Statcato can only calculate for less than. If we want to know what percent of women in the data have a height above 64 inches, we first calculate less than and then subtract the answer from 100%. In this case, 100% − 61.6% = 38.4%.

Normal Distribution: mean = 63.195 stdev = 2.741

Input: 64.0

Type: Cumulative probability

X	P(<=X)
64.0	0.615502



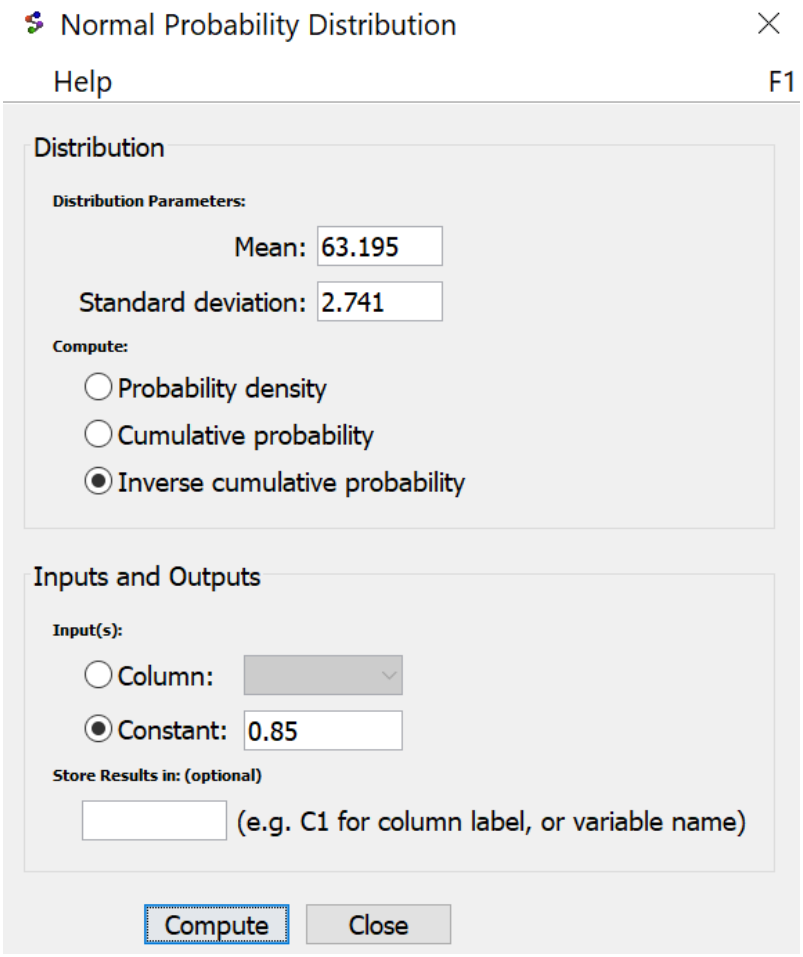
You can also use the "Inverse Cumulative Probability" function to calculate the height that 15% of women are taller than. Remember, Statcato only works with less than, so if 15% of women are greater than this height, than 85% of women are less than this same height. So we will enter 85% (0.85) into the constant box. We see the answer under "X". Therefore, 85% of women have a height less than 66.0358 inches. This also means that 15% of women have a height above 66.0358 inches.

Normal Distribution: mean = 63.195 stdev = 2.741

Input: 0.85

Type: Inverse cumulative probability

P(<=X)	X
0.85	66.0358



Calculating between is challenging with Statcato. Statcato does not have a between button, so we must use the percentages less than an X-value. If we want to find the two values that the middle 40% are in between, we have to think about the percentages less than each X-value. If 40% is in the middle, then the remaining 60% is divided into the two tails. Therefore, each tail must be 30%. So the X-value on the left will have 30% (0.3) less than. The X-value on the right will have 70% (0.7) less than. Put 0.3 into the "Constant" box and press inverse cumulative. Then put 0.7 into the "Constant" box and press inverse cumulative. For women's heights, we would get that the middle 40% of women's heights are between 61.7576 inches and 64.6324 inches.

Normal Distribution: mean = 63.195 stdev = 2.741
Input: 0.3
Type: Inverse cumulative probability

P(<=X)	X
0.3	61.7576

Normal Distribution: mean = 63.195 stdev = 2.741
Input: 0.7
Type: Inverse cumulative probability

P(<=X)	X
0.7	64.6324

Note on Rounding Statistics for Quantitative Data

It is often best to not round if you are unsure. Data analysts usually prefer better accuracy and can round to their own specifications. Rounding too much interferes with accuracy. If you must round, here are some general guidelines.

Percentages and proportions are usually rounded to three significant figures. Proportions are rounded to the thousandths place and percentages are rounded to the tenths place.

Quantitative statistics like the mean or standard deviation are usually rounded to one more decimal place to the right than the original data has. Notice the women's heights data is rounded to the tenths place (one number to the right of the decimal). So statistics calculated from this data would usually be rounded to the hundredths place (two numbers to the right of the decimal).

Mean (women's height) = 63.195 ≈ 63.20 inches

Standard Deviation (women's height) = 2.741 ≈ 2.74 inches

B
Women Ht (in)
64.3
66.4
62.3
62.3
59.6
63.6

Practice Problems Section 1F

1. Answer the following questions:

    a) What is meant by saying that data is normally distributed or "normal"?

    b) Define the mean average and explain how it is calculated.

    c) Define the standard deviation and explain how it is calculated.

2. Answer the following questions:

    a) If a data set is normally distributed, what measure of average should we use?

    b) If a data set is normally distributed, what measure of spread should we use?

    c) If a data set is normally distributed, how many standard deviations from the mean is considered typical?

    d) If a data set is normally distributed, what is the formula for finding typical values?

    e) If a data set is normally distributed, approximately what percentage is typical?

    f) If a data set is normally distributed, how many standard deviations from the mean is considered unusual?

    g) If a data set is normally distributed, approximately what percentage of the data is unusually high?

    h) If a data set is normally distributed, approximately what percentage of the data is unusually low?

(#3-4) Directions: The following graphs and statistics were calculated with Statcato and the "Bear" data from the website www.matt-teachout.org . Use the dot plot, histogram and summary statistics to answer the following questions. Here are the formulas for typical and unusual values.

Mean – Standard Deviation ≤ Typical Values for Normal Data ≤ Mean + Standard Deviation

Unusual Low Cutoff for Normal Data = Mean − (2 × Standard Deviation)

Unusual High Cutoff for Normal Data = Mean + (2 × Standard Deviation)

3. Bear neck circumference (inches)

    a) What is the data measuring and what are the units?
    b) How many numbers are in the data set?
    c) Is the data set normally distributed? (Yes or No)
    d) What is the minimum value?
    e) What is the maximum value?
    f) What is the average (center)? *(Give the number and the name of the statistic used.)*
    g) How much typical spread does the data set have?
    (Give the number and the name of the statistic used.)
    h) Find two numbers that typical values fall in between.
    i) What is the unusual high (high outlier) cutoff for this data?
    j) What is the unusual low (low outlier) cutoff for this data?
    k) List all high outliers in this data set. If there are no high outliers, put "none".
    l) List all low outliers in this data set. If there are no high outliers, put "none".

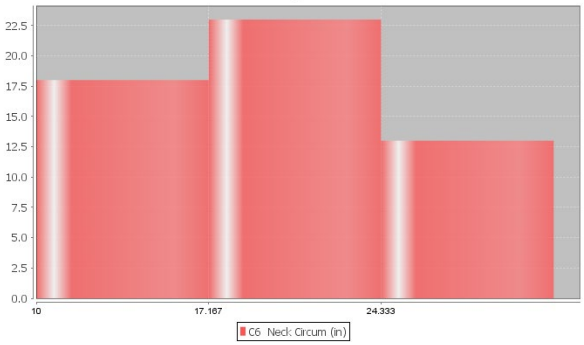
Descriptive Statistics

Variable	Mean	Standard Deviation
C6 Neck Circum (in)	20.556	5.641

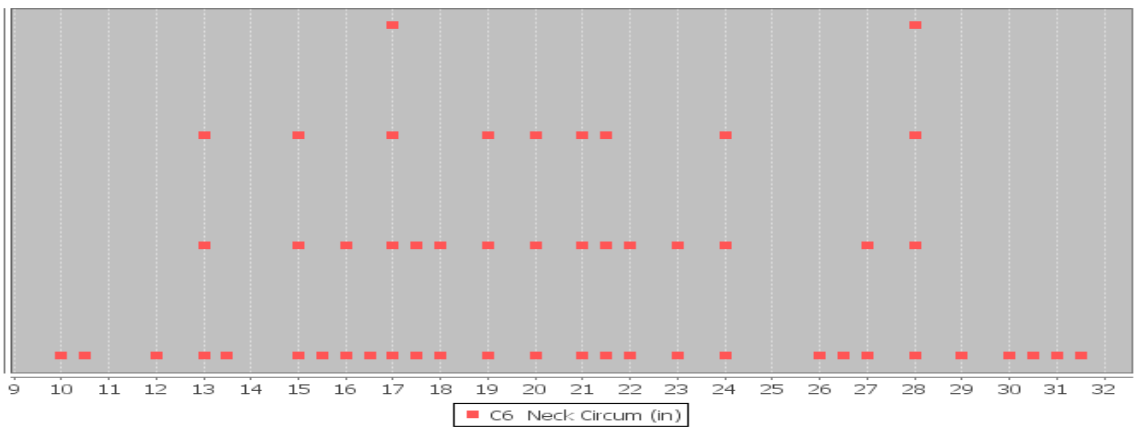
Variable	Min	Max
C6 Neck Circum (in)	10.0	31.5

Variable	N total
C6 Neck Circum (in)	54

Histogram



Dot Plot



4. Bear Chest Size (inches)

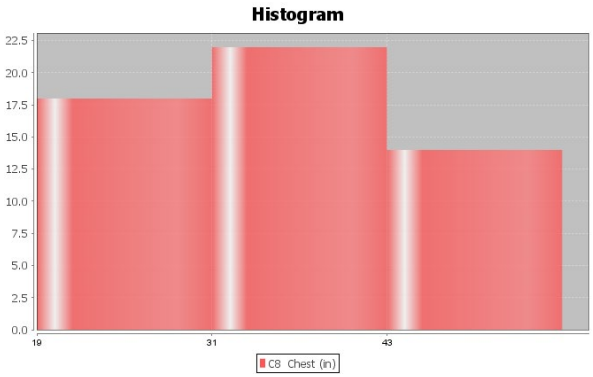
a) What is the data measuring and what are the units?
b) How many numbers are in the data set?
c) Is the data set normally distributed? (Yes or No)
d) What is the minimum value?
e) What is the maximum value?
f) What is the average (center)? *(Give the number and the name of the statistic used.)*
g) How much typical spread does the data set have?
(Give the number and the name of the statistic used.)
h) Find two numbers that typical values fall in between.
i) What is the unusual high (high outlier) cutoff for this data?
j) What is the unusual low (low outlier) cutoff for this data?
k) List all high outliers in this data set. If there are no high outliers, put "none".
l) List all low outliers in this data set. If there are no high outliers, put "none".

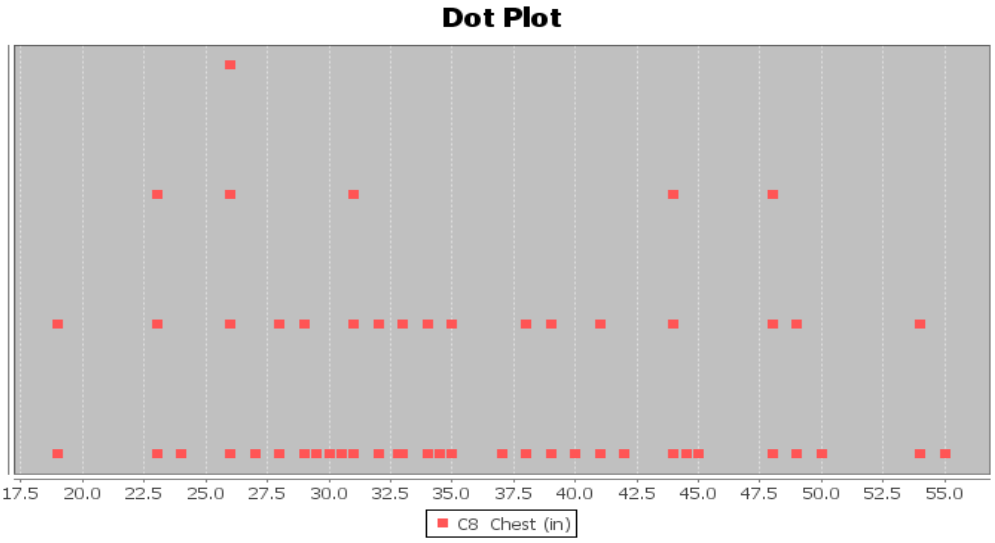
Descriptive Statistics

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





(#5-8) Directions: Open the "Health" data from the website www.matt-teachout.org . (Look under "Statistics" tab and then click the "data sets" tab.) Go to www.lock5stat.com and open StatKey. Under the "Descriptive Statistics and Graphs" menu, click on "One Quantitative Variable". Paste the indicated column of quantitative data into StatKey, create a dot plot and histogram, and find the summary statistics. Then answer the following questions. Here are the formulas for typical and unusual values.

Mean – Standard Deviation ≤ Typical Values for Normal Data ≤ Mean + Standard Deviation

Unusual Low Cutoff for Normal Data = Mean − (2 × Standard Deviation)

Unusual High Cutoff for Normal Data = Mean + (2 × Standard Deviation)

5. Women's Diastolic Blood Pressure (Millimeters of Mercury (mm of Hg))

a) What is the data measuring and what are the units?
b) How many numbers are in the data set?
c) Is the data set normally distributed? (Yes or No)
d) What is the minimum value?
e) What is the maximum value?
f) What is the average (center)? *(Give the number and the name of the statistic used.)*
g) How much typical spread does the data set have?
(Give the number and the name of the statistic used.)
h) Find two numbers that typical values fall in between.

i) What is the unusual high (high outlier) cutoff for this data?
j) What is the unusual low (low outlier) cutoff for this data?
k) List all high outliers in this data set. If there are no high outliers, put "none".
l) List all low outliers in this data set. If there are no high outliers, put "none".

6. Women's Wrist Circumference (Inches)

a) What is the data measuring and what are the units?
b) How many numbers are in the data set?
c) Is the data set normally distributed? (Yes or No)
d) What is the minimum value?
e) What is the maximum value?
f) What is the average (center)? *(Give the number and the name of the statistic used.)*
g) How much typical spread does the data set have?
(Give the number and the name of the statistic used.)
h) Find two numbers that typical values fall in between.
i) What is the unusual high (high outlier) cutoff for this data?
j) What is the unusual low (low outlier) cutoff for this data?
k) List all high outliers in this data set. If there are no high outliers, put "none".
l) List all low outliers in this data set. If there are no high outliers, put "none".

7. Men's Height (Inches)

a) What is the data measuring and what are the units?
b) How many numbers are in the data set?
c) Is the data set normally distributed? (Yes or No)
d) What is the minimum value?
e) What is the maximum value?
f) What is the average (center)? *(Give the number and the name of the statistic used.)*
g) How much typical spread does the data set have?
(Give the number and the name of the statistic used.)
h) Find two numbers that typical values fall in between.
i) What is the unusual high (high outlier) cutoff for this data?
j) What is the unusual low (low outlier) cutoff for this data?
k) List all high outliers in this data set. If there are no high outliers, put "none".
l) List all low outliers in this data set. If there are no high outliers, put "none".

8. Men's Weight (Pounds)

a) What is the data measuring and what are the units?
b) How many numbers are in the data set?
c) Is the data set normally distributed? (Yes or No)
d) What is the minimum value?
e) What is the maximum value?
f) What is the average (center)? *(Give the number and the name of the statistic used.)*
g) How much typical spread does the data set have?
(Give the number and the name of the statistic used.)
h) Find two numbers that typical values fall in between.
i) What is the unusual high (high outlier) cutoff for this data?
j) What is the unusual low (low outlier) cutoff for this data?
k) List all high outliers in this data set. If there are no high outliers, put "none".
l) List all low outliers in this data set. If there are no high outliers, put "none".

(#9-18) Directions: Use the following formula when needed and answer the following questions about Z-scores.

$Z = \frac{(Amount - Mean)}{Standard\ Deviation}$

9. Write the definition of a Z-score.

10. Explain how we can use Z-scores to tell if a number is typical in normal data?

11. Explain how we can use Z-scores to tell if a number is unusual in normal data?

12. A random sample of IQ tests is normally distributed with a mean of 99.8 and a standard deviation of 15.3. Bud has an IQ of 143. Use this information to answer the following Z-score questions.

a) Calculate the Z-score for Bud's IQ.

b) Write a sentence to explain the Z-score in context.

c) Is Buds' IQ unusually high compared to other people in the data set? Explain your answer.

13. A random sample of IQ tests is normally distributed with a mean of 99.8 and a standard deviation of 15.3. Jan has an IQ of 89. Use this information to answer the following Z-score questions.

a) Calculate the Z-score for Jan's' IQ.

b) Write a sentence to explain the Z-score in context.

c) Is Jan's' IQ unusually low compared to other people in the data set? Explain your answer.

14. A clothing store wants to study the amount of money spent in their store by customers. Census data indicated that the data is normally distributed with a mean of $46.89 and a standard deviation of $12.44. Maria spent $105.12 on merchandise in the store. Use this information to answer the following Z-score questions.

a) Calculate the Z-score for the amount Maria spent.

b) Write a sentence to explain the Z-score in context.

c) Is the amount Maria spent unusually high compared to other people in the data set? Explain your answer.

15. A clothing store wants to study the amount of money spent in their store by customers. Census data indicated that the data is normally distributed with a mean of $46.89 and a standard deviation of $12.44. Julie spent $13.61 on merchandise in the store. Use this information to answer the following Z-score questions.

a) Calculate the Z-score for the amount Julie spent.

b) Write a sentence to explain the Z-score in context.

c) Is the amount Julie spent unusually low compared to other people in the data set? Explain your answer.

16. Neck circumferences of bears are normally distributed with a mean circumference of 20.556 inches and a standard deviation of 5.641 inches. A bear has a neck circumference of 13.7 inches.

a) Calculate the Z-score for this bears neck circumference.

b) Write a sentence to explain the Z-score in context.

c) Is this bears' neck circumference unusually low compared to other bears in the data set? Explain your answer.

17. Chest sizes of bears was normally distributed with a mean chest size of 35.663 inches and a standard deviation of 9.352 inches. A bear has a chest size of 57 inches.

a) Calculate the Z-score for this bears chest size.

b) Write a sentence to explain the Z-score in context.

c) Is this bears' chest size unusually large compared to other bears in the data set? Explain your answer.

18. The diastolic blood pressure of a random sample of women had a mean of 67.425 mm of Hg and a standard deviation of 11.626. A woman in the data has a diastolic blood pressure of 72 mm of Hg.

a) Calculate the Z-score for this woman's diastolic blood pressure.

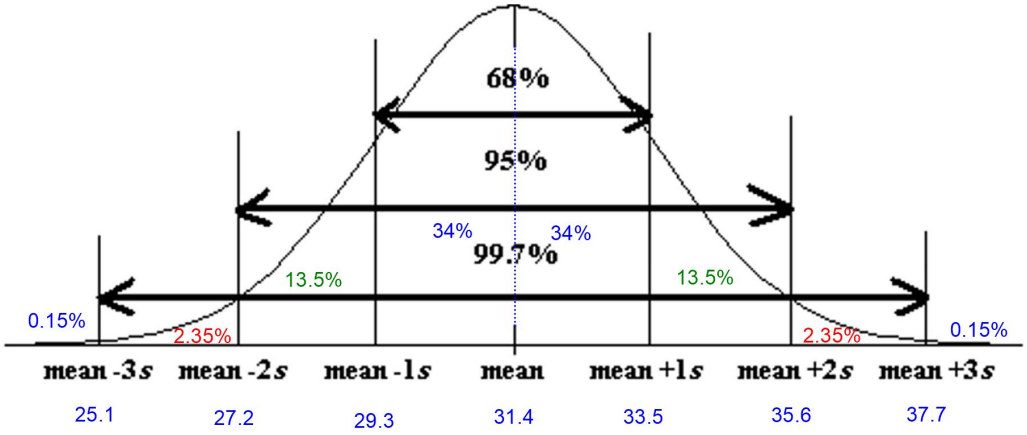
b) Write a sentence to explain the Z-score in context.

c) Is this woman's diastolic blood pressure unusually high compared to other women in the data set? Explain your answer.

(#19-25) Answer the following questions about the empirical rule.

19. Draw that standard normal curve. Label the mean and the values for one, two and three standard deviations above and below the mean. Also, label the percentages that make up the empirical rule.

20. The salaries of employees at a company are normally distributed with a mean of 31.4 thousand dollars and a standard deviation of 2.1 thousand dollars. Use the Empirical Rule graph below to answer the following questions.



Salaries in Thousands of Dollars

a) What percentage of the employees have a salary between 27.2 thousand dollars and 35.6 thousand dollars?

b) What percentage of the employees have a salary between 29.3 thousand dollars and 33.5 thousand dollars?

c) What percentage of the employees have a salary between 25.1 thousand dollars and 37.7 thousand dollars?

d) What percentage of the employees have a salary greater than 33.5 thousand dollars?

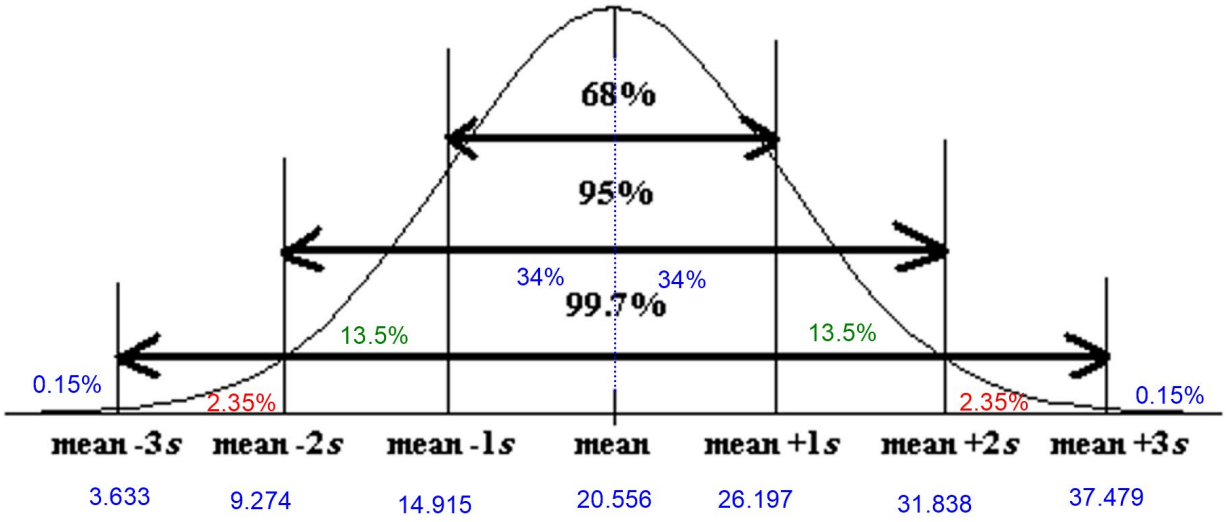
e) What percentage of the employees have a salary less than 27.2 thousand dollars?

f) Typical values for a normal curve are one standard deviation from the mean. Find two salaries that typical employee salaries fall in between?

g) The unusual high cutoff is two standard deviations above the mean. What salary represents the unusual high cutoff, which is the salary that 2.5% of the employees are greater than?

h) The unusual low cutoff is two standard deviations below the mean. What salary represents the unusual low cutoff, that is the salary that 2.5% of the employees are less than?

21. Neck circumferences of a sample of bears are normally distributed with a mean circumference of 20.556 inches and a standard deviation of 5.641 inches. Use the Empirical Rule graph below to answer the following questions.



a) What percent of the bears have a neck circumference between 14.915 inches and 31.838 inches?

b) What percent of the bears have a neck circumference less than 26.197 inches?

c) Typical bears have a neck circumference between what two amounts?

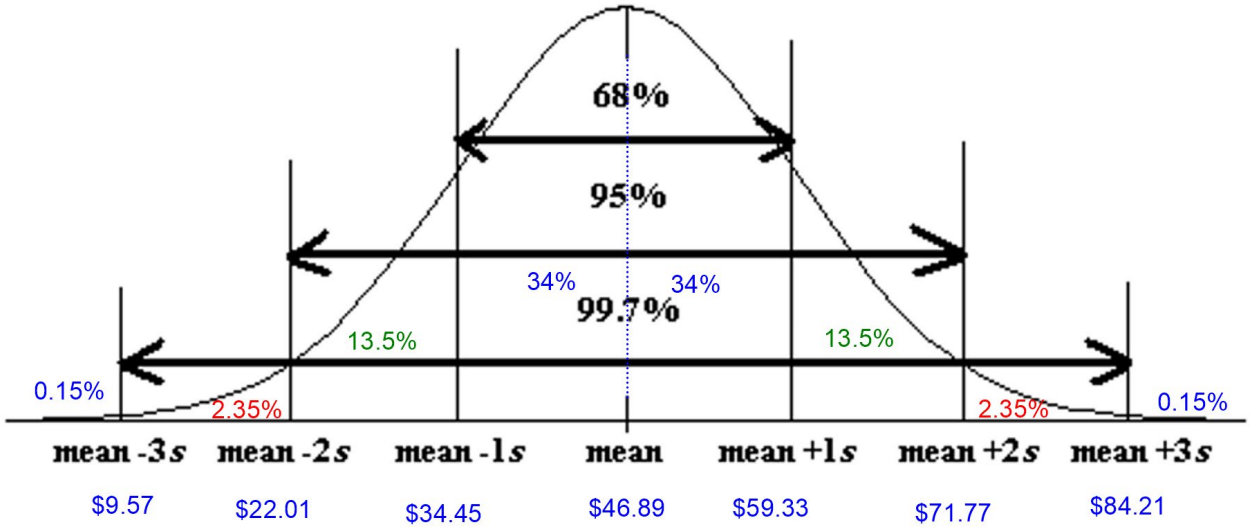
d) What is the unusual high cutoff, that is the bear neck circumference that 2.5% of bears are more than.

e) What is the unusual low cutoff, that is the bear neck circumference that 2.5% of bears are less than.

f) What is the bear neck circumference that 84% of bear neck circumferences are more than?

g) What percent of the bears have a neck circumference less than 14.915 inches?

22. A clothing store wants to study the amount of money spent in their store by customers. Census data indicated that the data is normally distributed with a mean of $46.89 and a standard deviation of $12.44. Use the Empirical Rule graph below to answer the following questions.



a) What percent of customers spent between $71.77 and $84.21 in the store?

b) What percent of customers spent less than $34.45 in the store?

c) Typical customers spent between what two amounts?

d) What is the unusual high cutoff, that is the dollar amount that 2.5% of customers spent more than.

e) What is the unusual low cutoff, that is the dollar amount that 2.5% of customers spent less than.

f) What is the dollar amount that 16% of customers spent more than?

g) What percent of customers spent less than $71.77?

23. A random sample of IQ tests is normally distributed with a mean of 99.8 and a standard deviation of 15.3. Use this information to answer the following questions. Go to www.lock5stat.com and open StatKey. Under the "Theoretical Distributions" menu, click on "Normal".

a) Use StatKey to calculate what percent of people in the IQ sample data that have an IQ greater than 77.
b) Use StatKey to calculate what percent of people in the IQ sample data that have an IQ less than 108.
c) Use StatKey to calculate what percent of people in the IQ sample data that have an IQ between 95 and 120.
d) Use StatKey to find the IQ score that 60% of people are less than.
e) Use StatKey to find the IQ score that 85% of people are greater than.
f) Use StatKey to find two IQ scores that the middle 40% of people are in between.

24. A clothing store wants to study the amount of money spent in their store by customers. Census data indicated that the data is normally distributed with a mean of $46.89 and a standard deviation of $12.44. Go to www.lock5stat.com and open StatKey. Under the "Theoretical Distributions" menu, click on "Normal".

a) Use StatKey to calculate the percent of people that spent more than $25.
b) Use StatKey to calculate the percent of people that spent less than $50.
c) Use StatKey to calculate the percent of people spent between $35 and $60.
d) Use StatKey to find the amount of money spent that 37% of people are less than.
e) Use StatKey to find the amount of money spent that 15% of people are more than.
f) Use StatKey to find two amounts that the middle 60% of people are in between.

25. The diastolic blood pressure of a random sample of women had a mean of 67.425 mm of Hg and a standard deviation of 11.626. Go to www.lock5stat.com and open StatKey. Under the "Theoretical Distributions" menu, click on "Normal".

a) Use StatKey to calculate the percent of women that have a diastolic blood pressure below 75 mm of Hg.
b) Use StatKey to calculate the percent of women that have a diastolic blood pressure above 50 mm of Hg.
c) Use StatKey to calculate the percent of women that have a diastolic blood pressure between 60 and 70 mm of Hg.
d) Use StatKey to find the diastolic blood pressure that 80% of women are lower than.
e) Use StatKey to find the diastolic blood pressure that 45% of women are higher than.
f) Use StatKey to find the two diastolic blood pressures that the middle 75% of women are in between.

Section 1G – Quantitative Data Analysis for Non-Normal Data and Summary Statistics

Vocabulary

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.

Normal Data: Data that is bell shaped, symmetric and unimodal.

Skewed Right Data: Also called positively skewed. Data where the center is on the far left and has a long tail to the right.

Skewed Left Data: Also called negatively skewed. Data where the center is on the far right and has a long tail to the left.

Sample Size: Also called the total frequency. The number of values are in a data set.

Median Average: The center of the data when the numbers are put in order. Also called the "50th Percentile" (P_{50}). Since about 50% of the numbers in the data set are less than the median. It is also called the "Second Quartile" (Q_2). The average for a data set that is not normal.

First Quartile (Q_1): The number that about 25% of the data values are less than. Used for typical values for data that is not normal.

Third Quartile (Q_3): The number that about 75% of the data values are less than. Used for typical values for data that is not normal.

Interquartile Range (IQR): The distance between the middle 50% of the numbers in a data set. Calculated by subtracting the 1st and 3rd quartiles. The measure of typical spread for a data set that is not normal.

Maximum: The largest number in a data set.

Minimum: The smallest number in a data set.

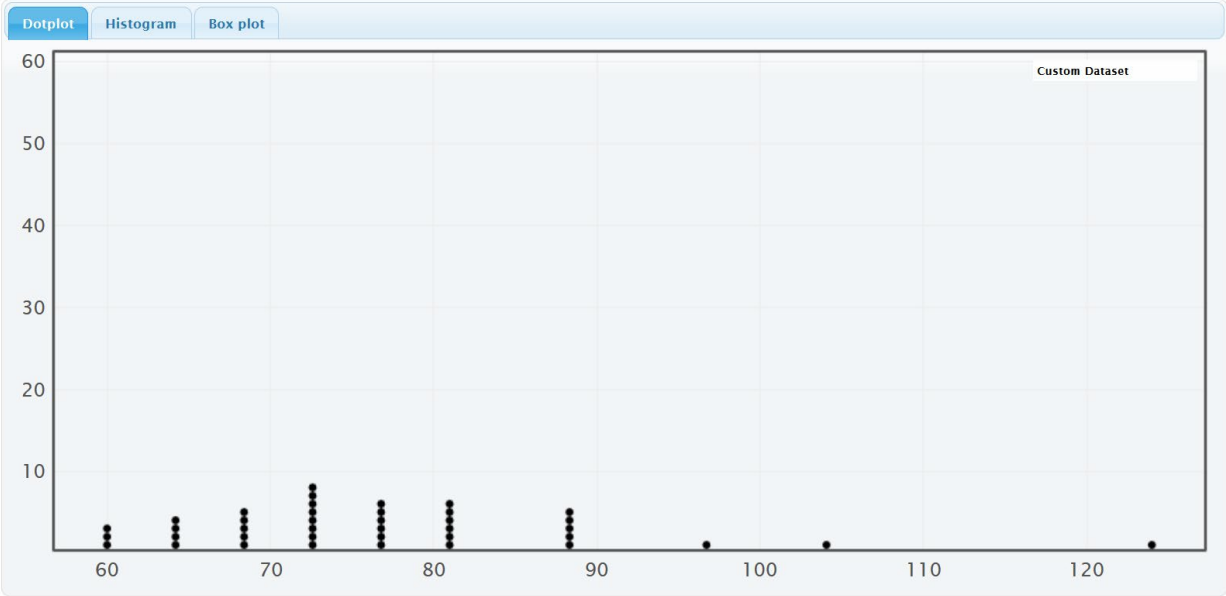
Range: A quick measure of total spread. Calculated by subtracting the minimum and maximum values in a data set.

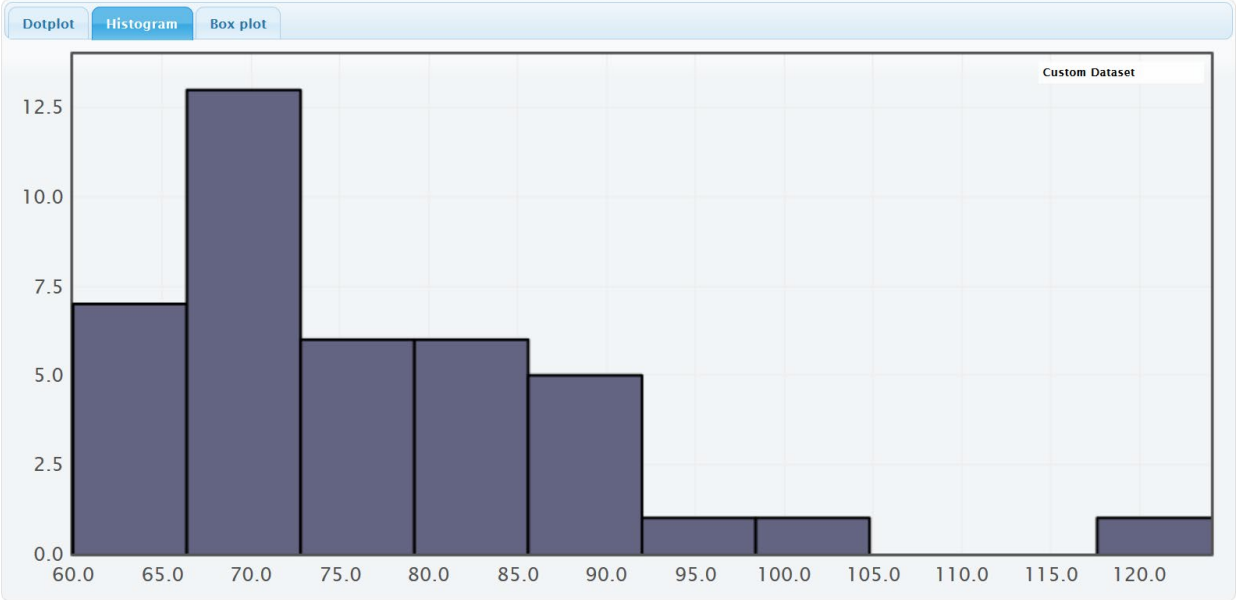
Outliers: Unusual values in the data set.

Introduction

When a data set is normal (or bell-shaped), we use the mean as our average and the standard deviation as our measure of typical spread. Not all data sets are normal though. Let us explore some data that is not normally distributed.

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). Go to www.matt-teachout.org and click on the "Statistics" tab and then the "Data Sets" tab. Open the health data in Excel. Copy the women's pulse rate data. Now go to www.lock5stat.com and click on "StatKey". Under the "Descriptive Statistics and Graphs" menu, click on "One Quantitative Variable". Under "Edit Data", paste the women's pulse rate data into StatKey. Uncheck the box that says, "First column is identifier". Check the box that says, "Data has header row". Push "OK". Here are the graphs and summary statistics.





Notice first that this is not normal data. The highest bar (center) is on the far left. The graph has a short tail to the left of the highest bar and a long tail to the right of the highest bar. This shape is called “skewed right” or “positively skewed”. We can adjust the number of bars (buckets) by using the slider on the right of the graph.

Summary Statistics

Statistic	Value
Sample Size	40
Mean	76.300
Standard Deviation	12.499
Minimum	60
Q_1	68.000
Median	74.000
Q_3	80.000
Maximum	124

Remember the mean and standard deviation are only accurate if the data is normal. Therefore, for this data set, we should not use the mean as the average and we should not use the standard deviation as our typical spread.

So what statistics should we use? Here is the general rule for skewed data or any data that is not normal.

Summary statistics for non-normal data

Average: Median

Typical Spread: Interquartile Range (IQR)

Typical Values: Between the first quartile (Q_1) and the third quartile (Q_3)

Outliers: Boxplot will indicate if there are outliers.

Quartiles are based on the numbers in order, so are much more accurate for data that is not normally distributed. The median is also called the 2nd quartile or the 50th percentile. It is the center of the data when the numbers are in order. About 50% of the numbers will be less than the median and about 50% of the numbers will be greater than the median. When a data set is not normally distributed, we use the median as our average. It is much closer to the center. Look at the histogram above. The summary statistics provided by StatKey show us that the mean was 76.3 beats per minute (bpm) and the median was 74 bpm. Notice 74 is closer to the highest bar in the data set. In other words, the median is closer to the center and a more accurate average than the mean. Mean averages are based on distances so will be pulled off the center in the direction of the skew.

The median is calculated by first putting the numbers in order from smallest to largest. If there is one number in the middle (sample size n is odd), then that is the median. If there are two numbers in the middle (sample size n is even), then the median will be half way between the two numbers in the middle.

The first quartile (Q_1) is also called the 25th percentile and is the number that about 25% of the data is less than. The third quartile (Q_3) is also called the 75th percentile and is the number that about 75% of the data is less than. The first and third quartiles are markers that mark the middle 50% of the data when it is in order. The middle 50% is considered "typical" in a data set that is not normally distributed. For normal data, we want the middle 68% (empirical rule) because there is more data in the middle.

The distance between the first and third quartiles is called the interquartile range (IQR). This is the best measure of typical spread for data that is not normally distributed. StatKey does not list the IQR in its summary statistics, but we can calculate it with the following formula.

IQR = $Q_3 - Q_1$

Since our women's pulse rate data was skewed right, we would use the following statistics.

Variable and Units: Women's pulse rates in beats per minute (bpm)

Minimum: The lowest pulse rate for these women was 60 bpm.

Maximum: The highest pulse rate for these women was 124 bpm.

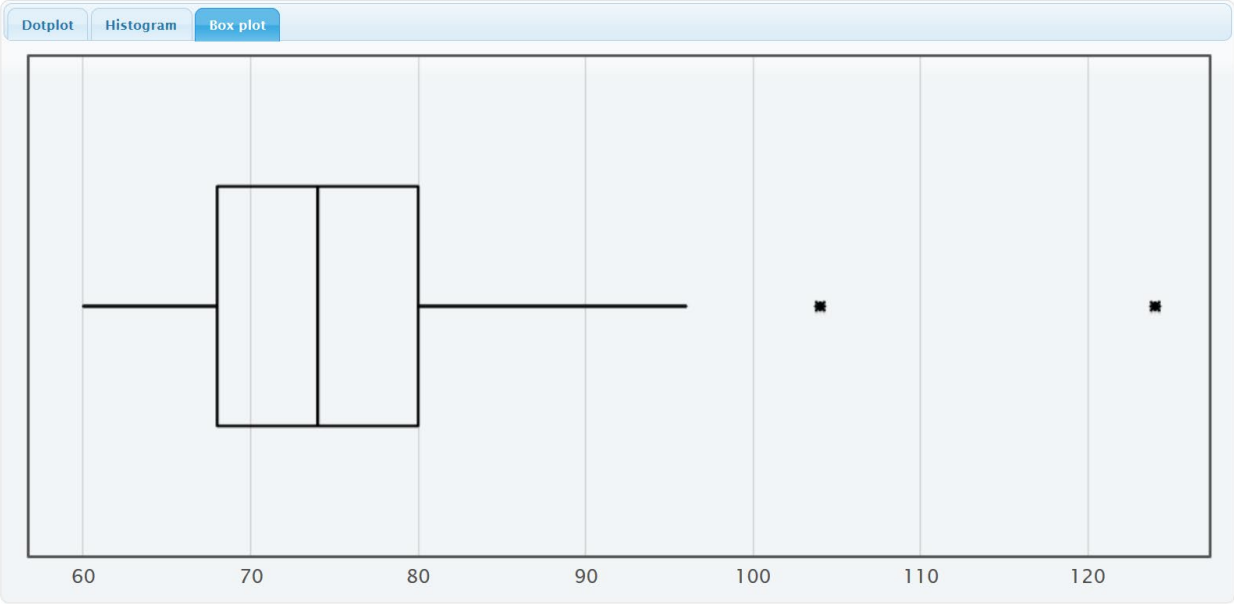
Average: The average pulse rate for these women is 74 bpm (median).

Typical spread: IQR = $Q_3 - Q_1$ = 80 – 68 = 12 bpm
Typical women in the data set had a pulse rate within 12 bpm of each other.

Typical Values: Typical pulse rates are between 68 bpm (Q_1) and 80 bpm (Q_3).

Finding outliers for non-normal data

To find outliers for data sets that are not normally distributed, we will introduce another graph. The graph is called a "box and whisker plot" or "box plot" for short.



A box plot is a graph of the first quartile, median, third quartile and outliers. It is the perfect graph to look at when a data set is not normal. The left of the box is Q_1(68 bpm) and far right of the box is Q_3(80 bpm). So the box represents the typical values (middle 50%). The line inside the box is the median average of 74 bpm. The lines that go to the left and right of the box are called whiskers. The whiskers go to the lowest and highest numbers in the data set that are not unusual (not outliers). The outliers are usually denoted by stars in StatKey and circles and triangles in Statcato. See the two stars the far right. Those are both outliers. There are two unusually high pulse rates in the data set. In StatKey, you can hold your curser over the stars and they will tell you what the numbers are. In this

case, the two high outliers are at 104 bpm and 124 pbm. There are no unusually low values since we do not see any stars on the left of the graph.

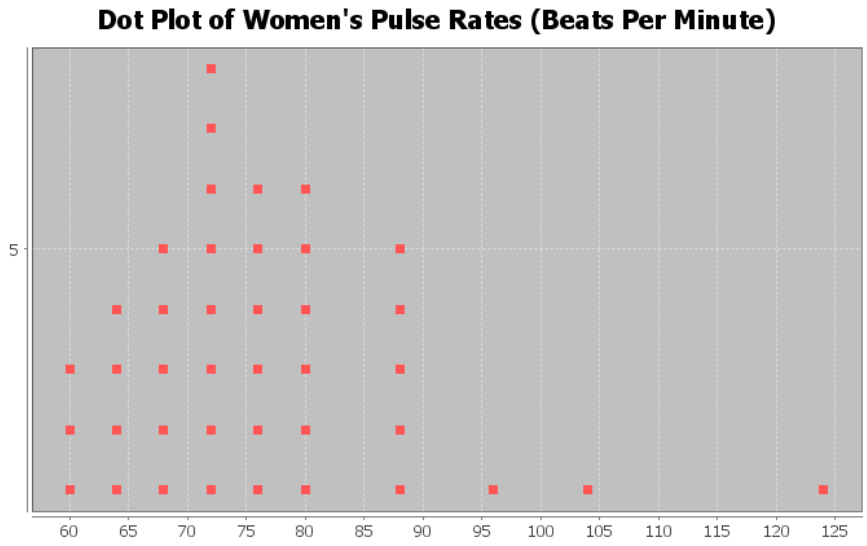
In case you are wondering, here are the formulas used by computer programs to determine outliers in a box plot. You do not need to calculate this yourself. The computer has already found your unusual values.

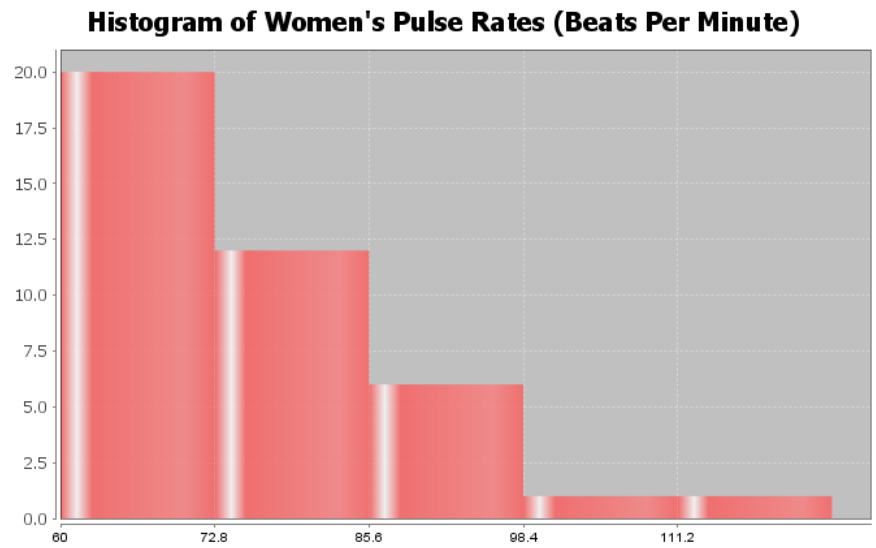
Unusual high (high outlier) cutoff: $Q_3 + (1.5IQR)$

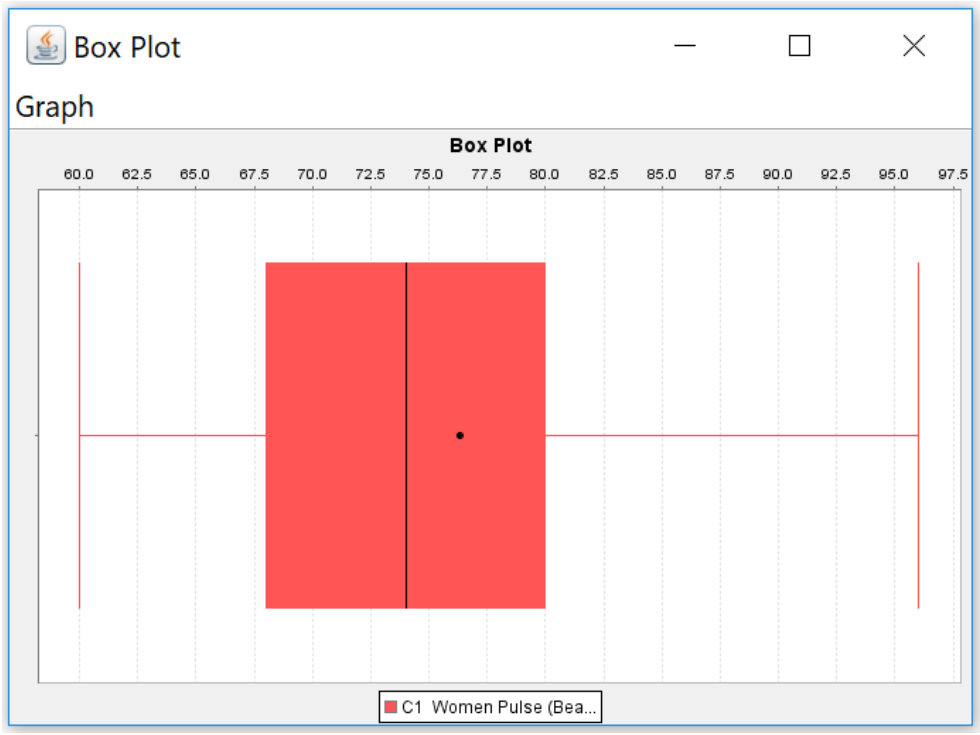
Unusual low (low outlier) cutoff: $Q_1 - (1.5IQR)$

Note about box plots and normal data: Remember, a box plot is a graph of the quartiles and the median. They work really well for data that is not normal. However, they do not show the mean or standard deviation, so it is important to be careful how you interpret box plots for normal data. Normal data has different characteristics than those shown on a box plot. For example, typical values for normal data are <u>not</u> between Q_1 and Q_3. In addition, the outlier cutoffs are different for normal data so there may be differences in what is considered an outlier.

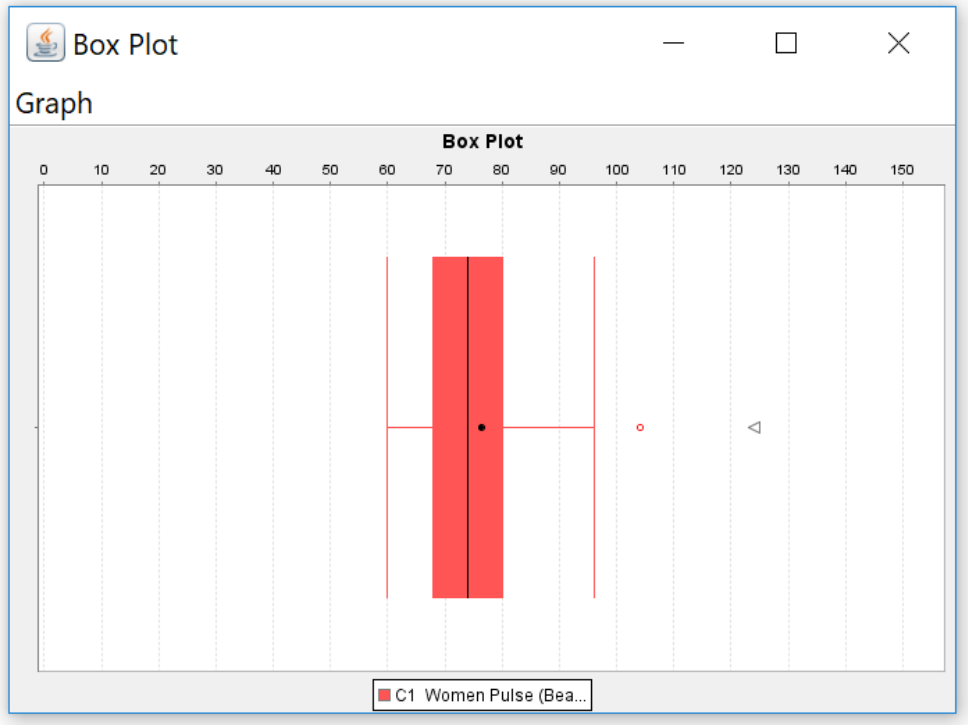
In the last section, we saw that we could also calculate dot plots, histograms, box plots and summary statistics with Statcato. Copy and paste the data into a column of Statcato. Then go to the graph menu and click on "dot plot", "histogram" or "box plot".







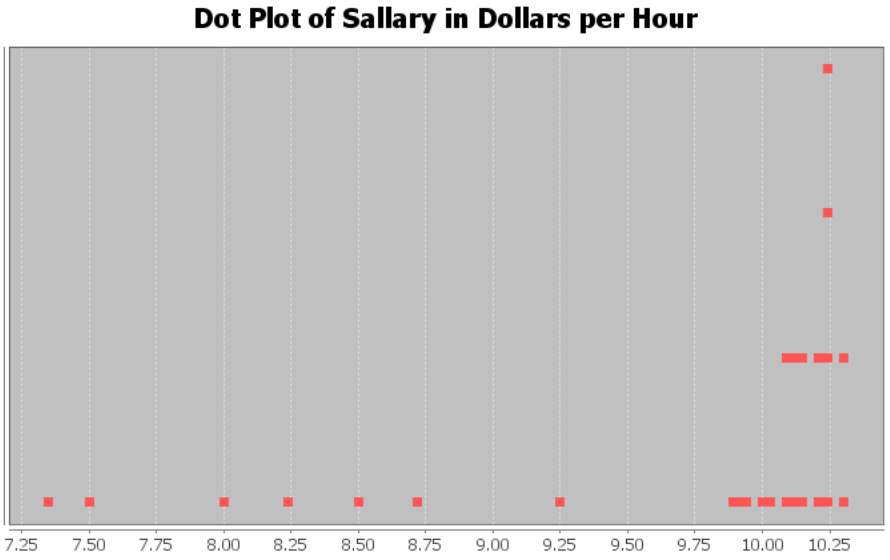
Notice that something is wrong with the Statcato box plot. The outliers have been left off. This is a common problem. To fix this, right click on the box-plot. Click on "zoom out" and "range axis". You may have to do this multiple times. You want to be able to see the minimum value (60 bpm) and maximum value (124 bpm) on the scale of the graph. Here is the correct box plot.

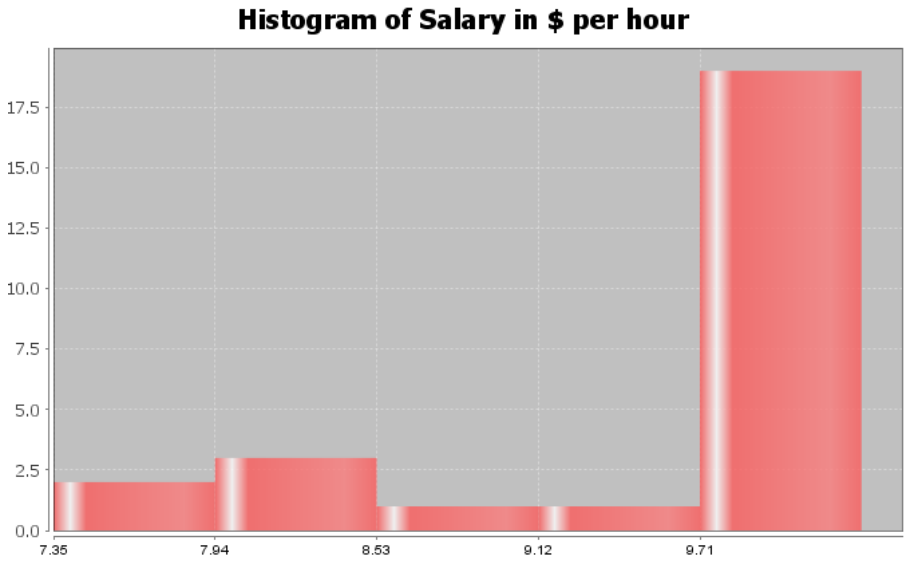


Notice Statcato designated 104 with a circle (regular outlier) and 124 with a triangle (far out outlier). The dot in the middle of the box plot is the mean. Most box plots do not have the mean, but Statcato puts it in so that you can compare it to the median.

Let us look at some other examples.

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.



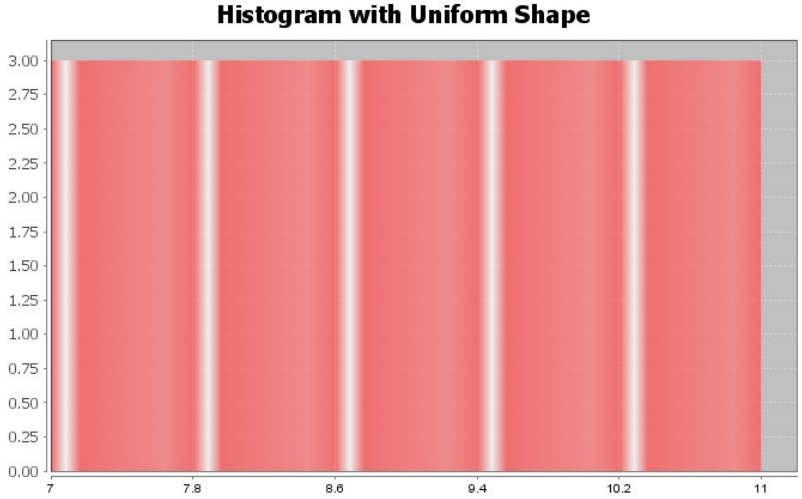


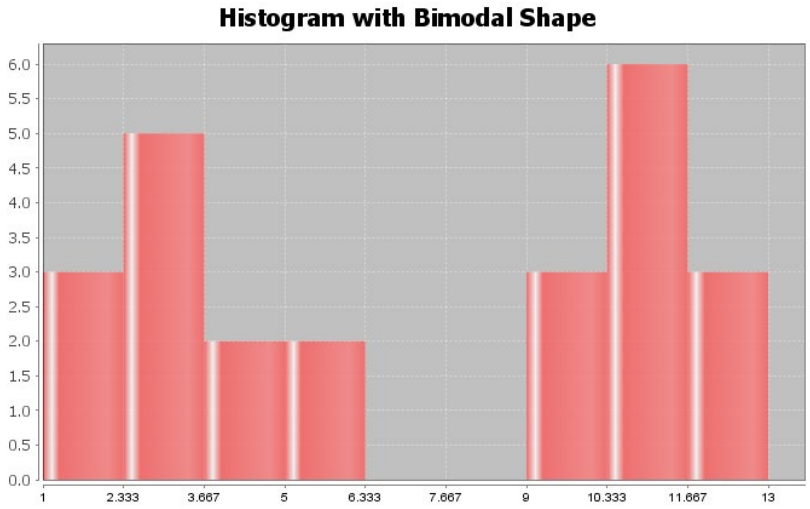
Notice the highest bar and most dots are on the far right, while there is a long tail to the left. Therefore, this is called "skewed left" or "negatively skewed".

Note: *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 two bars to the right and three 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 two bars to the right of the highest hill and seven 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".





Summary Statistics: Measures of Center, Spread and Position
Though the mean, median, standard deviation and IQR are used most often in data analysis, there are many different types of statistics that can 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.

Measures of Center

Mean Average: The balancing point in terms of distances. The measure of center or average used when a data set is bell shaped (normal).

Median Average: The center of the data in terms of order. Also called the second quartile (Q2) or the 50th 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. (Max + Min) / 2

Measures of Spread
Standard Deviation: How far typical values are from the mean in a bell shaped data set. It is the most accurate measure of spread for bell shaped 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. 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 bell shaped.

Range: A quick measure of spread that is not very accurate. It is based on unusual values and does not measure typical values in the data set. It can be calculated quickly without a computer. (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. Should not be used when data is bell shaped. (Q3-Q1)

Measures of Position

Minimum: The smallest number in the data set. Is sometimes classified as an unusual value (outlier).

Maximum: The largest number in the data set. Is sometimes classified as an unusual value (outlier).

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.

Frequency or Sample Size (n)
The 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.

Practice Problems Section 1G

1. Answer the following questions:

    a) Describe a skewed right shape?
    b) Describe a skewed left shape?
    c) Define the median average and explain how it is calculated.
    d) Define the first quartile (Q_1) and explain how it is calculated.
    e) Define the third quartile (Q_3) and explain how it is calculated.
    f) Define the interquartile range (IQR) and explain how it is calculated.

2. Answer the following questions:

    a) If a data set is not normally distributed, what measure of average should we use?

    b) If a data set is not normally distributed, what measure of typical spread should we use?

    c) If a data set is not normally distributed, what are the two statistics that typical values are in between?

    d) If a data set is not normally distributed, approximately what percentage is typical?

    e) If a data set is not normally distributed, how can we use a box plot to find high outliers in the data set?

    f) If a data set is not normally distributed, how can we use a box plot to find low outliers in the data set?

(#3-7) Directions: Analyze the following data sets. Go to www.matt-teachout.org, click on the "Statistics" tab, and then the "Data Sets" tab. Open the "Bear" data, the "Health" data, and the "Car" data. Go to www.lock5stat.com and copy and open StatKey. Under the "Descriptive Statistics and Graphs" menu, click on "One Quantitative Variable". Click on "Edit Data" and copy and paste the indicated data set. Use the graphs and summary statistics to answer the following questions.

3. Bear ages (months)

    a) What is the data measuring and what are the units?
    b) How many numbers are in the data set?
    c) What is the shape of the data set?
    d) What is the minimum value?
    e) What is the maximum value?

f) What is the average (center)? *(Give the number and the name of the statistic used.)*
g) How much typical spread does the data set have?
(Give the number and the name of the statistic used.)
h) Find two numbers that typical values fall in between.
i) List all high outliers in this data set. If there are no high outliers, put "none".
j) List all low outliers in this data set. If there are no high outliers, put "none".

4. Bear Weights (pounds)

a) What is the data measuring and what are the units?
b) How many numbers are in the data set?
c) What is the shape of the data set?
d) What is the minimum value?
e) What is the maximum value?
f) What is the average (center)? *(Give the number and the name of the statistic used.)*
g) How much typical spread does the data set have?
(Give the number and the name of the statistic used.)
h) Find two numbers that typical values fall in between.
i) List all high outliers in this data set. If there are no high outliers, put "none".
j) List all low outliers in this data set. If there are no high outliers, put "none".

5. Women's Systolic Blood Pressure in millimeters of mercury (mm of Hg)

a) What is the data measuring and what are the units?
b) How many numbers are in the data set?
c) What is the shape of the data set?
d) What is the minimum value?
e) What is the maximum value?
f) What is the average (center)? *(Give the number and the name of the statistic used.)*
g) How much typical spread does the data set have?
(Give the number and the name of the statistic used.)
h) Find two numbers that typical values fall in between.
i) List all high outliers in this data set. If there are no high outliers, put "none".
j) List all low outliers in this data set. If there are no high outliers, put "none".

6. Men's Diastolic Blood Pressure (mm of Hg)

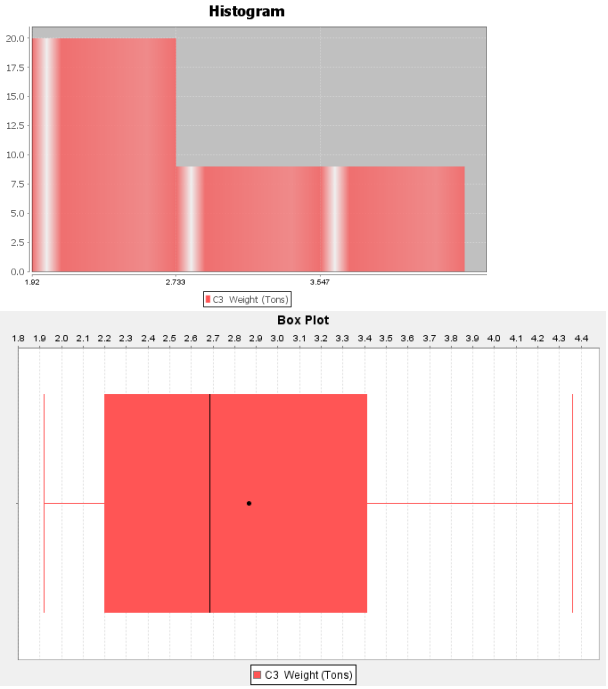
a) What is the data measuring and what are the units?
b) How many numbers are in the data set?
c) What is the shape of the data set?
d) What is the minimum value?
e) What is the maximum value?
f) What is the average (center)? *(Give the number and the name of the statistic used.)*
g) How much typical spread does the data set have?
(Give the number and the name of the statistic used.)
h) Find two numbers that typical values fall in between.
i) List all high outliers in this data set. If there are no high outliers, put "none".
j) List all low outliers in this data set. If there are no high outliers, put "none".

7. Women's Cholesterol in milligrams per deciliter (mg/dL)

a) What is the data measuring and what are the units?
b) How many numbers are in the data set?
c) What is the shape of the data set?
d) What is the minimum value?
e) What is the maximum value?
f) What is the average (center)? *(Give the number and the name of the statistic used.)*
g) How much typical spread does the data set have?
(Give the number and the name of the statistic used.)
h) Find two numbers that typical values fall in between.
i) List all high outliers in this data set. If there are no high outliers, put "none".
j) List all low outliers in this data set. If there are no high outliers, put "none".

(#8-12) The following graphs and summary statistics were created from the "Car" data at www.matt-teachout.org and Statcato. Use the Statcato graphs and summary statistics to answer the following questions.

8. Weight of various cars in tons.



Descriptive Statistics

Variable	Mean	Standard Deviation
Weight (Tons)	2.864	0.706

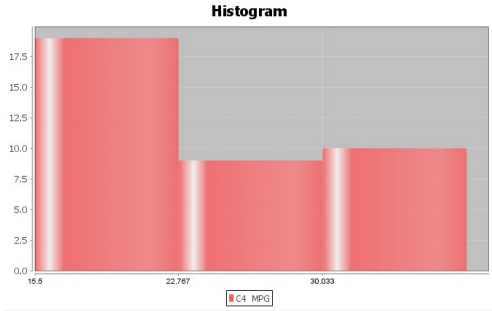
Variable	Q1	Median	Q3	IQR
Weight (Tons)	2.198	2.685	3.46	1.262

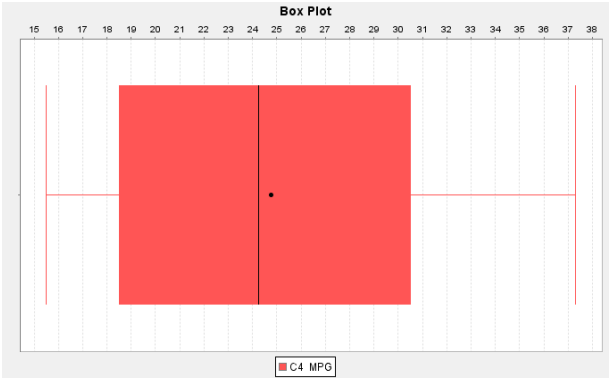
Variable	Min	Max	Range
Weight (Tons)	1.92	4.36	2.440

Variable	N total
Weight (Tons)	38

a) What is the data measuring and what are the units?
b) How many numbers are in the data set?
c) What is the shape of the data set?
d) What is the minimum value?
e) What is the maximum value?
f) What is the average (center)? *(Give the number and the name of the statistic used.)*
g) How much typical spread does the data set have?
(Give the number and the name of the statistic used.)
h) Find two numbers that typical values fall in between.
i) List all high outliers in this data set. If there are no high outliers, put "none".
j) List all low outliers in this data set. If there are no high outliers, put "none".

9. Gas mileage of various cars in miles per gallon (mpg).





Descriptive Statistics

Variable	Mean	Standard Deviation
MPG	24.761	6.547

Variable	Q1	Median	Q3	IQR
MPG	18.425	24.25	30.6	12.175

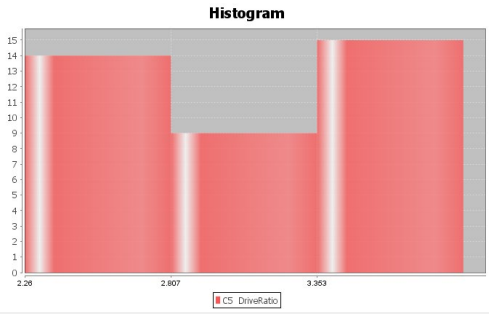
Variable	Min	Max	Range
MPG	15.5	37.3	21.800

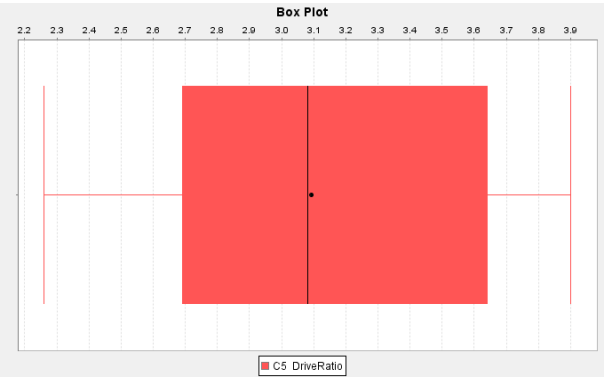
Variable	N total
MPG	38

a) What is the data measuring and what are the units?
b) How many numbers are in the data set?
c) What is the shape of the data set?
d) What is the minimum value?
e) What is the maximum value?
f) What is the average (center)? *(Give the number and the name of the statistic used.)*

g) How much typical spread does the data set have?
(Give the number and the name of the statistic used.)
h) Find two numbers that typical values fall in between.
i) List all high outliers in this data set. If there are no high outliers, put "none".
j) List all low outliers in this data set. If there are no high outliers, put "none".

10. The drive ratio of various cars.





Descriptive Statistics

Variable	Mean	Standard Deviation
DriveRatio	3.093	0.518

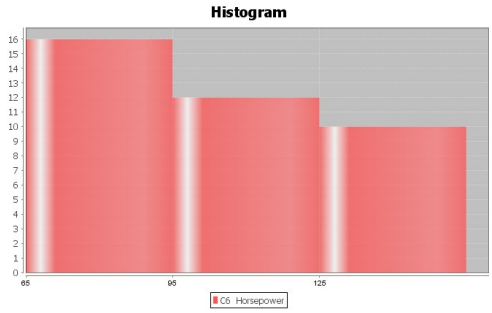
Variable	Q1	Median	Q3	IQR
DriveRatio	2.69	3.08	3.655	0.965

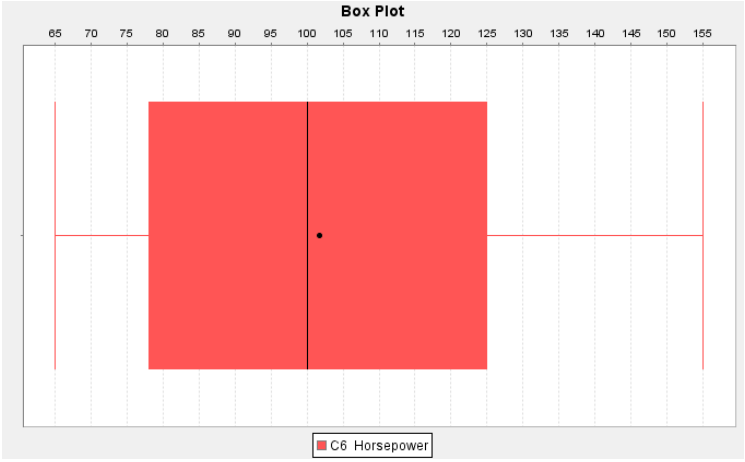
Variable	Min	Max	Range
DriveRatio	2.26	3.9	1.640

Variable	N total
DriveRatio	38

a) What is the data measuring and what are the units?
b) How many numbers are in the data set?
c) What is the shape of the data set?
d) What is the minimum value?
e) What is the maximum value?
f) What is the average (center)? *(Give the number and the name of the statistic used.)*
g) How much typical spread does the data set have?
(Give the number and the name of the statistic used.)
h) Find two numbers that typical values fall in between.
i) List all high outliers in this data set. If there are no high outliers, put "none".
j) List all low outliers in this data set. If there are no high outliers, put "none".

11. The horsepower of various cars.





Descriptive Statistics

Variable	Mean	Standard Deviation
Horsepower	101.737	26.445

Variable	Q1	Median	Q3	IQR
Horsepower	77.25	100.0	125.0	47.75

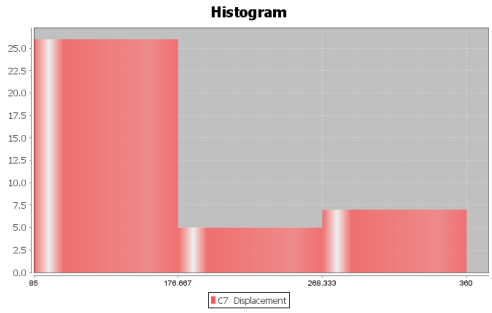
Variable	Min	Max	Range
Horsepower	65.0	155.0	90.0

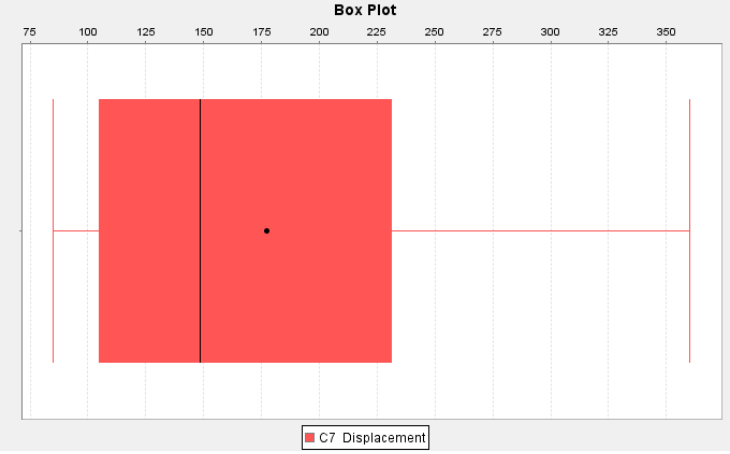
Variable	N total
Horsepower	38

a) What is the data measuring and what are the units?
b) How many numbers are in the data set?
c) What is the shape of the data set?
d) What is the minimum value?

e) What is the maximum value?
f) What is the average (center)? *(Give the number and the name of the statistic used.)*
g) How much typical spread does the data set have?
(Give the number and the name of the statistic used.)
h) Find two numbers that typical values fall in between.
i) List all high outliers in this data set. If there are no high outliers, put "none".
j) List all low outliers in this data set. If there are no high outliers, put "none".

12. The measure of displacement for various cars.





Descriptive Statistics

Variable	Mean	Standard Deviation
Displacement	177.289	88.877

Variable	Q1	Median	Q3	IQR
Displacement	103.25	148.5	237.75	134.5

Variable	Min	Max	Range
Displacement	85.0	360.0	275.0

Variable	N total
Displacement	38

a) What is the data measuring and what are the units?
b) How many numbers are in the data set?
c) What is the shape of the data set?
d) What is the minimum value?
e) What is the maximum value?
f) What is the average (center)? *(Give the number and the name of the statistic used.)*
g) How much typical spread does the data set have?
(Give the number and the name of the statistic used.)
h) Find two numbers that typical values fall in between.
i) List all high outliers in this data set. If there are no high outliers, put "none".
j) List all low outliers in this data set. If there are no high outliers, put "none".

13. Classify each of the following statistics as a measure of center, spread or position.

    a) Q1
    b) Mean
    c) Variance
    d) Standard Deviation
    e) Minimum
    f) Q3
    g) Mode
    h) IQR
    i) Median
    j) Range
    k) Maximum
    l) Midrange

14. Define each of the following statistics and describe when that statistic should be used.

    a) Q1
    b) Mean
    c) Variance
    d) Standard Deviation
    e) Minimum
    f) Q3
    g) Mode
    h) IQR
    i) Median
    j) Range
    k) Maximum
    l) Midrange

Chapter 1 Review

Key Vocabulary Terms

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.

Statistic: A number calculated from sample data in order to understand the characteristics of the data. For example, a sample mean average, a sample standard deviation, or a sample percentage.

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.

Cluster Sample: Sample data that collects data from groups of people in a population instead of one at a time. The groups should be chosen randomly to avoid sampling bias.

Stratified Sample: Sample data used to compare two or more groups or compare two or more populations. The individuals from each group should be chosen randomly to avoid sampling bias. For example, we may take a random sample of people living in Palmdale, CA and another random sample of people living in Valencia, CA and use the data to compare the average salaries.

Systematic Sample: Sample data that is collected with some type of system like choosing every twentieth person on a list.

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.

Percentage: A statistic calculated from categorical data that measures the part out of 100.

Proportion: The decimal equivalent to a percentage.

Sample Size (n): Also called the sample frequency or sample count. This is the number of people or objects represented in the sample data. If we collected data from 35 people, then the sample size would be n = 35.

Mean: A measure of center or average for quantitative data that balances the distances. The mean average is only accurate if the quantitative data is normal (bell shaped). Hence, the mean average is the center or average used for normal quantitative data.

Median: A measure of center or average for quantitative data that is found by finding the center of the data when the data values are in order. The median is the most accurate center or average when the data is skewed left, skewed right, or not normal.

Mode: The number or numbers that appear most often in a quantitate data set. The mode is used as a measure of center or average.

Midrange: A quick measure of center or average that is half way between the min and max of a quantitative data set. It is generally not very accurate, but easy to calculate.

Standard Deviation: The most accurate measure of typical spread for normal (bell shaped) quantitative data. The standard deviation measures how far typical values are from the mean on average. The standard deviation is only accurate if the quantitative data is normal (bell shaped).

Variance: A measure of spread used in ANOVA testing. The variance is the square of the standard deviation and is only accurate when data is normal (bell shaped).

Interquartile Range (IQR): The most accurate measure of spread for skewed or non-normal data. The IQR measures how far typical values are from each other in skewed or non-normal data. IQR is calculated by subtracting the Third Quartile (Q3) minus the First Quartile (Q1).

Range: A quick measure of spread that measures the distance between the max and min of a quantitative data set. It is easy to calculate (Max – Min), but is not an accurate measure of typical spread, since it does not involve typical values in the data.

First Quartile (Q1): The divider that approximately 25% of the quantitative data values are less than. Q1 is the bottom range of typical values for skewed or non-normal data. Typical values are between Q1 and Q3 in skewed or non-normal data. Q1 is considered a measure of position.

Third Quartile (Q3): The divider that approximately 75% of the quantitative data values are less than. Q3 is the top range of typical values for skewed or non-normal data. Typical values are between Q1 and Q3 in skewed or non-normal data. Q3 is considered a measure of position.

Max: The largest number in a quantitative data set. Considered a measure of position.

Min: The smallest number in a quantitative data set. Considered a measure of position.

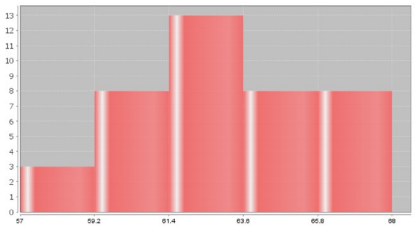
Categorical Data Analysis

- **To convert a decimal proportion into a percentage => Multiply by 100 and put on the % symbol.** (This will move the decimal two places to the right.)
- **To convert a percentage into a decimal proportion => Remove the % symbol and Divide by 100.** (This will move the decimal two places to the left.)
- **To calculate the proportion for each categorical variable: Proportion =** $\frac{x}{n} = \frac{Amount\ (\#\ of\ successes)}{Total\ Frequency\ (Sample\ Size)}$ (StatKey calculate counts and proportions for you.)
- **Round Proportions to the thousandths place.** (Three numbers to the right of decimal point. StatKey round to the thousandths place.)

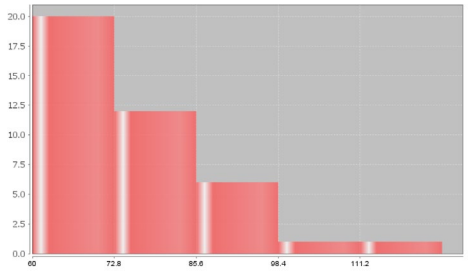
Quantitative Data Analysis

Shapes

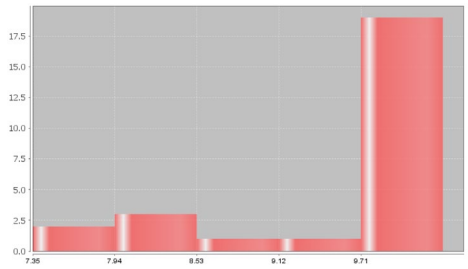
Normal (Bell Shaped, Unimodal and Symmetric)



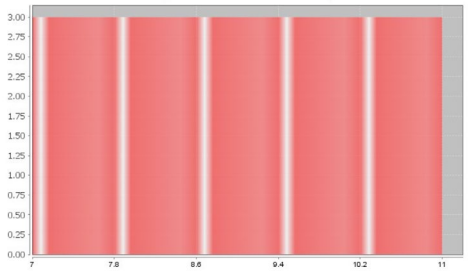
Skewed Right (Positively Skewed)



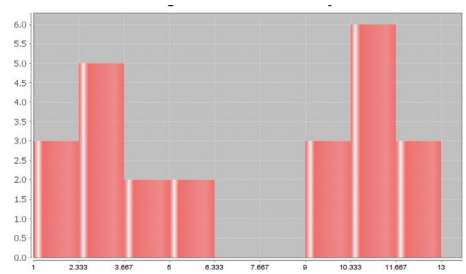
Skewed Left (Negatively Skewed)



Uniform



Bimodal



Shape determine what statistics are accurate!

Normal Quantitative Data

Center (Average): Mean

Typical Spread: Standard Deviation

Typical Values Between: Mean – Standard Deviation and Mean + Standard Deviation
(One Standard Deviation above and below the mean, Middle 68% of the data)

High Outliers: Data Values ≥ Mean + (2 x Standard Deviation)
(Two standard deviations above the mean, Top 2.5% of the data)

Low Outliers: Data Values ≤ Mean − (2 x Standard Deviation)
(Two standard deviations below the mean, Bottom 2.5% of the data)

Skewed or Non-normal Quantitative Data

Center (Average): Median

Typical Spread: Interquartile Range (IQR)

Typical Values Between: 1st quartile (Q1) and 3rd quartile (Q3)
(Middle 50% of data values)

High Outliers: Data Values ≥ Q3 + (1.5 x IQR) *Automatically calculated in Box-Plot!*

Low Outliers: Data Values ≥ Q1 − (1.5 x IQR) *Automatically calculated in Box-Plot!*

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.

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?

7. Explain the following.

a) Explain how to round a decimal to a given place value.
b) Explain how to convert a decimal proportion into a percentage.
c) Explain how to convert a percentage into a decimal proportion.
d) Explain how to calculate a percentage by using an amount and a total from categorical data.
e) Explain how to calculate an estimated amount by using a percentage and a total from categorical data.

8. Convert the following proportions into percentages. Do not round your answer.

a) 0.0722
b) 0.0041
c) 0.563
d) 0.0005

9. Convert the following percentages into decimal proportions. Do not round your answer.

    a) 35.9%
    b) 4.823%
    c) 0.026%
    d) 0.389%

10. A company has 74 employees. Of those employees 11 are managers, 27 are full-time employees and 36 are part-time employees. Use this information to answer the following questions.

    a) What proportion of the employees are managers? *(Round your answer to the thousandths place.)*
    b) What percentage of the employees are managers? *(Round your answer to the tenths place.)*
    c) What proportion of the employees are full-time employees? *(Round your answer to the thousandths place.)*
    d) What percentage of the employees are full-time employees? *(Round your answer to the tenths place.)*
    e) What proportion of the employees are part-time employees? *(Round your answer to the thousandths place.)*
    f) What percentage of the employees are part-time employees? *(Round your answer to the tenths place.)*
    g) Calculate the percent of increase between managers and full-time employees. Is there a significant difference between the percentages? Explain why.
    h) Calculate the percent of increase between full-time and part-time employees. Is there a significant difference between the percentages? Explain why.

11. According to an online article, approximately 60% of the voting population in the U.S. votes during a presidential election year. According to a census, there are approximately 41,743 people living in Saugus, CA. If 60% of them vote in the next presidential election, how many people do we expect to vote in Saugus?

12. Describe and draw a histogram for each of the following shapes.

    a) Normal
    b) Skewed Right
    c) Skewed Left
    d) Uniform
    e) Bimodal

13. Classify each of the following quantitative statistics as a measure of center, spread or position. Also, describe when that statistic should be used.

    a) Q1
    b) Mean
    c) Variance
    d) Standard Deviation
    e) Minimum
    f) Q3
    g) Mode
    h) IQR
    i) Median
    j) Range
    k) Maximum
    l) Midrange

14. Answer each of the following questions about quantitative data analysis.

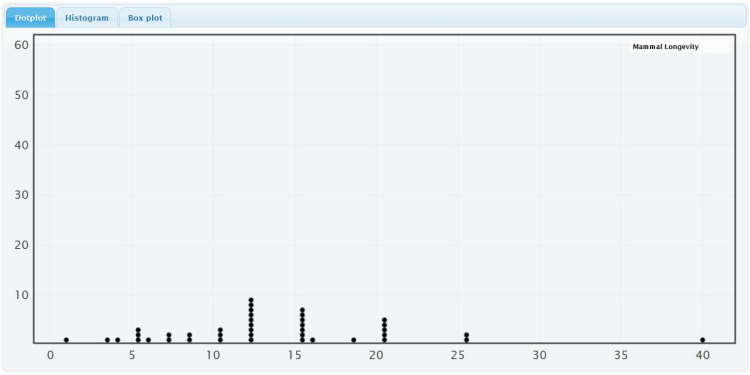
    a) What measure of center (average) should we use if the data is normal?
    b) What measure of center (average) should we use if the data is not normal?
    c) What measure of spread (variability) should we use if the data is normal?
    d) What measure of spread (variability) should we use if the data is not normal?

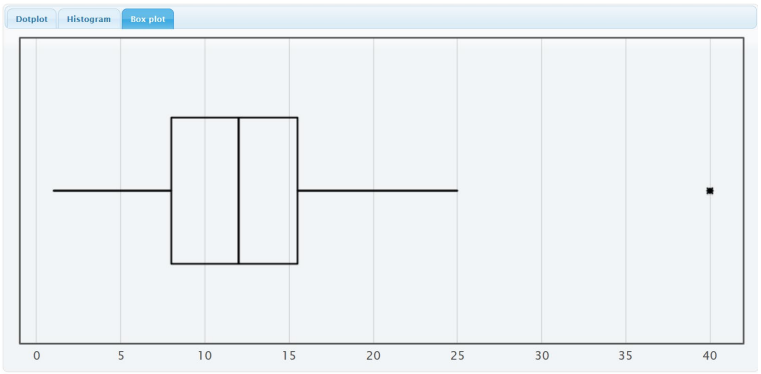
e) How do we find two numbers that typical values fall in between if the data is normal?
f) How do we find two numbers that typical values fall in between if the data is not normal?
g) What is the formula for the high outlier cutoff if the data is normal?
h) What is the formula for the high outlier cutoff if the data is not normal?
i) What is the formula for the low outlier cutoff if the data is normal?
j) What is the formula for the low outlier cutoff if the data is not normal?
k) How do we determine if a data value is an outlier when the data is normal?
l) How do we determine if a data value is an outlier when the data is not normal?

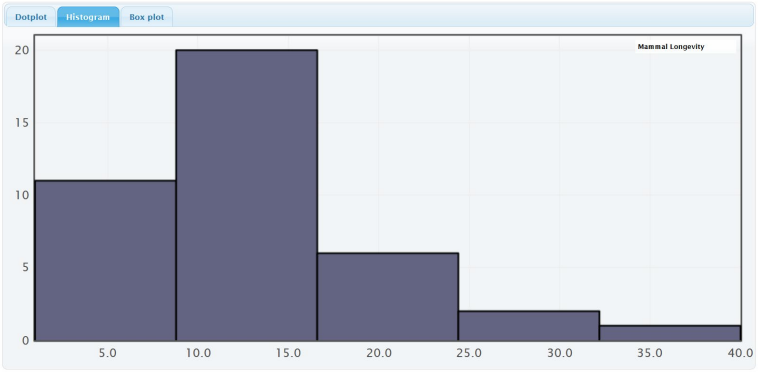
15. The following graphs and statistics were calculated with StatKey and describe the number of years mammals live. Use the graphs and statistics to answer the following questions.

Summary Statistics

Statistic	Value
Sample Size	40
Mean	13.150
Standard Deviation	7.245
Minimum	1
Q_1	8.000
Median	12.000
Q_3	15.500
Maximum	40







a) What is the data measuring and what are the units?
b) How many mammals are in the data set?
c) What is the shape of the data set?
d) What is the minimum value?
e) What is the maximum value?
f) What is the average (center)? *(Give the number and the name of the statistic used.)*
g) How much typical spread does the data set have?
(Give the number and the name of the statistic used.)
h) Find two numbers that typical values fall in between.
i) List all high outliers in this data set. If there are no high outliers, put "none".
j) List all low outliers in this data set. If there are no high outliers, put "none".

16. The following graphs and statistics were calculated with Statcato and describe the number of years employees have been employed at a company. Use the graphs and statistics to answer the following questions.

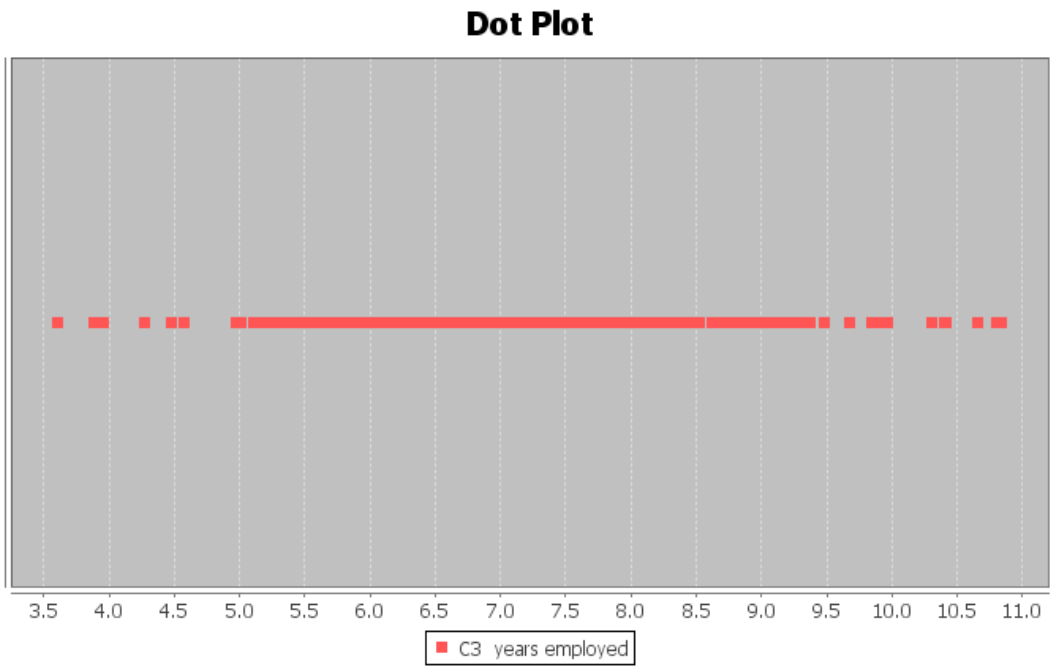
Descriptive Statistics

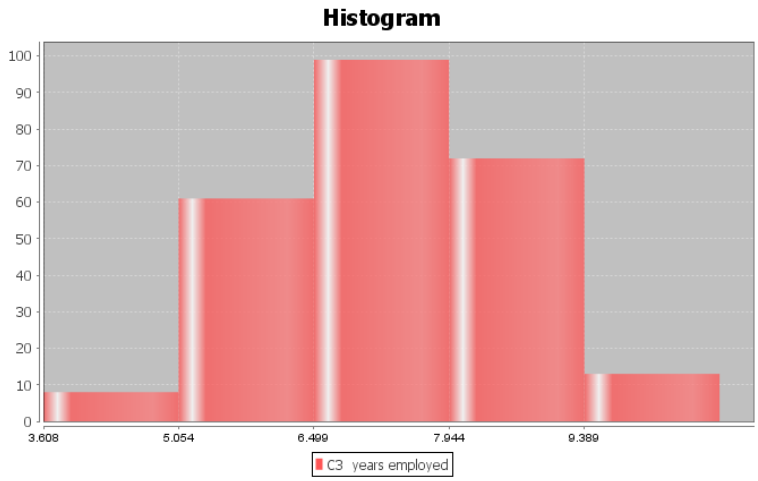
Variable	Mean	Standard Deviation
years employed	7.345	1.376

Variable	Q1	Median	Q3	IQR
years employed	6.4	7.35	8.3	1.9

Variable	Min	Max	Range
years employed	3.6	10.8	7.2

Variable	N total
years employ	253

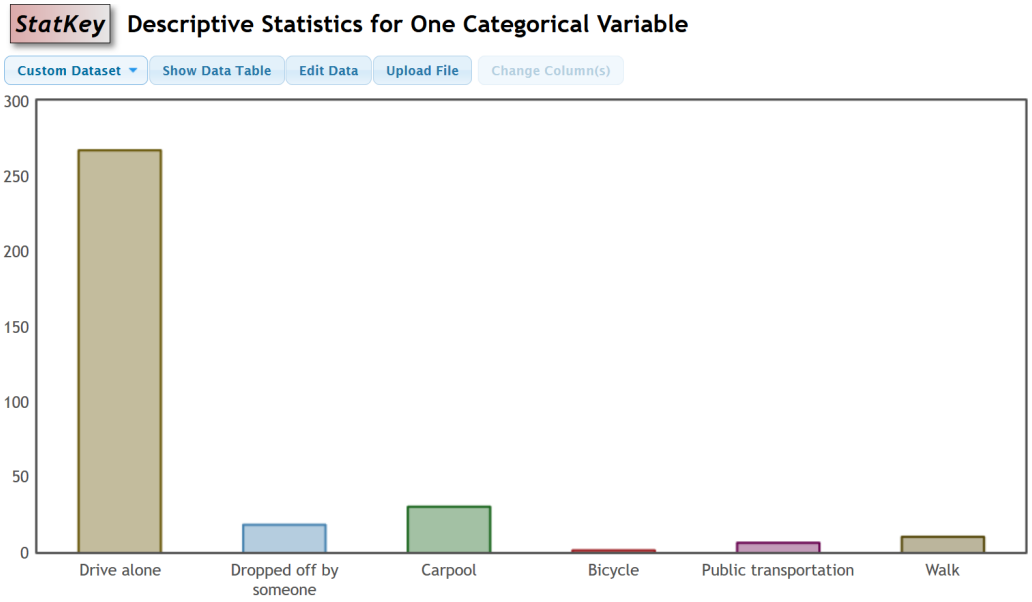




a) What is the data measuring and what are the units?
b) How many employees are in the data set?
c) What is the shape of the data set?
d) What is the minimum value?
e) What is the maximum value?
f) What is the average (center)? *(Give the number and the name of the statistic used.)*
g) How much typical spread does the data set have?
(Give the number and the name of the statistic used.)
h) Find two numbers that typical values fall in between.
i) Calculate the high-outlier cutoff. Give approximate values of the high outliers in this data set. If there are no high outliers, put "none".
i) Calculate the low-outlier cutoff. Give approximate values of the low outliers in this data set. If there are no low outliers, put "none".

17.

Statistics students were asked what mode of transportation they take to get to school. Use the following bar chart and statistics to answer the following.



Summary Statistics

	Count	Proportion
Drive alone	267	0.804
Dropped off by someone	18	0.054
Carpool	30	0.09
Bicycle	1	0.003
Public transportation	6	0.018
Walk	10	0.03
Total	332	1.000

a) What was the most population mode of transportation?

b) What was the least population mode of transportation?

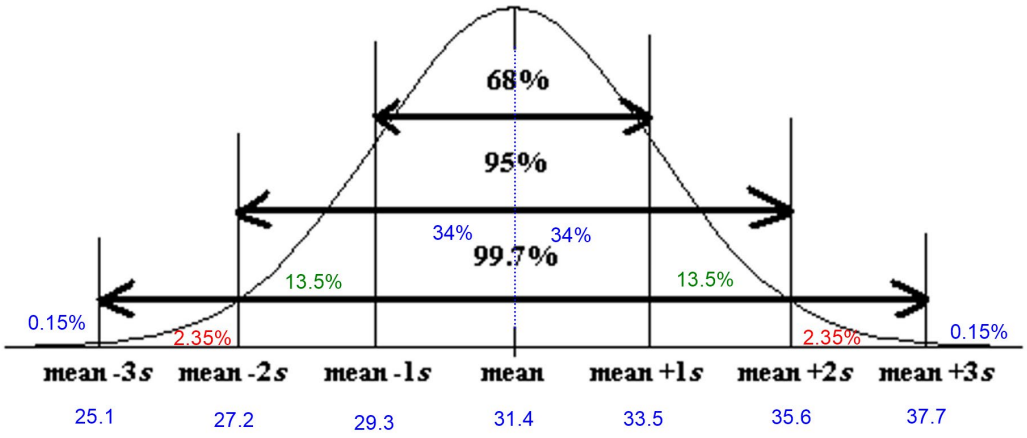
c) How many statistics students walked to school?

d) What proportion of statistics students were dropped off by someone? Do not calculate the answer. Use the table provided. Do not round the answer.

e) What percentage of the statistics students use public transportation? Do not calculate the answer. Use the table provided and convert the answer into a percentage. Do not round the answer.

18.

The salaries of employees at a company are normally distributed with a mean of 31.4 thousand dollars and a standard deviation of 2.1 thousand dollars. Use the Empirical Rule graph below to answer the following questions about this normal quantitative data.



a) What percent of the salaries are between 29.3 thousand dollars and 31.4 thousand dollars?

b) What percent of the salaries are 33.5 thousand dollars or more?

c) Typical salaries are between what two values?

d) What is the high outlier cutoff?

e) What is the low outlier cutoff?

f) What is the average salary for employees of this company?

Appendix A: Answers to Selected Exercises Chapter 1

Section 1A Answers to Odd Problems

1. Bear Ages: Quantitative, Months
   Bear Month Data Taken: Categorical, 8 options (May-November)
   Bear Gender: Categorical, 2 options
   Head Length: Quantitative, Inches
   Head Width: Quantitative, Inches
   Neck Circumference: Quantitative, Inches
   Length: Quantitative, Inches
   Chest: Quantitative, Inches
   Weight: Quantitative, Pounds

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 to Odd Problems

1. Population of Interest: All students at the college.
   Method: Voluntary Response
   Will not represent the population very well. There is sampling Bias, since the individuals were not chosen randomly.

3. Population of Interest: All students at the high school.
   Method: Convenience
   Will not represent the population very well. There is sampling Bias, since the individuals were not chosen randomly.

5. Population of Interest: All people in Rachael's home town.
   Method: Systematic
   Might represent the population. There is sampling bias, since the individuals were not chosen randomly, but it may be representative since the whole population was on the list. This data is not as biased as convenience or voluntary response, but not as good as a random sample.

7. Population of Interest: All employees at the company.
   Method: Census
   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.

9. Population of Interest: All people in Toronto.
Method: Simple Random Sample
Will represent the population well as long as there is no other types of bias present. No sampling bias.

11. Population of Interest: All people that use smart phones.
Method: Voluntary Response
Will not represent the population very well. Sampling bias, since the individuals were not chosen randomly.

13. Population of Interest: All teenagers and adults.
Method: Stratified since they are comparing groups.
Will represent the population well as long as there is no other types of bias present. Individuals were chosen randomly, so no sampling bias.

15. Population of Interest: All adults in North Carolina
Method: Systematic and Convenience
Will not represent the population very well. There is sampling bias since the individuals were not selected randomly. This data is particularly bad since most of the population has no opportunity to enter the store.

Section 1C Answers to Odd Problems

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.

3. 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.
Non-response Bias: There will be many people that randomly selected, but refuse to answer the question.

5. 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.

7. 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.

9. Population of interest: All pills made by the company.

Deliberate Bias: They deleted data that poorly reflected the pharmaceutical company.

Section 1D Answers to Selected Problems

1. Explanatory Variable: Having a cell phone or not.
   Response Variable: Ruler catch length in inches or "drop".

2. We needed a control group the measures the classes ability to catch a ruler in general. We can then compare the cell phone data to the control group.

3. The two groups were people with the cell phone (treatment group) and those without a cell phone (control group.) They were perfectly alike in all confounding variables since they were the same exact people measured twice.

4. Answers may vary. Confounding Variables: Age, hand-eye coordination, distractions besides the phone, hand size, ability to text one handed, position of the ruler when dropped, …

Since the same people were measured twice, the two groups had the exact same ages, hand-eye coordination and ability to text one-handed. The amount of distraction was relatively the same with or without the phone. The instructor gave a demonstration so that everyone would hold and drop the ruler the same way.

5. Neither. The explanatory variable was having a cell phone or not. The person knew whether they had a cell phone or not. Not knowing when the ruler would be dropped does not constitute blind since it is the response variable.

6. Answers will vary from class to class. The average catch distance was lower for the no cell phone group, but it is difficult to determine if it is significant at this point. We will learn that later. The number of drops was significantly greater in the cell phone group. Since confounding variables were controlled, we have proven that texting does cause you to drop the ruler more often. Whether this experiment applies to texting while driving is debatable. Some people have said that dropping the ruler may be equivalent to not hitting the breaks in time. Again, this is debatable. The experiment does prove that texting slows reflexes and you do need reflexes when you drive.

7. 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.

8. Experiment: A scientific method for controlling confounding variables and proving cause and effect.

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

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

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

12. 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.

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

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

15. Single Blind: When only the person receiving the treatment does not know if it is real or a placebo.

16. 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.

17. This is an experiment, since they need to prove that the medicine has the desired effect. They must control confounding variables like the amount of motion, genetics, age, diet, pregnancy, etc. If they control all of the confounding variables and the medicine (treatment) group has significantly less motion sickness, they will have succeeded in proving cause and effect.

19. This is an observational study since they just collected data without thought to controlling confounding variables. This can show the number of cases of tuberculosis is related to or associated with the low income, crowed cities, but it will not be able to prove cause and effect.

21. This is an observational study since they just collected data without thought to controlling confounding variables. This can show that obesity is related to or associated with having diabetes, but they will not be able to prove cause and effect. There are many variables involved in determining why someone has diabetes.

Section 1E Answers to Odd Problems and #2

1.

a) 0.75
c) 0.00664
e) 0.397
g) 0.00189
i) 0.0316
k) 0.961
m) 0.00007
o) 0.662
q) 1

2.

a) 5.7%
c) 0.33%
e) 6.13%
g) 0.045%
i) 4.6%
k) 0.27%
m) 0.58%
o) 100%
q) 2.04%

3.

15% = 0.15

Estimated Amount = $0.15 \times 78300 = 11{,}745$

We estimate that approximately 11,745 people in Chino Hills are without health insurance.

5.

9.3% = 0.093

Estimated Amount = $0.093 \times 18400 = 1711.2 \approx 1711$

We estimate that approximately 1,711 students at COC have diabetes.

7.

1.47% = 0.0147

Estimated Amount = $0.0147 \times 136400 = 2005.08 \approx 2005$

We estimate that approximately 2,005 people in Van Nuys have autism.

9.

14.8% = 0.148

Estimated Amount = $0.148 \times 305700 = 45243.6 \approx 45244$

We estimate that approximately 45,244 people in Stockton live below the poverty line.

11.

Athletic Wear: $139/213 \approx 0.653 = 65.3\%$

Traditional Jeans: $74/213 \approx 0.347 = 34.7\%$

Percent of Increase = $(0.653 - 0.347)/0.347 = 0.882 = 88.2\%$ of increase.

The percent of women that prefer athletic wear does seem to be significantly higher than the percent that prefer traditional jeans. It is also practically significant since there was 65 more women in the sample that preferred athletic wear.

13.

Med/Surg: $57/350 \approx 0.163 = 16.3\%$

Telemetry: $49/350 \approx 0.14 = 14\%$

Percent of Increase = $(0.163 - 0.14)/0.14 \approx 0.164 = 16.4\%$ of increase.

The percent of patients admitted to telemetry and med/surge seem very close. The percent of increase is very small. There were also only 8 more patients in Med/Surg than telemetry. These indicate there is no significant difference, practically or statistically.

15.

Medicine: $13/57 \approx 0.228 = 22.8\%$

Placebo: $11/61 \approx 0.180 = 18.0\%$

Percent of Increase = $(0.228 - 0.18)/0.18 \approx 0.267 = 26.7\%$ of increase.

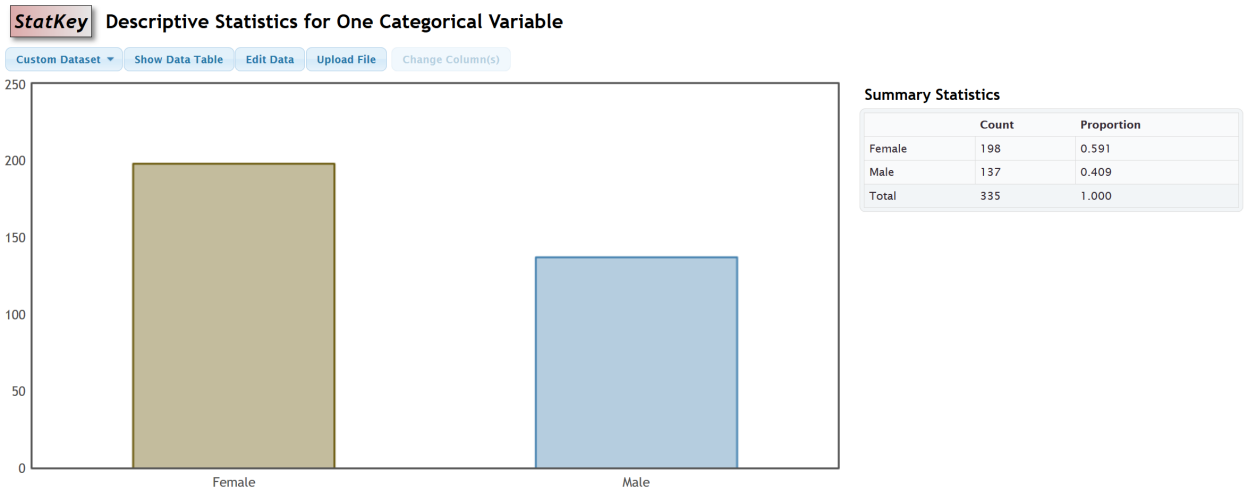
The percent of patients that improved on the medicine is only slightly higher than the placebo group. Practically there is not much difference. Only two more patients on the medicine showed improvement. That is not practically significant.

17.

Proportion of female ≈ 0.591

Proportion of male ≈ 0.409

Percent of Increase = (0.591 – 0.409)/0.409 ≈ 0.445 = 44.5% of increase. This indicates that the percentage of female COC statistics students is significantly higher than the percentage of male students. It is also practically significant since there were 61 more female students than male.

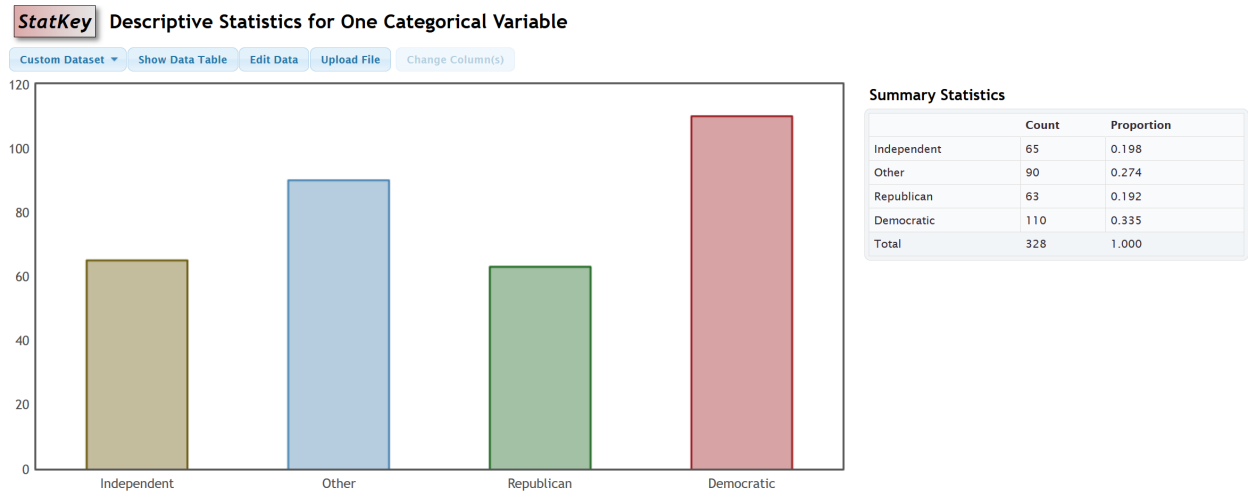


19.

Proportion of Democrat ≈ 0.335

Proportion of Republican ≈ 0.192

Percent of Increase = (0.335 – 0.192)/0.192 ≈ 0.745 = 74.5% of increase. This indicates that the percentage of COC statistics students that are democrat is significantly higher than republican. It is practically significant also since there are 47 more democratic statistics students than republican.



21.

Percentage of cars made in France ≈ 3%

Number of cars made in U.S. = 22

Proportion of cars made in Sweden ≈ 0.05

Proportion of cars made in Japan ≈ 0.18

Proportion of cars made in Germany ≈ 0.13

Percent of Increase (Japan & Germany) ≈ (0.18 – 0.13) / 0.13 ≈ 0.385 = 38.5%

The percent of increase does seem to indicate that the percent of cars made in Japan is significantly higher than Germany. However does not seem to be practically significant since the number of cars made in Japan was only 2 more than Germany. Overall, I would say it is not significant.

23.

Percentage of cereals made by Quaker ≈ 17%

Number of cereals made by Ralston = 3

Proportion of cereals made by General ≈ 0.29

Proportion of cereals made by Kelloggs ≈ 0.33

Proportion of cereals made by Quaker ≈ 0.17

Percent of Increase (Kelloggs & Quaker) ≈ (0.33 – 0.17) / 0.17 ≈ 0.941 = 94.1%

The percent of increase indicates that the percent of cereals made by Kelloggs is significantly higher than Quaker. However does not seem to be practically significant since it was a small sample size and the number of cereals made by Kelloggs was only 4 more than Quaker.

25.

Percentage of cereals on top shelf ≈ 33%

Number of cereals on bottom shelf = 8

Proportion of cereals on middle shelf ≈ 0.33

Percent of Increase (top and bottom shelves) ≈ (0.33-0.33) / 0.33 ≈ 0 = 0%

There is no significant difference between the percentages of cereals put on the top and bottom shelves. They appear to be about the same.

27.

a) 0.122 = 12.2%
b) 0.113 = 11.3%
c) 0.796 = 79.6%
d) 1 – 0.796 = 0.204 = 20.4%
e) 0.110 = 11.0%
f) 1 – 0.110 = 0.89 = 89%

Binomial Distribution: n=84, p=0.12
Input: 11.0
Type: Probability density

X	P(X)
11.0	0.122219

Binomial Distribution: n=84, p=0.12
Input: 8.0
Type: Probability density

X	P(X)
8.0	0.113128

Binomial Distribution: n=84, p=0.12
Input: 12.0
Type: Cumulative probability

X	P(<=X)
12.0	0.796146

Binomial Distribution: n=84, p=0.12
Input: 6.0
Type: Cumulative probability

X	P(<=X)
6.0	0.109721

29.

a) 0 = 0%

b) 1 – 0 = 1 = 100%

Binomial Distribution: n=57, p=0.845
Input: 9.0
Type: Cumulative probability

X	P(<=X)
9.0	0

Section 1F Answers to Odd Problems and #2

1.

a) The shape of the data is relatively bell shaped or unimodal and symmetric. In a histogram, the tallest bars are relatively in the middle and the left and right tail are about the same length.

b) The mean average is the average that we use when the data is normal. It also balances the distances. It is calculated by adding all of the data values and dividing the sum by the sample size n.

c) The standard deviation calculates the average distance that data values are from the mean. It is the measure of spread used when data is normal. To calculate the standard deviation for one quantitative data set, take every number in the data set and subtract the mean from it. Then square are the differences. Add up all the squares and divide by n – 1 where n is the sample size. Now take the square root of the answer. Always have a computer calculate the standard deviation for you.

2.

a) Mean Average

b) Standard Deviation

c) One Standard Deviation or less.

d) Mean ± Standard Deviation

e) 68%

f) Two or more standard deviations.

g) 2.5%

h) 2.5%

3.

a) The data measures the neck circumference of bears. The units are inches.

b) 54 total bears

c) Yes. The data is nearly normal (almost bell shaped).

d) 10 inches

e) 31.5 inches

f) Center = Mean Average = 20.556 inches

g) Typical Spread = Standard Deviation = 5.641 inches

h)

20.556 – 5.641 = 14.915 inches

20.556 + 5.641 = 26.197 inches

Typical bears have a neck circumference between 14.915 inches and 26.197 inches.

i)

Unusual High Cutoff: $20.556 + (2 \times 5.641) = 20.556 + 11.282 = 31.838$ inches.

Any neck circumference of 31.838 inches or more would be considered unusually high (high outlier).

j)

Unusual Low Cutoff: $20.556 - (2 \times 5.641) = 20.556 - 11.282 = 9.274$ inches.

Any neck circumference of 9.274 inches or less would be considered unusually low (low outlier).

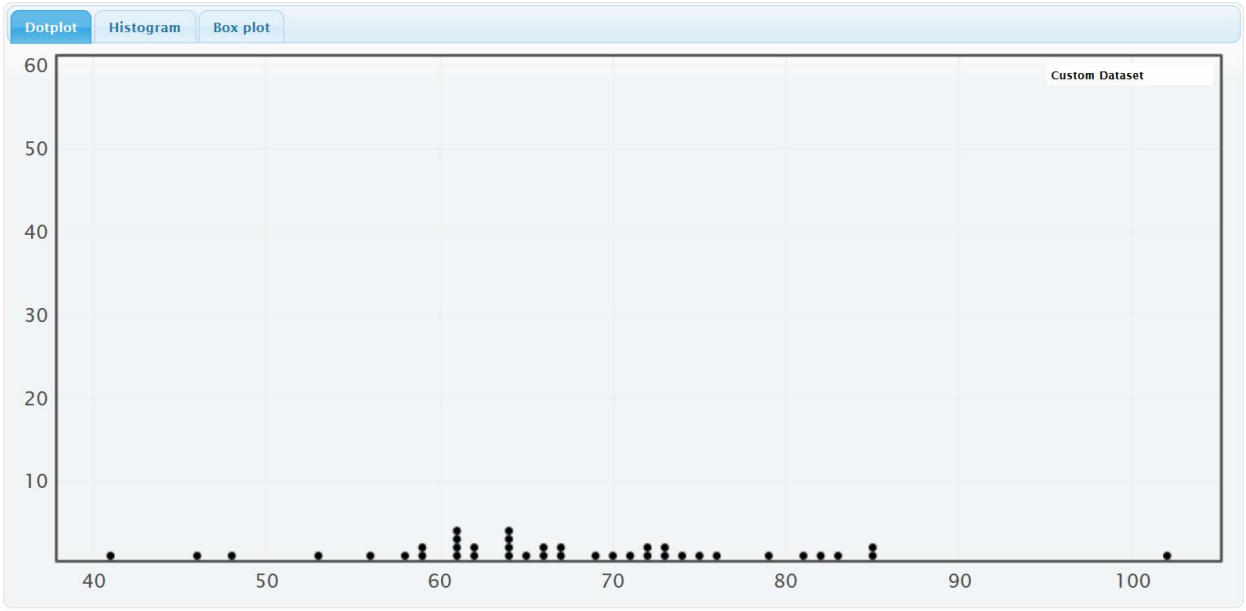
k)

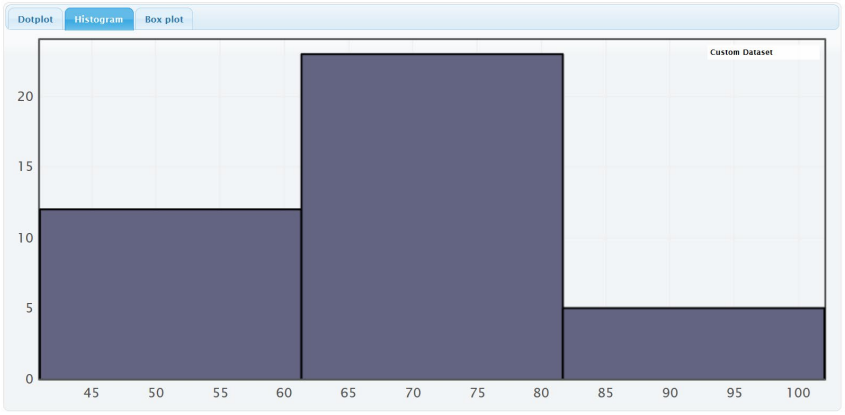
There are no unusually large bear neck sizes. The largest neck size is 31.5 inches which is not higher than the unusually high cutoff of 31.838 inches.

l)

There are no unusually small bear neck sizes. The smallest neck size is 10 inches which is not lower than the unusually low cutoff of 9.274 inches.

5.





Summary Statistics

Statistic	Value
Sample Size	40
Mean	67.425
Standard Deviation	11.626
Minimum	41
Q_1	61.000
Median	66.000
Q_3	73.500
Maximum	102

Histogram Controls Set Limits

Number of buckets: 3

lower bound | upper bound | frequency

a) The data is measuring the diastolic blood pressure of women. The units are millimeters of mercury (mm of Hg).

b) There were 40 total women in the data set.

c) The data is nearly normal (almost bell shaped).

d) Min = 41 mm of Hg

e) Max = 102 mm of Hg

f) Center = Mean Average = 67.425 mm of Hg

g) Typical Spread = Standard Deviation = 11.626 mm of Hg

h)

Mean – Standard Deviation = 67.425 – 11.626 = 55.799 mm of Hg

Mean + Standard Deviation = 67.425 + 11.626 = 79.051 mm of Hg

Typical women in this data have a diastolic blood pressure between 79.051 mm of Hg and 55.799 mm of Hg.

i)

Unusual High Cutoff: $67.425 + (2 \times 11.626) = 67.425 + 23.252 = 90.677$ mm of Hg.

Any woman in the data with a diastolic blood pressure of 90.677 mm of Hg or higher would be considered unusually high (high outlier).

j)

Unusual Low Cutoff: $67.425 - (2 \times 11.626) = 67.425 - 23.252 = 44.173$ mm of Hg.

Any woman in the data with a diastolic blood pressure of 44.173 mm of Hg or lower would be considered unusually low (low outlier).

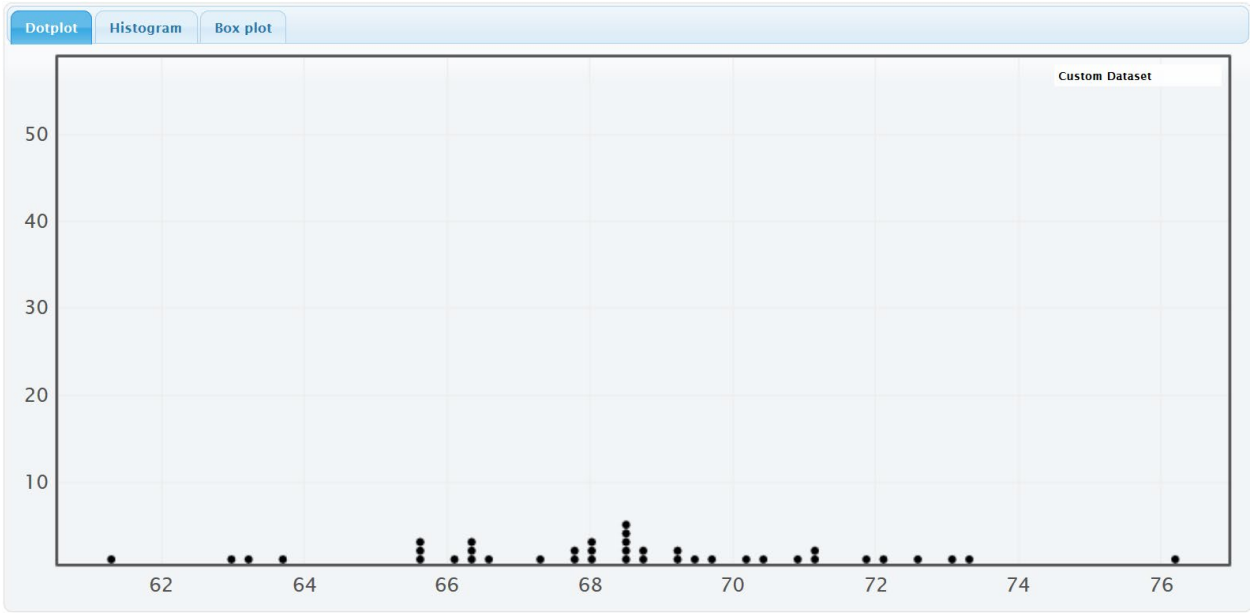
k)

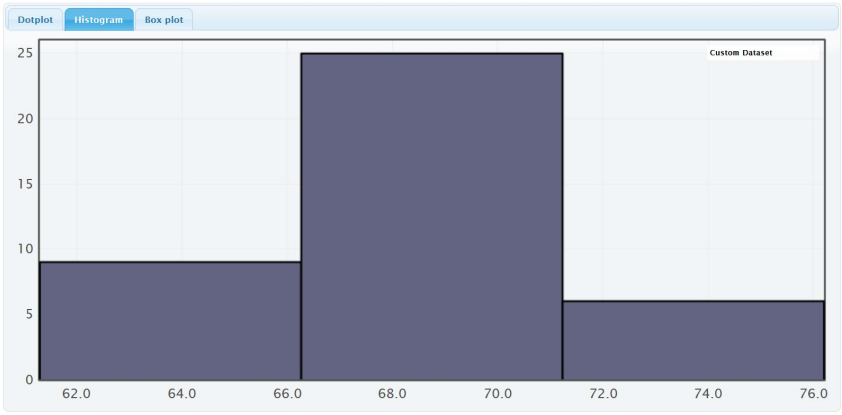
The dot plot shows only one dot above the unusual high cutoff of 90.677 mm of Hg. It is the maximum value of 102 mm of Hg. So the woman with a diastolic blood pressure of 102 mm of Hg is considered unusually high or a high outlier.

l)

The dot plot shows only one dot below the unusual low cutoff of 44.173 mm of Hg. It is the minimum value of 41 mm of Hg. So the woman with a diastolic blood pressure of 41 mm of Hg is considered unusually low or a low outlier.

7.





Summary Statistics

Statistic	Value
Sample Size	40
Mean	68.335
Standard Deviation	3.020
Minimum	61.3
Q_1	66.300
Median	68.300
Q_3	70.150
Maximum	76.2

Histogram Controls Set Limits

Number of buckets: 3

lower bound | upper bound | frequency

a) The data is measuring the heights of men. The units are inches.

b) There were 40 total men in the data set.

c) The data is nearly normal (almost bell shaped).

d) Min = 61.3 inches

e) Max = 76.2 inches

f) Center = Mean Average = 68.335 inches

g) Typical Spread = Standard Deviation = 3.020 inches

h)

Mean – Standard Deviation = 68.335 – 3.020 = 65.315 inches

Mean + Standard Deviation = 68.335 + 3.020 = 71.355 inches

Typical men in this data have a height between 65.315 inches and 71.355 inches.

i)

Unusual High Cutoff: $68.335 + (2 \times 3.020) = 68.335 + 6.040 = 74.375$ inches.

Any man in the data with a height of 74.375 inches or more would be considered unusually high (high outlier).

j)

Unusual High Cutoff: $68.335 - (2 \times 3.020) = 68.335 - 6.040 = 62.295$ inches.

Any man in the data with a height of 62.295 inches or less would be considered unusually low (low outlier).

k)

The dot plot shows only one dot above the unusual high cutoff of 74.375 inches. It is the maximum value of 76.2 inches. So the man with a height of 76.2 inches is considered unusually tall or a high outlier.

l)

The dot plot shows only one dot below the unusual low cutoff of 62.295 inches. It is the minimum value of 61.3 inches. The next largest dot was 62.9 inches which is not below the cutoff. So the man with a height of 61.3 inches is considered unusually short or a low outlier.

9.

A Z-score is the number of standard deviations that a value is from the mean.

10.

Typical values have a Z-score between −1 and +1 inclusively.

11.

For normal data, a Z-score of +2 or higher would indicate that the data value is unusually high or a high outlier.

For normal data, a Z-score of −2 or less would indicate that the data value is unusually low or a low outlier.

13.

a) $Z = (89 - 99.8) / 15.3 \approx -0.71$

b) Jan's IQ score was only 0.71 standard deviations below the mean.

c) Jan's IQ is not unusual, since the Z-score is not +2 or above or −2 or below. In fact, she has a very typical IQ, since the Z-score since it is between −1 and +1.

15.

a) $Z = (13.61 - 46.89) / 12.44 \approx -2.68$

b) The amount of money that Julie spent is 2.68 standard deviations below the mean.

c) The amount that Julie spent was unusually low (low outlier) since the Z-score was below −2.

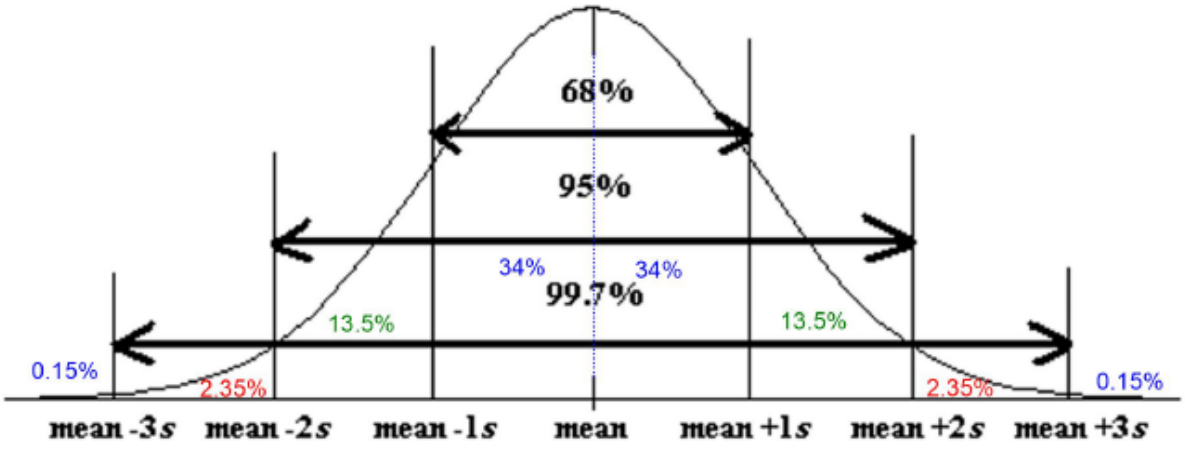
17.

a) Z = (57 – 35.663) / 9.352 ≈ +2.28

b) This bears chest size is 2.28 standard deviations above the mean.

c) This bears chest size is unusually large (high outlier) since the Z-score is greater than +2.

19.



21.

a) 34% + 34% + 13.5% = 81.5%

b) 34% + 34% + 13.5% + 2.35% + 0.15% = 84%

c) One standard deviation from the mean is typical. So typical bear neck circumferences are between 14.915 inches and 26.197 inches.

d) The unusual high cutoff is two standard deviations above the mean which is 31.838 inches. So any bear with a neck circumference of 31.838 inches or more is considered unusually large.

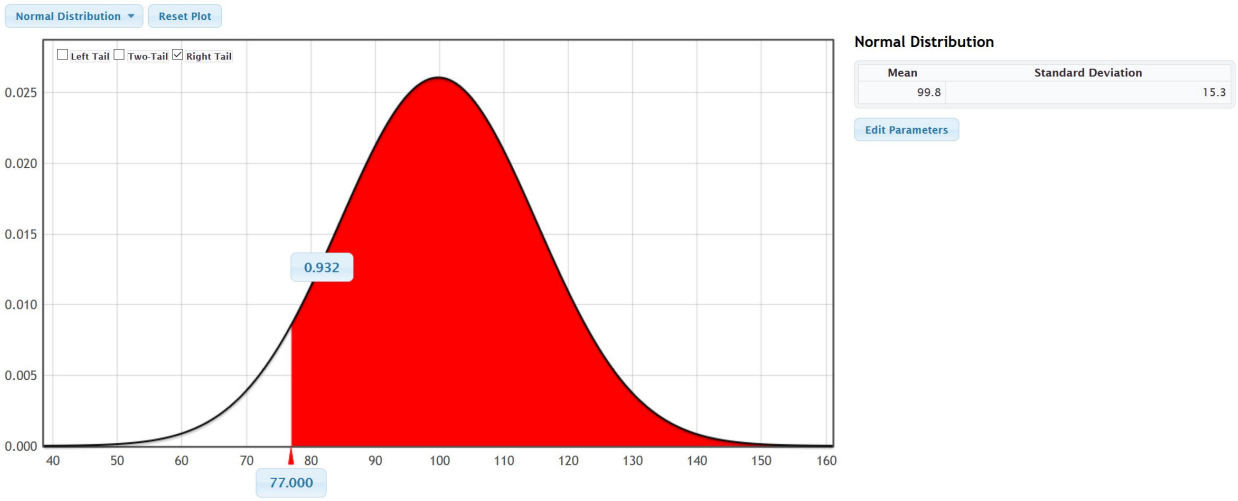
e) The unusual low cutoff is two standard deviations below the mean which is 9.274 inches. So any bear with a neck circumference of 9.274 inches or less is considered unusually small.

f) One standard deviation below the mean has 84% above. So 84% of the bears have a neck circumference greater than 14.915 inches.

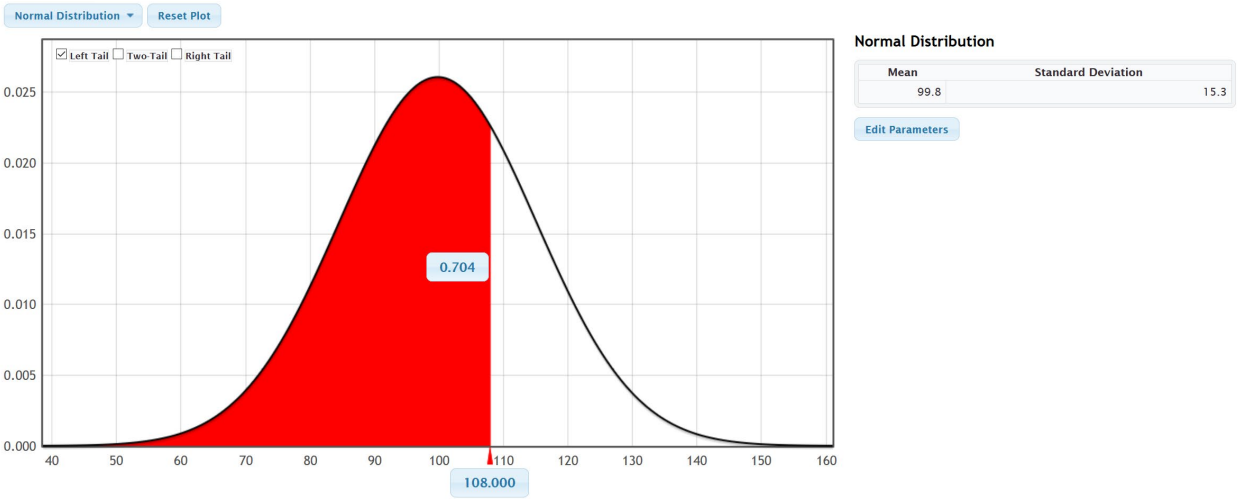
g) 13.5% + 2.35% + 0.15% = 16%

23.

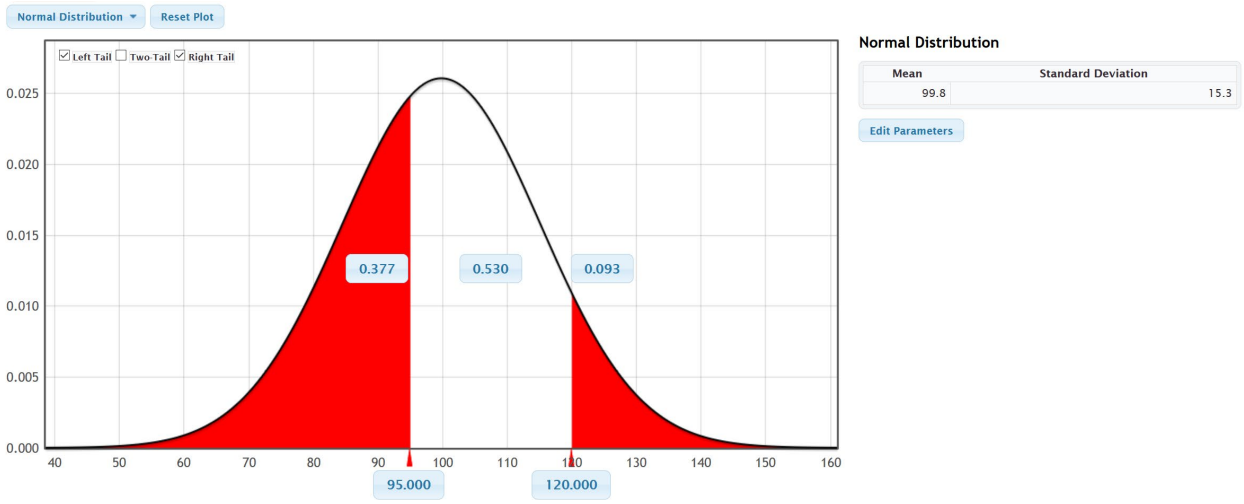
a) Based on this data, about 93.2% of people have an IQ above 77.



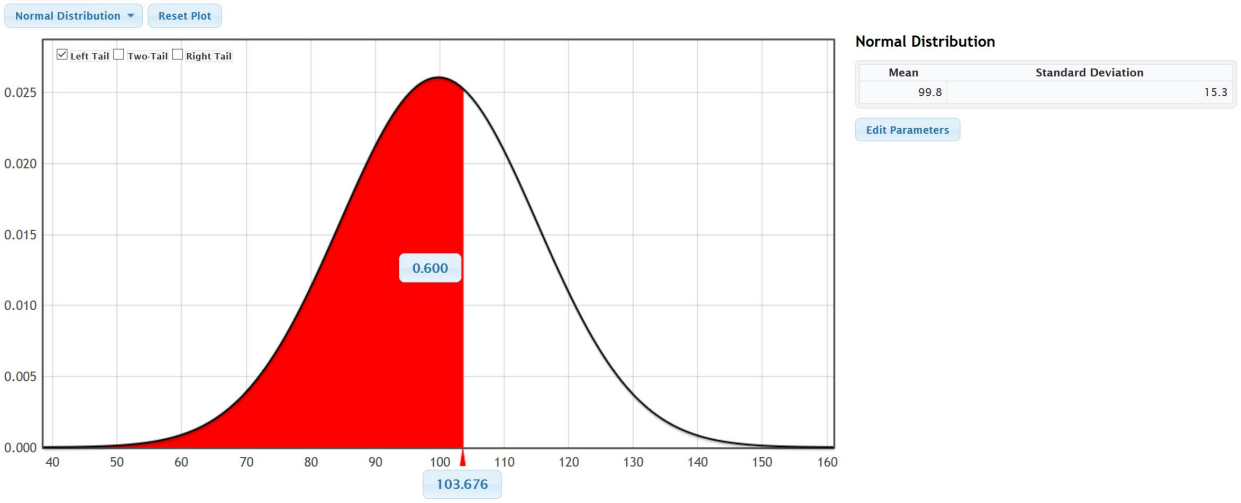
b) Based on this data, about 70.4% of people have an IQ below 108.



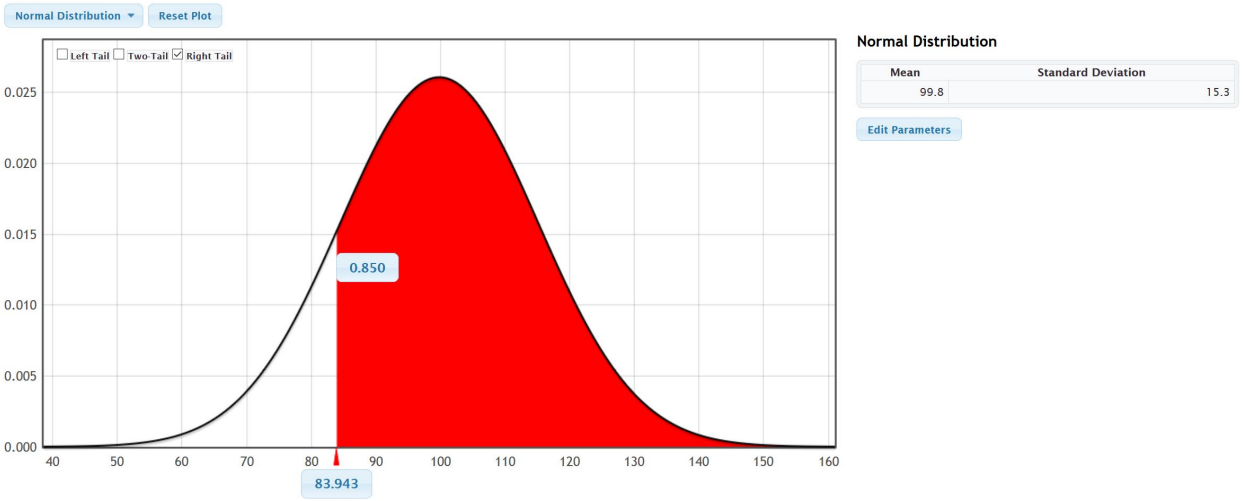
c) Based on this data, about 53.0% of people have an IQ between 95 and 120.



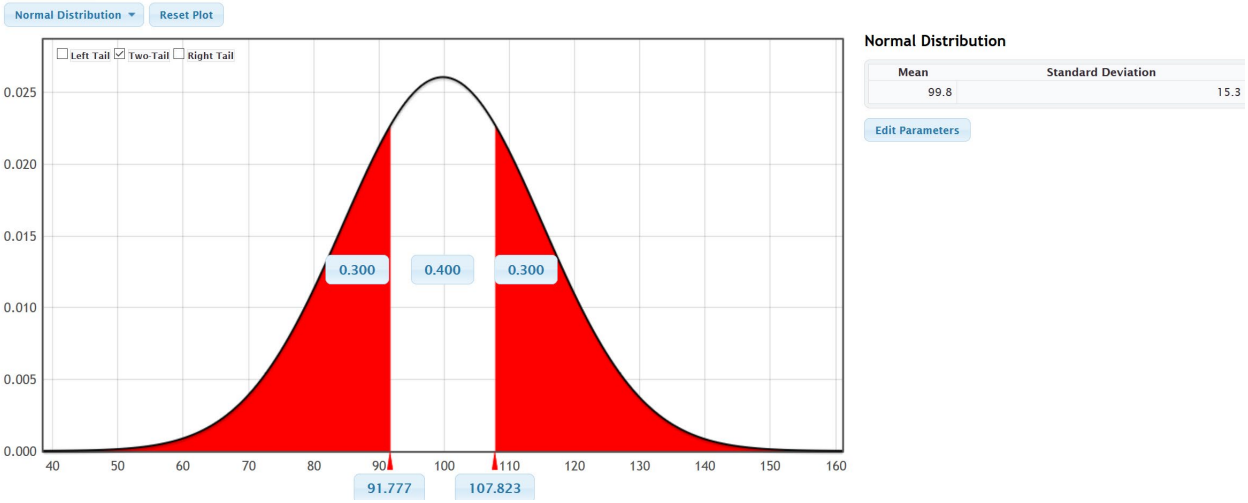
d) About 60% of people have an IQ below 103.676.



e) Based on this data, about 85% of people have an IQ above 83.943.

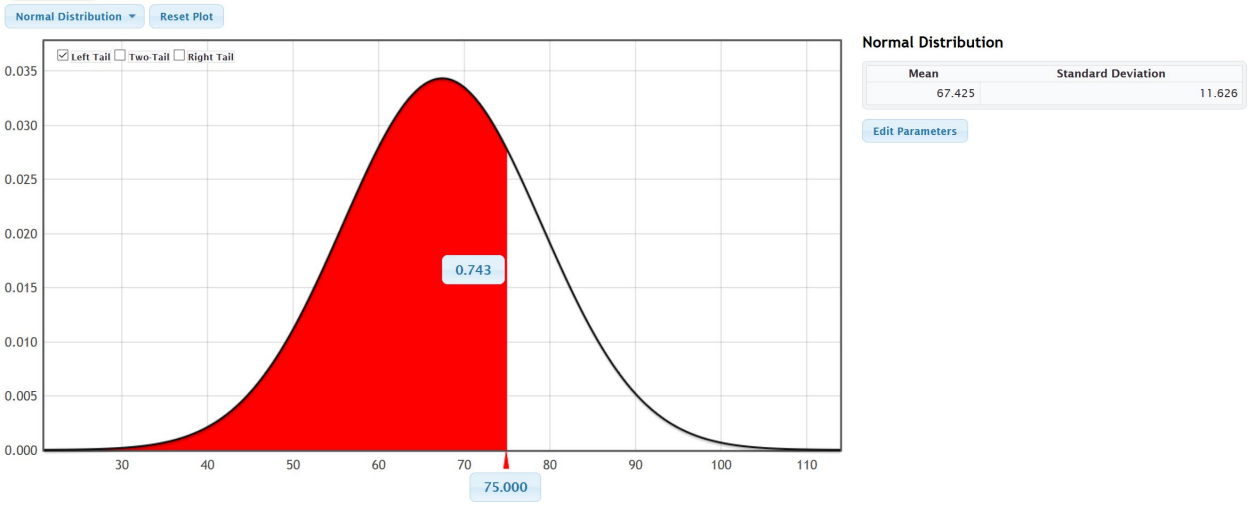


f) Based on this data, the middle 40% of people have an IQ between 91.777 and 107.823.

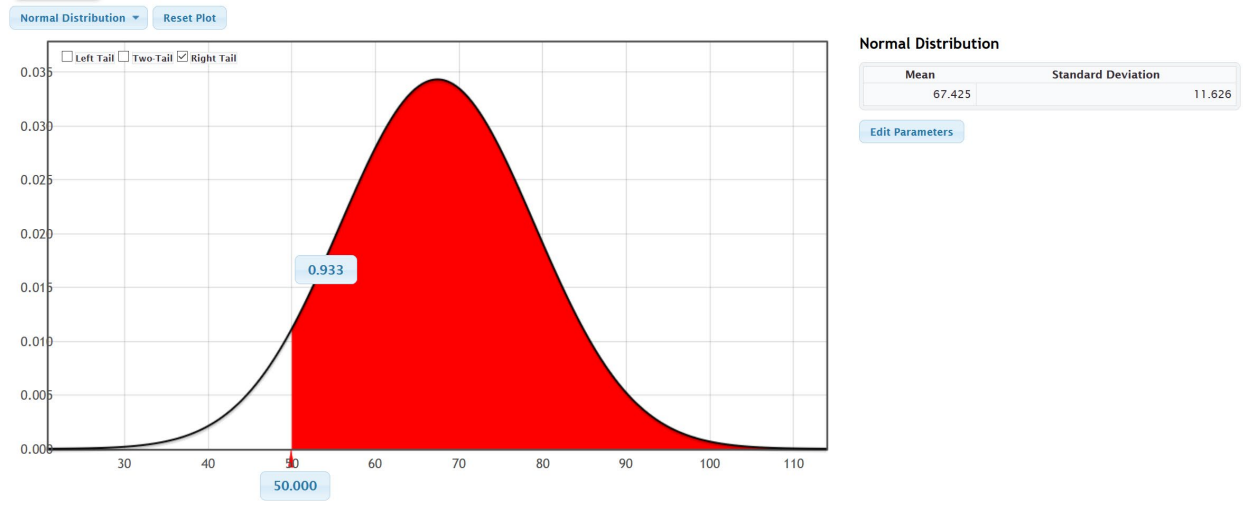


25.

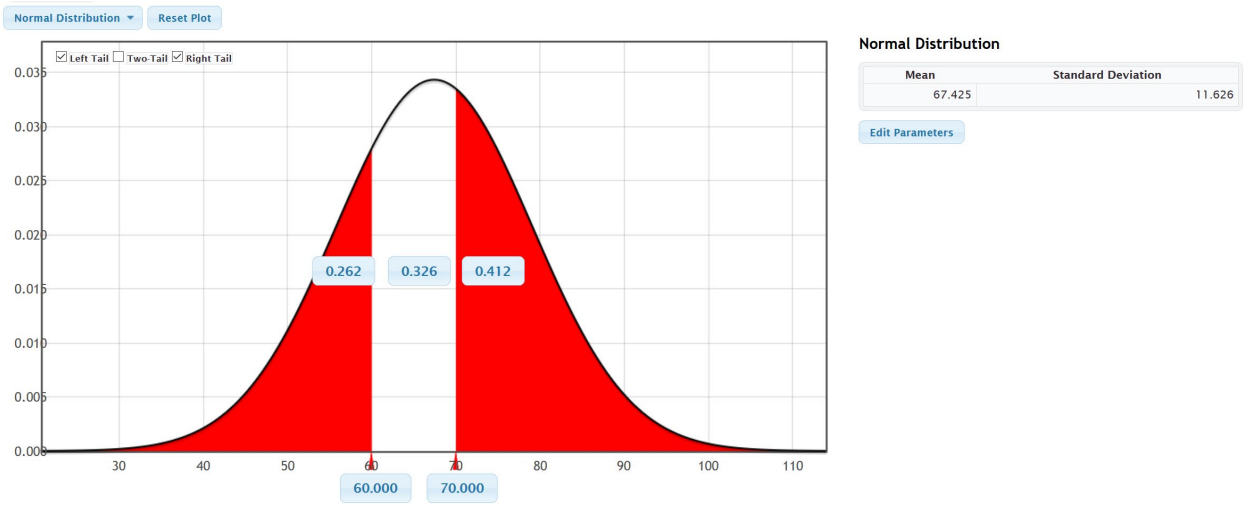
a) Based on this data, about 74.3% of women have a diastolic blood pressure below 75 mm of Hg.



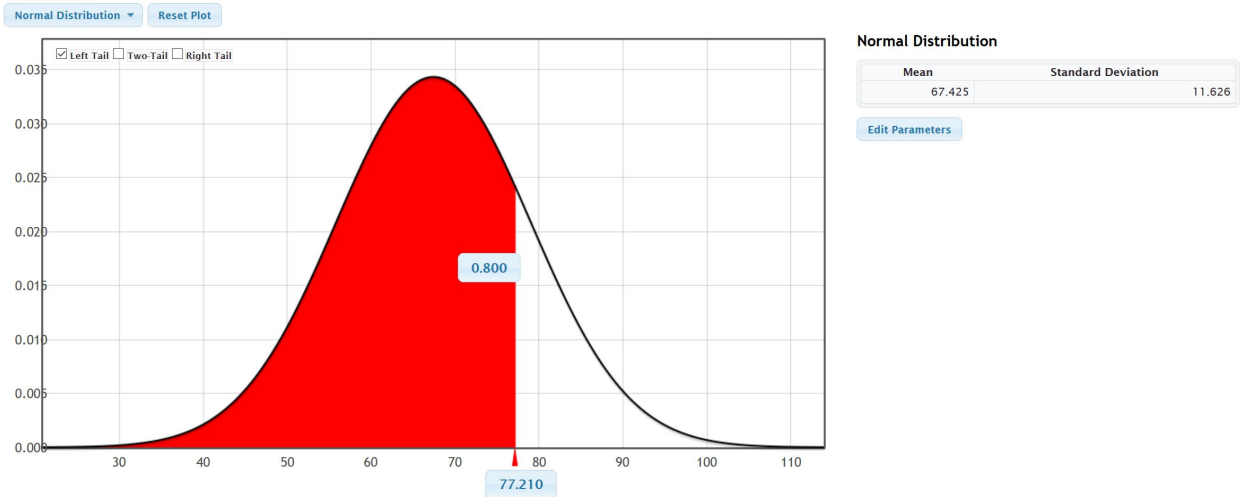
b) Based on this data, about 93.3% of women have a diastolic blood pressure above 50.



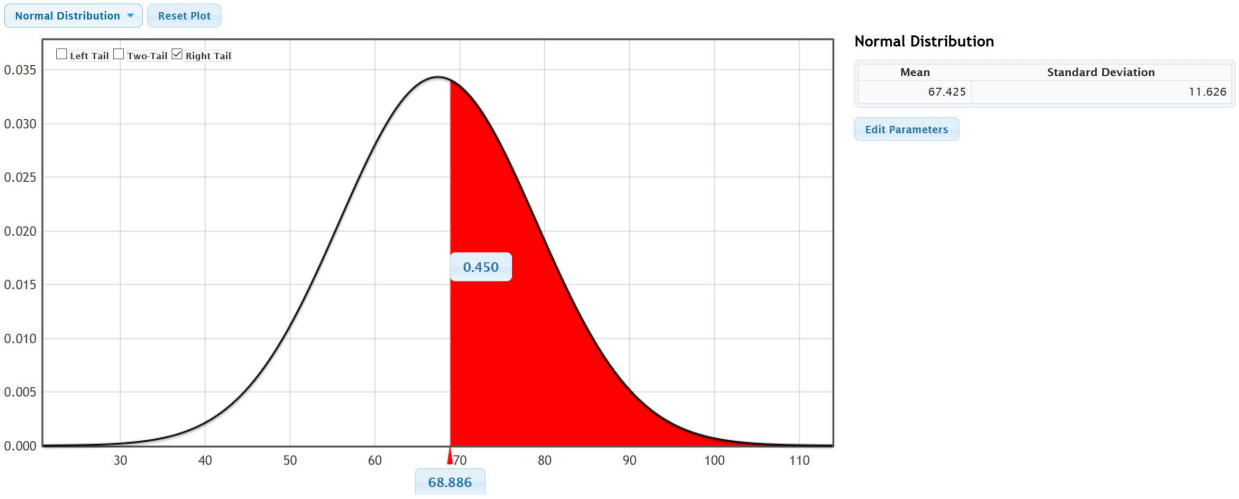
c) Based on this data, about 32.6% of women have a diastolic blood pressure between 60 and 70.



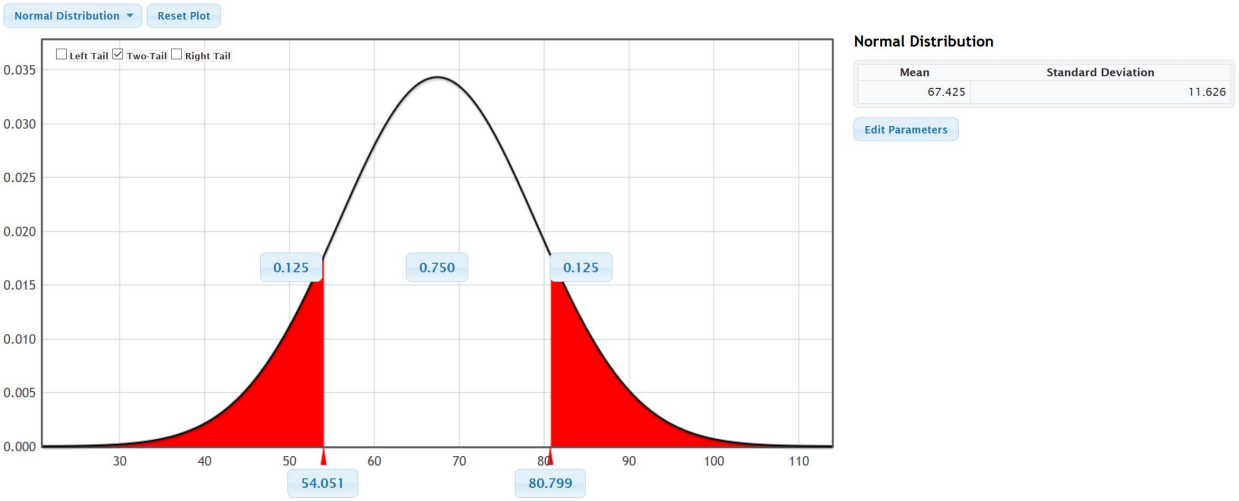
d) Based on this data, about 80% of women have a diastolic blood pressure less than 77.21 mm of Hg.



e) Based on this data, about 45% of women have a diastolic blood pressure above 68.886 mm of Hg.



f) Based on this data, the middle 75% of women's diastolic blood pressures fall between 54.051 mm of Hg and 80.799 mm of Hg.



Section 1G Answers to Odd Problems and #2

1.

a) A skewed right shape has the center on the far left and a long tail to the right. A histogram would have the highest bars on the far left with a short left tail and a long right tail.

b) A skewed left shape has the center on the far right and a long tail to the left. A histogram would have the highest bars on the far right with a short right tail and a long left tail.

c) The median average is the average or center when the data values are put in order. When data sets are not normal, we prefer to use the median as our average. The median also slits the data so that approximately 50% of the data is above the median and 50% of the data is below the median. To calculate the median, first put the numbers in order. If there is one number in the middle, then that is the median. If there are two numbers in the middle, then the median will be half way between the two numbers in the middle.

d) The first quartile is the number that approximately 25% of the data values are less than and 75% of the data values are greater than. To calculate the first quartile, simply calculate the median of the bottom half of the data when the data values are in order.

e) The third quartile is the number that approximately 75% of the data values are less than and 25% of the data values are greater than. To calculate the third quartile, simply calculate the median of the top half of the data when the data values are in order.

f) The interquartile range is the best measure of typical spread for non-normal data. It measures the distance between the middle 50% of the data values. It can also be thought of as the maximum distance between typical values in a non-normal data set. To calculate IQR, subtract the third quartile minus the first quartile.

2.

a) If the data is not normal, we should use the median as our average or center.

b) If the data is not normal, we should use the IQR as the best measure of typical spread.

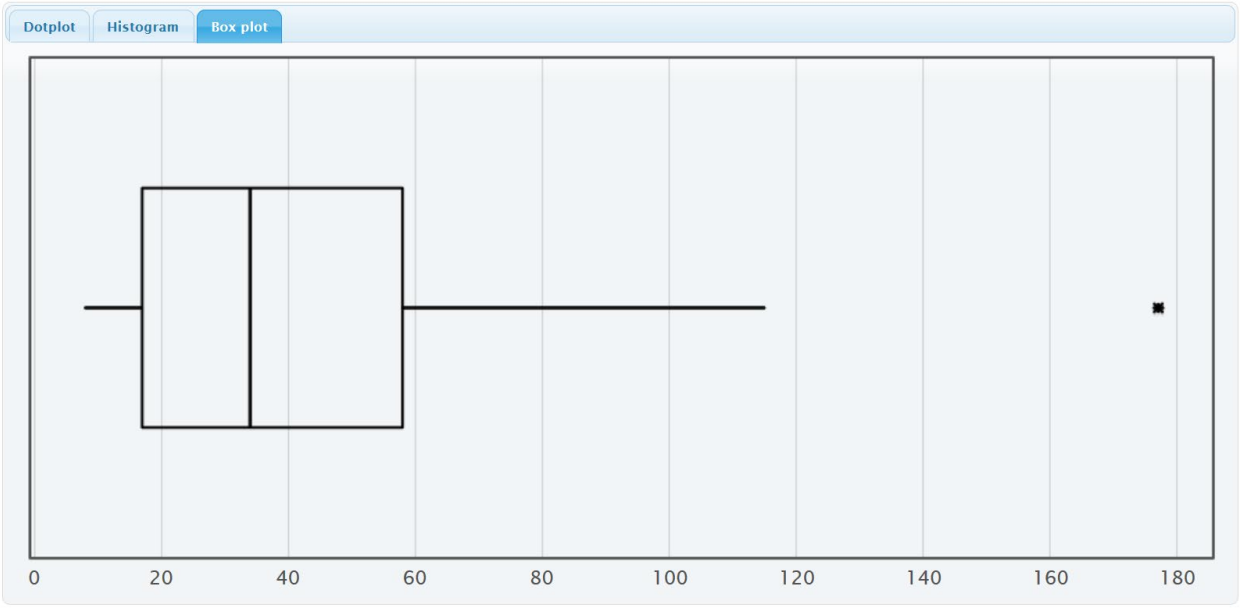
c) If the data is not normal, then typical values will fall between the 1st quartile and the 3rd quartile.

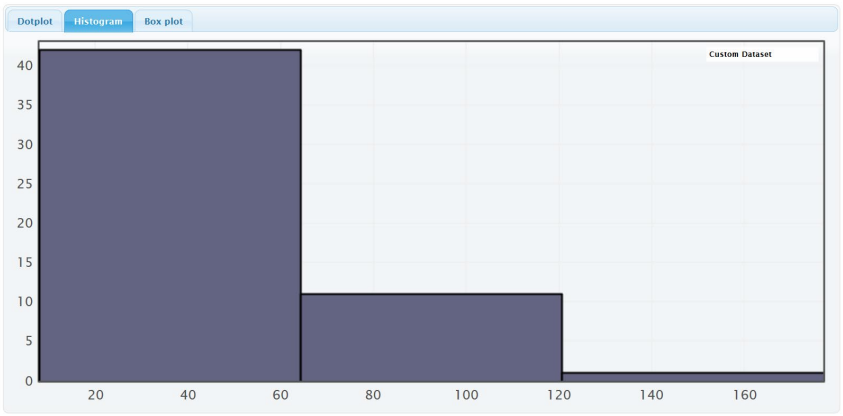
d) The middle 50% is typical for data that is not normal.

e) If the data is not normal, then you can use the box plot to identify unusually high values (high outliers). For horizontal box plots, look for circles, triangles or stars to the far right of the right whisker. For vertical box plots, look for circles, triangles or stars above the top whisker.

f) If the data is not normal, then you can use the box plot to identify unusually low values (low outliers). For horizontal box plots, look for circles, triangles or stars to the far left of the left whisker. For vertical box plots, look for circles, triangles or stars below the bottom whisker.

3.





Summary Statistics

Statistic	Value
Sample Size	54
Mean	43.519
Standard Deviation	33.721
Minimum	8
Q_1	17.000
Median	34.000
Q_3	58.000
Maximum	177

Histogram Controls Set Limits

Number of buckets: 3

lower bound | upper bound | frequency

a) The data is measuring the ages of bears. The units are months.

b) There are 54 bears in the data set. (Sample size)

c) The histogram shows that the data is skewed right.

d) The youngest bear is 8 months old.

e) The oldest bear is 177 months old.

f) Since the data is not normal, we should use the median as our average or center. The median average is 34 months, so the average age of the bears is 34 months old.

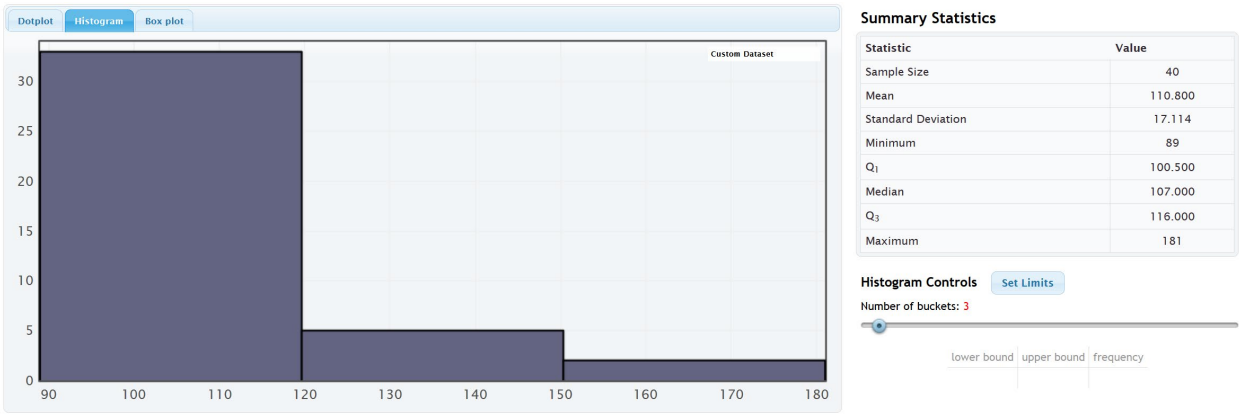
g) Since the data is not normal we should use the IQR as our best measure of typical spread. The IQR is the 3rd quartile minus the 1st quartile = 58 – 17 = 41 months. So typical bear ages are within 41 months of each other.

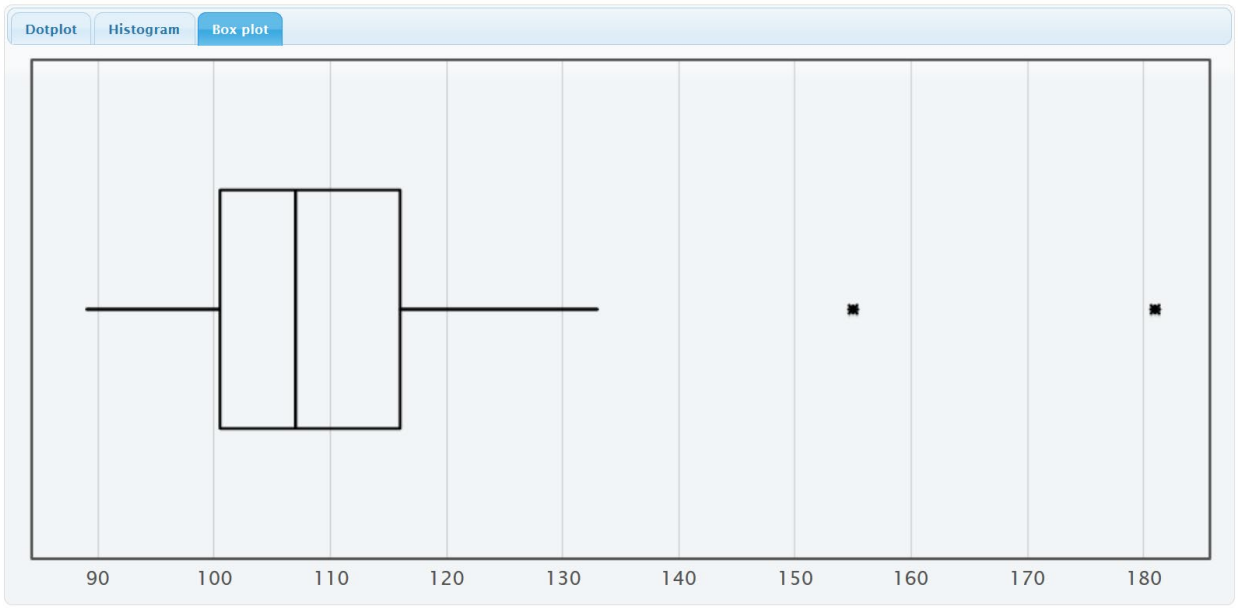
h) Since the data is not normal, typical values will fall between the 1st quartile (17 months) and the 3rd quartile (58 months). So typical bear are between 17 months old and 58 months old.

i) The box plot has one star to the far right. This corresponds to the maximum value of 177 months. So there is only one high outlier in the data set at 177 months old.

j) The box plot does not have any stars to the far left, so there is no low outliers in this data.

5.





a) The data is measuring the systolic blood pressures of women. The units are millimeters of mercury (mm of Hg).

b) There are 40 women in the data set. (Sample size)

c) The histogram shows that the data is skewed right.

d) The lowest systolic blood pressure for these women was 89 mm of Hg.

e) The highest systolic blood pressure for these women was 181 mm of Hg.

f) Since the data is not normal, we should use the median as our average or center. The median average is 107 mm of Hg, so the average systolic blood pressure for these women is 107 mm of Hg.

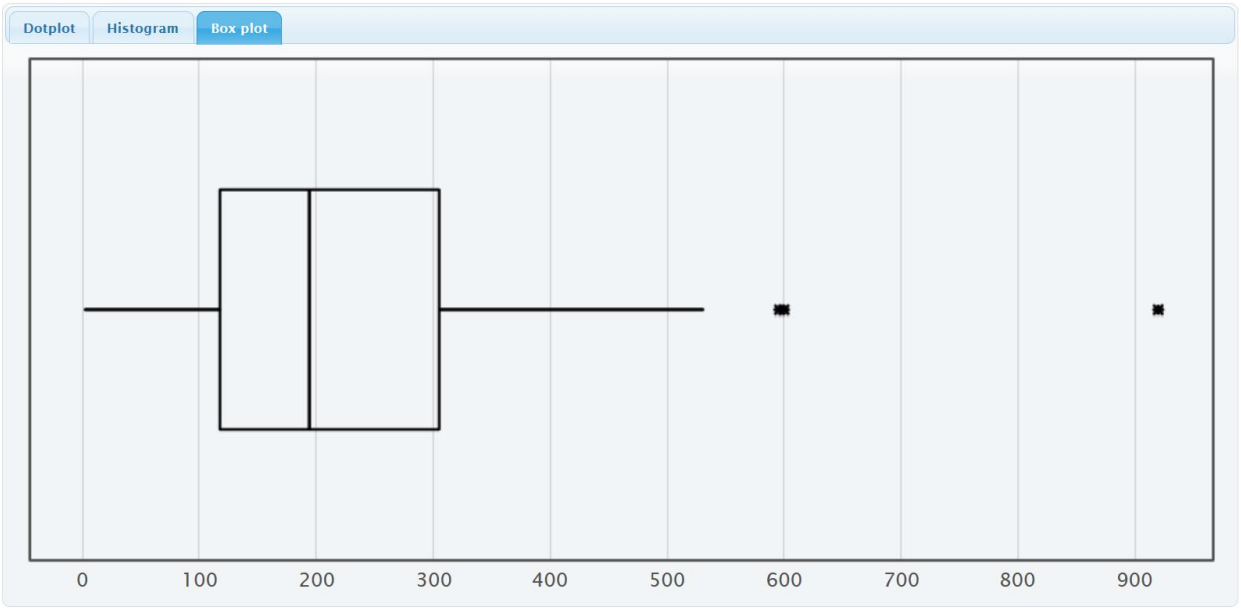
g) Since the data is not normal we should use the IQR as our best measure of typical spread. The IQR is the 3rd quartile minus the 1st quartile = 116 – 100.5 = 15.5 mm of Hg. So typical women have a systolic blood pressure within 15.5 mm of Hg of each other.

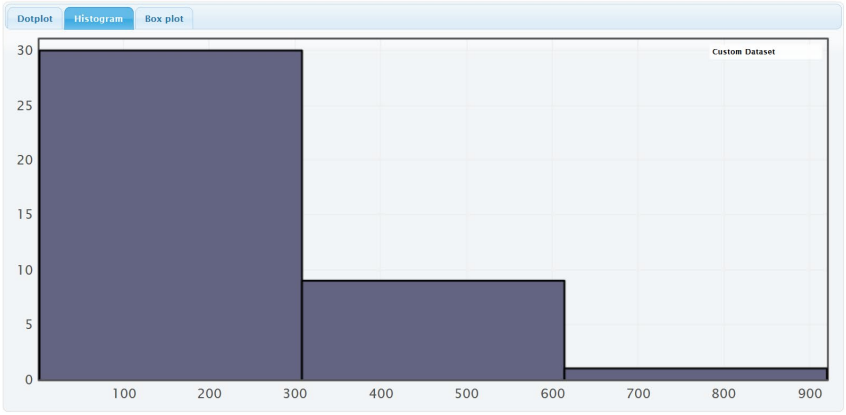
h) Since the data is not normal, typical values will fall between the 1st quartile (100.5 mm of Hg) and the 3rd quartile (116 mm of Hg). So typical women in this data had a systolic blood pressure between 100.5 mm of Hg and 116 mm of Hg.

i) The box plot has two stars to the far right. This corresponds to 155 mm of Hg and the maximum value of 181 mm of Hg. So there is only two high outliers in the data set at 155 mm of Hg and 181 mm of Hg.

j) The box plot does not have any stars to the far left, so there is no low outliers in this data.

7.





Summary Statistics

Statistic	Value
Sample Size	40
Mean	240.875
Standard Deviation	185.982
Minimum	2
Q_1	117.500
Median	194.000
Q_3	305.000
Maximum	920

Histogram Controls Set Limits

Number of buckets: 3

lower bound | upper bound | frequency

a) The data is measuring the cholesterol of women. The units are milligrams per deciliter (mg/dL).

b) There are 40 women in the data set. (Sample size)

c) The histogram shows that the data is skewed right.

d) The lowest cholesterol for these women was 2 mg/dL. This may be a mistake in the data. Doesn't sound possible.

e) The highest cholesterol for these women was 920 mg/dL.

f) Since the data is not normal, we should use the median as our average or center. The median average is 194 mg/dL, so the average cholesterol for these women is 194 mg/dL.

g) Since the data is not normal we should use the IQR as our best measure of typical spread. The IQR is the 3rd quartile minus the 1st quartile = 305 – 117.5 = 187.5 mg/dL. So typical women have a cholesterol within 187.5 mg/dL of each other.

h) Since the data is not normal, typical values will fall between the 1st quartile (117.5 mg/dL) and the 3rd quartile (305 mg/dL). So typical women in this data had a cholesterol between 117.5 mg/dL and 305 mg/dL.

i) The box plot has three stars to the far right. So there are three unusually high cholesterols for these women. They corresponds to the cholesterols of 596 mg/dL, 600 mg/dL and 920 mg/dL.

j) The box plot does not have any stars to the far left, so there is no low outliers in this data.

9.

a) The data is measuring the gas mileage for various cars. The units are in miles per gallon (mpg).

b) There are 38 cars in the data set. ("N Total")

c) The histogram shows that the data is skewed right.

d) The lowest miles per gallon was 15.5 mpg.

e) The highest miles per gallon was 37.3 mpg.

f) Since the data is not normal, we should use the median as our average or center. The median average is 24.25 mpg, so the average gas mileage for these cars is 24.25 mpg.

g) Since the data is not normal we should use the IQR as our best measure of typical spread. The IQR is given as 12.175 mpg. So typical cars in this data are within 12.175 mpg of each other.

h) Since the data is not normal, typical values will fall between the 1st quartile (18.425 mpg) and the 3rd quartile (30.6 mpg). So typical cars in this data have a gas mileage between 18.425 mpg and 30.6 mpg.

j) The box plot does not have any stars to the far right, so there is no high outliers in this data.

j) The box plot does not have any stars to the far left, so there is no low outliers in this data.

11.

a) The data is measuring the horsepower for various cars. The units are the number of horsepower the car has.

b) There are 38 cars in the data set. ("N Total")

c) The histogram shows that the data is skewed right.

d) The smallest horsepower for these cares was 65.

e) The largest horsepower for these cars was 155.

f) Since the data is not normal, we should use the median as our average or center. The median average is 100 horsepower, so the average for these cars is 100 horsepower.

g) Since the data is not normal we should use the IQR as our best measure of typical spread. The IQR is given as 47.75 horsepower. So typical cars in this data are within 47.75 horsepower of each other.

h) Since the data is not normal, typical values will fall between the 1st quartile (77.25 horsepower) and the 3rd quartile (125 horsepower). So typical cars in this data have a horsepower between 77.25 and 125.

j) The box plot does not have any stars to the far right, so there is no high outliers in this data.

j) The box plot does not have any stars to the far left, so there is no low outliers in this data.

13.

a) Q1 is a measure of position.

b) Mean is a measure of center.

c) Variance is a measure of spread.

d) Standard deviation is a measure of spread.

e) Minimum value is a measure of position.

f) Q3 is a measure of position.

g) The mode is a measure of center.

h) The IQR is a measure of spread.

i) The median is a measure of center.

j) The range is a measure of spread.

k) The maximum is a measure of position.

l. The midrange is a measure of center.

Chapter 1 Review Sheet Answers to All Problems

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 5th 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.

7.

a) Identify the place value you wish to round. Look at the number to the right of the place value. If the number is 5 or above, add 1 to the place value and cut off the rest of the decimal. If the number is 4 or less, leave the place value alone and cut off the rest of the decimal.

b) To convert a decimal proportion into a percentage, simply multiply the decimal by 100 and add on the "%" sign.

c) To convert a percentage into a decimal proportion, remove the "%" sign, and divide the percentage by 100.

d) To calculate a percentage divide the amount by the total.

e) To estimate an amount, convert the percentage into a decimal proportion and multiply the proportion by the total. Round the answer to the ones place.

8.

a) 7.22%

b) 0.41%

c) 56.3%

d) 0.05%

9.

a) 0.359

b) 0.04823

c) 0.00026

d) 0.00389

10.

a) 11/74 ≈ 0.149

b) 0.149 x 100% = 14.9%

Approximately 14.9% of the company are managers.

c) 27/74 ≈ 0.365

d) 0.365 x 100% = 36.5%

Approximately 36.5% of the company are full-time employees.

e) 36/74 ≈ 0.486

f) 0.486 x 100% = 48.6%

Approximately 48.6% of the company are part-time employees.

g) Percent of Increase = (0.365 – 0.149) / 0.149 ≈ 145.0% increase. This seems to be a significantly large percent of increase so the percentage of full-time employees seems significantly higher than the percentage of managers. The difference also seems to be practically significant since there are 16 more full time employees than managers and the whole company is 74 total.

h) Percent of Increase = (0.486 – 0.365) / 0.365 ≈ 33.2% increase. This seems to be a large percent of increase so the percentage of part-time employees seems significantly higher than the percentage of full-time employees. The difference may not be practically significant since there are only 9 more part-time employees than full-time.

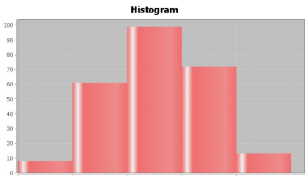
11.

60% = 0.6

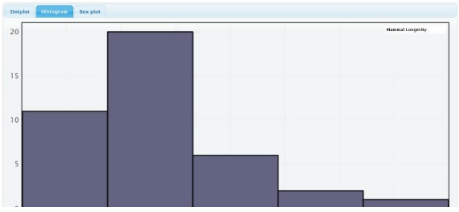
Estimated Amount = 0.6 x 41743 ≈ 25,046 voters in Saugus.

12.

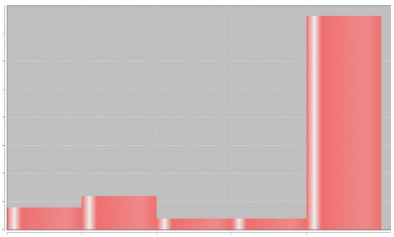
a) A normal or normally distributed histogram is unimodal and symmetric. This means that we expect the highest bar or bars to be in the middle with smaller and smaller bars as we go away from the middle. The left and right tails will be approximately the same length.



b) A skewed right or positively skewed histogram will have the highest bar or bars on the far left of the graph. It will have very few bars to the left of the center and many bars to the right of the center. Therefore the right tail will look much longer than the left tail.



c) A skewed left or negatively skewed histogram will have the highest bar or bars on the far right of the graph. It will have very few bars to the right of the center and many bars to the left of the center. Therefore, the left tail will look much longer than the right tail.



13.

a) The first quartile (Q1) is a measure of position. It is used to analyze typical values when data is skewed or not normal.

b) The mean is a measure of center. It is the primary center or average when the data is normal.

c) The variance is a measure of spread. It is used when the data is normal.

d) The standard deviation is a measure of spread. It is the primary measure of spread when the data is normal.

e) The minimum value is a measure of position. It can sometimes be an outlier and is used in all quantitative data regardless of shape.

f) The third quartile (Q3) is a measure of position. It is used to analyze typical values when data is skewed or not normal.

g) The mode is a measure of center. It is often used in business applications or any time we wish to know the value or values that appear most often.

h) The interquartile range (IQR) is a measure of spread. It is the primary measure of spread when the data is skewed or not normal.

i) The median or 50th percentile or 2nd quartile (Q2) is a measure of center. It is the primary measure of center or average when the data is skewed or not normal.

j) The range is a measure of spread. It is usually used when someone wants a quick easy to calculate measure of spread. It does not represent typical spread.

k) The maximum value is a measure of position. It can sometimes be an outlier and is used in all quantitative data regardless of shape.

l) The midrange is a measure of center. It is usually used when someone wants a quick easy to calculate center or average. It may not be a very accurate average since it is often based on outliers.

14.

a) We should use the mean average when the data is normal.

b) We should use the median average when the data is not normal.

c) We should use the standard deviation as our main measure of typical spread when data is normal.

d) We should use the interquartile range (IQR) as our main measure of typical spread when data is not normal.

e) When the data is normal, add and subtract the mean and standard deviation. Typical values will fall between $\bar{x}$ − s and $\bar{x}$ + s .

f) When data is not normal, typical values will fall between Q_1 and Q_3.

g) To calculate the unusually high cutoff for normal data, multiply the standard deviation by two and then add it to the mean ($\bar{x}$ + 2s).

h) To calculate the unusually high cutoff for non-normal data, multiply the IQR by 1.5 and add it to Q_3. (Q_3 + (1.5xIQR)).

i) To calculate the unusually low cutoff for normal data, multiply the standard deviation by two and then subtract from the mean ($\bar{x}$ − 2s).

j) To calculate the unusually low cutoff for non-normal data, multiply the IQR by 1.5 and subtract it from Q_1. (Q_1 − (1.5xIQR)).

k) To find low outliers in a normal data set, calculate the unusual low cutoff $\bar{x}$ − 2s and look for any data values that are lower than the low cutoff. To find high outliers in a normal data, calculate the unusual high cutoff $\bar{x}$ + 2s and look for any data values that are higher than the high cutoff.

l) To find low and high outliers in a skewed or non-normal data set, create a boxplot and look for any stars, circles or triangles outside of the whiskers.

15.

a) The data is measuring the age of mammals. The units are in years.

b) Sample size = 40. There are 40 mammals in the data set.

c) The data is skewed right.

d) The youngest mammal was 1 year old.

e) The oldest mammal was 40 years old.

f) Since the data was not normal, we will use the median average. The average age of the mammals is 12 years.

g) Since the data was not normal, we will use the interquartile range to measure the typical spread. IQR = Q3 – Q1 = 15.5 – 8 = 7.5 years. So typical mammal ages in this data set are within 7.5 years of each other.

h) Since the data was not normal, typical values will fall between Q1 and Q3. So typical mammal ages in this data set are between 8 years and 15.5 years.

i) The box plot shows that there is one high outlier at 40 years.

j) The box plot shows that there is no low outliers.

16.

a) The data is measuring the amount of time employees have been with the company. The units are years.

b) N total = 253. There are 253 employees in the data set.

c) The data appears normal.

d) The employee that has been with the company the shortest amount of time is 3.6 years.

e) The employee that has been with the company the longest amount of time is 10.8 years.

f) Since the data is normal, we will use the mean average. The average time that employees have been with this company is 7.345 years.

g) Since the data is normal, we will use the standard deviation as our measure of typical spread. So typical employee times with the company are within 1.376 years of the mean.

h) Since the data is normal, the high outlier cutoff is the mean + (2 x standard deviation) = 7.345 + (2 x 1.376) = 10.097. So any employees that have been with the company 10.097 years or moreis considered an unusually large amount of time. The dot plot shows five employees that have been with the company from approximately 10.5 years to 10.8 years. All of these times are unusually long.

i) Since the data is normal, the low outlier cutoff is the mean − (2 x standard deviation) = 7.345 − (2 x 1.376) = 4.593. So any employees that have been with the company 4.593 years or less is considered an unusually small amount of time. The dot plot shows five employees that have been with the company from approximately 3.6 years to 4.5 years. All of these times are unusually short.

17.

a) Driving alone was most popular.

b) Biking was least popular.

c) 10 of the stat students walked to school.

d) 0.054

e) 0.018 x 100% = 1.8%

1.8% of the stat students used public transportation.

18.

a) 34%

b) 16%

c) Typical salaries are between 29.3 thousand dollars and 33.5 thousand dollars.

d) Salaries above 35.6 thousand dollars are considered unusually high (high outliers).

e) Salaries below 27.2 thousand dollars are considered unusually low (low outliers).

f) The average salary is 31.4 thousand dollars.
