Chapter 1: Statistics: Part 1

Section 1.1: Statistical Basics

Data are all around us. Researchers collect data on the effectiveness of a medication for lowering cholesterol. Pollsters report on the percentage of Americans who support gun control. Economists report on the average salary of college graduates. There are many other areas where data are collected. In order to be able to understand data and how to summarize it, we need to understand statistics.

Suppose you want to know the average net worth of a current U.S. Senator. There are 100 Senators, so it is not that hard to collect all 100 values, and then summarize the data. If instead you want to find the average net worth of all current Senators and Representatives in the U.S. Congress, there are only 435 members of Congress. So even though it will be a little more work, it is not that difficult to find the average net worth of all members. Now suppose you want to find the average net worth of everyone in the United States. This would be very difficult, if not impossible. It would take a great deal of time and money to collect the information in a timely manner before all of the values have changed. So instead of getting the net worth of every American, we have to figure out an easier way to find this information. The net worth is what you want to measure, and is called a variable. The net worth of every American is called the population. What we need to do is collect a smaller part of the population, called a sample. In order to see how this works, let's formalize the definitions.

Variable: Any characteristic that is measured from an object or individual.
Population: A set of measurements or observations from all objects under study
Sample: A set of measurements or observations from some objects under study (a subset of a population)

Example 1.1.1: Stating Populations and Samples

Determine the population and sample for each situation.

a. A researcher wants to determine the length of the lifecycle of a bark beetle. In order to do this, he breeds 1000 bark beetles and measures the length of time from birth to death for each bark beetle.

   Population: The set of lengths of lifecycle of all bark beetles
   Sample: The set of lengths of lifecycle of 1000 bark beetles

b. The National Rifle Association wants to know what percent of Americans support the right to bear arms. They ask 2500 Americans whether they support the right to bear arms.

   Population: The set of responses from all Americans to the question, "Do you support the right to bear arms?"
   Sample: The set of responses from 2500 Americans to the question, "Do you support the right to bear arms?"

c. The Pew Research Center asked 1000 mothers in the U.S. what their highest attained education level was.

   Population: The set of highest education levels of all mothers in the U.S.
   Sample: The set of highest education level of 1000 mothers in the U.S.

It is very important that you understand what you are trying to measure before you actually measure it. Also, please note that the population is a set of measurements or observations, and not a set of people. If you say the population is all Americans, then you have only given part of the story. More important is what you are measuring from all Americans. The question is, do you want to measure their race, their eye color, their income, their education level, the number of children they have, or other variables? Therefore, it is very important to state what you measured or observed, and from whom or what the measurements or observations were taken. Once you know what you want to measure or observe, and the source from which you want to take measurements or observations, you need to collect the data.

A **data set** is a collection of values called data points or data values. **N** represents the number of data points in a population, while **n** represents the number of data points in a sample. A data value that is much higher or lower than all of the other data values is called an **outlier**. Sometimes outliers are just unusual data values that are very interesting and should be studied further, and sometimes they are mistakes. You will need to figure out which is which.

In order to collect the data, we have to understand the types of variables we can collect. There are actually two different types of variables. One is called qualitative and the other is called quantitative.

Qualitative (Categorical) Variable: A variable that represents a characteristic. Qualitative variables are not inherently numbers, and so they cannot be added, multiplied, or averaged, but they can be represented graphically with graphs such as a bar graph.

Examples: gender, hair color, race, nationality, religion, course grade, year in college, etc.

Quantitative (Numerical) Variable: A variable that represents a measurable quantity. Quantitative variables are inherently numbers, and so can they be added, multiplied, averaged, and displayed graphically.

Examples: Height, weight, number of cats owned, score of a football game, etc.

Quantitative variables can be further subdivided into other categories – continuous and discrete.

Continuous Variable: A variable that can take on an uncountable number of values in a range. In other words, the variable can be any number in a range of values. Continuous variables are usually things that are measured.

Examples: Height, weight, time to take a test, length, etc.

Discrete Variable: A variable that can take on only specific values in a range. Discrete variables are usually things that you count.

Examples: IQ, shoe size, family size, number of cats owned, score in a football game, etc.

Example 1.1.2: Determining Variable Types

Determine whether each variable is quantitative or qualitative. If it is quantitative, then also determine if it is continuous or discrete.

a. Length of run
   Quantitative and continuous, since this variable is a number and can take on any value in an interval.

b. Opinion of a person about the President
   Qualitative, since this variable is not a number.

c. House color in a neighborhood
   Qualitative, since this variable is not a number.

d. Number of houses that are in foreclosure in a state
   Quantitative and discrete, since this variable is a number but can only be certain values in an interval.

e. Weight of a baby at birth
Quantitative and continuous, since this variable is a number and can take on any value in an interval.

f. Highest education level of a mother
Qualitative, since the variable is not a number.

Section 1.2: Random Sampling

Now that you know that you have to take samples in order to gather data, the next question is how best to gather a sample? There are many ways to take samples. Not all of them will result in a representative sample. Also, just because a sample is large does not mean it is a good sample. As an example, you can take a sample involving one million people to find out if they feel there should be more gun control, but if you only ask members of the National Rifle Association (NRA) or the Coalition to Stop Gun Violence, then you may get biased results. You need to make sure that you ask a cross-section of individuals. Let's look at the types of samples that can be taken. Do realize that no sample is perfect, and may not result in a representation of the population.

Census: An attempt to gather measurements or observations from all of the objects in the entire population.

A true census is very difficult to do in many cases. However, for certain populations, like the net worth of the members of the U.S. Senate, it may be relatively easy to perform a census. We should be able to find out the net worth of each and every member of the Senate since there are only 100 members. But, when our government tries to conduct the national census every 10 years, you can believe that it is impossible for them to gather data on each and every American.

The best way to find a sample that is representative of the population is to use a random sample. There are several different types of random sampling. Though it depends on the task at hand, the best method is often simple random sampling which occurs when you randomly choose a subset from the entire population.

Simple Random Sample: Every sample of size n has the same chance of being chosen, and every individual in the population has the same chance of being in the sample.

An example of a simple random sample is to put all of the names of the students in your class into a hat, and then randomly select five names out of the hat.

Stratified Sampling: This is a method of sampling that divides a population into different groups, called strata, and then takes random samples inside each strata.

An example where stratified sampling is appropriate is if a university wants to find out how much time their students spend studying each week; but they also want to know if different majors spend more time studying than others. They could divide the student body into the different majors (strata), and then randomly pick a number of people in each major to ask them how much time they spend studying. The number of people asked in each major (strata) does not have to be the same.

Systematic Sampling: This method is where you pick every kth individual, where k is some whole number. This is used often in quality control on assembly lines.

For example, a car manufacturer needs to make sure that the cars coming off the assembly line are free of defects. They do not want to test every car, so they test every 100^{th} car. This way they can periodically see if there is a problem in the manufacturing process. This makes for an easier method to keep track of testing and is still a random sample.

Cluster Sampling: This method is like stratified sampling, but instead of dividing the individuals into strata, and then randomly picking individuals from each strata, a cluster sample separates the individuals into groups, randomly selects which groups they will use, and then takes a census of every individual in the chosen groups.

Cluster sampling is very useful in geographic studies such as the opinions of people in a state or measuring the diameter at breast height of trees in a national forest. In both situations, a cluster sample reduces the traveling distances that occur in a simple random sample. For example, suppose that the Gallup Poll needs to perform a public opinion poll of all registered voters in Colorado. In order to select a good sample using simple random sampling, the Gallup Poll would have to have all the names of all the registered voters in Colorado, and then randomly select a subset of these names. This may be very difficult to do. So, they will use a cluster sample instead. Start by dividing the state of Colorado up into categories or groups geographically. Randomly select some of these groups. Now ask all registered voters in each of the chosen groups. This makes the job of the pollsters much easier, because they will not have to travel over every inch of the state to get their sample but it is still a random sample.

Quota Sampling: This is when the researchers deliberately try to form a good sample by *creating* a cross-section of the population under study.

For an example, suppose that the population under study is the political affiliations of all the people in a small town. Now, suppose that the residents of the town are 70% Caucasian, 25% African American, and 5% Native American. Further, the residents of the town are 51% female and 49% male. Also, we know information about the religious affiliations of the townspeople. The residents of the town are 55% Protestant, 25% Catholic, 10% Jewish, and 10% Muslim. Now, if a researcher is going to poll the people of this town about their political affiliation, the researcher should gather a sample that is representative of the entire population. If the researcher uses quota sampling, then the researcher would try to artificially create a cross-section of the town by insisting that his sample should be 70% Caucasian, 25% African American, and 5% Native American. Also, the researcher would want his sample to be 51% female and 49% male. Also, the researcher would want his sample to be 55% Protestant, 25% Catholic, 10% Jewish, and 10% Muslim. This sounds like an admirable attempt to create a good sample, but this method has major problems with selection bias.

The main concern here is when does the researcher stop profiling the people that he will survey? So far, the researcher has cross-sectioned the residents of the town by race, gender, and religion, but are those the only differences between individuals? What about socioeconomic status, age, education, involvement in the community, etc.? These are all influences on the political affiliation of individuals. Thus, the problem with quota sampling is that to do it right, you have to take into account all the differences among the people in the town. If you cross-section the town down to every possible difference among people, you end up with single individuals, so you would have to survey the whole town to get an accurate result. The whole point of creating a sample is so that you do not have to survey the entire population, so what is the point of quota sampling?

Note: The Gallup Poll did use quota sampling in the past, but does not use it anymore.

Convenience Sampling: As the name of this sampling technique implies, the basis of convenience sampling is to use whatever method is easy and convenient for the investigator. This type of sampling technique creates a situation where a random sample is not achieved. Therefore, the sample will be biased since the sample is not representative of the entire population.

For example, if you stand outside the Democratic National Convention in order to survey people exiting the convention about their political views. This may be a convenient way to gather data, but the sample will not be representative of the entire population.
Of all of the sampling types, a random sample is the best type. Sometimes, it may be difficult to collect a perfect random sample since getting a list of all of the individuals to randomly choose from may be hard to do.

Example 1.2.1: Which Type of Sample?

Determine if the sample type is simple random sample, stratified sample, systematic sample, cluster sample, quota sample, or convenience sample.

a. A researcher wants to determine the different species of trees that are in the Coconino National Forest. She divides the forest using a grid system. She then randomly picks 20 different sections and records the species of every tree in each of the chosen sections.

   This is a cluster sample, since she randomly selected some of the groups, and all individuals in the chosen groups were surveyed.

b. A pollster stands in front of an organic foods grocery store and asks people leaving the store how concerned they are about pesticides in their food.

   This is a convenience sample, since the person is just standing out in front of one store. Most likely the people leaving an organic food grocery store are concerned about pesticides in their food, so the sample would be biased.

c. The Pew Research Center wants to determine the education level of mothers. They randomly ask mothers to say if they had some high school, graduated high school, some college, graduated from college, or advance degree.

   This is a simple random sample, since the individuals were picked randomly.

d. Penn State wants to determine the salaries of their graduates in the majors of agricultural sciences, business, engineering, and education. They randomly ask 50 graduates of agricultural sciences, 100 graduates of business, 200 graduates of engineering, and 75 graduates of education what their salaries are.

   This is a stratified sample, since all groups were used, and then random samples were taken inside each group.

e. In order for the Ford Motor Company to ensure quality of their cars, they test every 130^{th} car coming off the assembly line of their Ohio Assembly Plant in Avon Lake, OH.

   This is a systematic sample since they picked every 130^{th} car.

f. A town council wants to know the opinion of their residents on a new regional plan. The town is 45% Caucasian, 25% African American, 20% Asian, and 10% Native American. It also is 55% Christian, 25% Jewish, 12% Islamic, and 8% Atheist. In addition, 8% of the town did not graduate from high school, 12% have graduated from high school but never went to college, 16% have had some college, 45% have obtained bachelor's degree, and 19% have obtained a post-graduate degree. So the town council decides that the sample of residents will be taken so that it mirrors these breakdowns.

   This is a quota sample, since they tried to pick people who fit into these subcategories.

Section 1.3: Clinical Studies

Now you know how to collect a sample, next you need to learn how to conduct a study. We will discuss the basics of studies, both observational studies and experiments.

Observational Study: This is where data is collected from just observing what is happening. There is no treatment or activity being controlled in any way. Observational studies are commonly conducted using surveys, though you can also collect data by just watching what is happening such as observing the types of trees in a forest.

Survey: Surveys are used for gathering data to create a sample. There are many different kinds of surveys, but overall, a survey is a method used to ask people questions when interested in the responses. Examples of surveys are Internet and T.V. surveys, customer satisfaction surveys at stores or restaurants, new product surveys, phone surveys, and mail surveys. The majority of surveys are some type of public opinion poll.

Experiment: This is an activity where the researcher controls some aspect of the study and then records what happens. An example of this is giving a plant a new fertilizer, and then watching what happens to the plant. Another example is giving a cancer patient a new medication, and monitoring whether the medication stops the cancer from growing. There are many ways to do an experiment, but a clinical study is one of the more popular ways, so we will look at the aspects of this.

Clinical Study: This is a method of collecting data for a sample and then comparing that to data collected for another sample where one sample has been given some sort of treatment and the other sample has not been given that treatment (control). *Note: There are occasions when you can have two treatments, and no control. In this case you are trying to determine which treatment is better.*

Example 1.3.1: Clinical Study Examples

Here are examples of clinical studies.

a. A researcher may want to study whether or not smoking increases a person's chances of heart disease.
b. A researcher may want to study whether a new antidepressant drug will work better than an old antidepressant drug.
c. A researcher may want to study whether taking folic acid before pregnancy will decrease the risk of birth defects.

Clinical Study Terminology:

Treatment Group: This is the group of individuals who are given some sort of treatment. The word treatment here does not necessarily mean medical treatment. The treatment is the cause, which may produce an effect that the researcher is interested in.

Control Group: This is the group of individuals who are not given the treatment. Sometimes, they may be given some old treatment, or sometimes they will not be given anything at all. Other times, they may be given a placebo (see below).

Example 1.3.2: Treatment/Control Group Examples

Determine the treatment group, control group, treatment, and control for each clinical study in Example 1.3.1.

a. A researcher may want to study whether or not smoking increases a person's chances of heart disease.
   The treatment group is the people in the study who smoke and the treatment is smoking. The control group is the people in the study who do not smoke and the control is not smoking.

b. A researcher may want to study whether a new antidepressant drug will work better than an old antidepressant drug.

The treatment group is the people in the study who take the new antidepressant drug and the treatment is taking the new antidepressant drug. The control group is the people in the study who take the old antidepressant drug and the control is taking the old antidepressant drug. *Note: In this case the control group is given some treatment since you should not give a person with depression a non-treatment.*

c. A researcher may want to study whether taking folic acid before pregnancy will decrease the risk of birth defects.

The treatment group is the women who take folic acid before pregnancy and the treatment is taking folic acid. The control group is the women who do not take folic acid before pregnancy and the control is not taking the folic acid. *Note: In this case, you may choose to do an observational study of women who did or did not take folic acid during pregnancy so that you are not inducing women to avoid folic acid during pregnancy which could be harmful to their baby.*

Confounding Variables: These are other possible causes that may produce the effect of interest rather than the treatment under study. Researchers minimize the effect of confounding variables by comparing the results from the treatment group versus the control group.

Controlled Study: Any clinical study where the researchers compare the results of a treatment group versus a control group.

Placebo: A placebo is sometimes used on the control group in a study to mimic the treatment that the treatment group is receiving. The idea is that if a placebo is used, then the people in the control group and in the treatment group will all think that they are receiving the treatment. However, the control group is merely receiving something that looks like the treatment, but should have no effect on the outcome. An example of a placebo could be a sugar pill if the treatment is a drug in pill form.

Example 1.3.3: Placebo Examples

For each situation in Example 1.3.1, identify if a placebo is necessary to use.

a. A researcher may want to study whether or not smoking increases a person's chances of heart disease.

In this example, it is impossible to use a placebo. The treatment group is comprised of people who smoke and the control group is comprised of people who do not smoke. There is no way to get the control group to think that they are smoking as well as the treatment group.

b. A researcher may want to study whether a new antidepressant drug will work better than an old antidepressant drug.

   In this example, a placebo is not needed since we are comparing the results of two different antidepressant drugs.

c. A researcher may want to study whether taking folic acid before pregnancy will decrease the risk of birth defects.

   In this example, the control group could be given a sugar pill instead of folic acid. However, they may think that they are taking folic acid and so the psychological effect on a person's health can be measured. This way, when we compare the results of taking folic acid versus taking a sugar pill, we can see if there were any dramatic differences in the results.

Blind Study: Usually, when a placebo is used in a study, the people in the study will not know if they received the treatment or the placebo until the study is completed. In other words, the people in the study do not know if they are in the treatment group or in the control group. This type of study is called a blind study. *Note: When researchers use a placebo in a blind study, the people in the study are told ahead of time that they may be getting the actual treatment, or they may be getting the placebo.*

Double-Blind Study: Sometimes when researchers are conducting a very extensive study using many healthcare workers, the researchers will not tell the people in the study or the healthcare workers which patients will receive the treatment and which patients will receive the placebo. In other words, the healthcare workers who are administering the treatment or placebo to the people in the study do not know which people are in the treatment group and which people are in the control group. This type of study is called a double-blind study.

Randomized Controlled Study: Any clinical study in which the treatment group and the control group are selected randomly from the population.

Parameter and Statistic: Whether you are doing an observational study or an experiment, you need to figure out what to do with the data. You will have many data

values that you collected, and it sometimes helps to calculate numbers from these data values. Whether you are talking about the population or the sample, determines what we call these numbers.

Parameter: A numerical value calculated from a population
Statistic: A numerical value calculated from a sample, and used to estimate the parameter

Some examples of parameters that can be estimated from statistics are the percentage of people who strongly agree to a question and mean net worth of all Americans. The statistic would be the percentage of people asked who strongly agree to a question, and the mean net worth of a certain number of Americans.

Notation for Parameter and Statistics:
Parameters are usually denoted with Greek letters. This is not to make you learn a new alphabet. It is because there just are not enough letters in our alphabet. Also, if you see a letter you do not know, then you know that the letter represents a parameter. Examples of letters that are used are μ (mu), σ (sigma), ρ (rho), and p (yes this is our letter because there is not a good choice in the Greek alphabet).
Statistics are usually denoted with our alphabet, and in some cases we try to use a letter that would be equivalent to the Greek letter. Examples of letter that are used are $\bar{x}$ (x-bar), s, r, and $\hat{p}$ (p-hat, since we already used p for the parameter).
In addition, N is used to denote the size of the population and n is used to denote the size of the sample.

Sampling Error: This is the difference between a parameter and a statistic. There will always be some error between the two since a statistic is an estimate of a parameter. Sampling error is attributed to chance error and sample bias. This comes from the fact that two different samples from the same population will likely give two different statistics.

Chance Error: The error is because what you see in your population or sample could just be a coincidence. It could just happen because it happened and not because of anything at all. The question in statistics is, "Are parameters or differences in parameters due to chance or due to something else?"

Sample Bias: The error from using a sample that does not represent the population. To avoid this, use some sort of random sample.
Sampling Rate: The fraction of the total population that is in the sample. This can be denoted by n/N.

Section 1.4: Should You Believe a Statistical Study?

Now we have looked at the basics of a statistical study, but how do you make sure that you conduct a good statistical study? You need to use the following guidelines.

Guidelines for Conducting a Statistical Study:

1. State the goal of your study precisely. Make sure you understand what you actually want to know before you collect any data. Determine exactly what you would like to learn about.
2. State the population you want to study and state the population parameter(s) of interest.
3. Choose a sampling method. A simple random sample is the best type of sample, though sometimes a stratified or cluster sample may be better depending on the question you are asking.
4. Collect the data for the sample and summarize these data by finding sample statistics.
5. Use the sample statistics to make inferences about the population parameters.
6. Draw conclusions: Determine what you learned and whether you achieved your goal.

The mistake that most people make when doing a statistical study is to collect the data, and then look at the data to see what questions can be answered. This is actually backwards. So, make sure you know what question you want to answer before you collect any data. However, in the age of "big data," more and more often data is collected and then the data is analyzed to determine if there are conclusions that can be made. Because of this new trend, the field of data science or data analytics is emerging. This concept of understanding your question before collecting the data is evolving.

Even if you do not conduct your own study, you will be looking at studies that other people have conducted. Every day you hear and see statistics on the news, in newspapers and magazines, on the Internet and other places. Some of these statistics may be legitimate and beneficial, but some may be inaccurate and misleading. Here are some steps to follow when evaluating whether or not a statistical study is believable.

Steps for Determining whether a Statistical Study is Believable:

1. Are the population, goal of the study, and type of study clearly stated?
You should be able to answer the following questions when reading about a statistical study:

- Does the study have a clear goal? What is it?

- Is the population defined clearly? What is it?
- Is the type of study used clear and appropriate?

2. Is the source of the study identified? Are there any concerns with the source?
A study may not have been conducted fairly if those who funded the study are biased.

Example 1.4.1: Source of Study 1

Suppose a study is conducted to find out the percentage of United States college professors that belong to the Libertarian party. If this study was paid for by the Libertarian party, or another political party, then there may have been bias involved with conducting the study. Usually an independent group is a good source for conducting political studies.

Example 1.4.2: Source of Study 2

There was once a full-page ad in many of the newspapers around the U.S. that said that global warming was not happening. The ad gave some reasons why it was not happening based on studies conducted. At the bottom of the page, in small print, were the words that the study and ad were paid for by the oil and gas industry. So, the study may have been a good study, but since it was funded by the industry that would benefit from the results, then you should question the validity of the results.

3. Are there any confounding variables that could skew the results of the study?
Confounding variables are other possible causes that may produce the effect of interest besides the variable under study. In a scientific experiment researchers may be able to minimize the effect of confounding variables by comparing the results from a treatment group versus a control group.

Example 1.4.3: Confounding Variable

A study was done to show that microwave ovens were dangerous. The study involved plants, where one plant was given tap water and one plant was given water that was boiled in the microwave oven. The plant given the water that was boiled died. So the conclusion was that microwaving water caused damage to the water and thus caused the plant to die. However, it could easily have been the fact that boiling water was poured onto the plant that caused the plant to die.

4. Could there be any bias from the sampling method that was used?
Sometimes researchers will take a sample from the population and the results may be biased.
Selection Bias: This occurs when the sample chosen from the population is not representative of the population.
Participation Bias (or Nonresponse Bias): This occurs when the intended objects in the sample do not respond for many different reasons. Those who feel strongly about an issue will be more likely to participate.

Example 1.4.4: Bias

The 1936 *Literary Digest* Poll. The *Literary Digest* was a magazine that was founded in 1890. Starting with the 1916 U.S. presidential election, the magazine had predicted the winner of each election. In 1936, the *Literary Digest* predicted that Alfred Landon would win the election in a landslide over Franklin Delano Roosevelt with fifty-seven percent of the popular vote. The process for predicting the winner was that the magazine sent out ten million mock ballots to its subscribers and names of people who had automobiles and telephones. Two million mock ballots were sent back. In reality, Roosevelt won the election with 62% of the popular vote. ("Case Study 1: The 1936 *Literary Digest* Poll," n.d.)

A side note is that at the same time that the *Literary Digest* was publishing its prediction, a man by the name of George Gallup also conducted a poll to predict the winner of the election. Gallup only polled about fifty thousand voters using random sampling techniques, yet his prediction was that Roosevelt would win the election. His polling techniques were shown to be the more accurate method and have been used to present-day.

Selection Bias: Because of the people whom the *Literary Digest* polled, they created something called a selection bias. The poll asked ten million people who owned cars, had telephones, and subscribed to the magazine. Today, you would probably think that this group of people would be representative of the entire U.S. However, in 1936 the country was in the midst of the Great Depression. So the people polled were mostly in the upper middle to upper class. They did not represent the entire country. It did not matter that the sample was very large. The most important part of a sample is that it is representative of the entire population. If the sample is not, then the results could be wrong, as demonstrated in this case. It is important to collect data so that it has the best chance of representing the entire population.

Nonresponse Bias: When looking at the number of ballots returned, two million appears to be a very large number. However, ten million ballots were sent out. So that means that only about one-fifth of all the ballots were actually returned. This is known as a nonresponse bias. The only people who probably took the time to fill out and return the ballot were those who felt strongly about the issue. So when you send out a survey, you have to pay attention to what percentage of surveys are actually returned. If at all possible, it is better to conduct the survey in person or through the telephone. Most credible polls conducted today, such as Gallup, are conducted either in person or over the telephone. Do be careful though, just because a polling group conducts the poll in person or on the telephone does not mean that it is necessarily credible.

5. Are there any problems with the setting of a survey?
The setting of the survey can create bias. So you want to make sure that the setting is as neutral as possible, so that someone does not answer based on where the survey is conducted or who is giving the survey.

Example 1.4.5: Setting Example

Suppose a survey is being conducted to learn more about illegal drug use among college students. If a uniformed police officer is conducting the survey, then the results will very likely be biased since the college students may feel uncomfortable telling the truth to the police officer.

6. Are there any problems with the wording of a survey?
How a question is worded can elicit a particular response. Also the order of the questions may affect a person's answers. So make sure that the questions are worded in a way that would not lead a person to a particular answer.

Example 1.4.6: Wording Example

A question regarding the environment may ask "Do you think that global warming is the most important world environmental issue, or pollution of the oceans?" Alternatively, the question may be worded "Do you think that pollution of the oceans is the most important world environmental issue, or global warming?" The answers to these two questions will vary greatly simply because of how they are worded. The best way to handle a question like this is to present it in multiple choice format as follows:
What do you think is the most important world environmental issue?
a. Global warming

b. Pollution of the oceans
c. Other

7. Are the results presented fairly?
Be sure that any concluding statements accurately represent the data and statistics that were calculated from the data. Many times people make conclusions that are beyond the scope of the study, or are beyond the results of the data.

Example 1.4.7: Wrong Conclusion

Many studies have been done on cancer treatments using rats. The wrong conclusion is to say that because a treatment cured cancer in rats, then it will cure cancer in people. The fact that a treatment cured cancer in rats, means that there is a chance that it will cure cancer in people, but you would have to try it on people before making such claims. Rats and people have different physiology, so you cannot assume what works on one will work on the other.

8. Are there any misleading graphics?
Be sure that any graphics that are presented along with the results are not misleading. Some examples of misleading graphs are:

- The vertical axis does not start at zero. This means that any changes will look more dramatic than they really are.
- There is no title. This means that you do not know what the graph is actually portraying.
- There are missing labels or units. This means that you do not know what the variables are or what the units are.
- The wrong type of graph is used. Sometimes people use the wrong graph, like using a bar graph when a line graph would be more appropriate.

9. Final considerations
Ask yourself the following questions about the overall effectiveness of the statistical study.

- Do the conclusions of the study answer the initial goal of the study?
- Do the conclusions of the study follow from the data and statistics?
- Do the conclusions of the study indicate practical changes should be made?

Overall, you should follow these steps when analyzing the validity of any statistical study.

Section 1.5: Graphs

Once we have collected data, then we need to start analyzing the data. One way to analyze the data is using graphical techniques. The type of graph to use depends on the type of data you have. Qualitative data use graphs like bar graphs, pie graphs, and pictograms. Quantitative data use graphs such as histograms. In order to create any graphs, you must first create a summary of the data in the form of a frequency distribution. A frequency distribution is created by listing all the data values (or grouping of data values) and how often the data value occurs.

Frequency: Number of times a data value occurs in a data set.

Frequency Distribution: A listing of each data value or grouping of data values (called classes) with their frequencies.

Relative Frequency: The frequency divided by n, the size of the sample. This gives the percent of the total for each data value or class of data values.

Relative Frequency Distribution: A listing of each data value or class of data values with their relative frequencies.

How to create a frequency distribution depends on whether you have qualitative or quantitative variable. We will now look at how to create each type of frequency distribution according to the type of variable, and the graphs that go with them.

Qualitative Variable:
First let's look at the types of graphs that are commonly created for qualitative variables. Remember, qualitative variables are words, and not numbers.

Bar graph: A graph where rectangles represent the frequency of each data value or class of data values. The bars can be drawn vertically or horizontally. *Note: The bars do not touch and they are the same width.*

Pie Chart: A graph where the "pie" represents the entire sample and the "slices" represent the categories or classes. To find the angle that each "slice" takes up, multiply the relative frequency of that slice by 360°. *Note: The percentages in each slice of a pie chart must all add up to 100%.*

Pictograms: A bar graph where the bars are made up of icons instead of rectangles.

Pictograms are overused in the media and they are the same as a regular bar graph except more eye-catching. To be more professional, bar graphs or pie charts are better.

Example 1.5.1: Qualitative Variable Frequency Distribution and Graphs

Suppose a class was asked what their favorite soft drink is, and the following is the results:

Table 1.5.1: Favorite Soft Drink

Coke	Pepsi	Mt. Dew	Coke	Pepsi	Dr. Pepper	Sprite	Coke	Mt. Dew
Pepsi	Pepsi	Dr. Pepper	Coke	Sprite	Mt. Dew	Pepsi	Dr. Pepper	Coke
Pepsi	Mt. Dew	Coke	Pepsi	Pepsi	Dr. Pepper	Sprite	Pepsi	Coke
Dr. Pepper	Mt. Dew	Sprite	Coke	Coke	Pepsi			

a. Create a frequency distribution for the data.

To do this, just list each drink type, and then count how often each drink comes up in the list. Notice Coke comes up nine times in the data set. Pepsi comes up 10 times. And so forth.

Table 1.5.2: Frequency Distribution of Favorite Soft Drink

Drink	**Coke**	**Pepsi**	**Mt Dew**	**Dr. Pepper**	**Sprite**
Frequency	9	10	5	5	4

b. Create a relative frequency distribution for the data.

To do this, just divide each frequency by 33, which is the total number of data values. Round to three decimal places.

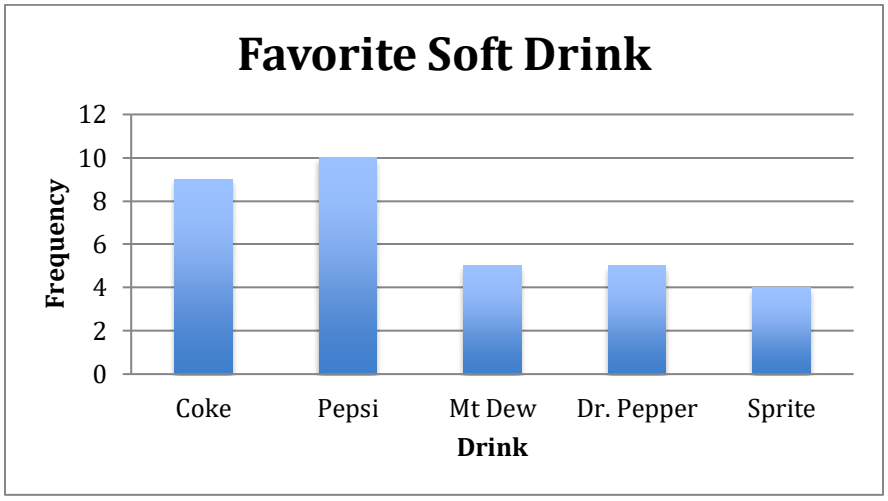
Table 1.5.3: Relative Frequency Distribution of Favorite Soft Drink

Drink	**Coke**	**Pepsi**	**Mt Dew**	**Dr. Pepper**	**Sprite**
Frequency	9	10	5	5	4
Relative Frequency	9/33 =0.273 =27.3%	10/33 =0.303 =30.3%	5/33 =0.152 =15.2%	5/33 =0.152 =15.2%	4/33 =0.121 =12.1%

c. Draw a bar graph of the frequency distribution.

Along the horizontal axis you place the drink. Space these equally apart and allow space to draw a rectangle above it. The vertical axis contains the frequencies. Make sure you create a scale along that axis in which all the frequencies will fit. Notice that the highest frequency is 10, so you want to make sure the vertical axis goes to at least 10, and you may want to count by two for every tick mark. Using Excel, this is what your graph will look like.

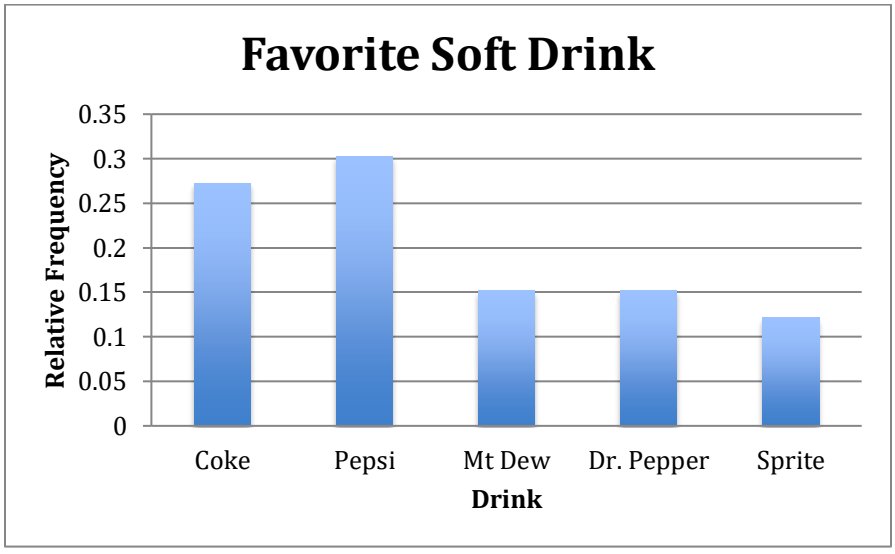
Graph 1.5.4: Bar Graph of Favorite Soft Drink



d. Draw a bar graph of the relative frequency distribution.

This is similar to the bar graph for the frequency distribution, except that you use the relative frequencies instead. Notice that the graph does not actually change except the numbers on the vertical scale.

Graph 1.5.5: Relative Frequency Bar Graph of Favorite Soft Drink



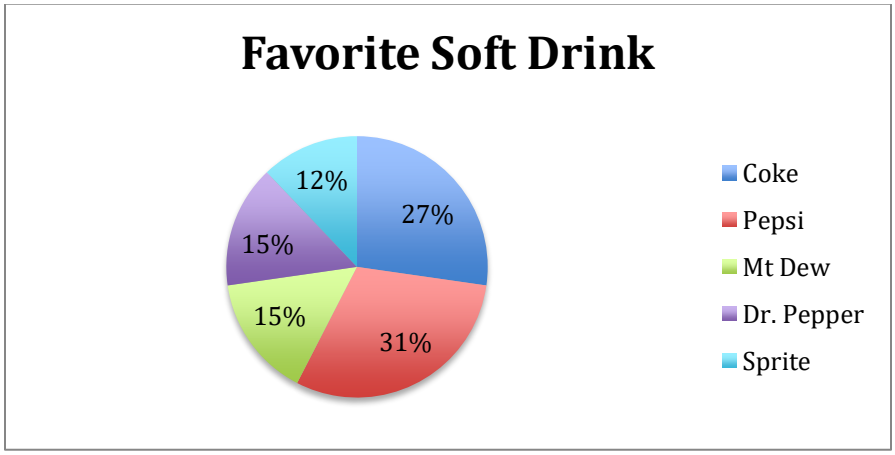
e. Draw a pie chart of the data.

To draw a pie chart, multiply the relative frequencies by 360°. Then use a protractor to draw the corresponding angle. Or it is easier to use Excel, or some other spreadsheet program to draw the graph.

Table 1.5.6: Angles for the Pie Chart of Favorite Soft Drink

Drink	Coke	Pepsi	Mt Dew	Dr. Pepper	Sprite
Frequency	9	10	5	5	4
Relative Frequency	0.273	0.303	0.152	0.152	0.121
Angles	(9/33)*360 =98.2°	(10/33)*360 =109.1°	(5/33)*360 =54.5°	(5/33)*360 =54.5°	(4/33)*360 =43.6°

Graph 1.5.7: Pie Chart for Favorite Soft Drink



f. Draw a pictograph for the favorite soft drink data.

Here you can get creative. One thing to draw would be glasses. Now you would not want to draw 10 glasses. So what you can do is let each glass be worth a certain number of data values, let's say one glass = frequency of two. So this means that you will need to draw half of a glass for some of the frequencies. So for the first drink, with a frequency of nine, you need to draw four and a half glasses. For the second drink, with a frequency of 10, you need to draw five glasses. And so on.

Graph 1.5.8: Pictograph for Favorite Soft Drink

Pictographs are not really useful graphs. The makers of these graphs are trying to use graphics to catch a person's eye, but most of these graphs are missing labels, scaling, and titles. Additionally, it can sometimes be unclear what ½ or ¼ of an icon represents. It is better to just do a bar graph and use color to catch a person's eye.

Quantitative Variable:
Quantitative variables are numbers, so the graph you create is different from the ones for qualitative data. First, the frequency (or relative frequency) distribution is created by dividing the interval containing the data values into equally spaced subintervals. Then you count how many data values fall into each subinterval. Since the subintervals do not overlap, but do touch, then the graph you create has the bars touching.

Histogram: A graph of a quantitative variable where rectangles are used for each subinterval, the height of the rectangle represents the frequency (or relative frequency) of the data values in the subinterval, and there are no gaps in between the rectangles. Sometimes the midpoint of each subinterval is graphed instead of the endpoints of the subinterval.

Example 1.5.2: Quantitative Variable Frequency Distribution and Graphs

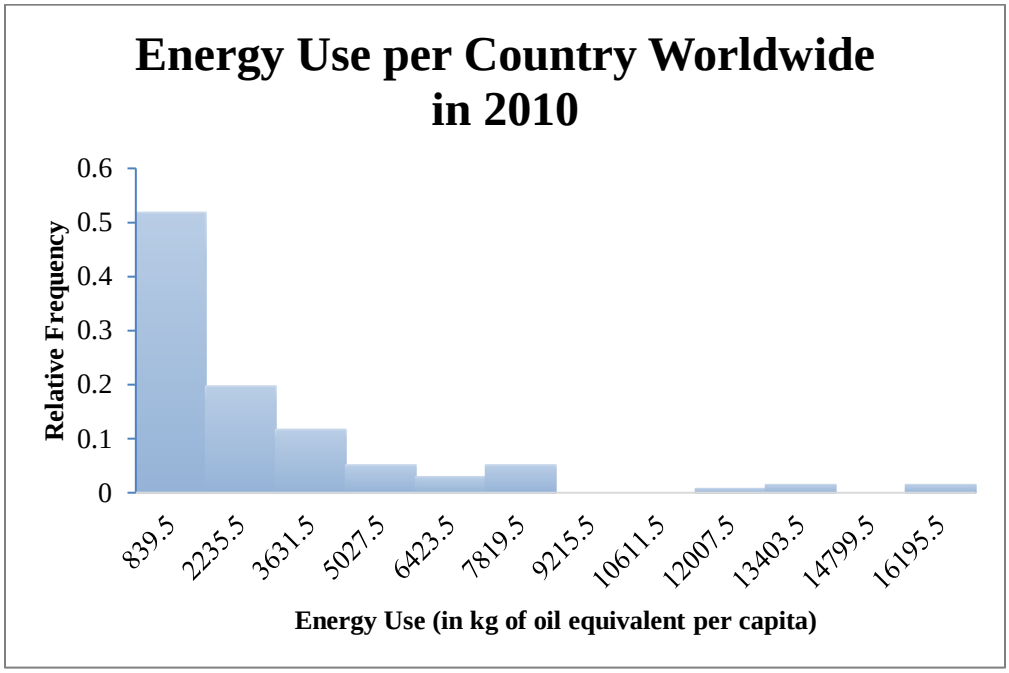
The energy used (in kg of oil equivalent per capita) in 2010 of 137 countries around the world is summarized in the following relative frequency distribution. Use this distribution draw a histogram. (World Bank, 2010).

This relative frequency distribution was created by dividing the range of the data into 12 equally spaced subintervals, sometimes called classes.

Table 1.5.9: Relative Frequency Distribution for Energy Used

Lower limit	Upper limit	Midpoint	Frequency	Relative Frequency
142	1537	839.5	71	0.5182
1538	2933	2235.5	27	0.1971
2934	4329	3631.5	16	0.1168
4330	5725	5027.5	7	0.0511
5726	7121	6423.5	4	0.0292
7122	8517	7819.5	7	0.0511
8518	9913	9215.5	0	0
9914	11309	10611.5	0	0
11310	12705	12007.5	1	0.0073
12706	14101	13403.5	2	0.0146
14102	15497	14799.5	0	0
15498	16893	16195.5	2	0.0146

Graph 1.5.10: Histogram for Energy Used in 2010 for 137 Countries in the World

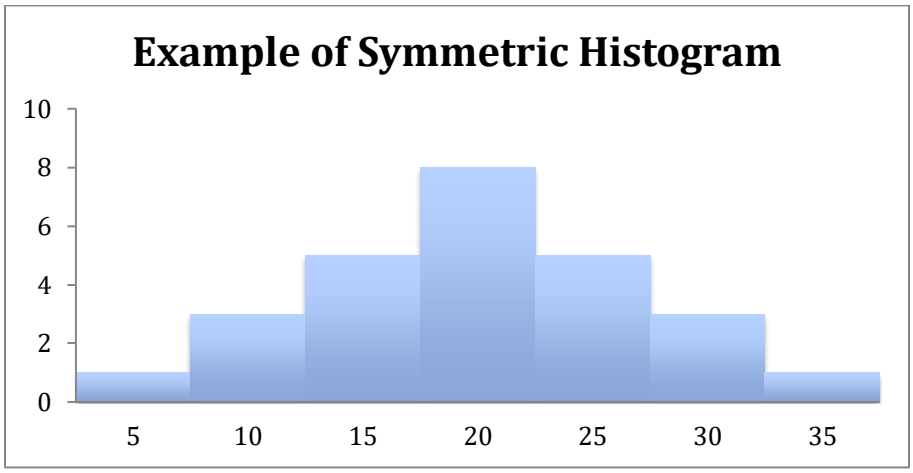


Notice that the vertical axis starts at 0, there is a title on the graph, the axes have labels, and the tick marks are labeled. This is a correct way to draw a graph and allows people to know what the data represents.

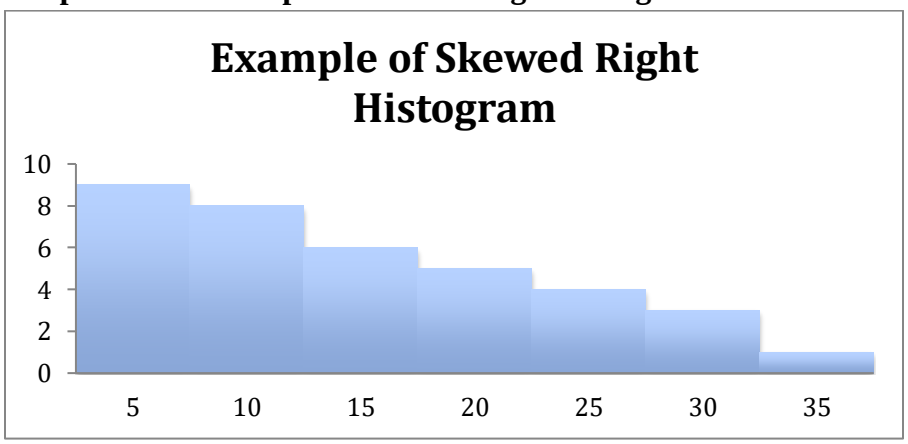
Interpreting graphs:

It is important to be able to interpret graphs. If you look at the graphs in Example 1.5.1, you can see that Pepsi is more popular than any of the other drinks. You can also see that Sprite is the least popular, and that Mt. Dew and Dr. Pepper are equally liked. If you look at the graph in Example 1.5.2, you can see that most countries use around 839.5 kg of energy per capita. You can also see that the graph is heavily weighted to the lower amounts of energy use, and that there is a gap between the bulk of the amounts and the higher ends. So there are very few countries that use over 9215.5 kg of energy per capita. Since the data is quantitative, we can talk about the shape of the distribution. This graph would be called skewed right, since the data on the right side of the graph is the unusual data, and if it was not there, then the graph may look more symmetric. Some basic shapes of histograms are shown below.

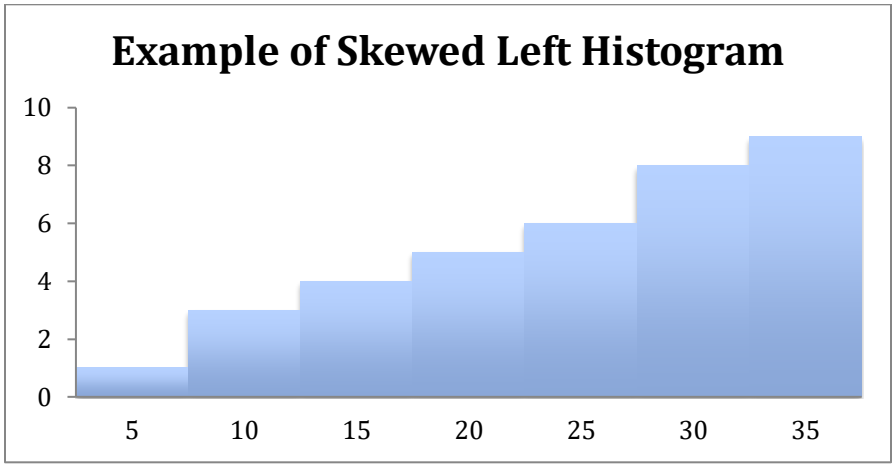
Graph 1.5.11: Example of Symmetric Histogram



Graph 1.5.12: Example of Skewed Right Histogram



Graph 1.5.13: Example of Skewed Left Histogram



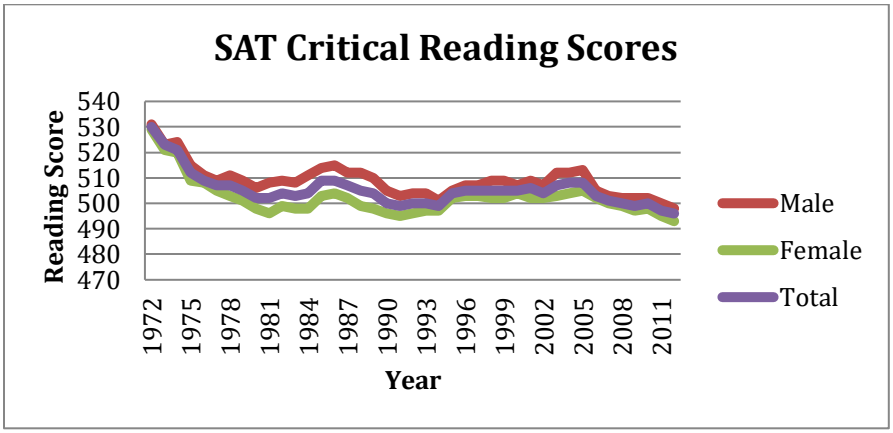
Section 1.6: Graphics in the Media

There are many other types of graphs you will encounter in the media.

Multiple Line Graphs:

A line graph is useful for seeing trends over time. Multiple line graphs are useful for seeing trends over time, and also comparing two or more data sets. As an example, suppose you want to examine the average SAT critical reading score over time for Arizona students, but you further want to compare the averages overall and between genders. So, a multiple line graph like the following may be used. As you can see, the average score for SAT critical reading has been going down over the years. You can also see that the average score for male students is higher than the average score for female students for all of the years. The other interesting aspect that you can see is that the average scores for male and female used to be closer to each other, then they separated, and look to be getting closer to each other again. One last comment is that even though there is a difference between male and female average score, the average scores seem to follow each other. In other words, when the male score was going down, so was the female score. Do be careful. Do not try to make up a reason for the scores to go down. You cannot say why the scores have decreased, since you did not run an experiment. The scores could have decreased because our education system is not teaching as well, funding has decreased for education, or the intelligence of Arizona students has decreased. Or it could be that the percentage of the overall student body that takes the SAT has increased over the years, meaning that more than just the highest ranked students have been taking the SAT in later years, which could lower the averages. Any one of these reasons, or other reasons, could be the right one, and you cannot determine which it is. Do not make unsubstantiated claims. *Note: A next step in analyzing this data could be to compare the AZ trends to national trends.*

Graph 1.6.1: Multiple Line Graph for SAT Critical Reading Scores in AZ



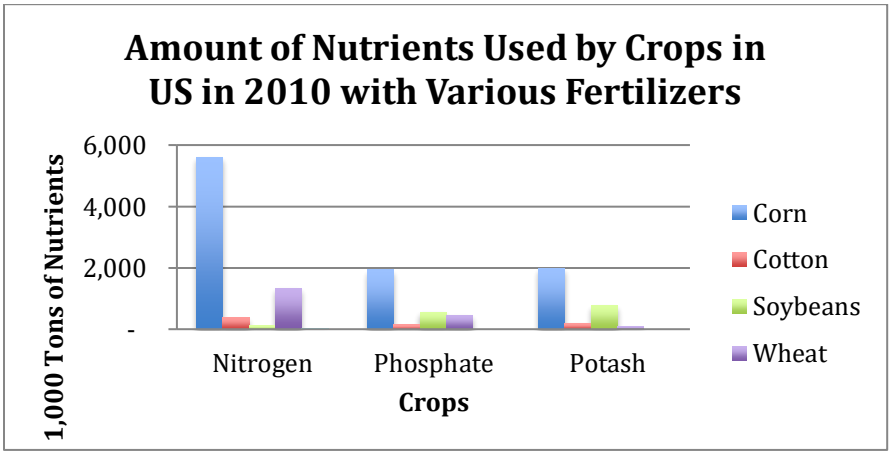
(College Board: Arizona, 2012)

Note: In this case the vertical axis did not start at zero, because if it did start at zero, the different lines would be very close together and difficult to see. When at all possible, the scale on the vertical axis should start at zero. Always carefully consider that if a vertical scale of a graph does not start at zero, the researchers could be attempting to exaggerate insignificant differences in the data.

Multiple Bar Graphs:

Sometimes you have information for multiple variables and instead of putting the information on different bar graphs, you can put them all on one so that you can compare the variables. The following is an example of where you might use this. The data is the amount of nutrients used by crops with various fertilizers. By creating a graph that has all of the fertilizers and all of the crops, you can see that corn with nitrogen uses the most nutrients, and soybeans with nitrogen uses the least amount of nutrients. You can also see that phosphate seems to use low amounts of nutrients for all of its crops. So, a multiple bar graph is useful to make all these observations.

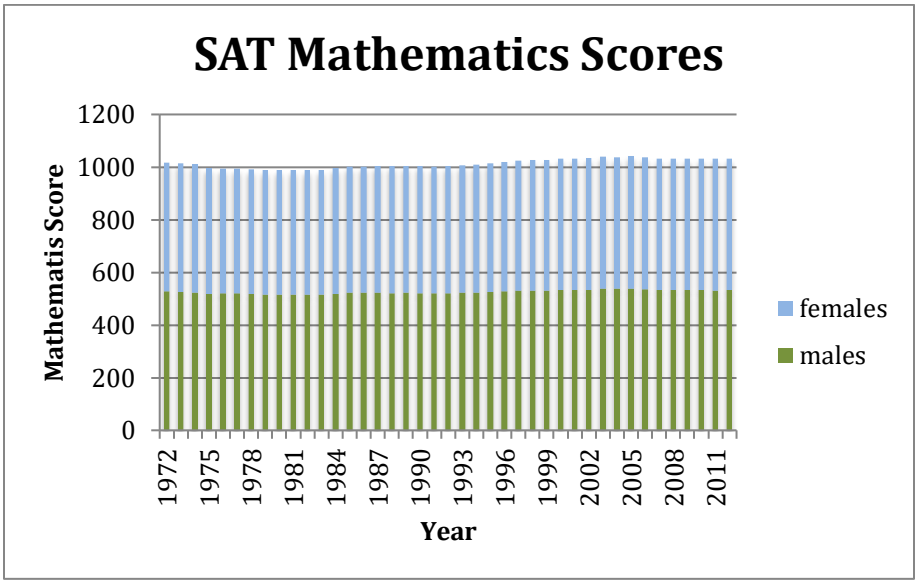
Graph 1.6.2: Multiple Bar Graph for the Amount of Nutrients Used by Crops: 2010



(United States Department of Agriculture [USDA], 2010)

Stack Plots:
A stack plot is basically a multiple line graph, but with the lines separated (or stacked) on top of each other instead of overlapping. This can be useful when it is difficult to interpret a multiple line graph since the lines are so close to one another. To read a given line on a stack plot, you must subtract that line from the line below it. In the example of a stack plot below, you can see that the SAT Mathematics score for males in 1972 is about 530 while the SAT Mathematics score for females in 1972 is about 1010 – 530 = 480.

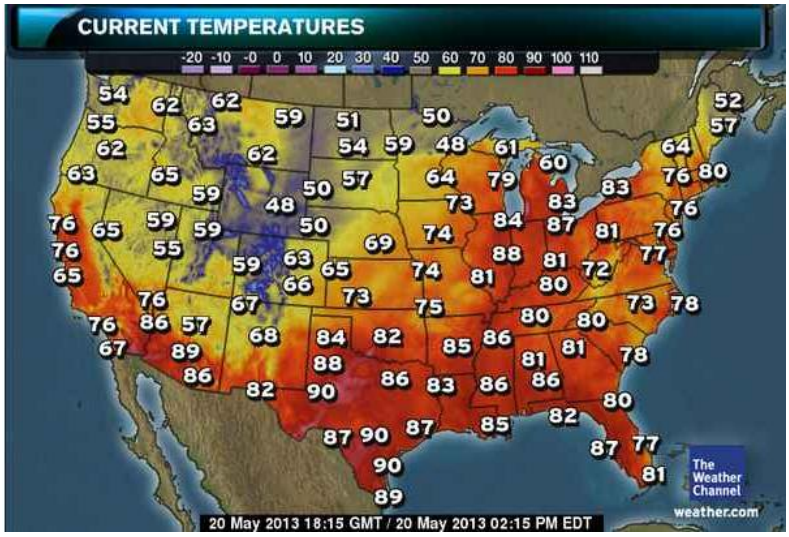
Graph 1.6.3: Stack Plot of SAT Mathematics Scores



(College Board: Arizona, 2012)

Geographical Graphs:
Weather maps, topographic maps, population distribution maps, gravity maps, and vegetation maps are examples of geographical graphs. They allow you to see a trend of information over a geographic area. The following is an example of a weather map showing temperatures. As you can see, the different colors represent certain temperature ranges. From this graph, you can see that on this date, the red in the south means that the temperature was in the 80s and 90s there, and the blue in the Rockies area means that the temperature was in the 40s there.

Graph 1.6.4: Geographical Graph

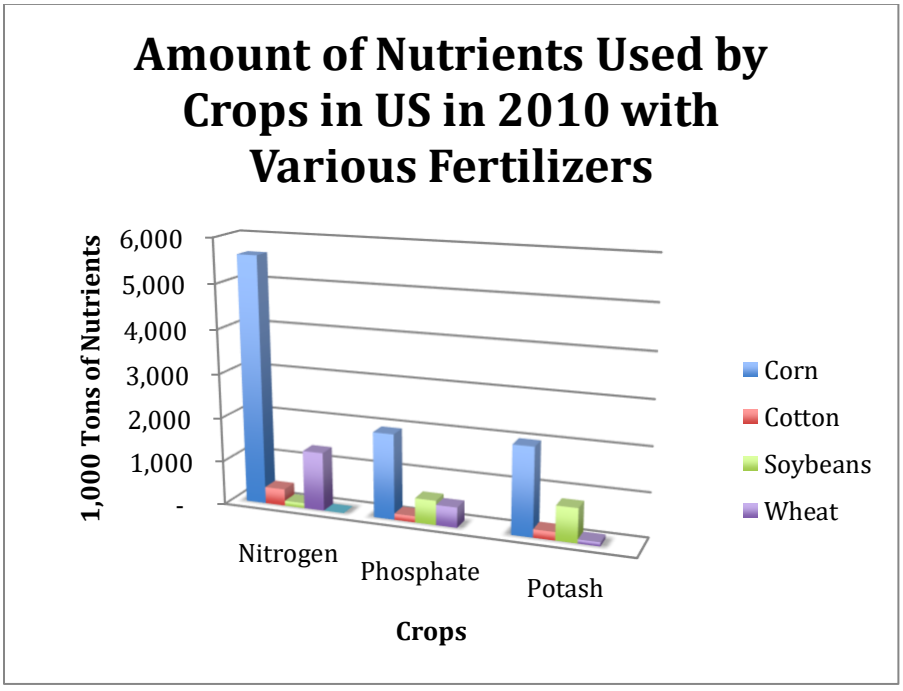


(Weather Channel, 2013)

Three-Dimensional Graphics

Some people like to show a bar graph in three-dimensions. Occasionally, a three-dimensional graph is used to graph three variables together on three axes, but this type of graph may be difficult to read. The following graph just represents two variables and so it is basically the same as a standard bar graph, but the three-dimensional look may add a bit more style.

Graph 1.6.5: Three-Dimensional Graph

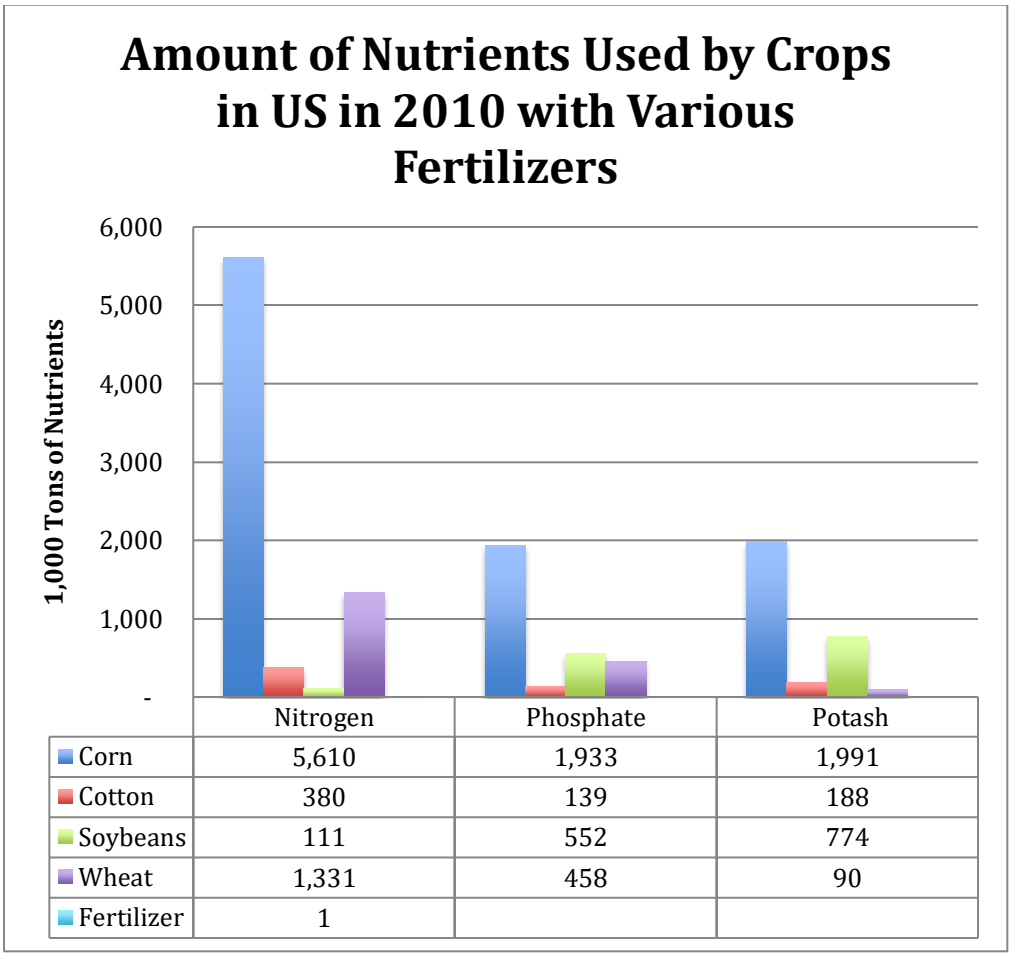


(USDA, 2010)

Combination Graphics

Some graphs are created so that they combine the data table and a graph or combine two types of graphs in one. The advantage is that you can see a graphical representation of the data, and still have the data to find exact values. The disadvantage is that they are busier, and usually people show graphical representation of data because people do not like looking at the data.

Graph 1.6.6: Combination Graph



	Nitrogen	Phosphate	Potash
Corn	5,610	1,933	1,991
Cotton	380	139	188
Soybeans	111	552	774
Wheat	1,331	458	90
Fertilizer	1		

(USDA, 2010)

These are just a few of the different types of graphs that exist in the world. There are many other ones. A quick Google search on statistical graphs will show you many more. Just open a newspaper, magazine, or website and you are likely to see others. The most important thing to remember is that you need to look at the graph objectively and interpret for yourself what it says. Also, do not read any cause and effect into what you see

Chapter 1 Homework

1. A study was conducted of schools across the U.S. about whether they require school uniforms. Two hundred ninety-six schools gave their response to the question, "Does your school require school uniforms?" State the population and sample.

2. The U.S. Department of Labor collects information on the average hourly earnings of professional and business services positions. Suppose that you look at 20 years and collect data on the average hourly earnings in those years. State the population and sample.

3. A person collects the gas prices at 25 gas stations in Phoenix, AZ. State the population and sample.

4. The Center for Disease Control collects data on the number of children with autism. They collect data on 32,601 children in the state of Arizona, and then look to see how many have autism. State the population and sample.

5. Determine if the variable is qualitative or quantitative. If quantitative, then also state if the variable is discrete or continuous.
   a. Height of buildings in a town
   b. Eye color of all students at a college
   c. Weight of cars
   d. Number of dogs in a household
   e. Religion of people in a town
   f. Number of fish caught daily

6. Determine if the variable is qualitative or quantitative. If quantitative, then also state if the variable is discrete or continuous.
   a. Letter grades of students in a class
   b. Distance a person runs every day
   c. Number of prairie dogs on a parcel of land
   d. Time that a task takes to complete
   e. Gender of a person
   f. Number of cars at a dealership on a given day

7. A study to estimate the average salary of workers at a university was conducted using the following designs. Categorize the sampling method as a simple random sample, stratified sample, cluster sample, systematic sample, or convenience sample.
   a. Each person who is employed by the university is first divided into groups of administrative professional, classified, faculty, and part-time. Then each person in those groupings is given a number and then random samples are taken inside each grouping.
   b. Each researcher asks the first 40 people he or she encounters on campus what their salary is.
   c. The researchers number all employees, and then start with the 34^{th} person. Then they record the salary of every 10^{th} person after that.
   d. The researchers number all employees, and then use a random number generator to determine which employees they will use.
   e. Each college on campus is given a number. Then five colleges are chosen at random and all employees' salaries in each college is recorded.

8. A study to determine the opinion of Americans about the use of marijuana for medical purposes is being conducted using the following designs. Categorize the sampling method as a simple random sample, stratified sample, cluster sample, systematic sample, or convenience sample.
   a. The researchers attend a festival in a town in Kansas and ask all the people they can what their opinions are.
   b. The researchers divide Americans into groups based on the person's race, and then take random samples from each group.
   c. The researchers number all Americans and call the 50^{th} person on the list. Then they call every $10{,}000^{th}$ person after the 50^{th} person.
   d. The researchers call every person in each of 10 area codes that were randomly chosen.
   e. The researchers number every American, and then call all randomly selected Americans.

9. A biologist is looking to see the effect of microgravity on plant growth. The researcher sends some seeds to the International Space Station and has the height of the plants measured on specific days. The researcher also plants the same number of seeds in a laboratory on Earth, using the same lighting, soil, and water conditions, and measures the height of the plants on specific days. Describe the treatment group and the control group.

10. To see if a new blood pressure medication works or not, volunteers are divided into two groups. One group is given the new medication and the other group is given an older medication. Describe the treatment group and the control group.

11. In each situation state if a placebo is needed or not and explain your reasoning.
    a. A new medication for treating heartworm in dogs is being investigated to see if it works.
    b. A new training method for managers is being investigated to see if it improves morale.
    c. A new medication is being tested to see if it reduces itching due to eczema.

12. In each situation state if a placebo is needed or not and explain your reasoning.
    a. A new drug is being developed to treat high blood pressure and needs to be tested to see if it works better than a previous type of drug.
    b. A new medication is being developed to treat headaches.
    c. A new drink flavor is being developed to see if it is marketable.

13. To see if a new medication works to reduce fevers, volunteers are divided into two groups. One group is given the new medication, and the other group is given a placebo. The volunteers do not know which group they are in, but the researchers do. Is this a blind or double-blind study?

14. To see if a new blood pressure medication works or not, volunteers are divided into two groups. One group is given the new medication and the other group is given the old medication. The volunteers do not know which group they are in, and neither do the researchers. Is this a blind or double-blind study?

15. In each situation, identify a potential source of bias and explain your reasoning.
    a. A study on teenage boys shows that a new drug works on curing acne. The company then concludes that the new drug will work on everyone.
    b. A radio station asks listeners to phone in their choice in a daily poll.
    c. The Beef Council releases a study stating that consuming red meat poses little cardiovascular risk.
    d. A poll asks, “Do you support a new transportation tax, or would you prefer to see our public transportation system fall apart?”

e. A study is conducted on whether artificial light is better for plants. A certain type of plant is grown under artificial light and its height is recorded. A different type of plant is grown under natural light and its height is recorded. Both plant groups get the same amount of water and soil.

16. In each situation, identify a potential source of bias and explain your reasoning.
    a. A study is conducted to see if grades improve if students are given tutoring. Students in a calculus class are given tutoring and compared to students in a statistics class who are not given tutoring.
    b. An organic grower conducts a study and shows that pesticides are harmful to people who eat food that used the pesticide.
    c. A poll asks, "If 53% of all people will need assisted living in the future, should you worry about needing assisted living insurance?"
    d. A website asks people to say if they support the President or not.
    e. A study of behavior modification therapy is tried on a specific breed of dogs. The study concludes that the behavior modification therapy works on all breeds of dogs.

17. The average SAT test scores for Michigan students in reading for the years 1971 to 2012 are given in the table below (College Board: Michigan, 2012). Create a table showing frequencies and relative frequencies for the data using the classes of 490 to 496, 497 to 503, 504 to 510, 511 to 517, 518 to 524, and 525 to 531.

496	500	502	504	505	507	509
497	500	503	504	505	507	512
499	500	503	505	505	508	521
499	500	504	505	506	508	523
499	501	504	505	507	509	530
500	502	504	505	507	509	509

18. The average hourly earnings of all employees in the U.S. in professional and business services for each month in the time period from March 1, 2006, to June 1, 2013, is given in the table below (Federal Reserve Bank of St. Louis, 2013). Create a table showing frequencies and relative frequencies for the data using the classes of 23 to 23.9, 24 to 24.9, 25 to 25.9, 26 to 26.9, 27 to 27.9, and 28 to 28.9.

23.40	23.81	24.85	25.72	26.73	26.84	27.19	27.23	27.74
23.87	24.22	24.44	25.46	26.74	26.98	27.14	27.39	27.48
23.23	24.42	24.91	25.47	26.89	26.98	27.43	28.19	27.67
23.14	24.56	24.65	25.72	27.37	26.88	26.96	27.81	28.09
23.68	24.45	24.80	25.59	27.32	27.27	27.06	27.55	27.79
23.31	24.84	25.33	25.66	27.02	27.19	27.31	27.57	27.97
23.64	24.46	25.33	25.89	26.84	27.32	27.18	27.84	28.55
23.97	24.45	25.48	26.05	26.74	27.49	27.21	27.45	28.17
28.03	28.30	27.85	27.76	28.24	27.68	28.27	27.90	28.04
28.58	28.56	28.61	28.43	28.38	28.24	28.57		

19. Create histogram of the data in problem 17. State the overall shape of this histogram.

20. Create a relative frequency histogram of the data in problem 17. State the overall shape of this histogram.

21. Create histogram of the data in problem 18. State the overall shape of this histogram.

22. Create a relative frequency histogram of the data in problem 18. State the overall shape of this histogram.

23. In Kenya, the government is interested in the number of health care facilities in each of the provinces. Below is a table showing this data in 2013 (Kenya Open Data, 2013).

Province: * The province's name was missing from the original data.	**Number of facilities**
RIFT VALLEY	1645
EASTERN	1093
NYANZA	962
NAIROBI	878
COAST	765
CENTRAL	745
WESTERN	541
N. EASTERN	131
*	9
PROV	3

a. Draw a bar graph of the data.
b. Draw a pie chart of the data.

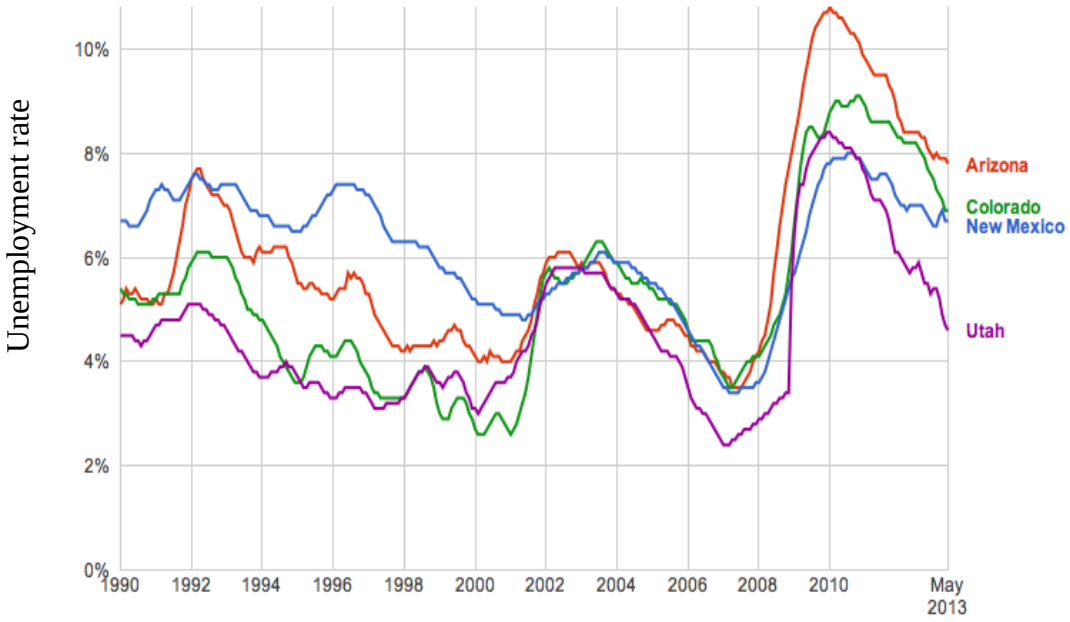
24. According to the Pew Research Center, the percentage of people who own a particular smartphone is given in the table (Smith, 2011).

Platform	**Percentage**
Android	35%
iPhone	24%
Blackberry	24%
Palm device	6%
Windows phone	4%
Unspecified	7%

a. Draw a bar graph of the data.
b. Draw a pie chart of the data.

25. The following is a multiple line graph of unemployment percentages for the four corner states (Arizona, Colorado, New Mexico, and Utah). ("Unemployment Rate," n.d.) Compare and contrast the data for the four states, describing in detail at least three observations you identify from the graph.

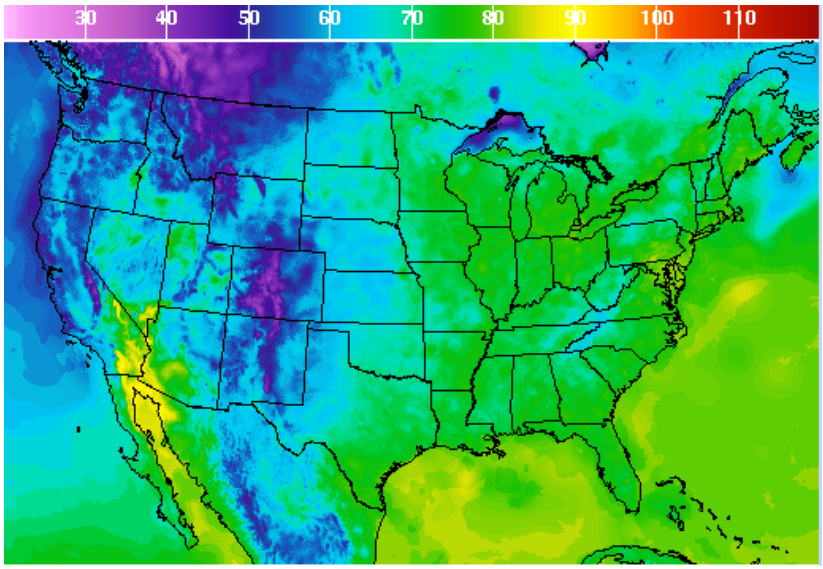
Unemployment Rate from 1990 to 2013 – Seasonally Adjusted



Time (year)

26. Below are the temperatures in the United States on July 16, 2013 at 8 am EDT (National Weather Service, 2013).

Temperature of U.S. on July 16, 2013 at 8 am EDT.



a. What temperature range did Arizona have on that day?
b. What temperature range did Montana have on that day?
c. What temperature range did Texas have on that day?

27. The following graph is the salaries and unemployment rate for different levels of education. ("Employment Projections," n.d.) Describe in detail at least three observations you identify from the graph.