# Section 1.1 Types of Data

**Get Started –** What is the difference between information and knowledge?

* How is qualitative data different from quantitative data?
* How is a population different from a sample?
* How do you tell a statistic from a parameter?
* What is the difference between discrete data and continuous data?

## Get Started – What is the difference between information and knowledge?

What is it that separates good students from poor students?

Are good students just born smarter?

They say that hard work can overcome talent, but many students claim to spend hours studying and still do poorly. Even worse, after spending all that time studying, such a student might get discouraged and give up. They think that they aren’t a “math person”. They must not be if they studied 100’s of hours for the test and still failed…

Well, it’s not JUST about hard work. It’s not even necessarily about working efficiently. What researchers have discovered is that what you are thinking about is extremely important for learning effectively.

Metacognition is thinking about your thinking. In our case, we will be thinking about our learning. We will be learning about our learning process. As we improve our process, we will improve as students (we learn better), which will hopefully improve our grades!

Most Math Students have the bad habit of only studying the day before a test. There are two serious disadvantages attached to this method of study.

The first reason is that this does not allow for regular practice of mathematics. Just like becoming a good athlete or musician or cook or artist, to become good at mathematics you must practice regularly.

One recent study found that it takes 10,000 hours of deliberate practice to achieve mastery in a field. Deliberate practice is a highly structured activity engaged in with the specific goal of improving performance.

The second reason cramming before a test isn’t a good idea has to do with short term vs long term memory. It has been found that within 24 hours - on average – you will forget up to 80% of what you just learned. If the study material is reviewed regularly, you remember more, and it takes longer to forget what you have learned. You turn the short-term memory (Snapchat) into a long-term memory (flash drive). You goal should be long term knowledge, not short-term success.

Building up your mathematical ability is like building a house. You need a strong foundation, because future math subjects are built upon the material that has come before.

When you cram for a test and get a decent grade, you may deceive yourself into mistakenly believing you have mastered an important skill.

You knew the material for a short time (until the test was over), but you won’t remember it three weeks from now when it comes up in class again. So, you really didn’t KNOW it.

According to the Forgetting Curve, you forget 80% of what you’ve learned in the first 24 hours. This is like trying to build a house with a cardboard foundation. After 24 hours, the cardboard gets wet and loses 80% of its strength. You try to build on top of this weak foundation and eventually the whole thing comes tumbling down.

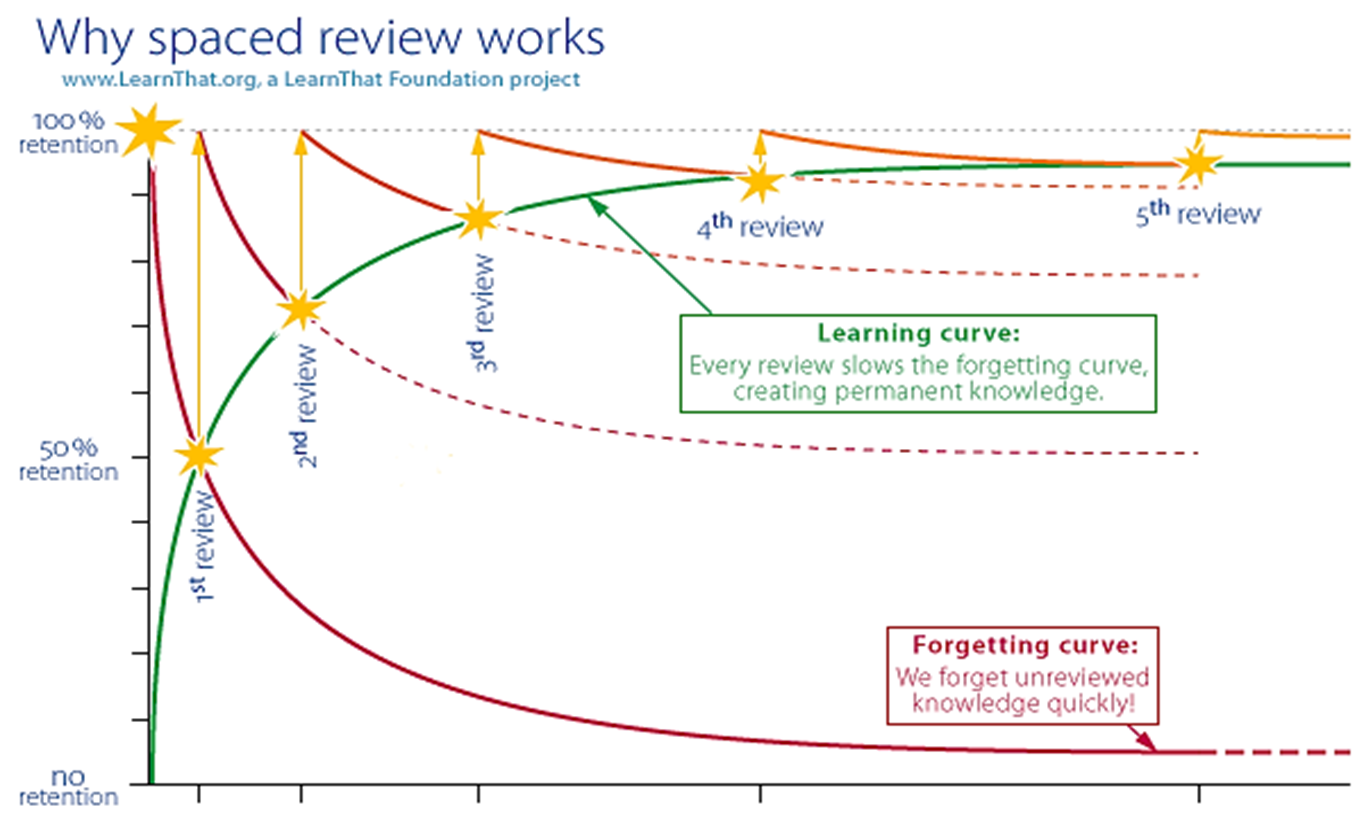
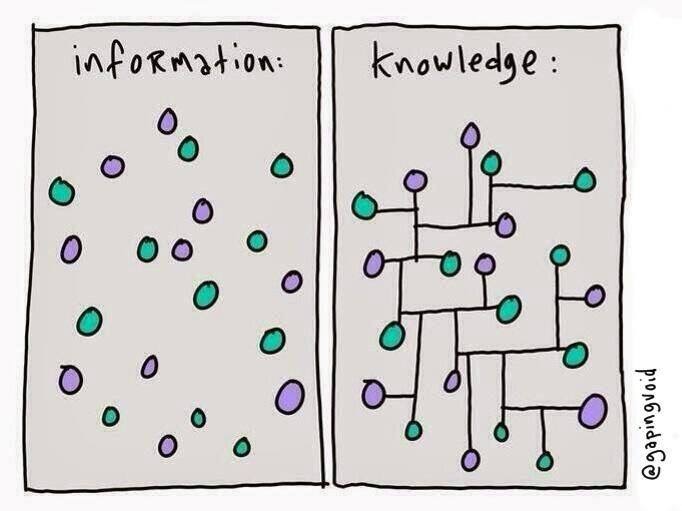


Figure 1- The Forgetful Curve

Believing you have mastered a skill that you have mostly forgotten is an example of poor metacognition (you do not have an accurate gauge of your true ability). There is a HUGE difference between knowing that you know it and believing that you know it (but forgetting most of it a week later).

The hardest part of learning is making the transition from knowing a bunch of individual facts to being able to put the pieces together in a sensible, useable way.

The challenge you will have is discovering how to organize the information in the chapter into knowledge. Near the end of this chapter, we’ll examine how concept maps can help you to organize the information in the chapter and prove to yourself that you understand the concepts in the chapter.

## How is qualitative data different from quantitative data?

Key Terms

Qualitative data Quantitative data Categorical data

Summary

Suppose I poll 60 elementary school students and ask them: “What is your favorite ice cream flavor?” Ice cream flavor is NOT a numerical quantity. We call data this type of data **qualitative data**. (Instead of thinking about “quantities” we are thinking about “qualities”. Qualities include things like gender, major, car brands, names, places, etc.)

In practice, one could have many, many qualities that result from a single poll question. For example, “What is your name?”, is a question that results in a lot of qualitative data. Unfortunately, analyzing this type of data is often difficult because almost every piece of data is different.

What we typically do is break qualitative data down into a manageable number of categories that we can analyze. For this reason, the terms “qualitative data” and **categorical data** are frequently used interchangeably in the context of statistics.

Suppose I poll 100 college students and ask: “How many times did you go to office hours to seek help when you had difficulty in class?” The type of data that results from this type of question is **quantitative data**. (“quant” is the root of the word “quantity” which measures “how much of something we have.” Quantitative data measures the amount of something and results in a number.

Sometimes, data made up of numbers is actually qualitative data. For example, “The 210 freeway” is qualitative data even though it has a number in it, because 210 is not measuring the amount of anything it is just serving as a name. Zip codes and social security numbers are also not quantitative data.

A good way for you to tell whether a number is acting as quantitative data is to ask yourself – “Is this number measuring or counting the amount of anything?” If it is, then the number is quantitative data.

Sometimes we can take qualitative data and turn it into quantitative data. For example, customer satisfaction ratings of “Poor”, “Fair”, “Good”, and “Excellent” could be represented with numbers as 1, 2, 3, and 4 respectively. By converting the qualitative data to quantitative data, we can apply various statistical techniques.

Notes

1. Practice

|  |  |
| --- | --- |
| Classify each of the following as quantitative or qualitative data.   1. The president of the US is tall.   **Solution** Tall is not a numerical quantity son this data is qualitative.   1. The fire fighter is 5 ft 10 inches tall.   **Solution** We can convert this data to inches, 70 inches. Since this data measures an amount of height and is a number, this is quantitative data.   1. The area code is 626.   **Solution** Although this data is a number, it does not measure an amount of something. So, this is qualitative data. | Classify each of the following as quantitative or qualitative data.   1. Sam Dean has blue eyes. 2. Dave Reynolds wears a size 13 shoe. 3. A student has completed 3 math classes. |

## How is a population different from a sample?

Key Terms

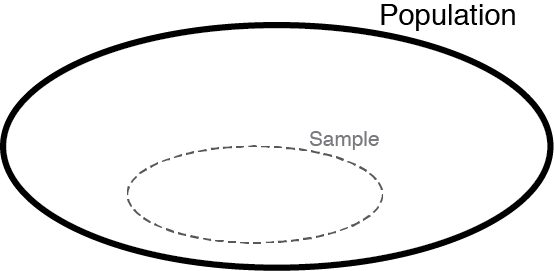
Population Sample

Summary

In statistics, we frequently want to understand data that comes from a particular group. For example, suppose I want to have a better understanding of the student body at a college.

The **population** is the entire group we are looking to study. In this case, our population is all students at the college.

Since it is generally not possible to collect data on an entire population (we don’t have the time to survey all students) we collect data from a smaller group or subset taken from the population. This smaller group is called the **sample**.



An example of a sample for a student population would be all the students in a particular math class. In general, we would like our sample to do a good job of representing our population. Using students from a particular math class as our sample probably won’t do a good job of representing all of the students at the college.

Notes

1. Practice

|  |  |
| --- | --- |
| For each pair determine which is the sample and which is the population.   1. The professional women’s basketball team Phoenix Mercury and all basketball teams.   **Solution** All basketball teams is the population. A portion of this population is the professional women’s basketball team the Phoenix Mercury. Since this is a subset of the population, the Phoenix Mercury is a sample.   1. People with college degrees and people who have attended college.   **Solution** The larger population is people who have attended college. A subset of this population is the sample people with college degrees. | For each pair determine which is the sample and which is the population.   1. Students enrolled in college and students enrolled at the University of Arizona. 2. Registered voters in the United States and registered voters likely to vote in the United States. |

## How do you tell a statistic from a parameter?

Key Terms

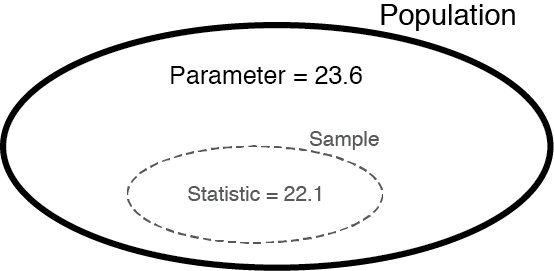
Statistic Parameter

Summary

A parameter is distinguished from a statistic on the basis of whether it corresponds to a population or sample. A number that represents a characteristic of the population is called a **parameter**. A number that represents a characteristic of the sample is called a **statistic**.

For example, we could define the population to be all college students. If we calculated the average age of all college students and found it to be 23.6 years old, this average would be a parameter, because it was calculated using the ages of ALL college students.

If I randomly selected 100 students at the University of Nevada and calculated that their average age was 22.1 this would be a statistic, because we only used the ages of a sample of the whole population.



Notes

Guided Example 3 Practice

|  |  |
| --- | --- |
| Determine which **bold** phrase is a statistic and which is a parameter.   1. In the Substance Abuse and Mental Health Services Administration survey, **13.2%** of respondents said they have driven under the influence of alcohol.   **Solution** A survey is administered to a group of people which is presumably the population. From this group, only a portion will respond so the respondents are a sample. This means the number 13.2% is a statistic.   1. In a Gallup poll of 1,023 renters, it was found that **58.2%** of them said that they had only wireless phones.   **Solution** This poll was conducted with 1023 renters which is a sample of the group all renters. This means that 58.2% is a statistic. | Determine which **bold** phrase is a statistic and which is a parameter.   1. In the 2014-2015 school year the retention rate in Math 150 was **89.6%** while the retention rate in all math courses at PCC was **79.4%.** 2. The **average height of Americans** is more than then **average height of women in America.** |

## What is the difference between discrete data and continuous data?

Key Terms

Discrete data Continuous data

Summary

Quantitative Data can be further broken two into two types: discrete or continuous.

**Discrete** data is data that you can count on your fingers (assuming you had an infinite supply of fingers). For instance, we can count the number of eggs a hen lays similar to how we count on our fingers. Or we could count the number of people in a room on our fingers. **Continuous** data is data generated from measurements. We don’t count how tall someone is, we measure how tall they are. Heights are continuous data.

**Discrete** data is quantitative data that is counted, not measured. We count the number of people or eggs. We don’t measure the number of eggs.

**Continuous** data is quantitative data that is measured, not counted. We measure heights, weights, and volume. We don’t count them.

For example, suppose I want to measure precisely how tall someone is. Our convention is to give our height in feet and inches, but this is really an **approximation**. It gives our height to the nearest inch. Two people who claim to be 5 feet 4 inches tall are probably not *exactly* the same height.

Notes

Guided Example 4 Practice

|  |  |
| --- | --- |
| Classify each of the following pieces of quantitative data as either continuous or discrete.   1. Number of math classes you have taken.   **Solution** Since you count the number of math classes you have taken, this is discrete data.   1. The speed of the Gold Line Metro train.   **Solution** The speed of the Gold Line Metro train is measured so it is continuous data. | Classify each of the following pieces of quantitative data as either continuous or discrete.   1. The length of your foot. 2. Number of children in a household. |

# Section 1.2 Sampling and Bias

**Get Started –** What is a simple linear equation and how do you solve it?

* How do you identify the type of sampling used in a statistical study?
* What is bias and how do you identify a potential source of bias in a statistical study?

## **Get Started –** What is a simple linear equation and how do you solve it?

Key Terms

Linear equation Solve an equation

Summary

The root of our word “Algebra” comes from the Arabic “al-jabr” which is often translated as: restoring. We think of solving equations as restoring them to their original state. We must undo the mathematical operations that have been performed. To undo addition, we will subtract. To undo subtraction, we will add. To undo multiplication, we will divide. To undo division, we will multiply. Equations typically involve more than one operation at a time. Since we are undoing the math, we will often follow the order of operations in reverse order: add or subtract before we multiply or divide.

Notes

1. Practice

|  |  |
| --- | --- |
| Solve and check .  **Solution** We will undo the + 5 first, then we will undo the multiplying by 2.  Original equation    Simplify  Divide both sides by 2  Simplify  Subtract 5 from both sides  To check the solution, substitute the solution into the original equation:  Original equation    Simplify  Substitute *x* = 7.35  When checking it is important to go back to the original equation, if we only checked from  we aren’t checking our first step of subtracting the five. | Solve and check . |

Most equations involve more than one operation and often operations occur more than once, so we need some guidelines for dealing with more complicated equations. The most important thing to remember is that we need to maintain balance. We need to treat each side of the equation equally. We will be using all our properties of equality. Our goal when solving linear equations is to get the variable to equal a single number: we need to isolate the variable. And we are trying to undo the existing math, so we will often be following the order of operations backwards.

## Guidelines for Solving Linear Equations

* Simplify the expressions on each side of the equal sign separately: use distributive property to remove any grouping symbols & combine like terms
* Use the addition/subtraction properties of equality to move all the variables to one side of the equal sign and the constant terms to the other side of the equal sign
* Use the multiplication/division properties of equality to solve for the variable.
* CHECK: plug in your solution to make sure it works.

Notes

1. Practice

|  |  |
| --- | --- |
| Solve and check  **Solution** We will distribute the 5 to remove the parentheses and then isolate the variable.  Original equation    Simplify.  Divide both sides by 5  Simplify.  Subtract 10 from both sides.  Remove the parentheses by distributing the 5.  To check the solution, substitute the solution into the original equation:  Original equation    Simplify  Substitute *a =* 13 | Solve and check . |

## How do you identify the type of sampling used in a statistical study?

Key Terms

Census Simple random sample Stratified sampling

Systematic sampling Cluster sampling Quota sampling

Convenience sampling

Summary

Now that you know that you must 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. This means that the results of the sample do not reflect the results of the population. 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 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 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 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 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 several 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 *k*th 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 100th 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. 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 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 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 the individuals to randomly choose from may be hard to do.

Notes

1. Practice

|  |  |
| --- | --- |
| Determine if the sample type is simple random sample, stratified sample, systematic sample, cluster sample, quota sample, or convenience sample.   1. 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.   **Solution** This is a cluster sample, since she randomly selected some of the groups, and all individuals in the chosen groups were surveyed.   1. 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.   **Solution** 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.   1. 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.   **Solution** This is a simple random sample, since the individuals were picked randomly.   1. 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.   **Solution** This is a stratified sample, since all groups were used, and then random samples were taken inside each group.   1. For the Ford Motor Company to ensure quality of their cars, they test every 130th car coming off the assembly line of their Ohio Assembly Plant in Avon Lake, OH.   **Solution** This is a systematic sample since they picked every 130th car.   1. 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.   **Solution** This is a quota sample, since they tried to pick people who fit into these subcategories. | Determine if the sample type is simple random sample, stratified sample, cluster sample, systematic sample, or convenience sample.  A study to determine the opinion of Americans about the use of marijuana for medical purposes is being conducted using the following designs.   1. The researchers attend a festival in a town in Kansas and ask all the people they can what their opinions are. 2. The researchers divide Americans into groups based on the person’s race, and then take random samples from each group. 3. The researchers number all Americans and call the 50th person on the list. Then they call every 10,000th person after the 50th person. 4. The researchers call every person in each of 10 area codes that were randomly chosen. 5. The researchers number every American, and then call all randomly selected Americans. |

## What is bias and how do you identify a potential source of bias in a statistical study?

Key Terms

Bias Selection bias Non-response bias

Voluntary response bias Self-interest study Response bias

Perceived lack of anonymity Loaded questions

Summary

When we collect data, we often sample a population to measure a statistic. We hope that the statistic from the sample matches the corresponding parameter from the population.

**Bias** is the tendency for a statistic from a sample to underestimate or overestimate a parameter from a population.

Two types of bias are commonly encountered when we collect data.

**Sample bias** (selection bias) occurs when the sample chosen from the population is not representative of the population.

**Nonresponse bias** 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.

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. 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 while 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.

Because of the people whom the Literary Digest polled, they created a sample 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 during the Great Depression. 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.

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 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 must pay attention to what percentage of surveys are returned. If 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.

There are many other types of bias that may be encountered when data is collected for a sample.

**Voluntary response bias** often occurs when the sample is volunteers. For example, suppose a survey is conducted among callers to a radio show to determine their attitudes towards vaccinations. The sample members are volunteers who might tend to have strong opinions regarding vaccinations. This overrepresentation might lead to statistics that do not represent the attitudes of the population.

**Self-interest bias** may occur when the researchers have an interest in the outcome. Consider a recent study which found that chewing gum may raise math grades in teenagers. This study was conducted by the Wrigley Science Institute, a branch of the Wrigley chewing gum company. This is an example of a self-interest study; one in which the researches have a vested interest in the outcome of the study. While this does not necessarily ensure that the study was biased, it certainly suggests that we should subject the study to extra scrutiny.

**Response bias** may occur when the responder gives inaccurate responses for any reason. Suppose a survey asks people “when was the last time you visited your doctor?” This might suffer from response bias, since many people might not remember exactly when they last saw a doctor and give inaccurate responses. Sources of response bias may be innocent, such as bad memory, or as intentional as pressuring by the pollster.

**Perceived lack of anonymity** is possible when the responder fears giving an honest answer might negatively affect them. 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.

**Loaded questions** are questions where wording influences the responses. 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

**Non-response bias** may be an issue when people refusing to participate in the study can influence the validity of the outcome. If a telephone poll asks the question “Do you often have time to relax and read a book?”, and 50% of the people called refused to answer the survey. It is unlikely that the results will be representative of the entire population. When people refuse to participate, we can no longer be so certain that our sample is representative of the population.

Notes

1. Practice

|  |  |
| --- | --- |
| In each situation, identify a potential source of bias.   1. A survey asks how many sexual partners a person has had in the last year.   **Solution** This survey suffers from response bias where the responder might give inaccurate responses. In this case, men are likely to over-report the number of sexual partners and women are likely to under-report the number of sexual partners.   1. A radio station asks readers to phone in their choice in a daily poll.   **Solution** The readers are volunteers and will be more likely to respond with strong opinions. The survey has the potential to suffer from voluntary response bias.   1. High school students are asked if they have consumed alcohol in the last two weeks.   **Solution** Since students are asked a question whose response might impact them negatively, this is an example of perceived lack of anonymity.   1. The Beef Council releases a study stating that consuming red meat poses little cardiovascular risk.   **Solution** The Beef Council has an interest in the results of the study so the study might suffer from self-interest bias.   1. A poll asks “Do you support a new transportation tax, or would you prefer to see our public transportation system fall apart?”   **Solution** The question uses the words “fall apart” in describing the potential failure of the public transportation system. This choice of words might influence responses so this is an example of a loaded question. | In each situation, identify a potential source of bias.   1. A survey asks the following: Should the mall prohibit loud and annoying rock music in clothing stores catering to teenagers? 2. A survey asks people to report their actual income and the income they reported on their IRS tax form. 3. A survey asks the following: Should the death penalty be permitted if innocent people might die? 4. To determine opinions on voter support for a downtown renovation project, a surveyor randomly questions people working in downtown businesses. |

Notes

How Do You Conduct a Study?

**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 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. Most 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.

Here are examples of clinical studies.

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

Participants in a clinical study are broken into two groups.

**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).

Any clinical study where the researchers compare the results of a treatment group versus a control group is called a **controlled study**. Any clinical study in which the treatment group and the control group are selected randomly from the population is called a **randomized controlled study**.

Notes

1. Practice

|  |  |
| --- | --- |
| Determine the treatment group, control group, treatment, and control for each clinical study.   1. A researcher may want to study whether or not smoking increases a person's chances of heart disease.   **Solution** 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.   1. A researcher may want to study whether a new antidepressant drug will work better than an old antidepressant drug.   **Solution** 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.* | Determine the treatment group, control group, treatment, and control for each clinical study.  A researcher may want to study whether taking folic acid before pregnancy will decrease the risk of birth defects. |

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

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.

1. Practice

|  |  |
| --- | --- |
| For each situation, identify if a placebo is necessary to use.   1. A researcher may want to study whether smoking increases a person's chances of heart disease.   **Solution** 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.   1. A researcher may want to study whether a new antidepressant drug will work better than an old antidepressant drug.   **Solution** In this example, a placebo is not needed since we are comparing the results of two different antidepressant drugs. | For each situation, identify if a placebo is necessary to use.   1. A researcher may want to study whether taking folic acid before pregnancy will decrease the risk of birth defects. 2. A researcher wants to determine if morphine reduces pain during dental tooth extractions. |

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.*

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**.

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. As mentioned in an earlier section, a parameter is a numerical value calculated from a population. A statistics is 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 all 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.

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  (mu), and  (sigma).Statistics are usually denoted with our alphabet. In some cases, we try to use a letter that would be equivalent to the Greek letter. Examples of letter that are used are  (x-bar), s, and r.

# Section 1.3 Visualizing Data

**Get Started –** How can you use inequalities and interval notation to represent quantities on a number line?

* What is a frequency distribution?
* What are some of the ways you can represent data visually?

## **Get Started –** How can you use inequalities and interval notation to represent quantities on a number line?

Key Terms

Inequalities Interval notation

Summary

The **Greater Than or Equal To** symbol ≥ is a hybrid between the greater than symbol > and the equals sign =.

Left expression ≥ Right expression

means that what is on the left is “equal to or bigger than” the what is on the right. Thus, the thing on the left can be bigger than the thing on the right, or it can be equal to the thing on the right, but it cannot be smaller than the thing on the right. For example,

Number of units to be considered a fulltime student ≥ 12

The age you have to be to legally buy alcohol in the USA ≥ 21

Number of siblings you have ≥ 0

The phrase “**at least**” translates into the “greater than or equal to” symbol. Other phrases that translate into the “greater than or equal to” symbol are “No less than” or “x or more”.

|  |  |  |
| --- | --- | --- |
| “The number of times I will eat pizza this year is at least 5” | Translates to | “number of times I will eat  pizza this year is ≥ 5” |
| “At least 5 students will pass this class” | Translates to | “number of students who  will pass ≥ 5” |
| “5 or more people will adopt a Growth Mindset this semester.” | Translates to | “number of people who adopt Growth Mindset ≥ 5” |
| “No less than 5 students in this class are better at math than they think.” | Translates to | “number of students in this class are better at math than they think ≥ 5” |

The **Less Than or Equal To** symbol  is a hybrid between the less than symbol < and the equals sign =.

Left expression  Right expression

means that the expression on the left is “equal to or less than” the expression on the right. For example,

Your GPA at PCC  4.0

Number of days in a year  366

21  Age of someone legally allowed to buy alcohol

The phrase “**at most**” translates into the “less than or equal to” symbol. Other phrases that translate into the “less than or equal to” symbol are “No more than” or “x or fewer”.

|  |  |  |
| --- | --- | --- |
| “The number of times I will visit Las Vegas this year is at most 5.” | Translates to | “number of times I will visit Las Vegas this year is 5.” |
| “At most 5 students will fail this class.” | Translates to | “number of students will  fail this class 5.” |
| “5 or fewer teammates are allowed on the basketball court at one time.” | Translates to | “number of teammates are allowed on the basketball  court at one time 5.” |
| “Lyman will take no more than 5 classes this semester.” | Translates to | “number of classes Lyman will take this semester5.” |

Guided Example 1 Practice

|  |  |
| --- | --- |
| Write in the appropriate inequality symbol (≤ or ≥) in each box to make an accurate phrase.   1. 2 feet  10 inches   **Solution** Since 2 feet is the same as 24 inches, the quantity on the left is greater than the quantity on the right so  2 feet  10 inches   1. 100 years Your teacher’s age   **Solution** Your teacher is most likely younger than 100 years old,  100 years Your teacher’s age | Write in the appropriate inequality symbol (≤ or ≥) in each box to make an accurate phrase.   1. Your teacher’s age 0 2. 2 cups  1 pint |

We can use two inequalities to indicate a range of values. We can use two inequalities to indicate what scores result in a B grade:

80  score < 90

Notice how we want to include 80 since the lowest score you can get while still receiving a B letter grade is 80, but we don’t include 90 because getting a 90 would result in an A. When you write things in this way, we usually go from smaller to larger. It is perfectly legal, however, to write something like 90 > score ≥ 80, as they mean the same thing as above.

Another way we can describe a range of values is using **interval notation**.

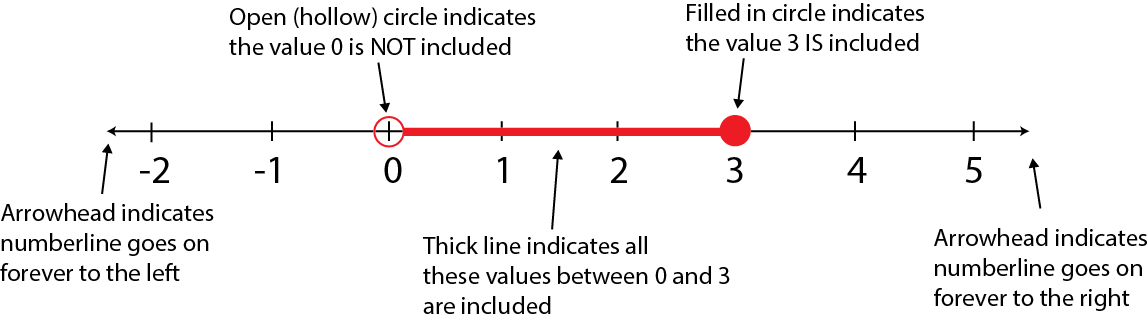
An interval is an ordered pair of numbers, that describes the set of all numbers bigger than the number on the left (the **left endpoint)** and less than the number on the right (the **right endpoint)**. The ordered pair is contained within some sort of brackets, and they type of bracket tells you whether you want to include the two end points or not.

If we want to include the left endpoint in the set we use the symbol “[“. If we don’t want to include it, we use the symbol “(”. Similarly, if we want to include the right endpoint we use the symbol “]”. If we don’t want to include the number on the right we use “)”. For example,

|  |  |  |
| --- | --- | --- |
| Interval |  | Inequality |
| (1,2) | Does not contain 1 or 2, but contains every number in between |  |
| [1,2] | Contains both 1 and 2 and contains every number in between |  |
| [1,2) | Contains 1 but not 2, and contains every number in between |  |
| (1,2] | Does not contain 1 but does contain 2, and contains every number in between |  |

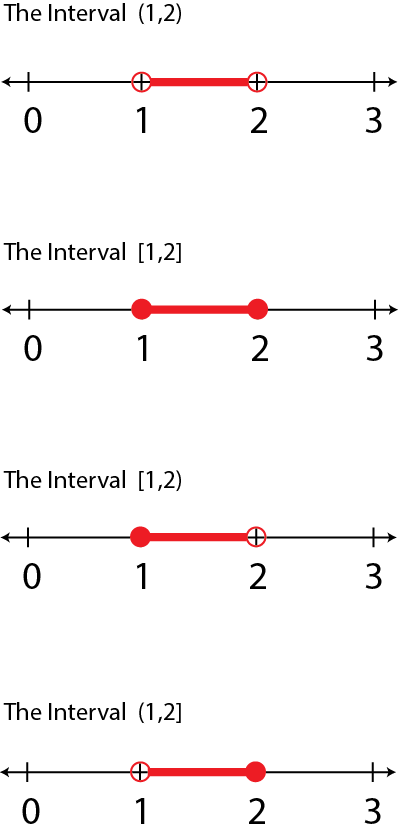
A nice way to visualize the infinite sets characterized by intervals is to use a **number line**.

Here is an example of a number line representation of the interval (0,3]:



We put arrows at the ends of the number line to indicate the line goes on forever in both directions. We use a closed (filled-in) circle to indicate that we want to include a point. We use an open (hollow) circle to indicate we don’t want to include the point. Any portion of the line that is “shaded” indicates that that point is included.

Here are number line representations of all the intervals we’ve used in the section above:



By using interval notation, we are implying that our data is continuous because we are shading all the values on the number line.

Notes

Guided Example 2 Practice

|  |  |
| --- | --- |
| Write each of the following in interval notation.      **Solution** The endpoint of the interval are 18 and 22 so these values make the left and right endpoints in interval notation. The value 18 is included so we need a [ on the left side of the interval notation. The value 22 is not included so the right side will have a ). This gives us the interval [18, 22).  5  3  4  1  2  0  -1  -2  -3  **Solution** The open circle indicates that the left endpoint is not included. The filled circle indicates that the right endpoint is included. This gives us the interval (-1, 4]. | Write each of the following in interval notation.      5  3  4  1  2  0  -1  -2  -3 |

## What is a frequency distribution?

Key Terms

Frequency Frequency distribution Relative frequency

Relative frequency distribution

Summary

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. 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.

The **frequency** is thenumber of times a data value occurs in a data set.

A **frequency distribution** is a listing of each data value or grouping of data values (called classes) with their frequencies.

The **relative frequency** is 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.

A **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.

Notes

Guided Example 3

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Suppose a class was asked what their favorite soft drink is with the following results:   |  |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | --- | | 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 |  |  |  |  1. Create a frequency distribution for the data.   **Solution** List each drink type and 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.   |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | | **Drink** | **Coke** | **Pepsi** | **Mt Dew** | **Dr. Pepper** | **Sprite** | | **Frequency** | 9 | 10 | 5 | 5 | 4 |  1. Create a relative frequency distribution for the data.   **Solution** Divide each frequency by 33, the total number of data values. Round to three decimal places.   |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | | **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% | |

Practice

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Suppose a class was asked what their eye color is with the following results:   |  |  |  |  |  | | --- | --- | --- | --- | --- | | blue | brown | brown | blue | blue | | brown | blue | brown | brown | hazel | | brown | brown | brown | brown | green | | brown | brown | blue | brown | blue |  1. Create a frequency distribution for the data. 2. Create a relative frequency distribution for the data. |

## What are some of the ways you can represent data visually?

Key Terms

Bar graph Pie chart Pictogram Histogram

Summary

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

A **bar graph** is 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.*

A **pie chart** is 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, multiple the relative frequency of that slice by 360°. *Note: The percentages in each slice of a pie chart must all add up to 100%.*

A **pictogram** is 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.

Notes

Guided Example 4

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Suppose a class was asked what their favorite soft drink is with the following results:   |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | | **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% |  1. Draw a bar graph of the frequency distribution.   **Solution** 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.   1. Draw a bar graph of the relative frequency distribution.   **Solution** This is like 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.   1. Draw a pie chart of the data.   **Solution** To draw a pie chart, multiply the relative frequencies by 360°. Then use a protractor to draw the corresponding angle.   |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | | **Drink** | **Coke** | **Pepsi** | **Mt Dew** | **Dr. Pepper** | **Sprite** | | **Frequency** | 9 | 10 | 5 | 5 | 4 | | **Relative Frequency** | 9/33 ≈ .27 | 10/33 ≈ .30 | 5/33 ≈ .15 | 5/33 ≈ .15 | 4/33 ≈ .12 | | **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° |  1. Draw a pictograph for the favorite soft drink data.   **Solution** Here you can get creative. One thing to draw would be glasses. Now you would not want to draw 10 glasses. To avoid this, et each glass be worth a certain number of data values, say one glass = frequency of two. 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. |

Practice

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Suppose a class was asked what their eye color is with the following results:   |  |  |  |  |  | | --- | --- | --- | --- | --- | | **Eye Color** | **Blue** | **Brown** | **Green** | **Hazel** | | **Frequency** | 6 | 12 | 1 | 1 | | **Relative Frequency** | 6/20 = .30 | 12/20 = .60 | 1/20 = .05 | 1/20 = .05 |  1. Draw a bar graph of the frequency distribution. 2. Draw a bar graph of the relative frequency distribution. 3. Draw a pie chart of the data. 4. Draw a pictograph for the eye color data. |

Pictographs are not 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 variables are numbers, so the graph you create is different from the ones for qualitative data. First, the 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.

A **histogram** is a graph of a quantitative variable where rectangles are used for each subinterval, the height of the rectangle represents the 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.

**In a proper graph, the vertical axis starts at 0, there is a title on the graph, the axes have labels, and the tick marks are labeled. This allows people to know what the data represents.**

Notes

Guided Example 5 Practice

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Suppose that we have collected weights from 100 male subjects as part of a nutrition study. These data are represented in the frequency distribution below.   |  |  | | --- | --- | | **Class** | **Frequency** | | [120, 135) | 4 | | [135, 150) | 14 | | [150, 165) | 16 | | [165, 180) | 28 | | [180, 195) | 12 | | [195, 210) | 8 | | [210, 225) | 7 | | [225, 240) | 6 | | [240, 255) | 2 | | [255, 270) | 3 |   **Solution** For each class, draw a bar where the base corresponds to the class and the height corresponds to the frequency.  Weights of Male Subjects | The total cost of textbooks for the term was collected from 36 students. Create a histogram for this data.  $140 $160 $160 $165 $180 $220  $235 $240 $250 $260 $280 $285  $285 $285 $290 $300 $300 $305  $310 $310 $315 $315 $320 $320  $330 $340 $345 $350 $355 $360  $360 $380 $395 $420 $460 $460  Assume first class for this data is [100, 150) and subsequent classes use the same width.  Make a frequency distribution for the data and use it to create a histogram. |

# Section 1.4 Measures of Central Tendency

**Get Started** – How are decimals rounded?

**Get Started –** How do you convert between a fraction, decimal and percent?

* How do you compute the mean, median, and mode of data?

## **Get Started** – How are decimals rounded?

Key Terms

Decimal Rounding

Summary

Because the calculations in this section involve multiplication and division of decimals, there will be times your calculation results in a number with many decimal places. In general, we will use the following **rounding** rule:

1. Determine which digit is the last one you want to keep (i.e. round to the nearest tenths place means you want to keep the digit to the right of the decimal point).
2. Leave it the same if the next digit to the right is less than 5and drop the rest of the digits to the right of the digit you want to keep.
3. Increase it by one if the next digit is 5 or more. Drop all digits to the right of the digit you want to keep.
4. Practice

|  |  |
| --- | --- |
| Round 14.123 to the nearest hundredth.  **Solution** Since we are rounding to the hundredths place, we look first at the digit in the hundredths place and find that 2. Since the number to the right of 2 is 3, we need to keep the 3 and drop the digits to the right of the 3.  Thus, our rounded value is 14.12. | Round 1002.143 to the nearest hundredth. |

1. Practice

|  |  |
| --- | --- |
| Round 123.456 to the nearest tenth.  **Solution** Since we are rounding to the tenth place, we look first at the digits in the tenths place and find that 4. Since the number to the right of 4 is 5 (or more), we need to round up.  Thus, our rounded value is 123.5. | Round 13.167 to the nearest tenth. |

## Get Started – How do you convert between a fraction, decimal and percent?

Key Terms

Fraction Decimal Percent

Summary

We will be using percentages throughout this course. Percentages are used to describe interest rates, describe probabilities, and express quantities in relation to the whole. This section helps your practice some of the techniques we use when working with percentages.

## **Converting Decimals & Fractions to Percentages**

To convert any number into an equivalent percent all we need to do is use our old trick of multiplying by the right form of 1. This time the right form is **100% = 1**

1. Practice

|  |  |
| --- | --- |
| Convert 0.065 to a percent  **Solution** All we need to do is multiply by 100%,    We can see that our answer is reasonable by using 0.06 = 6% as an under estimate and 0.07 = 7% as an over estimate we see our answer is reasonable. | Convert 0.174 to a percent. |

Guided Example 4 Practice

|  |  |
| --- | --- |
| Convert  to a percent.  **Solution** First we remember that the fraction bar indicates division,  . Then all we need to do is multiply by 100%    When we recall that = “one quarter” = 25 cents our answer checks. | Convert  to a percent |

Guided Example 5 Practice

|  |  |
| --- | --- |
| Convert  to a percent.  **Solution** All we need to do is multiply by 100%:      Note that we are not rounding our answer off. | Convert  to a percent. |

### Converting Percents to Decimals and Fractions

We will use the fact that the percent symbol % means “per 100” or “out of 100” or better yet “divided by 100.”

Guided Example 6 Practice

|  |  |
| --- | --- |
| Write an equivalent decimal and fraction for each percent.   1. 8%   **Solution**     1. 12.5%   **Solution** | Write an equivalent decimal and fraction for each percent.   1. 20% 2. 4% |

Notes

## How do you compute the mean, median, and mode of data?

Key Terms

Mean Median Mode

Summary

Sometimes we want to find a single number to represent a set of data. Usually, we want to use a number that represents the “center” of the data. We call these numbers measures of center.

The **mean** is the type of average that most people commonly call “the average.” You take all of the data values, find their sum, and then divide by the number of data values. Again, you will be using the sample statistic to estimate the population parameter, so we need formulas and symbols for each of these.

**Population Mean:**

****

where *N* is the size of the population and  are data values.

*Note: is a short cut way to write adding a bunch of numbers together.*

**Sample Mean:**

****

where *n* is the size of the sample and  are data values.

The **median** is the value that is found in the middle of the ordered data set. Most books give a long explanation of how to find the median. The easiest thing to do is to put the numbers in order and then count from both sides in, one data value at a time, until you get to the middle. If there is one middle data value, then that is the median. If there are two middle data values, then the median is the mean of those two data values. If you have a really large data set, then you will be using technology to find the value. There is no symbol or formula for median, neither population nor sample.

The **mode** is the data value that occurs most often. The mode is the only average that can be found on qualitative variables, since you are just looking for the data value with the highest frequency. The mode is not used very often otherwise. There is no symbol or formula for mode, neither population nor sample. Unlike the other two averages, there can be more than one mode or there could be no mode. If you have two modes, it is called bimodal. If there are three modes, then it is called trimodal. If you have more than three modes, then there is no mode. You can also have a data set where no values occur most often, in which case there is no mode.

When presented with a data set with many numbers and no specific round off instructions, the standard practice is to round one decimal place beyond what’s given in the data.

Notes

Guided Example 7 Practice

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| The first 11 days of May 2013 in Flagstaff, AZ, had the following high temperatures (in °F)   |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | | 71 | 59 | 69 | 68 | 63 | 57 | | 57 | 57 | 57 | 65 | 67 |  |  (Weather Underground, n.d.)  1. Find the mean high temperature   **Solution** Since there are only 11 days, then this is a sample.     1. Find the median high temperature   **Solution** First put the data in order from smallest to largest.  57, 57, 57, 57, 59, 63, 65, 67, 68, 69, 71  Now work from the outside in, until you get to the middle number.    So, the median is 63°F   1. Find the mode of the high temperature.   **Solution** From the ordered list it is easy to see that 57 occurs four times and no other data values occur that often. So the mode is 57°F.  We can now say that the expected high temperature in early May in Flagstaff, Arizona is around 63°F. | In a recent class, the overall percentages at the end of the semester were recorded.   |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | | 73 | 71 | 72 | 73 | 64 | 71 | 77 | 38 | | 58 | 75 | 85 | 36 | 86 | 39 | 61 |  |  1. Find the mean of this data. 2. Find the median of this data. 3. Find the mode of the data. |

Guided Example 8 Practice

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| The first 12 days of May 2013 in Flagstaff, AZ had the following high temperatures (in °F)   |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | | 71 | 59 | 69 | 68 | 63 | 57 | | 57 | 57 | 57 | 65 | 67 | 73 |  (Weather Underground, n.d.)  1. Find the mean high temperature.   **Solution** Since there are only 12 days, then this is a sample.     1. Find the median high temperature.   **Solution** This set of data has an even number of data values so there will be two middle numbers. First put the data in order from smallest to largest.  57, 57, 57, 57, 59, 63, 65, 67, 68, 69, 71, 73  Now work from the outside in, until you get to the middle numbers.    The median is the mean of the middle two numbers,     1. Find the mode of the high temperature   **Solution** From the ordered list it is easy to see that 57 occurs 4 times and no other data values occurs that often. So, the mode is 57°F. | In a recent class, the overall percentages at the end of the semester were recorded.   |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | | 73 | 71 | 72 | 73 | 64 | 71 | 77 | 38 | | 58 | 75 | 85 | 36 | 86 | 39 | 61 | 95 |  1. Find the mean of this data. 2. Find the median of this data. 3. Find the mode of the data. |

Guided Example 9

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Suppose a class was asked what their favorite soft drink is and the following is the results:   |  |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | --- | | 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 |  |  |  |   Find the average.  **Solution** The mean, median, and mode are all examples of averages. However, since the data is qualitative, you cannot find the mean and the median. The only average you can find is the mode. Notice, Coke was preferred by 9 people, Pepsi was preferred by 10 people, Mt Dew was preferred by 5 people, Dr. Pepper was preferred by 5 people, and Sprite was preferred by 4 people. So, Pepsi has the highest frequency, so Pepsi is the mode. If one more person came in the room and said that they preferred Coke, then Pepsi and Coke would both have a frequency of 10. In that situation, both Pepsi and Coke would be the modes, and we would call this bimodal. |

Practice

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Suppose a class was asked what their eye color is and the following is the results:   |  |  |  |  |  | | --- | --- | --- | --- | --- | | blue | brown | brown | blue | blue | | brown | blue | brown | brown | hazel | | brown | brown | brown | brown | green | | brown | brown | blue | brown | blue |   Find the average. |

Guided Example 10 Practice

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| The frequency distribution below shows the distribution of scores on a biology quiz.   |  |  | | --- | --- | | *x* | *f* | | 12 | 1 | | 13 | 3 | | 14 | 8 | | 15 | 4 | | 16 | 4 |  1. Find the mean of the distribution.   **Solution** To find the mean, we need to add all of the scores and divide by the total number of scores. The table indicates that certain scores occur many times. For instance, the score 14 occurs 8 times. By adding the frequencies, we see that there is a total of 1 + 3 + 8 + 4 + 4 or 20 scores. Instead of adding all 20 scores individually, use multiplication to account for the frequencies:    Notice that the mean is approximately in the middle of the scores.   1. Find the median of the distribution.   **Solution** There are 20 scores in this set of data. If we order the scores, there are two middle scores in position 10 and 11. Position 10 and 11 scores are both 14 so the median is 14.   1. Find the mode of the distribution.   **Solution** The most frequently occurring score is 14 (occurring 8 times) so the mode is 14. | The frequency distribution below shows the distribution of weights in the contents of bags of chips.   |  |  | | --- | --- | | *x* | *f* | | 7.5 | 3 | | 8 | 12 | | 8.5 | 10 | | 9 | 1 | | 9.5 | 1 |  1. Find the mean of the distribution. 2. Find the median of the distribution. 3. Find the mode of the distribution. |

Guided Example 11 Practice

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| The fuel consumption of a random selection of vehicles is collected. From this selection, the following frequency table is calculated:   |  |  | | --- | --- | | Class (miles per gallon) | Frequency | | (10, 18] | 6 | | (18, 26] | 26 | | (26, 34] | 20 | | (34, 42] | 1 | | (42, 50] | 2 |   Find the mean fuel consumption data.  **Solution** The data consists of classes and frequencies. To compute the mean, we need to find a representative data value from each class. This is done by computing the mean of the class end values:   |  |  | | --- | --- | | Class Representative | Frequency | | 14 | 6 | | 22 | 26 | | 30 | 20 | | 38 | 1 | | 46 | 2 |   To find the mean, we need add up all 6 + 26 + 20 + 1 + 2 or 55 data values and divide by 55. This is most easily done by adding each class representative multiplied by its corresponding frequency,    Multiplying the class representative by the frequency allows us to avoid adding 55 data values. | The grade point averages of a random section of students is collected. From this selection, the following frequency table is collected:   |  |  | | --- | --- | | Class (grade points) | Frequency | | (1, 1.5] | 1 | | (1.5, 2] | 2 | | (2, 2.5] | 10 | | (2.5, 3] | 20 | | (3, 3.5] | 12 | | (3.5, 4] | 5 |   Find the mean of the grade point data. |

Guided Example 12

|  |
| --- |
| Students were asked how many days per week they log into their online math. The results are graphed below.     1. What number of days per week did students pick least frequently?   **Solution** There is no bar at 0 times per week so this is the number of days picked least frequently.   1. What is the mean of this distribution?   **Solution** To calculate the mean, we need to add the data values times the corresponding frequency and divide by the total number of data values:    The total number of data values is found by adding the frequencies from each bar.   1. What is the median of this distribution?   **Solution** Since there are 43 data values, the middle value will be the one in position 22 since there will be 21 values on either side. If the data values are ordered, the 22nd value will be a 5 so the median is 5.   1. What is the mode of this distribution?   **Solution** The most frequently occurring data value is 4 with a frequency of 10. This means the mode is 4. |

Practice

|  |
| --- |
| 1. Students in a second-grade class were asked how many pets their family kept in their household. The results are graphed below.      1. What number of pets did students pick least frequently? 2. What is the mean of this distribution? 3. What is the median of this distribution? 4. What is the mode of this distribution? |

# Section 1.5 Measures of Spread

**Get Started –** How do you solve a formula for a variable?

* How do you compute the range of data?
* What does the standard deviation tell you about data?
* How can we use the coefficient of variation to compare the standard deviations of different sets of data?
* How do you interpret quartiles and percentiles of data?
* How do you assemble a five-number summary of data?
* How are stem and leaf plots used to compare data?

## **Get Started –** How do you solve a formula for a variable?

Key Terms

Equivalent form

Summary

In an earlier section, we examined how to solve an equation for a solution. The solution to an equation is a number or several numbers that when substituted for the variables results in a true statement.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | solves |  | since |  |
|  | solves |  | since |  |
|  | solves |  | since |  |

In each case above, replacing the variable with its corresponding value leads to a true statement.

Equations such as  are often encountered in math. These equations may have 2 or more variables in them. If we want to solve them for a particular variable, we need to get that variable by itself on one side of the equation. It should also only appear on that side of the equation and nowhere else.

To solve for a variable, we need to rewrite the equation in an **equivalent form**. This means that the solutions to the equation and its equivalent form are the same. There are two rules that help us to get the equivalent form.

1. Adding an expression to both sides of an equation or or subtracting an expression from both sides of an equation gives an equivalent equation.
2. Multiplying both sides of an equation or dividing both sides of an equation by an expression gives an equivalent equation. Take care that what you multiply or divide by is not equal to zero.

Notes

1. Practice 1

|  |  |
| --- | --- |
| Solve each of the equations for the indicated variable.   1. for *y*   **Solution** Start by isolating the term that contains the variable you are solving for. Subtracting  from both sides leads to    Subtract 2*x* from both sides  To get y by itself, divide both sides of the equation by 4:  Divide both sides by 4    Divide both terms on the right by 4 and simplify  Simplify the left side  The equation is solved for *y* in the second step. However, simplifying the right side often yields a simpler equation.   1. for *t*   **Solution** To solve for *t*, Subtract P from both sides of the equation and then divide to get *t* by itself:    Divide both sides by *Pr* and simplify  Subtract *P* from both sides and simplify   1. for *P*   **Solution** In solving for *P*, we need to notice that *P* appears in two places in the equation. Before we start isolating *P*, we need to factor *P* from the right side so that *P* appears in one place:    Factor *P* from both terms on the right  You can check this step by distributing the P inside the parentheses. To get P by itself, divide both sides of the equation by  :  Divide both sides by 1 + *rt* and simplify     1. for μ   **Solution** It is easier to work with this equation if we eliminate the fraction on the right side. Multiply both sides by σ:    Multiply both sides by σ and simplify  To get μ by itself, subtract *x* from both sides and then multiply both sides by -1:  Subtract *x* from both sides and simplify    Multiply both sides by -1 | Solve each of the equations for the indicated variable.   1. for *x* 2. for *h* 3. for *b* 4. for *x* |

## How do you compute the range of data?

Key Terms

Range

Summary

Consider these three sets of quiz scores:

Section A: 5 5 5 5 5 5 5 5 5 5

Section B: 0 0 0 0 0 10 10 10 10 10

Section C: 4 4 4 5 5 5 5 6 6 6

All three of these sets of data have a mean of 5 and median of 5, yet the sets of scores are clearly quite different. In section A, everyone had the same score; in section B half the class got no points and the other half got a perfect score, assuming this was a 10-point quiz. Section C was not as consistent as section A, but not as widely varied as section B.

In addition to the mean and median, which are measures of the "typical" or "middle" value, we also need a measure of how "spread out" or varied each data set is.

There are several ways to measure this "spread" of the data. The first is the simplest and is called the range.

The **range** is the difference between the maximum value and the minimum value of the data set.

Notes

1. Practice 2

|  |  |
| --- | --- |
| In a recent class, the overall percentages at the end of the semester were recorded.  82 86 80 77 83 92 81 61 68 73  86 83 84 94 86 78 65 90  Find the range of the data.  **Solution** Examining the data, notice that the smallest value is 61 and the highest value is 94. Since the range is the difference between the maximum and minimum values, | In a recent class, the overall percentages at the end of the semester were recorded.  73 71 72 73 64 71 77 38 58 75  85 36 86 39 61  Find the range of the data. |

## What does the standard deviation tell you about data?

Key Terms

Deviation from the mean Squared deviation from the mean

Sample variance Sample standard deviation

Population variance Population standard deviation

Summary

**Instead of looking at the difference between highest and lowest to find the range, let’s look at the difference between each data value and the center. The center we will use is the mean. The difference between the data value and the mean is called the deviation from the mean.**

Mean

Data Value





To see how this works, let’s use the data set from the data below describing temperatures in Flagstaff, Arizona.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 71 | 59 | 69 | 68 | 63 | 57 |
| 57 | 57 | 57 | 65 | 67 |  |

The mean was about 62.7°F.

To find the deviations from the mean, subtract 62.7 from each data value:

|  |  |
| --- | --- |
|  |  |
| 71 | 8.3 |
| 59 | -3.7 |
| 69 | 6.3 |
| 68 | 5.3 |
| 63 | 0.3 |
| 57 | -5.7 |
| 57 | -5.7 |
| 57 | -5.7 |
| 57 | -5.7 |
| 65 | 2.3 |
| 67 | 4.3 |
| **Sum** | **0.3** |

Notice that the sum of the deviations is around zero. If there is no rounding of the mean, then this should add up to exactly zero. So, what does that mean? Does this imply that on average the data values are zero distance from the mean? No. It just means that some of the data values are above the mean and some are below the mean. The negative deviations are for data values that are below the mean and the positive deviations are for data values that are above the mean. We need to get rid of the sign. How do we get rid of a negative sign? Squaring a number is a widely accepted way to make all the numbers positive.

Let’s square all of the deviations. To find these values, **square the deviations from the mean**.



Also, you can think of this as being the squared distance from the mean.

|  |  |  |
| --- | --- | --- |
|  |  |  |
| 71 | 8.3 | 68.89 |
| 59 | -3.7 | 13.69 |
| 69 | 6.3 | 39.69 |
| 68 | 5.3 | 28.09 |
| 63 | 0.3 | 0.09 |
| 57 | -5.7 | 32.49 |
| 57 | -5.7 | 32.49 |
| 57 | -5.7 | 32.49 |
| 57 | -5.7 | 32.49 |
| 65 | 2.3 | 5.29 |
| 67 | 4.3 | 18.49 |
| **Sum** | **0.3** | **304.19** |

**Now that we have the sum of the squared deviations, we should find the mean of these values. However, since this is a sample, the normal way to find the mean, summing and dividing by *n*, does not estimate the true population value correctly. It would underestimate the true value. So, to calculate a better estimate, we will divide by a slightly smaller number, *n* - 1. This strange average is known as the sample variance.**

The **sample variance** is the sum of the squared deviations from the mean divided by . The symbol for sample variance is  and the formula for the sample variance is

****

For this data set, the sample variance is

The variance measures the average squared distance from the mean. Since we want to know the average distance from the mean, we will need to take the square root at this point.

The **sample standard deviation** is the square root of the variance. The standard deviation is a measure of the average distance the data values are from the mean. The symbol for sample standard deviation is and the formula for the sample standard deviation is



For this data set, the sample standard deviation is . Note that the units are the same as the original data.

Since the sample variance and the sample standard deviation are used to estimate the population variance and population standard deviation, we should define the symbols and formulas for those as well.

**Population Variance**: 

**Population Standard Deviation**: 

Notes

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| An instructor is comparing the performance of students in to classes. He records the final percentages students have earned.  Class A: 82 86 80 77 83 92 81 61 68 73 86 83 84 94 86 78 65 90  Class B: 73 71 72 73 64 71 77 38 58 75 85 36 86 39 61   1. Without computing the mean, explain which class has the higher mean and why you think this is so.   **Solution** Start by ordering each set of data from smallest to largest:  Class A: 61 65 68 73 77 78 80 81 82 83 83 84 86 86 86 90 92 94  Class B: 36 38 39 58 61 64 71 71 72 73 73 75 77 85 86  Examining Class A, the center of the data is in the low 80’s. For class B, the middle of the data is in the low 70’s. However, Class B has a number of very small data values which could draw the mean down. So, class A has the higher mean.   1. Compute the mean for each class and verify your answer to part a.   **Solution** For each set of data, add the data values and divide by the total number of data,  Class A :  Class B :  These numbers are consistent with the estimates in part a.   1. Without computing the standard deviation for each class, explain which class has the higher standard deviation and why you think this is so.   **Solution** The standard deviation measures the spread in the data. Class A has scores that vary from a low of 61 to a high of 94 and Class B varies from a low of 36 to a high of 86. The range for Class A is 33 and for Class B has a range of 50. Since range is also a measure of spread and Class B has a larger range, the standard deviation for Class B is larger.   1. Compute the standard deviation for each class and verify your answer to part c.   **Solution** To calculate the standard deviation for each class, fill out the table below:     |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | | Class A *x* |  |  |  | Class B *x* |  |  | | 82 | 1.5 | 2.25 |  | 73 | 7.7 | 59.29 | | 86 | 5.5 | 30.25 |  | 71 | 5.7 | 32.49 | | 80 | -0.5 | 0.25 |  | 72 | 6.7 | 44.89 | | 77 | -3.5 | 12.25 |  | 73 | 7.7 | 59.29 | | 83 | 2.5 | 6.25 |  | 64 | -1.3 | 1.69 | | 92 | 11.5 | 132.25 |  | 71 | 5.7 | 32.49 | | 81 | 0.5 | 0.25 |  | 77 | 11.7 | 136.89 | | 61 | -19.5 | 380.25 |  | 38 | -27.3 | 745.29 | | 68 | -12.5 | 156.25 |  | 58 | -7.3 | 53.29 | | 73 | -7.5 | 56.25 |  | 75 | 9.7 | 94.09 | | 86 | 5.5 | 30.25 |  | 85 | 19.7 | 388.09 | | 83 | 2.5 | 6.25 |  | 36 | -29.3 | 858.49 | | 84 | 3.5 | 12.25 |  | 86 | 20.7 | 428.49 | | 94 | 13.5 | 182.25 |  | 39 | -26.3 | 691.69 | | 86 | 5.5 | 30.25 |  | 61 | -4.3 | 18.49 | | 78 | -2.5 | 6.25 |  |  |  |  | | 65 | -15.5 | 240.25 |  |  |  |  | | 90 | 9.5 | 90.25 |  |  |  |  | | **SUM** |  | **1374.5** |  | **SUM** |  | **3644.95** |   Use the formula for the standard deviation:    Based on these standard deviations, Class A has the larger standard deviation and is more spread out. This is consistent with the answer from Part c. |

Practice 3

|  |
| --- |
| 1. A tennis promoter is comparing two brands of tennis ball to determine which one gives a faster serve. The following data represents the top speeds (in mph) clocked by comparable players using each ball.   Ball A: 68, 76, 94, 74, 93, 70, 71, 86, 66, 92, 64, 62, 76, 85, 83, 69  Ball B: 98, 86, 84, 92, 96, 89, 73, 71, 91, 58, 87, 93, 73, 89, 93, 69   1. Without computing the mean, explain which ball has the higher mean and why you think this is so. 2. Compute the mean for each ball and verify your answer to part a. 3. Without computing the standard deviation for each ball, explain which class has the higher mean and why you think this is so. 4. Compute the standard deviation for each ball and verify your answer to part c. |

## How do you interpret quartiles and percentiles of data?

Key Terms

Five number summary Quartile Percentile

Box and whisker plot Interquartile range

Summary  
There are other calculations that we can do to look at spread. One of those is called **percentile**. This looks at what data value has a certain percent of the data at or below it.

A percentile is a value with that value of percent of the data at or below this value.

For example, if a data value is in the 80th percentile, then 80% of the data values fall at or below this value.

We see percentiles in many places in our lives. If you take any standardized tests, your score is given as a percentile. If you take your child to the doctor, their height and weight are given as percentiles. If your child is tested for gifted or behavior problems, the score is given as a percentile. If your child has a score on a gifted test that is in the 92nd percentile, then that means that 92% of all children who took the same gifted test scored the same or lower than your child. That also means that 8% scored the same or higher than your child. This may mean that your child is gifted.

1. Practice 4

|  |  |
| --- | --- |
| Suppose you took the SAT mathematics test and received your score as a percentile.   1. What does a score in the 90th percentile mean?   **Solution** This means that 90 percent of the scores were at or below your score (You did the same as or better than 90% of the test takers.)   1. What does a score in the 70th percentile mean?   **Solution** This tells you that 70% of the scores were at or below your score.   1. If the test was out of 800 points and you scored in the 80th percentile, what was your score on the test?   **Solution** You do not know! All you know is that you scored the same as or better than 80% of the people who took the test. If all the scores were low, you could have still failed the test. On the other hand, if many of the scores were high you could have gotten a 95% on the test. | Suppose you took the GRE general test and received your score as a percentile.   1. What does a score in the 50th percentile mean? 2. What does a score in the 99th percentile mean? 3. If your score was in the 95th percentile, does that mean you passed the test? |

There are three percentiles that are commonly used. They are the first, second, and third quartiles, where the quartiles divide the data into 25% sections.

**First Quartile (Q1):** 25th percentile (25% of the data falls at or below this value.)  
**Second Quartile (Q2 or M):** 50th percentile, also known as the median (50% of the data falls at or below this value.).  
**Third Quartile (Q3):** 75th percentile (75% of the data falls at or below this value.)

To find the quartiles of a data set:

1. Sort the data set from the smallest value to the largest value.
2. Find the median (M or Q2).
3. Find the median of the lower 50% of the data values. This is the first quartile (Q1).
4. Find the median of the upper 50% of the data values. This is the third quartile (Q3).

If we put the three quartiles together with the maximum and minimum values, then we have five numbers that describe the data set. This is called the **five-number summary**.

**Five-Number Summary:** Lowest data value known as the minimum (Min), the first quartile (Q1), the median (M or Q2), the third quartile (Q3), and the highest data value known as the maximum (Max).

Also, since we have the quartiles, we can talk about how much spread there is between the 1st and 3rd quartiles. This is known as the **interquartile range** (IQR).

IQR = Q3 – Q1

There are times when we want to look at the five-number summary in a graphical representation. This is known as a **box-and-whiskers plot** or a **box plot**.

A box plot is created by first setting a scale (number line) as a guideline for the box plot. Then, draw a rectangle that spans from Q1 to Q3 above the number line. Mark the median with a vertical line through the rectangle. Next, draw dots for the minimum and maximum points to the sides of the rectangle. Finally, draw lines from the sides of the rectangle out to the dots.

Notes

1. Practice 5

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| The first 11 days of May 2013 in Flagstaff, AZ, had the following high temperatures (in °F):   |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | | 71 | 59 | 69 | 68 | 63 | 57 | | 57 | 57 | 57 | 65 | 67 |  |  (Weather Underground, n.d.)  1. Find the five-number summary for this data.   **Solution** Notice that there is an odd number of data we will need to consider as we calculate medians.To find the five-number summary, you must first put the numbers in order from smallest to largest.  57, 57, 57, 57, 59, 63, 65, 67, 68, 69, 71  Now find the median. The number 63 is in the middle of the data set, so the median is 63°F.    To find Q1, look at the numbers below the median. Since 63 is the median, you do not include that in the listing of the numbers below the median. To find Q3, look at the numbers above the median. Since 63 is the median, you do not include that in the listing of the numbers above the median.  Looking at the numbers below the median, the median of those is 57so Q1 = 57°F. Looking at the numbers above the median, the median of those is 68 so Q3 = 68°F.  Now find the minimum and maximum. The minimum is 57°F and the maximum is 71°F.  Thus, the five-number summary is:  Min = 57°F  Q1 = 57°F  Med = Q2 = 63°F  Q3 = 68°F  Max = 71°F.   1. Find the interquartile range.   Solution The interquartile range is the difference between the third quartile and the first quartile,  IQR = Q3 – Q1 = 68 – 57 = 11°F   1. Draw a box plot for this data.   **Solution** Draw a box plot for this data set as follows:  Temperatures in °F in Flagstaff, AZ, in early May 2013    Notice that the median is basically in the center of the box, which implies that the data is not skewed. However, the minimum value is the same as Q1, so that implies there might be a little skewing, though not much. | In a recent class, the overall percentages at the end of the semester were recorded.   |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | | 73 | 71 | 72 | 73 | 64 | 71 | 77 | 39 | | 38 | 58 | 75 | 85 | 36 | 86 | 61 |  |  1. Find the five-number summary for this data. 2. Find the interquartile range. 3. Draw a box plot for this data. |

1. Practice 6

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| The first 12 days of May 2013 in Flagstaff, AZ, had the following high temperatures (in °F):   |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | | 71 | 59 | 69 | 68 | 63 | 57 | | 57 | 57 | 57 | 65 | 67 | 73 |    (Weather Underground, n.d.)  1. Find the five-number summary for this data.   **Solution** This set of data contains 12 values. To find the five-number summary, you must first put the data values in order from smallest to largest.  57, 57, 57, 57, 59, 63, 65, 67, 68, 69, 71, 73  Then find the median. The numbers 63 and 65 are in the middle of the data set, so the median is    To find Q1, look at the numbers below the median. Since the number 64 is the median, you include all the numbers below 64, including the 63 that you used to find the median.  Looking at the numbers below the median 57, 57, 57, 57, 59, 63, the median of those is  so Q1 = 57°F.  To find Q3, look at the numbers above the median. Since the number 64 is the median, you include all the numbers above 64, including the 65 that you used to find the median.  Looking at the numbers above the median 65, 67, 68, 69, 71, 73, the median of those is  so Q3 = 68.5°F.  Now find the minimum and maximum. The minimum is 57°F and the maximum is 73°F.Thus, the five-number summary is:  Min = 57°F  Q1 = 57°F  Med = Q2 = 64°F  Q3 = 68.5°F  Max = 73°F.   1. Find the interquartile range.   **Solution** The interquartile range is  IQR = Q3 – Q1 = 68.5 – 57 = 11.5°F   1. Draw a box plot for this data.   **Solution** Draw a box plot for this data set as follows:  Temperatures in °F in Flagstaff, AZ, in early May 2013    **Notice that the median is basically in the center of the box, so that implies that the data is not skewed. However, the minimum value is the same as Q1, so that implies there might be a little skewing, though not much.** | In a recent class, the overall percentages at the end of the semester were recorded.   |  |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | --- | | 82 | 86 | 80 | 77 | 83 | 92 | 81 | 61 | 68 | | 73 | 86 | 83 | 84 | 94 | 86 | 78 | 65 | 90 |  1. Find the five-number summary for this data. 2. Find the interquartile range. 3. Draw a box plot for this data. |

# Section 1.6 The Normal Distribution

**Get Started** – What is a concept map?

* How do I apply the 68-95-99.7 rule to a set of data?
* What is the relationship between the area under a normal curve and z-scores?

## **Get Started** – What is a concept map?

Key Terms

Concept map

Summary

When making a concept map there is no right way to do it. You want to make connections that make sense to you. You might want to include examples or pictures, anything that helps you make sense of what you are studying.

Flash cards can be a great way to learn facts, but they don’t usually lead to knowledge or understanding. Taking those flash cards and turning them into a concept map will help you obtain knowledge.

Steps for making a Concept Map

1. **Make a list of all the topics you are studying.** Looking through your text book for words in bold type, looking at the heading of the pages, thinking about what you’ve been studying are all good ways to start. Reviewing homework and quizzes (in-class and Canvas quizzes) will also help.
2. **Put each topic on a separate piece of paper.** Index cards work well but aren’t required. Any scraps of paper will work.
3. **Start by trying to group them in just a few general categories.** It is OK to make a category for topics you don’t know what to do with. But don’t get lazy and put every topic in that category. In the example on the previous page our categories were Types of Data and Measures of Center.
4. **Pick one group and organize just those topics.** How do they connect to each other?
5. **Repeat step 4 with each group.**
6. **Try to connect each subgroup with the other groups.** You’ll be surprised how many connections you can find when you look for them.
7. **Take a picture of your concept map.**

When studying for big exams making, and remaking, concept maps are a great way to review the material.

**As we work through the semester you should make a concept map for each chapter and then try to link the maps for each chapter into one big map**. If you can do this, you will be READY for the final exam.

Notes

## How do I apply the 68-95-99.7 rule to a set of data?

Key Terms

Normal distribution Empirical rule (68-95-99.7 rule

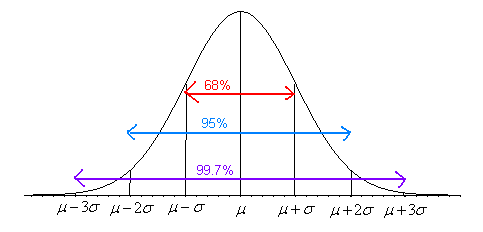
Summary

There are many different types of distributions (shapes) of quantitative data. In section 1.5 we looked at different histograms and described the shapes of them as symmetric, skewed left, and skewed right. There is a special symmetric shaped distribution called the normal distribution. It is high in the middle and then goes down quickly and equally on both ends. It looks like a bell, so sometimes it is called a bell curve. One property of the normal distribution is that it is symmetric about the mean. Another property has to do with what percentage of the data falls within certain standard deviations of the mean. This property is defined as the empirical Rule.

**The Empirical Rule:** Given a data set that is approximately normally distributed:

* Approximately 68% of the data is within one standard deviation of the mean.
* Approximately 95% of the data is within two standard deviations of the mean.
* Approximately 99.7% of the data is within three standard deviations of the mean.

To visualize these percentages, see the following figure.



*Note: The empirical rule is only true for approximately normal distributions.*

Notes

|  |
| --- |
| Suppose that your class took a test and the mean score was 75% and the standard deviation was 5%. If the test scores follow an approximately normal distribution, answer the following questions using the empirical rule (68-95-99.7 rule).   1. What percentage of the students had scores between 65 and 85?   **Solution** To solve this problem, it would be helpful to draw the normal curve that follows this situation. The mean is 75, so the center is 75. The standard deviation is 5, so for each line above the mean add 5 and for each line below the mean subtract 5. The graph looks like the following:    99.7%  95%  68%  From the graph we can see that 95% of the students had scores between 65 and 85.   1. What percentage of the students had scores between 65 and 75?   **Solution** Using the graph above, the scores of 65 to 75 are half of the area of the graph from 65 to 85. Because of symmetry, that means that the percentage for 65 to 85 is ½ of the 95%, which is 47.5%.   1. What percentage of the students had scores between 70 and 80?   **Solution** From the graph we can see that 68% of the students had scores between 70 and 80.   1. What percentage of the students had scores above 85?   **Solution** For this problem we need a bit of math. If you looked at the entire curve, you would say that 100% of all the test scores fall under it. So, because of symmetry 50% of the test scores fall in the area above the mean and 50% of the test scores fall in the area below the mean. We know from part b that the percentage from 65 to 75 is 47.5%. Because of symmetry, the percentage from 75 to 85 is also 47.5%. So, the percentage above 85 is 50% - 47.5% = 2.5%. |

Practice 1

|  |
| --- |
| A normal distribution has a mean of 25 and a standard deviation of 3. Use the​ empirical rule (68-95-99.7) rule to answer the questions below.   1. What is the percentage of values between 22 and 28? 2. What is the percentage of values between 22 and 25? 3. What is the percentage of values between 16 and 34? 4. What is the percentage of values below 22? |

1. Practice 2

|  |  |
| --- | --- |
| When consumer goods are packaged, the amount of goods in each package is not exactly what is labeled on the package. Suppose the weights (in ounces) of chips in a bag are normally distributed with a mean of 12 ounces and a standard deviation of 1 ounce. Use the empirical rule (68-95-99.7 rule) to answer each question below.   1. What percentage of bags have less than 12 ounces?   **Solution** Since 12 ounces is the mean, the percentage of bags below the mean will be 50%. This also means the percentage of bags above the mean is also 50% since the mean splits the normal distribution into two symmetrical halves.   1. What percentage of bags have less than 10 ounces?   **Solution** A weight of 10 ounces is two standard deviations below the mean. From the empirical rule, we know that 95% of bags will be between two standard deviations below the mean and two standard deviations above the mean. So,  or 47.5% of bags will be between two standard deviations below the mean and the mean. The bags below two standard deviations is calculated by taking all of the bags below the mean (50%) and subtracting the bags between two standard deviation below the mean and the mean (47.5%). This tells us that 50% - 47.5% or 2.5% have less than 10 ounces. | When consumer goods are packaged, the amount of goods in each package is not exactly what is labeled on the package. Suppose the weights (in ounces) of cookies in a bag are normally distributed with a mean of 20 ounces and a standard deviation of 2 ounces. Use the empirical rule (68-95-99.7 rule) to answer each question below.   1. What percentage of bags have less than 20 ounces? 2. What percentage of bags have less than 18 ounces? |

## What is the relationship between the area under a normal curve and z-scores?

Key Terms

Raw score z score

Summary

When we look at Guided Examples 1, we realize that the numbers on the scale are not as important as how many standard deviations a number is from the mean. As an example, the number 80 is one standard deviation from the mean. The number 65 is 2 standard deviations from the mean. However, 80 is above the mean and 65 is below the mean. Suppose we wanted to know how many standard deviations the number 82 is from the mean. How would we do that? The other numbers were easier because they were a whole number of standard deviations from the mean. We need a way to quantify this. We will use a z-score (also known as a z-value or standardized score) to measure how many standard deviations a data value is from the mean. This is defined as:



where *x* = data value (raw score)

*z* = standardized value (z-score or z-value)

μ = population mean

σ = population standard deviation

*Note: Remember that the z-score is always how many standard deviations a data value is from the mean of the distribution.*

Suppose a data value has a z-score of 2.13. This tells us two things. First, it says that the data value is above the mean, since it is positive. Second, it tells us that you must add more than two standard deviations to the mean to get to this value. Since most data (95%) is within two standard deviations, then anything outside this range would be considered a strange or unusual value. A z-score of 2.13 is outside this range so it is an unusual value. As another example, suppose a data value has a z-score of -1.34. This data value must be below the mean, since the z-score is negative, and you need to subtract more than one standard deviation from the mean to get to this value. Since this is within two standard deviations, it is an ordinary value.

An **unusual value** has a z-score < -2 or a z-score > 2

A **usual value** has a z-score between -2 and 2, that is .

You may encounter standardized scores on reports for standardized tests or behavior tests as mentioned previously.

Notes

1. Practice 3

|  |  |
| --- | --- |
| Suppose that your class took a test the mean score was 75% and the standard deviation was 5%. If test scores follow an approximately normal distribution, answer the following questions:   1. If a student earned 87 on the test, what is that student’s z-score and what does it mean?   **Solution** The problem tells us that ,  , and . Put these values into the formula for z-scores and we get    This means that the score of 87 is more than two standard deviations above the mean, and so it is considered to be an unusual score.   1. If a student earned 73 on the test, what is that student’s z-score and what does it mean?   **Solution** The problem tells us that ,  , and . Put these values into the formula for z-scores and we get    This means that the score of 73 is less than one-half of a standard deviation below the mean. It is considered to be a usual or ordinary score. | Suppose that your class took a test the mean score was 65% and the standard deviation was 10%. If test scores follow an approximately normal distribution, answer the following questions:   1. If a student earned 92 on the test, what is that student’s z-score and what does it mean? 2. If a student earned 52 on the test, what is that student’s z-score and what does it mean? |

1. Practice 4

|  |  |
| --- | --- |
| Suppose that your class took a test the mean score was 75% and the standard deviation was 5%. If test scores follow an approximately normal distribution, answer the following questions:   1. If a student has a z-score of 1.43, what actual score did she get on the test?   **Solution** This problem involves a little bit of algebra. Do not worry, it is not that hard. Since you are now looking for *x* instead of *z*, rearrange the equation solving for *x* as follows:    Add μ to both sides and simplify. Rearrange to get x on the left side.  Multiply both sides by σ and simplify    Now, you can use this formula to find *x* when you are given *z*.          Thus, the *z*-score of 1.43 corresponds to an actual test score of 82.15%.   1. If a student has a z-score of , what actual score did he get on the test?   **Solution** Use the formula for *x* from part d of this problem:          Thus, the z-score of -2.34 corresponds to an actual test score of 63.3%. | Suppose that your class took a test the mean score was 65% and the standard deviation was 10%. If test scores follow an approximately normal distribution, answer the following questions:   1. If a student has a z-score of 1.2, what actual score did she get on the test? 2. If a student has a z-score of -1.9, what actual score did he get on the test? |

Looking at the Empirical Rule, 99.7% of all data is within three standard deviations of the mean. This means that an approximation for the minimum value in a normal distribution is the mean minus three times the standard deviation, and for the maximum is the mean plus three times the standard deviation. In a normal distribution, the mean and median are the same. Lastly, the first quartile can be approximated by subtracting 0.67448 times the standard deviation from the mean, and the third quartile can be approximated by adding 0.67448 times the standard deviation to the mean. All these together give the five-number summary.

In mathematical notation, the five-number summary for the normal distribution with mean μ and standard deviation σ is as follows:

**Five-Number Summary for a Normal Distribution**



1. Practice 5

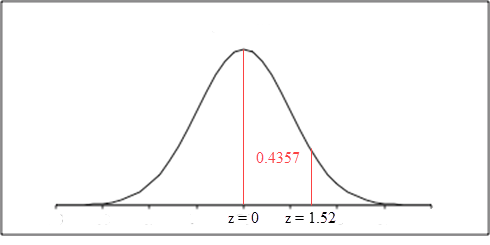
|  |  |
| --- | --- |
| Suppose that your class took a test and the mean score was 75% and the standard deviation was 5%. If the test scores follow an approximately normal distribution, find the five-number summary.  **Solution** The mean is  and the standard deviation is . Thus, the five-number summary for this problem is: | Suppose that your class took a test and the mean score was 65% and the standard deviation was 10%. If the test scores follow an approximately normal distribution, find the five-number summary. |

The empirical rule helps us to calculate the percentage of data values when the data values fall on one, two, or three standard deviations above or below the mean. These data values correspond to z = ±1, z = ±2, and z = ±3. However, what if a data value falls between these z-scores? In these cases, we use a table of normal curve areas to find percentages. A copy of this table is shown below:

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **z** | **0** | **0.01** | **0.02** | **0.03** | **0.04** | **0.05** | **0.06** | **0.07** | **0.08** | **0.09** |
| **0** | 0 | 0.0040 | 0.0080 | 0.0120 | 0.0160 | 0.0199 | 0.0239 | 0.0279 | 0.0319 | 0.0359 |
| **0.1** | 0.0398 | 0.0438 | 0.0478 | 0.0517 | 0.0557 | 0.0596 | 0.0636 | 0.0675 | 0.0714 | 0.0753 |
| **0.2** | 0.0793 | 0.0832 | 0.0871 | 0.0910 | 0.0948 | 0.0987 | 0.1026 | 0.1064 | 0.1103 | 0.1141 |
| **0.3** | 0.1179 | 0.1217 | 0.1255 | 0.1293 | 0.1331 | 0.1368 | 0.1406 | 0.1443 | 0.1480 | 0.1517 |
| **0.4** | 0.1554 | 0.1591 | 0.1628 | 0.1664 | 0.1700 | 0.1736 | 0.1772 | 0.1808 | 0.1844 | 0.1879 |
| **0.5** | 0.1915 | 0.1950 | 0.1985 | 0.2019 | 0.2054 | 0.2088 | 0.2123 | 0.2157 | 0.2190 | 0.2224 |
| **0.6** | 0.2257 | 0.2291 | 0.2324 | 0.2357 | 0.2389 | 0.2422 | 0.2454 | 0.2486 | 0.2517 | 0.2549 |
| **0.7** | 0.2580 | 0.2611 | 0.2642 | 0.2673 | 0.2704 | 0.2734 | 0.2764 | 0.2794 | 0.2823 | 0.2852 |
| **0.8** | 0.2881 | 0.2910 | 0.2939 | 0.2967 | 0.2995 | 0.3023 | 0.3051 | 0.3078 | 0.3106 | 0.3133 |
| **0.9** | 0.3159 | 0.3186 | 0.3212 | 0.3238 | 0.3264 | 0.3289 | 0.3315 | 0.3304 | 0.3365 | 0.3389 |
| **1.0** | 0.3413 | 0.3438 | 0.3461 | 0.3485 | 0.3508 | 0.3531 | 0.3554 | 0.3577 | 0.3599 | 0.3621 |
| **1.1** | 0.3643 | 0.3665 | 0.3686 | 0.3708 | 0.3729 | 0.3749 | 0.3770 | 0.3790 | 0.3810 | 0.3830 |
| **1.2** | 0.3849 | 0.3869 | 0.3888 | 0.3907 | 0.3925 | 0.3944 | 0.3962 | 0.3980 | 0.3997 | 0.4015 |
| **1.3** | 0.4032 | 0.4049 | 0.4066 | 0.4082 | 0.4099 | 0.4115 | 0.4131 | 0.4147 | 0.4162 | 0.4177 |
| **1.4** | 0.4192 | 0.4207 | 0.4222 | 0.4236 | 0.4251 | 0.4265 | 0.4279 | 0.4292 | 0.4306 | 0.4319 |
| **1.5** | 0.4332 | 0.4345 | 0.4357 | 0.4370 | 0.4382 | 0.4394 | 0.4406 | 0.4418 | 0.4429 | 0.4441 |
| **1.6** | 0.4452 | 0.4463 | 0.4474 | 0.4484 | 0.4495 | 0.4505 | 0.4515 | 0.4525 | 0.4535 | 0.4545 |
| **1.7** | 0.4554 | 0.4564 | 0.4573 | 0.4582 | 0.4591 | 0.4599 | 0.4608 | 0.4616 | 0.4625 | 0.4633 |
| **1.8** | 0.4641 | 0.4649 | 0.4656 | 0.4664 | 0.4671 | 0.4678 | 0.4686 | 0.4693 | 0.4699 | 0.4706 |
| **1.9** | 0.4713 | 0.4719 | 0.4726 | 0.4732 | 0.4738 | 0.4744 | 0.4750 | 0.4756 | 0.4761 | 0.4767 |
| **2.0** | 0.4772 | 0.4778 | 0.4783 | 0.4788 | 0.4793 | 0.4798 | 0.4803 | 0.4808 | 0.4812 | 0.4817 |
| **2.1** | 0.4821 | 0.4826 | 0.483 | 0.4834 | 0.4838 | 0.4842 | 0.4846 | 0.4850 | 0.4854 | 0.4857 |
| **2.2** | 0.4861 | 0.4864 | 0.4868 | 0.4871 | 0.4875 | 0.4878 | 0.4881 | 0.4884 | 0.4887 | 0.4890 |
| **2.3** | 0.4893 | 0.4896 | 0.4898 | 0.4901 | 0.4904 | 0.4906 | 0.4909 | 0.4911 | 0.4913 | 0.4916 |
| **2.4** | 0.4918 | 0.4920 | 0.4922 | 0.4925 | 0.4927 | 0.4929 | 0.4931 | 0.4932 | 0.4934 | 0.4936 |
| **2.5** | 0.4938 | 0.4940 | 0.4941 | 0.4943 | 0.4945 | 0.4946 | 0.4948 | 0.4949 | 0.4951 | 0.4952 |
| **2.6** | 0.4953 | 0.4955 | 0.4956 | 0.4957 | 0.4959 | 0.4960 | 0.4961 | 0.4962 | 0.4963 | 0.4964 |
| **2.7** | 0.4965 | 0.4966 | 0.4967 | 0.4968 | 0.4969 | 0.4970 | 0.4971 | 0.4972 | 0.4973 | 0.4974 |
| **2.8** | 0.4974 | 0.4975 | 0.4976 | 0.4977 | 0.4977 | 0.4978 | 0.4979 | 0.4979 | 0.4980 | 0.4981 |
| **2.9** | 0.4981 | 0.4982 | 0.4982 | 0.4983 | 0.4984 | 0.4984 | 0.4985 | 0.4985 | 0.4986 | 0.4986 |
| **3.0** | 0.4987 | 0.4987 | 0.4987 | 0.4988 | 0.4988 | 0.4989 | 0.4989 | 0.4989 | 0.4990 | 0.4990 |
| **3.1** | 0.4990 | 0.4991 | 0.4991 | 0.4991 | 0.4992 | 0.4992 | 0.4992 | 0.4992 | 0.4993 | 0.4993 |
| **3.2** | 0.4993 | 0.4993 | 0.4994 | 0.4994 | 0.4994 | 0.4994 | 0.4994 | 0.4995 | 0.4995 | 0.4995 |
| **3.3** | 0.4995 | 0.4995 | 0.4995 | 0.4996 | 0.4996 | 0.4996 | 0.4996 | 0.4996 | 0.4996 | 0.4997 |
| **3.4** | 0.4997 | 0.4997 | 0.4997 | 0.4997 | 0.4997 | 0.4997 | 0.4997 | 0.4997 | 0.4997 | 0.4998 |
| **3.5** | 0.4998 | 0.4998 | 0.4998 | 0.4998 | 0.4998 | 0.4998 | 0.4998 | 0.4998 | 0.4998 | 0.4998 |
| **3.6** | 0.4998 | 0.4998 | 0.4999 | 0.4999 | 0.4999 | 0.4999 | 0.4999 | 0.4999 | 0.4999 | 0.4999 |
| **3.7** | 0.4999 | 0.4999 | 0.4999 | 0.4999 | 0.4999 | 0.4999 | 0.4999 | 0.4999 | 0.4999 | 0.4999 |
| **3.8** | 0.4999 | 0.4999 | 0.4999 | 0.4999 | 0.4999 | 0.4999 | 0.4999 | 0.4999 | 0.4999 | 0.4999 |

This table indicates the percentage of data values between the mean and a *z*-value in the table.

For instance, the area in the table colored red tells us two things. The row the colored cell is in indicates the z-score begins with 1.5. The second decimal comes from the column the colored cell is in, .02. Putting this together with the number in the colored cell tells us that a z-score of 1.52 corresponds to the percentage 0.4357.



This tells us the 43.57% of data values are between the mean and *z* = 1.52.

Note that since the normal distribution is symmetric, the normal curve areas can be used for positive and negative *z*-scores.

Guided Example 6 Practice 6

|  |  |
| --- | --- |
| Use a Standard Normal Distribution table to find the percentage of values from *z* = 0 to *z* = -1.82.  **Solution** Using the table, locate the row for *z* = 1.8. Now locate the column for .02. The row and column intersect at z = 1.82. Recalling that the table also corresponds to negative z values enables us to read the percentage of values from *z* = 0 to *z* = -1.82 as 0.4656 or 46.56%. | Use a Standard Normal Distribution table to find the percentage of values from *z* = 0 to *z* = 0.5. |

Practice 7

|  |  |
| --- | --- |
| Use a Standard Normal Distribution table to find the percentage of values from *z* = 0.4 to *z* = 1.2.  Solution Start by finding the percentage of values using the table for *z* = 1.2. This percentage is 38.49%. This is the percentage of values from the mean to *z* = 1.2.  Now find the percentage for z = 0.4. This s read from the table as 0.1554 or 15.54%. The difference between these percentages, 38.49% - 15.54% or 22.95% is the percentage of values between *z* = 0.4 and *z* = 1.2. | Use a Standard Normal Distribution table to find the percentage of values from *z* = 0.5 to *z* = 0.8. |

Practice 8

|  |  |
| --- | --- |
| Use a Standard Normal Distribution table to find the percentage of values under the standard normal curve that is above *z* = 1.59.  Solution The percentage of the values above the mean is 50%. If we subtract the percentage of values from the mean to z = 1.59 from 50%, we will get the percentage above *z* = 1.59. The table gives the percentage between the mean and *z* = 1.59 as 44.41%. So, the percentage above z = 1.59 is 50% - 44.41% or 5.59%. | Use a Standard Normal Distribution table to find the percentage of area under the standard normal curve that is below *z* = -1.25. |

**Chapter 1 Practice Solutions**

Section 1.1

1. a. Qualitative, b. Quantitative , c. Quantitative
2. a. Population is students enrolled in college. Sample is students enrolled at the University of Arizona, b. Population is registered voters in the United States. Sample is registered voters likely to vote in the United States
3. a. The statistic is 89.6% and the parameter is 79.4%. b. The average height of Americans is a parameter and the average height of women in America is a statistic.
4. a. continuous, b. discrete

Section 1.2

1. 
2. 
3. a. Convenience sampling. b. Stratified sampling. c. Systematic sampling. d. Cluster sampling. e. Simple random sampling
4. a. Loaded question bias. b. Perceived lack of anonymity. c. Loaded question. d. Self-interest bias
5. Treatment group is pregnant women who take folic acid. Treatment is taking folic acid. Control group is pregnant women who do not take folic acid.
6. a. Yes, an inactive pill that appears to look like folic acid could be administered. b. No, you cannot hold pain medication for a dental extraction.

Section 1.3

1. a. > b. <
2. a. [-5, 2), b. [-3, 2)
3. a.

|  |  |
| --- | --- |
| Color | Frequency |
| Blue | 6 |
| Brown | 12 |
| Hazel | 1 |
| Green | 1 |

b.

|  |  |
| --- | --- |
| Color | Frequency |
| Blue | 0.3 |
| Brown | 0.6 |
| Hazel | 0.05 |
| Green | 0.05 |

1. a.

b.

c.

d.

|  |  |
| --- | --- |
| Blue |  |
| Brown |  |
| Hazel |  |
| Green |  |

 = 2 people



Section 1.4

1. 1002.14
2. 13.2
3. 17.4%
4. 40%
5. 
6. a. 0.2 =  , b. 0.04 = 
7. a. approximately 65.3, b. 71, c. 71, 73
8. a. 67.125, b. 71.5, c. 71, 73
9. mode = brown
10. a. approximately 8.2, b. 8, c. 8
11. 2.8
12. a. 4 or 5, b. approximately 1.5, c. 1, d. 0

Section 1.5

1. a.  , b.  or  , c.  , d. 
2. 50
3. a. Ball B has higher values so mean is higher, b. Mean of Ball A is 76.8125 and mean of ball B is 83.875, c. Ball B values are more spread out so the standard deviation is higher, d. The standard deviation of Ball A is 10.67 and the standard deviation of Ball B is 11.51.
4. No, you only know that 95% of scores were the same or lower.
5. a. Minimum = 36, Q1 = 58, Median = 71, Q3 = 75, Maximum = 86, b. 17, c.

30 40 50 60 70 80 90

1. a. Minimum = 61, Q1 = 77, Median = 82.5, Q3 = 86, Maximum = 94, b. 9,

c.

60 70 80 90 100

Section 1.6

1. a. 68%, b. 34%, c. 99.7%, d. 16%
2. a. 50%, b. 16%
3. a. 2.7, the score is 2.7 standard deviations above the mean, b. -1.3, the score is 1.3 standard deviations below the mean.
4. a. 77, b. 46
5. Minimum = 35, Q1 ≈ 58.3, Median = 65, Q3 ≈ 71.7, Maximum = 95
6. 19.15%
7. 9.66%
8. 10.56%