Chapter 3: Statistics

Student Outcomes for this Chapter

Section 3.1: Overview of the Statistical Process

Students will be able to:

- Define and identify the population, parameter, sample and statistic
- Identify four sampling methods: simple random sample (SRS), stratified, systematic and convenience
- Identify and discuss types of bias association with sampling
- Distinguish between experimental and observational studies
- Explain margin of error and confidence intervals

Section 3.2: Describing Data

Students will be able to:

- Define and identify categorical and quantitative data
- Read and construct frequency tables and relative frequency tables
- Make bar charts and pie charts for categorical variables by hand and/or using technology
- Identify elements of misleading graphs: 3-dimensional graphs, perceptual distortion, misleading scales, stacked bar graphs
- Make histograms for quantitative variables by hand and/or using technology
- Identify the number of modes in a distribution and whether it is symmetric, skewed to the left, or skewed to the right

Section 3.3: Summary Statistics: Measures of Center

Students will be able to:

- Calculate and describe the measures of center: mean and median
- Analyze the relationship of the mean and median to the shape of the data

Section 3.4: Summary Statistics: Measures of Variation

Students will be able to:

- Calculate and describe the measures of variation: standard deviation, range and interquartile range (IQR)
- Calculate the 5-number summary and construct boxplots by hand and/or using technology (boxplots using technology may be modified or not)
- Compare distributions with side-by-side boxplots and percentiles
- Calculate and apply Z-scores

Section 3.1 Overview of the Statistical Process

Introduction to Statistics

We are bombarded by information and statistics every day. But if we cannot distinguish credible information from misleading information, then we are vulnerable to manipulation and making decisions that are not in our best interest. Statistics provides tools for us to evaluate information critically. In this sense, statistics is one of the most important things to know about.

Statistics are often presented to add credibility to an argument. To give some examples, here are some claims that we have heard on several occasions. (We are *not* saying that each one of these claims is true!)

- 4 out of 5 dentists recommend Dentyne.
- Almost 85% of lung cancers in men and 45% in women are tobacco-related.
- Condoms are effective 94% of the time.
- People tend to be more persuasive when they look others directly in the eye and speak loudly and quickly.
- Women make 75 cents to every dollar a man makes when they work the same job.
- A surprising new study shows that eating egg whites can increase one's life span.
- People predict that it is very unlikely there will ever be another baseball player with a batting average over 400.
- There is an 80% chance that in a room full of 30 people that at least two people will share the same birthday.
- 79.48% of all statistics are made up on the spot.

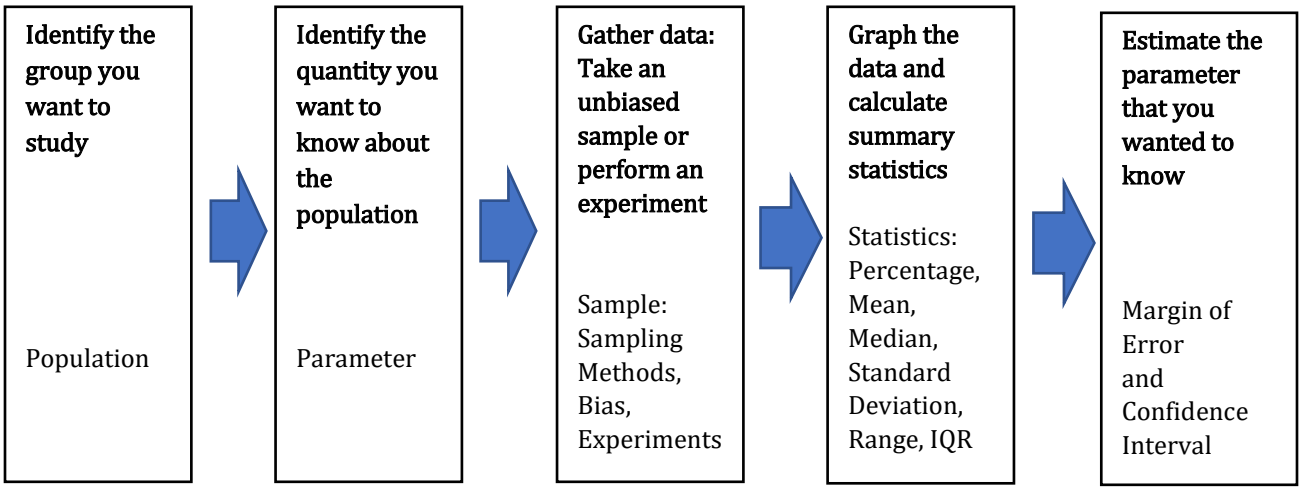
All these claims are statistical in character. We suspect that some of them sound familiar; if not, you have probably heard other claims like them. Notice how diverse the examples are; they come from psychology, health, law, sports, business, etc. Data and data-interpretation show up in virtually every facet of contemporary life.

Many of these numbers do not represent careful statistical analysis. They can be misleading and push you into decisions that you might regret. This chapter will help you learn the skills to be a critical consumer of statistical claims.

Statistical Process

To give you an overview, here is a diagram of the steps taken in a poll or other statistical study, and the elements in each step that we will discuss in this chapter. We will use many examples to illustrate the whole process.

Overview of the Statistical Process



Population

Before we begin gathering any data to analyze, we need to identify the population we are studying. The **population** of a study is the group we want to know something about. The population could be people, auto parts or tomato plants.

If we want to know the amount of money spent on textbooks by a typical college student, our population might be all students at Portland Community College. Or it might be:

- All community college students in the state of Oregon.
- All students at public colleges and universities in the state of Oregon.
- All students at all colleges and universities in the state of Oregon.
- All students at all colleges and universities in the entire United States.
- And so on.

The intended population is also called the **target population**, since if we design our study badly, the collected data might not actually be representative of the intended population.

> Example 1: A newspaper website contains a poll asking people their opinion on a recent news article. What is the population?
>
> While the target (intended) population may have been all people, the real population of the survey is readers of the website.

Parameter

A **parameter** is the value (percentage, average, etc.) that we are interested in for the whole population. Since it is often too time-consuming, expensive and/or impossible to get data for the entire population, the parameter is usually a theoretical quantity that we are trying to estimate. For example, the typical amount of money spent per year on textbooks by students at your college in a year is a parameter.

Sample

To estimate the parameter, we select a **sample**, which is a smaller subset of the entire population. It is very important that we choose a **representative sample**, one that matches the characteristics of the population, to have a good estimate. If we survey 100 students at your college, those students would be our sample. We will talk about how to choose a representative sample later in this section.

Statistic

To get our **data**, we would then ask each student in the sample how much they spent on textbooks and record the answers, or **raw data**. Then we could calculate the average, which is our statistic. A **statistic** is a value (percentage, average, etc.) calculated using data from a sample.

> Example 2: A researcher wants to know how the citizens of Portland feel about a voter initiative. To study this, they go downtown to the Pioneer Place Mall and survey 500 shoppers. Sixty percent indicate they are supportive of the initiative. Identify the population, parameter, sample and statistic.
>
> *Population:* While the intended population of this survey is Portland citizens, the effective population is Pioneer Place Mall shoppers. There is no reason to assume that shoppers at this mall would be representative of all Portland citizens.
>
> *Parameter:* The parameter is what we want to know about the population, the percentage of Portland citizens that support the initiative.
>
> *Sample:* The sample is the subgroup of the population selected. The 500 shoppers questioned make up the sample, which, again, is probably not representative of the population.
>
> *Statistic:* The statistic is calculated using the data from the sample. The percentage of people sampled who support the initiative is 60%.

Sampling

As we mentioned in a previous section, the first thing we should do before conducting a survey is to identify the population that we want to study. Suppose we are hired by a politician to determine the amount of support they have among the electorate should they decide to run for another term. What population should we study? Every person in the district? Eligible voters might be better, but what if they don't register? Registered voters may not vote. What about "likely voters?"

This is the criteria used in a lot of political polling, but it is sometimes difficult to define a "likely voter." Here is an example of the challenges of political polling.

> Example 3: In November 1998, former professional wrestler Jesse "The Body" Ventura was elected governor of Minnesota. Up until right before the election,

most polls showed he had little chance of winning. There were several contributing factors to the polls not reflecting the actual intent of the electorate:

- Ventura was running on a third-party ticket and most polling methods are better suited to a two-candidate race.
- Many respondents to polls may have been embarrassed to tell pollsters that they were planning to vote for a professional wrestler.
- The mere fact that the polls showed Ventura had little chance of winning might have prompted some people to vote for him in protest to send a message to the major-party candidates.

But one of the major contributing factors was that Ventura recruited a substantial amount of support from young people, particularly college students, who had never voted before and who registered specifically to vote in that election. The polls did not deem these young people likely voters (since in most cases young people have a lower rate of voter registration and a lower turnout rate for elections) so the polling samples were subject to **sampling bias**: they omitted a portion of the electorate that was weighted in favor of the winning candidate.

So, identifying the population can be a difficult job, but once we have identified the population, how do we choose a good sample? We want our statistic to estimate the parameter we are interested in, so we need to have a representative sample. Returning to our hypothetical job as a political pollster, we would not anticipate very accurate results if we drew all of our samples from customers at a Starbucks, or the membership list of the local Elks club. How do we get a sample that resembles our population?

Sampling Methods

One way to get a representative sample is to use *randomness*. We will look at three types of sampling that use randomness and one that does not.

Simple random sample (SRS)

A **simple random sample**, abbreviated SRS, is one in which every member of the population has an equal probability of being chosen.

Example 4: If we could somehow identify all likely voters in the state, put each of their names on a piece of paper, toss the slips into a (very large) hat and draw 1000 slips out of the hat, we would have a simple random sample.

In practice, computers are better suited for this sort of endeavor. It is always possible, however, that even a random sample might end up not being totally representative of the population. If we repeatedly take samples of 1000 people from among the population of likely voters in the state of Oregon, some of these samples might tend to have a slightly higher percentage of Democrats (or Republicans) than does the general population; some samples might include more older people and some samples might include more younger people; etc. This is called **sampling variation**.

If there are certain groups that we want to make sure are represented, we might instead use a stratified sample.

Stratified sampling

In **stratified sampling**, a population is divided into a number of subgroups (or strata). Random samples are then taken from each subgroup. It is often desirable to make the sample sizes proportional to the size of each subgroup in the population.

> Example 5: Suppose that data from voter registrations in the state indicated that the electorate was comprised of 39% Democrats, 37% Republicans and 24% Independents. In a sample of 1000 people, they would then expect to get about 390 Democrats, 370 Republicans and 240 Independents. To accomplish this, they could randomly select 390 people from among those voters known to be Democrats, 370 from those known to be Republicans, and 240 from those with no party affiliation.

A way to remember stratified sampling is think about having a piece of layer cake. Each layer represents a stratum or subgroup, and a slice of the cake represents a sample of each layer.

Systematic sampling

In **systematic sampling**, every n^{th} member of the population is selected to be in the sample. The starting position is often chosen at random.

> Example 6: To select a systematic sample, Portland Community College could use their database to select a random student from the first 100 student ID numbers. Then they would select every 100^{th} student ID number after that.

Systematic sampling is not as random as a simple random sample (if your ID number is right next to your friend's because you applied at the same time, you could not both end up in the same sample) but it can yield acceptable samples. This method can be useful for people waiting in lines, parts on a manufacturing line, or plants in a row.

Convenience sampling

Convenience sampling is when samples are chosen by selecting whomever is convenient. This is the worst type of sampling because it does not use randomness.

> Example 7: A pollster stands on a street corner and interviews the first 100 people who agree to speak to them. This is a convenience sample.

Statistical Bias

There is no way to correct for biased data, so it is very important to think through the entire study and data analysis before we start. We talked about **sampling** or **selection bias**, which is when the sample is not representative of the population. One example of this is **voluntary response bias**, which is bias introduced by only collecting data from those who volunteer to participate. This can lead to bias if the people who volunteer

have different characteristics than the general population. Here is a summary of some additional sources of bias.

Types of bias

Sampling bias – when the sample is not representative of the population

> **Voluntary response bias** – the sampling bias that often occurs when the sample is made up of volunteers

Self-interest study – bias that can occur when the researchers have an interest in the outcome

Response bias – when the responder gives inaccurate responses for any reason

> **Perceived lack of anonymity** – when the responder fears giving an honest answer might negatively affect them
>
> **Loaded questions** – when the question wording influences the responses

Non-response bias – when people refuse to participate in a study or drop out of an experiment, we can no longer be certain that our sample is representative of the population

Sources of bias may be conscious or unconscious. They may be innocent or as intentional as pressuring by a pollster. Here are some examples of the types of bias.

> Example 8:
>
> a. Consider a recent study which found that chewing gum may raise math grades in teenagers[1]. 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 mean the study was biased, we should subject the study to extra scrutiny.
>
> b. Consider online reviews of products and businesses. Customers tend to leave reviews if they are very satisfied or very dissatisfied. While you can look for overall patterns and get useful information, these reviews suffer from voluntary response bias and likely capture more extreme views than the general population.
>
> c. A survey asks participants a question about their interactions with people of different ethnicities. This study could suffer from response bias. A respondent might give an untruthful answer to not be perceived as racist.

[1] Reuters. http://news.yahoo.com/s/nm/20090423/od_uk_nm/oukoe_uk_gum_learning. Retrieved 4/27/09

d. An employer puts out a survey asking their employees if they have a drug abuse problem and need treatment help. Here, answering truthfully might have serious consequences; responses might not be accurate if there is a perceived lack of anonymity and employees fear retribution.

e. A survey asks, "Do you support funding research on alternative energy sources to reduce our reliance on high-polluting fossil fuels?" This is an example of a loaded or leading question – questions whose wording leads the respondent towards a certain answer.

f. A poll was conducted by phone with the question, "Do you often have time to relax and read a book?" Fifty percent of the people who were called refused to participate in the survey (Probably because they didn't have the time). It is unlikely that the results will be representative of the entire population. This is an example of non-response bias.

Loaded questions can occur intentionally by pollsters with an agenda, or accidentally through poor question wording. Also of concern is question order, where the order of questions changes the results. Here is an example from a psychology researcher[2]:

> Example 9: *"My favorite finding is this: we did a study where we asked students, 'How satisfied are you with your life? How often do you have a date?' The two answers were not statistically related - you would conclude that there is no relationship between dating frequency and life satisfaction. But when we reversed the order and asked, 'How often do you have a date? How satisfied are you with your life?' the statistical relationship was a strong one. You would now conclude that there is nothing as important in a student's life as dating frequency."*

Observational Studies

So far, we have primarily discussed surveys and polls, which are types of **observational studies** – studies based on observations or measurements. These observations may be solicited, like in a survey or poll. Or, they may be unsolicited, such as studying the percentage of cars that turn right at a red light even when there is a "no turn on red" sign.

Experiments

In contrast, it is common to use **experiments** when exploring how subjects react to an outside influence. In an experiment, some kind of **treatment** is applied to the subjects and the results are measured and recorded. When conducting experiments, it is essential to isolate the treatment being tested. Here are some examples of treatments.

[2] Swartz, Norbert. http://www.umich.edu/~newsinfo/MT/01/Fal01/mt6f01.html. Retrieved 3/31/2009

Example 10:

a. A pharmaceutical company tests a new medicine for treating Alzheimer's disease by administering the drug to 50 elderly patients with recent diagnoses. The treatment here is the new drug.

b. A gym tests out a new weight loss program by enlisting 30 volunteers to try out the program. The treatment here is the new program.

c. A psychology researcher explores the effect of music on affect by measuring people's mood while listening to different types of music. The music is the treatment.

d. Suppose a middle school finds that their students are not scoring well on the state's standardized math test. They decide to run an experiment to see if a new curriculum would improve scores. To run the test, they hire a math specialist to come in and teach a class using the new curriculum. To their delight, they see an improvement in test scores.

The difficulty with the last scenario is that it is not clear whether the new curriculum or the math specialist is responsible for the improvement. This is called confounding and it is the downfall of many experiments, though it is often hidden.

Confounding

Confounding occurs when there are two or more potential variables that could have caused the outcome and it is not possible to determine which one actually caused the result.

Example 11:

a. A drug company study about a weight loss pill might report that people lost an average of 8 pounds while using their new drug. However, in the fine print you find a statement saying that participants were encouraged to also diet and exercise. It is not clear in this case whether the weight loss is due to the pill, to diet and exercise, or a combination of both. In this case confounding has occurred.

b. Researchers conduct an experiment to determine whether students will perform better on an arithmetic test if they listen to music during the test. They first give the student a test without music, then give a similar test while the student listens to music. In this case, the student might perform better on the second test, regardless of the music, simply because it was the second test and they were warmed up.

There are a number of measures that can be introduced to help reduce the likelihood of confounding. The primary measure is to use a control group.

Control group

In experiments, the participants are typically divided into a treatment group and a control group. The treatment group receives the treatment being tested; the **control group** does not receive the treatment.

Ideally, the groups are otherwise as similar as possible, isolating the treatment as the only potential source of difference between the groups. For this reason, the method of dividing groups is important. Some researchers attempt to ensure that the groups have similar characteristics (same number of each gender identity, same number of people over 50, etc.), but it is nearly impossible to control for every characteristic. Because of this, random assignment is very commonly used.

Example 12:

a. To determine if a two-day prep course would help high school students improve their scores on the SAT test, a group of students was randomly divided into two subgroups. The first group, the treatment group, was given a two-day prep course. The second group, the control group, was not given the prep course. Afterwards, both groups took the SAT test.

b. A company testing a new plant food grows two crops of plants in adjacent fields that typically produce the same amount of food. The treatment group receives the new plant food and the control group does not. The crop yields would then be compared. By growing the two crops at the same time in similar fields, they are controlling for weather and other confounding factors.

Sometimes not giving the control group anything does not completely control for confounding variables. For example, suppose a medicine study is testing a new headache pill by giving the treatment group the pill and the control group nothing. If the treatment group showed improvement, we would not know whether it was due to the medicine, or a response to have something. This is called a placebo effect.

Placebo effect

The **placebo effect** is when the effectiveness of a treatment is influenced by the patient's perception of how effective they think the treatment will be, so a result might be seen even if the treatment is ineffectual.

Example 13: A study found that when doing painful dental tooth extractions, patients told they were receiving a strong painkiller while actually receiving a saltwater injection found as much pain relief as patients receiving a dose of morphine.[3]

[3] Levine JD, Gordon NC, Smith R, Fields HL. (1981) Analgesic responses to morphine and placebo in individuals with postoperative pain. Pain. 10:379-89.

To control for the placebo effect, a **placebo**, or dummy treatment, is often given to the control group. This way, both groups are truly identical except for the specific treatment given.

Placebo and Placebo-controlled experiments

An experiment that gives the control group a placebo is called a **placebo-controlled** experiment.

Example 14:

a. In a study for a new medicine that is dispensed in a pill form, a sugar pill could be used as a placebo.

b. In a study on the effect of alcohol on memory, a non-alcoholic beer might be given to the control group as a placebo.

c. In a study of a frozen meal diet plan, the treatment group would receive the diet food, and the control group could be given standard frozen meals taken out of their original packaging.

In some cases, it is more appropriate to compare to a conventional treatment than a placebo. For example, in a cancer research study, it would not be ethical to deny any treatment to the control group or to give a placebo treatment. In this case, the currently acceptable medicine would be given to the second group, called a **comparison group**. In our SAT test example, the non-treatment group would most likely be encouraged to study on their own, rather than be asked to not study at all, to provide a meaningful comparison. It is very important to consider the ethical ramifications of any experiment.

Blind studies

A **blind study** is one that uses a placebo and the participants do not know whether they are receiving the treatment or a placebo. A **double-blind study** is one in which the subjects and those interacting with them don't know who is in the treatment group and who is in the control group.

Example 15: In a study about anti-depression medication, you would not want the psychological evaluator to know whether a patient is in the treatment or control group, as it might influence their evaluation. The experiment should be conducted as a double-blind study.

Margin of Error and Confidence Intervals

Even when a study or experiment has successfully avoided bias and has been well done, there is still an element of variation. If we took 5 different random samples of 100 college students and calculated their average textbook cost, we wouldn't expect to get the exact same average for each sample. This is due to sampling variation. To account for this, researchers publish their margin of error or a confidence interval for their

statistics. These numbers describe the precision of the estimate for a certain confidence level.

You've probably heard something like, "The candidate has 54 percent of the likely voters, plus or minus three percent." The 3% is called the **margin of error**, so the true percentage is somewhere between 51% and 57%, with a certain level of confidence. To write this as a **confidence interval**, we place the numbers in parentheses from smallest to largest, separated by a comma: (51%, 57%).

The most common **confidence level** is 95%, which means if the poll was conducted repeatedly, we would expect the true percentage, or parameter, to fall within our confidence interval 95 out of 100 times. You can learn more on how to calculate the margin of error for different confidence levels in a statistics class.

> Example 16: Let's say we asked a random sample of 100 students at Portland Community College and found that they spent an average of $451.32 on books their first year, plus or minus $85.63. Write this as a confidence interval, assuming a 95% confidence level.
>
> If the margin of error was calculated to be plus or minus $85.63, then with a confidence level of 95% we could say that the average amount spent by the population is somewhere between $424.69 and $478.26. We could also write this as a **confidence interval**: ($365.69, $536.95).

Now we have come full circle and seen how we can use data from a sample to estimate the parameter we were interested in for our population.

Exercises 3.1

1. A political scientist surveys 28 of the current 106 representatives in a state's congress. Of them, 14 said they were supporting a new education bill, 12 said there were not supporting the bill, and 2 were undecided.
    a. Who is the population of this survey?
    b. What is the size of the population?
    c. What is the size of the sample?
    d. Give the statistic for the percentage of representatives surveyed who said they were supporting the education bill.
    e. If the margin of error was 5%, give the confidence interval for the percentage of representatives we might we expect to support the education bill.

2. The city of Raleigh has 9,500 registered voters. There are two candidates for city council in an upcoming election: Brown and Feliz. The day before the election, a telephone poll of 350 randomly selected registered voters was conducted. 112 said they'd vote for Brown, 207 said they'd vote for Feliz, and 31 were undecided.
    a. Who is the population of this survey?

   b. What is the size of the population?
   c. What is the size of the sample?
   d. Give the statistic for the percentage of voters surveyed who said they'd vote for Brown.
   e. If the margin of error was 3.5%, give the confidence interval for the percentage of voters surveyed that we might we expect to vote for Brown.

3. To determine the average length of trout in a lake, researchers catch 20 fish and measure them. Describe the population and sample of this study.

4. A college reports that the average age of their students is 28 years old. Is this a parameter or a statistic?

5. Which sampling method is being described?
   a. In a study, the sample is chosen by separating all cars by size and selecting 10 of each size grouping.
   b. In a study, the sample is chosen by writing everyone's name on a playing card, shuffling the deck, then choosing the top 20 cards.
   c. Every 4th person on the class roster was selected.

6. Which sampling method is being described?
   a. A sample was selected to contain 25 people aged 18-34 and 30 people aged 35-70.
   b. Viewers of a new show are asked to respond to a poll on the show's website.
   c. To survey voters in a town, a polling company randomly selects 100 addresses from a database and interviews those residents.

7. Identify the most relevant source of bias in each situation.
   a. A survey asks the following: Should the mall prohibit loud and annoying rock music in clothing stores catering to teenagers?
   b. To determine opinions on voter support for a downtown renovation project, a surveyor randomly questions people working in downtown businesses.
   c. A survey asks people to report their actual income and the income they reported on their IRS tax form.
   d. A survey randomly calls people from the phone book and asks them to answer a long series of questions.
   e. The Beef Council releases a study stating that consuming red meat poses little cardiovascular risk.
   f. A poll asks, "Do you support a new transportation tax, or would you prefer to see our public transportation system fall apart?"

8. Identify the most relevant source of bias in each situation.
    a. A survey asks the following: Should the death penalty be permitted if innocent people might die?
    b. A study seeks to investigate whether a new pain medication is safe to market to the public. They test by randomly selecting 300 people who identify as men from a set of volunteers.
    c. A survey asks how many sexual partners a person has had in the last year.
    d. A radio station asks listeners to phone in their response to a daily poll.
    e. A substitute teacher wants to know how students in the class did on their last test. The teacher asks the 10 students sitting in the front row to state their latest test score.
    f. High school students are asked if they have consumed alcohol in the last two weeks.

9. Identify whether each situation describes an observational study or an experiment.
    a. The temperature on randomly selected days throughout the year was measured.
    b. One group of students listened to music and another group did not while they took a test and their scores were recorded.
    c. The weights of 30 randomly selected people are measured.

10. Identify whether each situation describes an observational study or an experiment.
    a. Subjects are asked to do 20 jumping jacks, and then their heart rates are measured.
    b. Twenty coffee drinkers and twenty tea drinkers are given a concentration test.
    c. The weights of potato chip bags are weighed on the production line before they are put into boxes.

11. A team of researchers is testing the effectiveness of a new vaccine for human papilloma virus (HPV). They randomly divide the subjects into two groups. Group 1 receives new HPV vaccine, and Group 2 receives the existing HPV vaccine. The patients in the study do not know which group they are in.
    a. Which is the treatment group?
    b. Which is the control group (if there is one)?
    c. Is this study blind, double-blind, or neither?
    d. Is this best described as an experiment, a controlled experiment, or a placebo-controlled experiment?

12. For the clinical trials of a weight loss drug containing *Garcinia Cambogia* the subjects were randomly divided into two groups. The first received an inert pill along with an exercise and diet plan, while the second received the test medicine along with the same exercise and diet plan. The patients do not know which group they are in, nor do the fitness and nutrition advisors.
    a. Which is the treatment group?
    b. Which is the control group (if there is one)?
    c. Is this study blind, double-blind, or neither?
    d. Is this best described as an experiment, a controlled experiment, or a placebo-controlled experiment?

13. A study is conducted to determine whether people learn better with routine or crammed studying. Subjects volunteer from an introductory psychology class. At the beginning of the semester 12 subjects volunteer and are assigned to the routine studying group. At the end of the semester 12 subjects volunteer and are assigned to the crammed studying group.
    a. Identify the target population and the sample.
    b. Is this an observational study or an experiment?
    c. This study involves two kinds of non-random sampling: 1. Subjects are not randomly sampled from a specified population and 2. Subjects are not randomly assigned to groups. Which problem is more serious? What effect on the results does each have?

14. A farmer believes that playing Barry Manilow songs to his peas will increase their yield. Describe a controlled experiment the farmer could use to test his theory.

15. A sports psychologist believes that people are more likely to be extroverted as an adult if they played team sports as a child. Describe two possible studies to test this theory. Design one as an observational study and the other as an experiment. Which is more practical?

16. To test a new lie detector, two groups of subjects are given the new test. One group is asked to answer all the questions truthfully. The second group is asked to tell the truth on the first half of the questions and lie on the second half. The person administering the lie detector test does not know what group each subject is in. Does this experiment have a control group? Is it blind, double-blind, or neither? Explain.

17. A poll found that 30%, plus or minus 5% of college freshmen prefer morning classes to afternoon classes.
    a. What is the margin of error?
    b. Write the survey results as a confidence interval.

18. A poll found that 38% of U.S. employees are engaged at work, plus or minus 3.5%.
    a. What is the margin of error?
    b. Write the survey results as a confidence interval.

19. Studies are often done by pharmaceutical companies to determine the effectiveness of a treatment. Suppose that a new cancer treatment is under study. Of interest is the average length of time in months patients live once starting the treatment. Two researchers each follow a different set of 40 cancer patients throughout this new treatment.
    a. What is the population of this study?
    b. Would you expect the data from the two researchers to be identical? Why or why not?
    c. If the first researcher collected their data by randomly selecting 10 nearby ZIP codes, then selecting 4 people from each, which sampling method did they use?
    d. If the second researcher collected their data by choosing 40 patients they knew, what sampling method did they use? What concerns would you have about this data set, based upon the data collection method?

20. Find a newspaper or magazine article, or the online equivalent, describing the results of a recent study (not a simple poll). Give a summary of the study's findings, then analyze whether the article provided enough information to determine the validity of the conclusions. If not, produce a list of things that are missing from the article that would help you determine the validity of the study. Look for the things discussed in the text: population, sample, randomness, blind, control, margin of error, etc.

21. Use a polling website such as www.pewresearch.org or www.gallup.com and search for a poll that interests you. Find the result, the margin of error and confidence level for the poll and write the confidence interval.

Section 3.2 Describing Data

Once we have collected data from an observational study or an experiment, we need to summarize and present it in a way that will be meaningful to our audience. The raw data is not very useful by itself. In this section we will begin with graphical presentations of data and in the rest of the chapter we will learn about numerical summaries of data.

Types of Data

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

Categorical (qualitative) data are pieces of information that allow us to classify the subjects into various categories.

> Example 1: We might conduct a survey to determine the name of the favorite movie that people saw in a movie theater. When we conduct such a survey, the responses would look like: *Finding Nemo*, *Black Panther*, *Titanic, etc*.
>
> We can count the number of people who give each answer, but the answers themselves do not have any numerical values: we cannot perform computations with an answer like *"Black Panther"* because it is categorical data.

Quantitative data are responses that are numerical in nature and with which we can perform meaningful calculations.

> Example 2: A survey could ask the number of movies you have seen in a movie theater in the past 12 months (0, 1, 2, 3, 4, ...). This would be quantitative data.

Other examples of quantitative data would be the running time of the movie you saw most recently (104 minutes, 137 minutes, 110 minutes, etc.) or the amount of money you paid for a movie ticket the last time you went to a movie theater ($5.50, $9.75, $10.50, etc.).

We cannot assume that all numbers are quantitative data, and sometimes it is not so clear-cut. Here are some examples to illustrate this.

> Example 3:
>
> a. Suppose we gather respondents' ZIP codes in a survey to track their geographical location. ZIP codes are numbers, but we can't do any meaningful calculations with them (it doesn't make sense to say that 98036 is "twice" 49018 — that's like saying that Lynnwood, WA is "twice" Battle Creek, MI, which doesn't make sense at all), so ZIP codes are really categorical data.
>
> b. A survey about the movie you most recently saw includes the question, "How would you rate the movie?" with these possible answers:

1 - It was awful.
2 - It was just okay.
3 - I liked it.
4 - It was great.
5 - Best movie ever!

Again, there are numbers associated with the responses, but these are really categories. A movie that rates a 4 is not necessarily twice as good as a movie that rates a 2, whatever that means; However, we often see that a movie got an average of 3.7 stars, which is an average of categorical ratings and it can give us important information.

Overall, it is important to look at the purpose of the study for any variables that could be classified as either categorical or quantitative. Another consideration is what you plan to do with the data. Next, we will talk about how to display each type of data.

Presenting Categorical Data

Since we can't do calculations with categorical data, we begin by summarizing the data in a frequency table or a relative frequency table.

Frequency Tables

A **frequency table** has one column for the categories, and another for the **frequency**, or number of times that category occurred.

Example 4: An insurance company determines vehicle insurance premiums based on known risk factors. If a person is considered a higher risk, their premiums will be higher. One potential factor is the color of your car. The insurance company believes that people with some color cars are more likely to get in accidents. To research this, they examine police reports for recent total-loss collisions. The data is summarized in this table.

Car Color	Frequency of Total-Loss Collisions
Blue	25
Green	52
Red	41
White	36
Black	39
Grey	23
Total	**216**

Relative Frequency Tables

Numbers are usually not as easy to interpret as percentages, so we will add a column for the relative frequencies. A **relative frequency** is the percentage for the category, found by dividing each frequency by the total and converting to a percentage. You'll notice the percentages may not add up to exactly 100% due to rounding.

Example 4 Continued:

Car Color	Frequency of Total-Loss Collisions	Relative Frequency of Total-Loss Collisions
Blue	25	25/216 = 0.116 or 11.6%
Green	52	52/216 = 0.241 or 24.1%
Red	41	41/216 = 0.190 or 19.0%
White	36	36/216 = 0.167 or 16.7%
Black	39	39/216 = 0.181 or 18.1%
Grey	23	23/216 = 0.107 or 10.7%
Total	**216**	216/216 = 1.0 or 100%

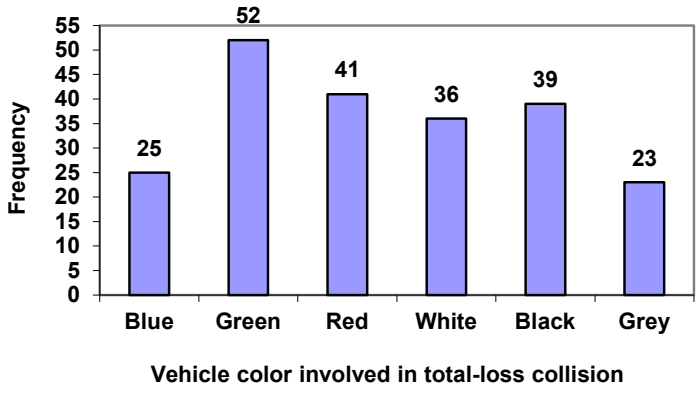
It would be even more useful to have a visual to see what is going on, and this is where charts and graphs come in. For categorical data we can display our data using bar graphs and pie charts.

Bar graphs

A **bar graph** is a graph that displays a bar for each category with the height of the bar indicating the frequency of that category. To construct a bar graph with vertical bars, we label the horizontal axis with the categories. The vertical axis will have a scale for the frequency or relative frequency.

The highest frequency in our car data is 52 collisions, so we will set our vertical axis to go from 0 to 55, with a scale of 5 units.

To draw bar graphs by hand graph paper is useful, or you can use technology. It is also very helpful to label each bar with the frequency or relative frequency.

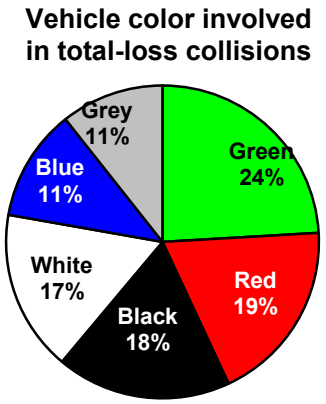


Pie Charts

A natural way to visualize relative frequencies is with a pie chart. A **pie chart** is a circle with wedges cut of varying sizes like slices of pizza or pie. The size of each wedge corresponds to the relative frequency of the category. The slices add up to 100%, just like relative frequencies.

Pie charts can often benefit from including frequencies or relative frequencies in the pie slices.

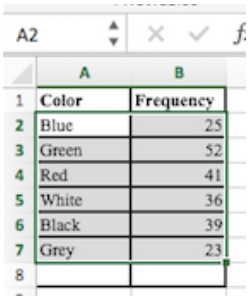
Pie charts look nice but are harder to draw by hand than bar charts since to draw them accurately we would need to compute the angle each wedge cuts out of the circle, then measure the angle with a protractor. A spreadsheet is much better suited to drawing pie charts.



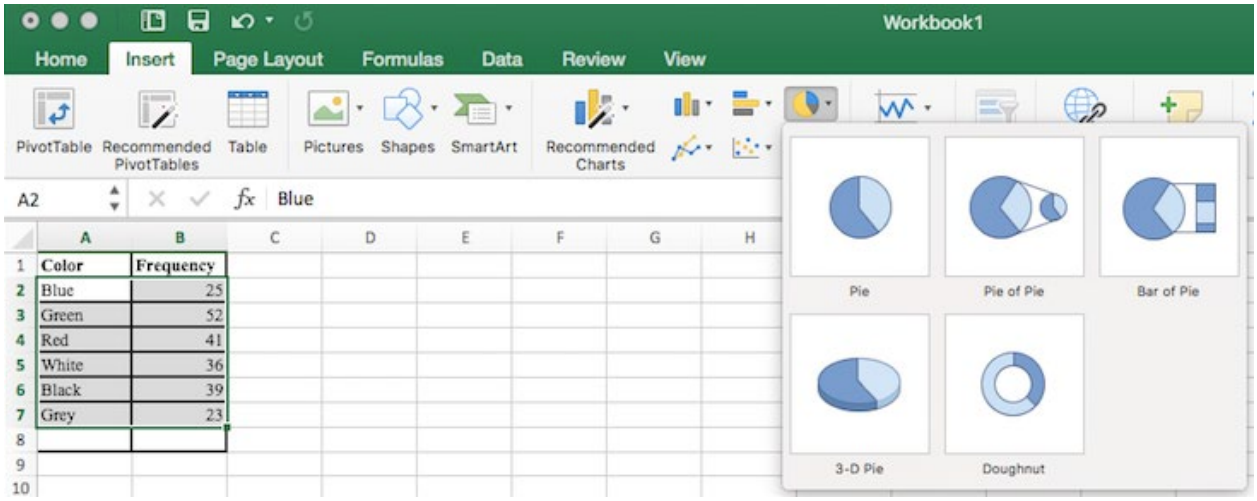
Using a Spreadsheet to Make Bar Charts and Pie Charts

To make a graph using a spreadsheet, place the data from the frequency table into the cells.

Then select the data, go to the Insert tab, and choose the bar graph or pie chart that you would like. For this example, we will choose a pie graph.

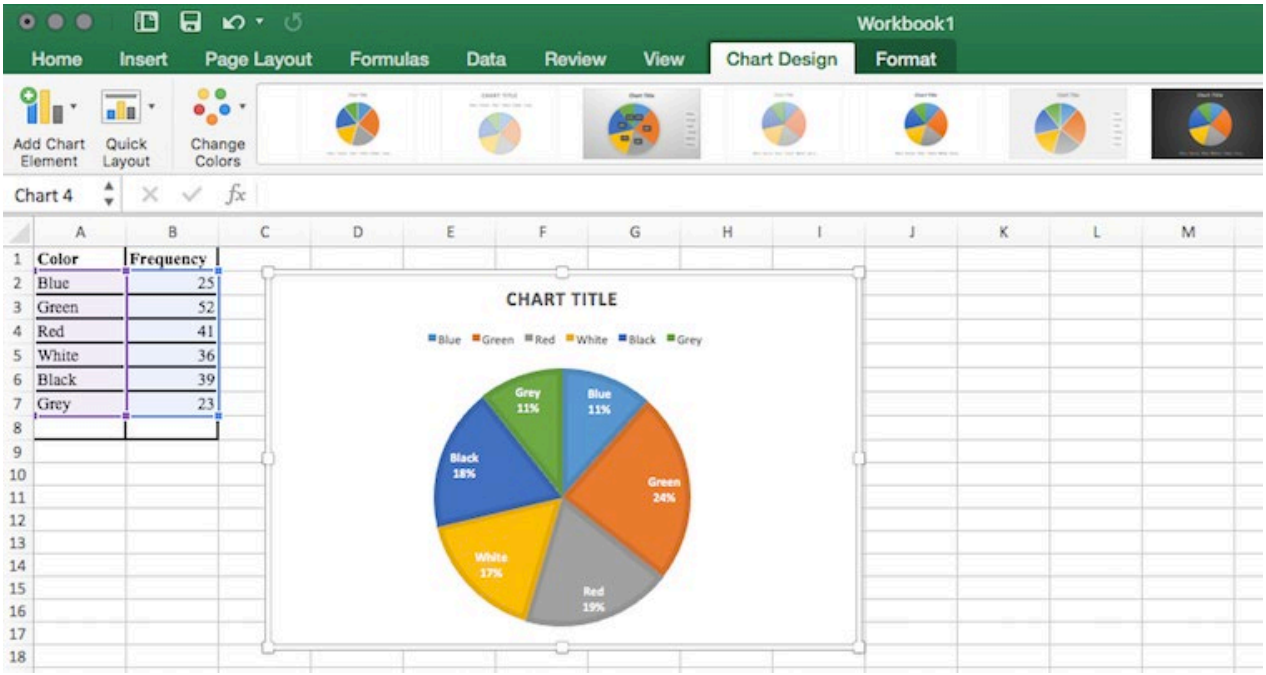


	A	B
1	Color	Frequency
2	Blue	25
3	Green	52
4	Red	41
5	White	36
6	Black	39
7	Grey	23
8		



After the spreadsheet has created your pie graph you can choose which design you prefer by clicking on the Chart Design tab. Since these pie pieces represent car colors, we matched the color of each wedge to the color of the car in our pie chart above.

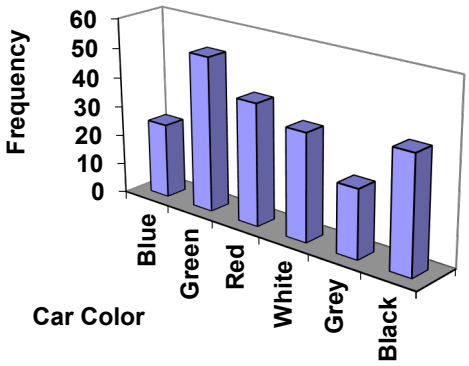
To give your graph a meaningful title, click on Chart Title. There are many other settings that you can experiment with.



Misleading Graphs

Graphs can be misleading intentionally or unintentionally. It's better to keep them simple, clear and well-labeled. People sometimes add features to graphs that don't help convey their information.

Example 5: A 3-dimensional bar chart like the one shown is usually not as effective as a 2-dimensional graph. The extra dimension does not add any useful information.



.

Here is another way that fanciness can sometimes lead to trouble. Instead of plain bars, it is tempting to substitute images. This type of graph is called a pictogram.

Perceptual Distortion

A **pictogram** is a statistical graphic in which the size of the picture is intended to represent the frequency or size of the values being represented. We need to be careful

with these, because our brains perceive the relationship between the areas, not the heights.

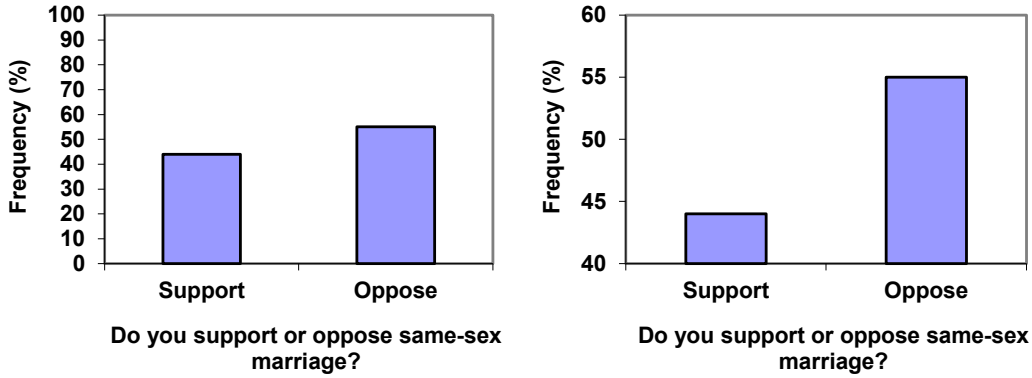
Example 6: A labor union might produce this graph to show the difference between the average manager salary and the average worker salary.

The average manager salary is twice as high as the average worker salary as in a bar graph, but the image is also twice as wide. That makes it look like the manager salary is 4 times as large as the worker salary. The area needs to accurately portray the relationship, otherwise we will have a perceptual distortion.

Misleading Scale

Another type of distortion in bar charts results from setting the baseline to a value other than zero. The baseline is the bottom of the vertical axis, representing the least number of cases that could have occurred in a category. Normally, this number should be zero.

Example 7: Compare the two graphs below showing support for same-sex marriage rights from a poll taken in December, 2008[4]. At a glance, the two graphs suggest very different stories. The second graph makes it look like more than three times as many people oppose marriage rights as support them. But when we look at the scale we can see that the difference is about 12%. By not starting at zero the difference looks enlarged.



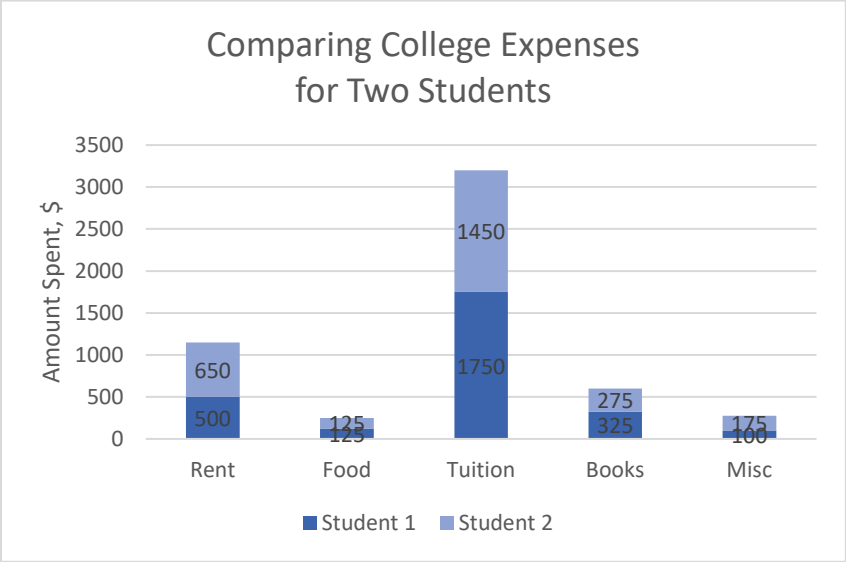
[4]CNN/Opinion Research Corporation Poll. Dec 19-21, 2008, from http://www.pollingreport.com/civil.htm

Stacked Bar Graphs

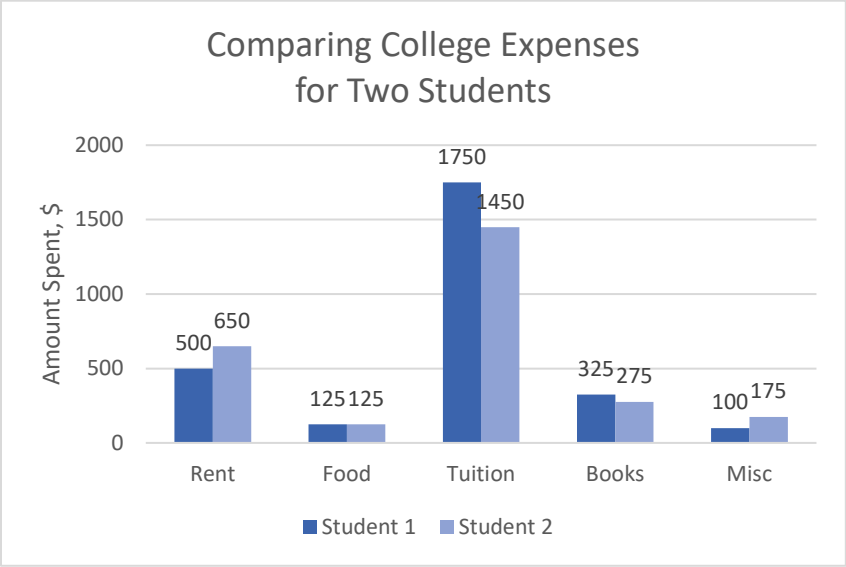
Another type of graph that can be hard to read and sometimes misleading is a stacked bar graph. In a **stacked bar graph**, the values we are comparing are stacked on top of each other vertically.

Example 8: The table lists college expenses for two different students and we want to compare them. A stacked bar graph shows the expenses stacked vertically, but we are interested in the differences, not the totals.

Expense	Student 1	Student 2
Rent	$500	$650
Food	$125	$125
Tuition	$1750	$1450
Books	$325	$275
Misc	$100	$175



It would be much easier to interpret the differences in a side-by-side bar chart.



Presenting Quantitative Data

With categorical data, the horizontal axis is the category, but with quantitative, or numerical, data we have numbers. If we have repeated values we can also make a frequency table.

Example 9: A teacher records scores on a 20-point quiz for the 30 students in their class. The scores are:

19, 20, 18, 18, 17, 18, 19, 17, 20, 18, 20, 16, 20, 15, 17, 12, 18, 19, 18, 19, 17, 20, 18, 16, 15, 18, 20, 5, 0 and 0.

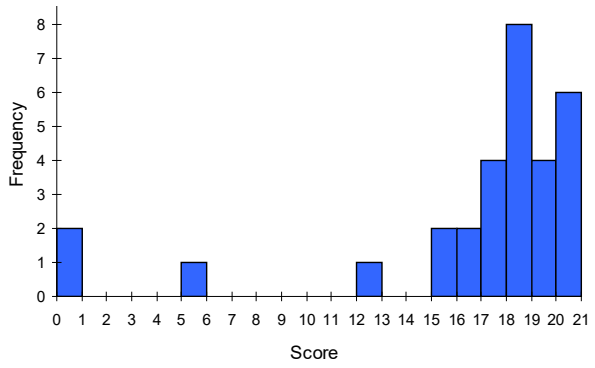
Here is a frequency table with the scores grouped and put in order.

Quiz Score	Frequency of Students
0	2
5	1
12	1
15	2
16	2
17	4
18	8
19	4
20	6

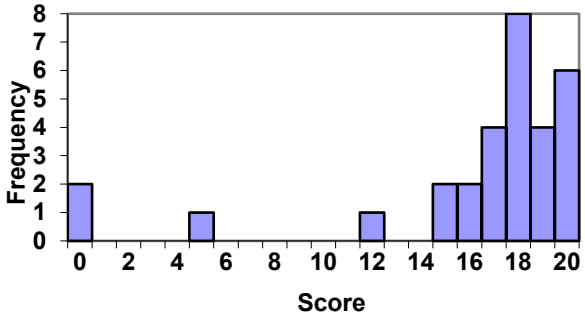
Using this table, it would be possible to create a standard bar chart from this summary, like we did for categorical data. However, since the scores are numerical values, this chart wouldn't make sense; the first and second bars would be five values apart, while the later bars would only be one value apart. Instead, we will to treat the horizontal axis as a number line. This type of graph is called a histogram.

Histograms

A **histogram** is like a bar graph, but the horizontal axis is a number line. Unlike a bar graph, there are no spaces between the bars. Here is a histogram for the data given above. Notice that in this histogram, the two scores of 15 are to the right of 15, or between 15 and 16.



The horizontal scales on histograms can be confusing for this reason. Some people choose to have bars start at ½ values to avoid this ambiguity, as in this second histogram.



If we have a large number of different data values, a frequency table listing every possible value would be way too long. There would be too many bars on the histogram to reveal any patterns. For this reason, it is common with quantitative data to group data into class intervals.

Class Intervals

Class intervals are groupings of the data. In general, we define class intervals so that:

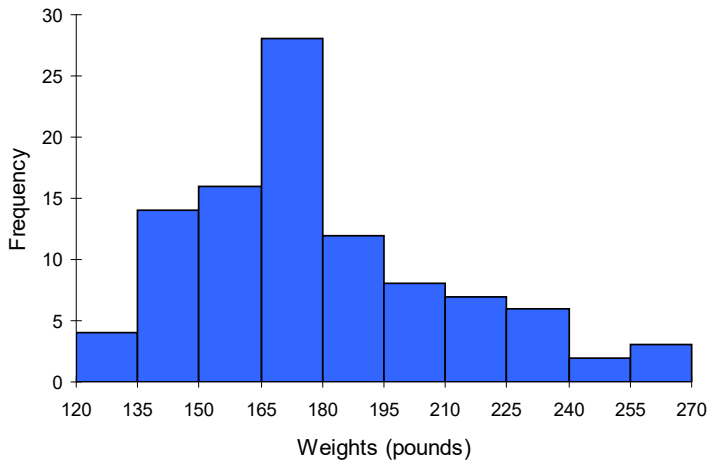
1. Each interval is equal in size. For example, if the first class contains values from 120-134, the second class should include values from 135-149.
2. We typically have somewhere between 5 and 20 classes, depending on the number of data values we're working with.

In the next example, we'll make a histogram using class intervals.

Example 10: Suppose we have collected weights from 100 subjects who identify as male, as part of a nutrition study. For our weight data, we have values ranging from a low of 121 pounds to a high of 263 pounds, giving a total span of 263-121 = 142. We could create 7 intervals with a width of around 20, 14 intervals with a width of around 10, or somewhere in between. We often have to experiment with a few possibilities to find something that represents the data well. We will try using a class width of 15. We could start at 121, or at 120 since it is a nice round number.

Interval	Frequency
120 - 134	4
135 – 149	14
150 – 164	16
165 – 179	28
180 – 194	12
195 – 209	8
210 – 224	7
225 – 239	6
240 – 254	2
255 - 269	3

Here is a histogram of this data:

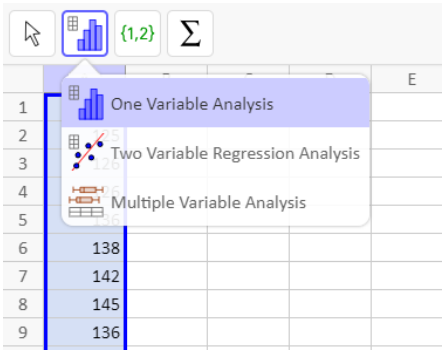


When using class intervals, it is much easier to use technology that was specifically designed to make histograms. GeoGebra is one program that lets you adjust the class widths to see which graph best displays the data.

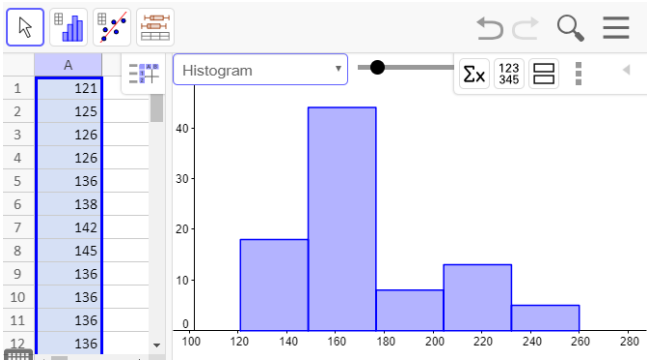
Histograms Using Technology

We will be using GeoGebra throughout this chapter to make graphs and calculate summary statistics. There is an online version and one you can download available at www.GeoGebra.org. The instructions are similar for both.

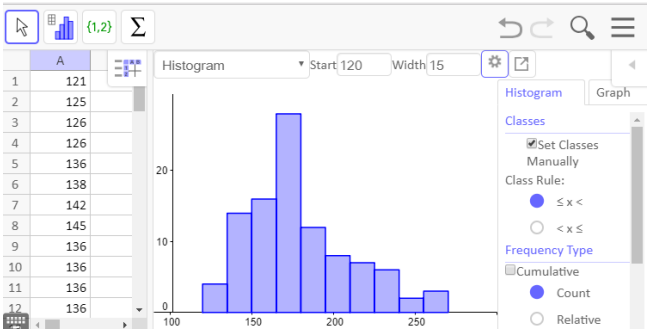
The first thing we need to do is enter the data in GeoGebra's spreadsheet. You can access the spreadsheet from Main Menu →View →Spreadsheet. Next, enter your data and select that column. Then click on the histogram icon in the menu bar on the left side and select One Variable Analysis.



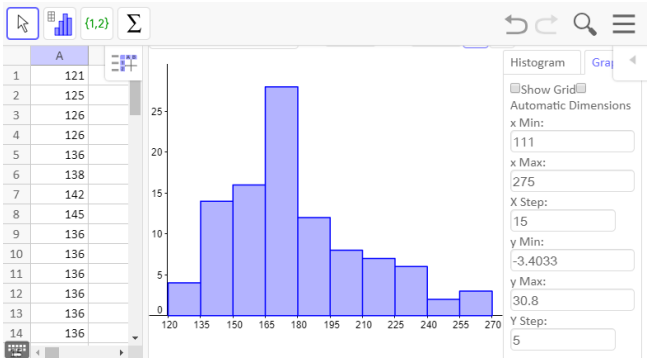
A new window will pop up showing a visual of the data. There is a drop-down menu for the type of graph, but histogram is the default. Notice that the bars are not lined up with the tick-marks at the bottom, so we want to edit this histogram. The slider bar at the top will let you see different class widths, but we want to choose our class widths manually.



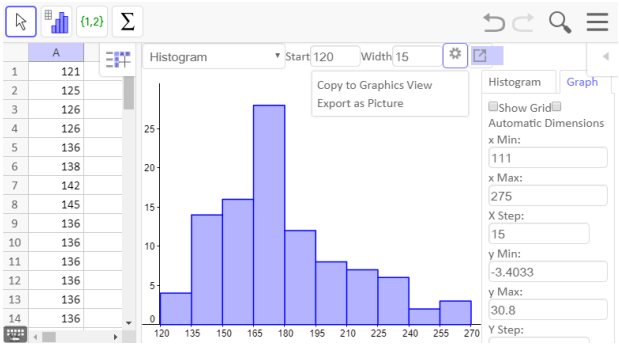
If you close the menu at the top right by clicking on the left pointing triangle, you will see a settings wheel. Click on the wheel and check the box for set classes manually. To match our previous histogram, we will start at 120 pounds and set a class width of 15 pounds.



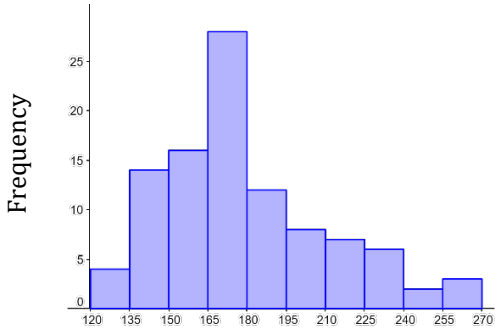
Now the bars of the histogram match our previous graph, but we need to edit the axis labels to match. Click on the graph tab on the right side and uncheck the box for automatic dimensions. We set the x min, x max, x step, y min, y max and y step as shown.



To put the graph in an assignment or a book such as this one, select the export icon and choose Export as Picture. The downloaded version also has a Copy to Clipboard option. Then insert the graph into any document and add axis labels.



Here is our finished histogram:



Weights (pounds)

The Shape of a Distribution

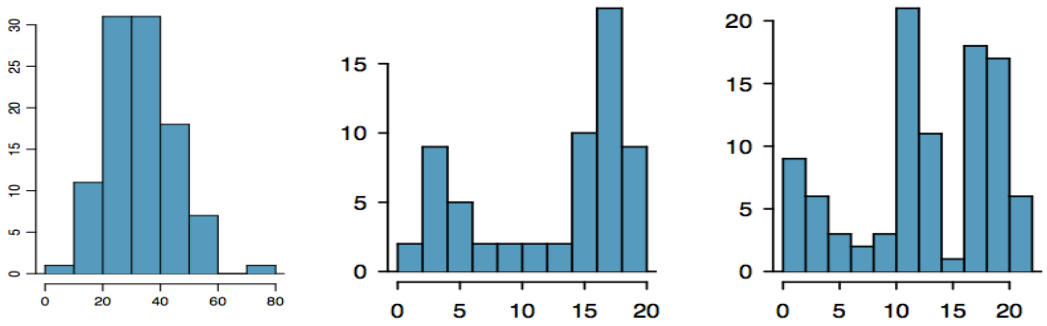
Once we have our histogram we can use it to determine the shape of the data or **distribution**. When describing distributions, we are going to look at three characteristics: modality, symmetry and skewness.

Modality

The **modality** of a distribution indicates the number of peaks or hills in its histogram.

- It is **unimodal** if it has one peak.
- It is **bimodal** if it has two peaks.
- It is **multimodal** if it has multiple peaks.

Example 11: The first graph is unimodal, the second is bimodal and the third is multimodal.

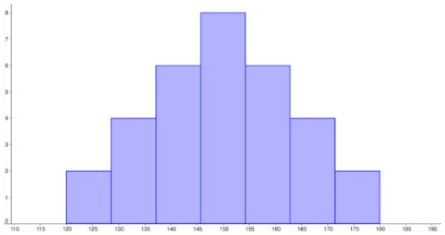


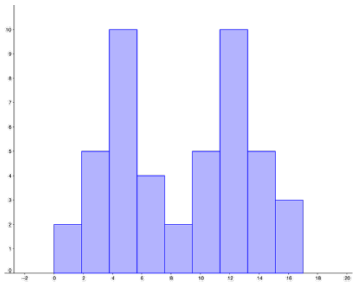
A bimodal distribution can result when two different populations have been grouped together and they are overlapping. It would be better to separate them into two separate graphs. For example, the grams of sugar per serving in sugar and non-sugar cereals.

Symmetry

A distribution is **symmetric** if the left side of the graph mirrors the right side.

> Example 12: The graph on the left is symmetric and unimodal while the graph on the right is roughly symmetric and bimodal.



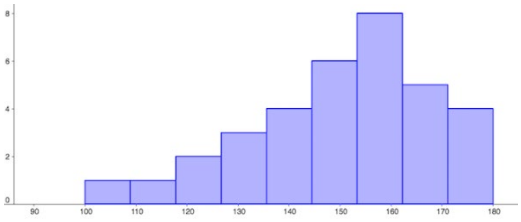


Skewness

If a distribution is not symmetric then we say it is skewed. A graph can be **skewed to the left** or **skewed to the right**. We say it is skewed in the direction of the longer tail.

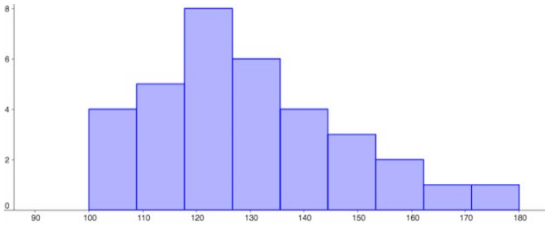
Skewed to the Left

A left skewed graph is also called a **negatively skewed** graph. The longer tail will be on the left or negative side.



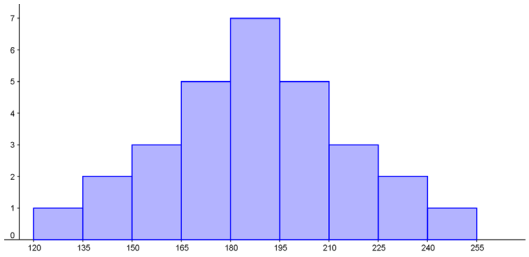
Skewed to the Right

A right skewed graph is also called a **positively skewed** graph. The longer tail will be on the right or positive side.



The Normal Distribution

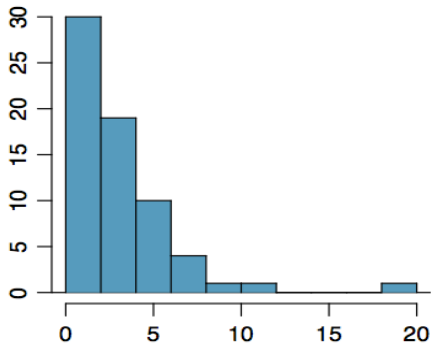
The **normal distribution** has a very specific shape. It is unimodal and symmetric with a bell-shaped graph.



Outliers

Outliers are data values that are unusually far away from the rest of the data. There is often a gap between the outlier and the rest of the graph. This visual determination of outliers is often subjective and depends on the situation.

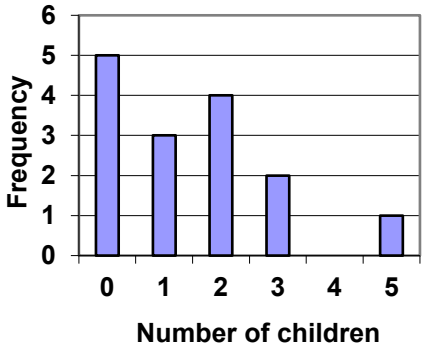
Example 13: In the graph to the right we have a unimodal distribution that is skewed to the right. There appears to be an outlier near 20.



Exercises 3.2

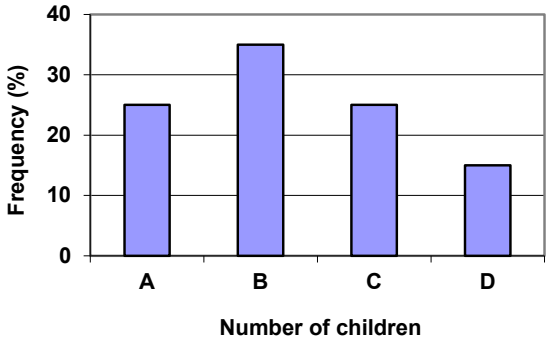
1. Is the data described categorial or quantitative?
   a. In a study, you ask the subjects their age in years.
   b. In a study, you ask the subjects their gender.
   c. In a study, you ask the subjects their ethnicity.
   d. The daily high temperature of a city over several weeks.
   e. A person's annual income.

2. A group of adults were asked how many children they have in their family. The bar graph to the right shows the number of adults who indicated each number of children.
   a. How many adults had 3 children?
   b. How many adults where questioned?
   c. What percentage of the adults questioned had 0 children?

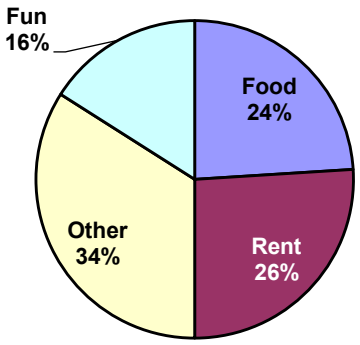


3. Jasmine was interested in how many days it would take a DVD order from Netflix to arrive at her door. The graph shows the data she collected.
   a. How many movies took 2 days to arrive?
   b. How many movies did she order in total?
   c. What percentage of the movies arrived in one day?

4. This relative frequency bar graph shows the percentage of students who received each letter grade on their last English paper. The class contains 20 students. What number of students earned an A on their paper?



5. Corey categorized his spending for this month into four categories: Rent, Food, Fun, and Other. The percentages he spent in each category are pictured here. If he spent a total of $2,600 this month, how much did he spend on rent?



6. In a survey[5], 1012 adults were asked whether they personally worried about a variety of environmental concerns. The number of people who indicated that they worried "a great deal" about some selected concerns is listed below.
   a. Is this categorical or quantitative data?
   b. Make a bar chart for this data.
   c. Why can't we make a pie chart for this data?

Environmental Issue	Frequency
Pollution of drinking water	597
Contamination of soil and water by toxic waste	526
Air pollution	455
Global warming	354

7. A group of adults were asked how many cars they had in their household.
   a. Is this categorical or quantitative data?
   b. Make a relative frequency table for the data.
   c. Make a bar chart for the data.
   d. Make a pie chart for the data.

1	4	2	2	1	2	3	3	1	4	2	2
1	2	1	3	2	2	1	2	1	1	1	2

[5] Gallup Poll. March 5-8, 2009. http://www.pollingreport.com/enviro.htm

8. The table below shows scores on a math test.
   a. Is this categorical or quantitative data?
   b. Make a relative frequency table for the data using a class width of 10.
   c. Construct a histogram of the data.

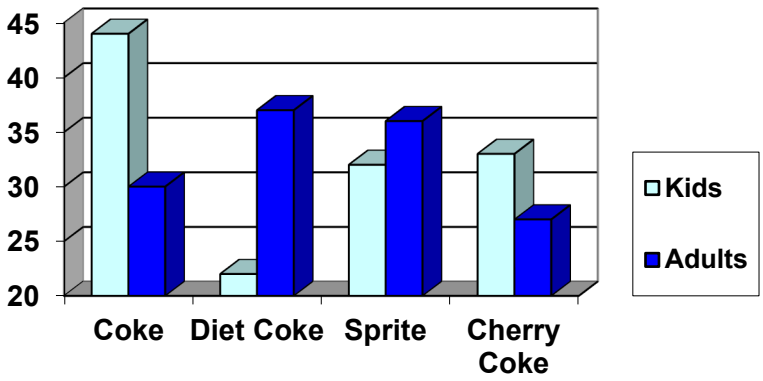
82	55	51	97	73	79	100	60	71	85	78	59
90	100	88	72	46	82	89	70	100	68	61	52

9. Studies are often done by pharmaceutical companies to determine the effectiveness of a treatment. Suppose that a new cancer drug is currently under study. Of interest is the average length of time in months patients live once starting the treatment. Two researchers each follow a different set of 40 cancer patients throughout their treatment. The following data (in months) are collected.
   a. Create a histogram for each dataset, using the same class intervals and scales so you can compare them.
   b. Compare and contrast the two distributions.

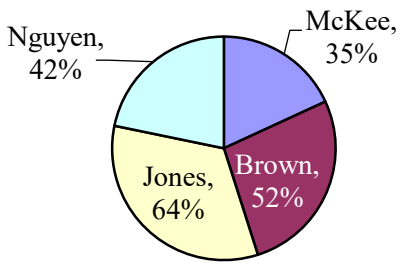
   Researcher 1: 3, 4, 11, 15, 16, 17, 22, 44, 37, 16, 14, 24, 25, 15, 26, 27, 33, 29, 35, 44, 13, 21, 22, 10, 12, 8, 40, 32, 26, 27, 31, 34, 29, 17, 8, 24, 18, 47, 33, 34

   Researcher 2: 3, 14, 11, 5, 16, 17, 28, 41, 31, 18, 14, 14, 26, 25, 21, 22, 31, 2, 35, 44, 23, 21, 21, 16, 12, 18, 41, 22, 16, 25, 33, 34, 29, 13, 18, 24, 23, 42, 33, 29

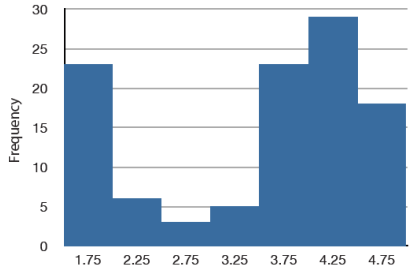
10. This graph shows the number of adults and kids who prefer each type of soda. There were 130 adults and kids surveyed. Discuss some ways in which the graph could be improved.



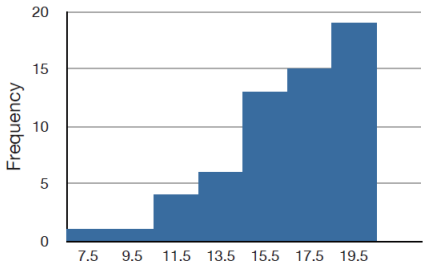
11. A poll was taken asking people if they agreed with the positions of the 4 candidates for a county office. Does this pie chart present a good representation of this data? Explain.



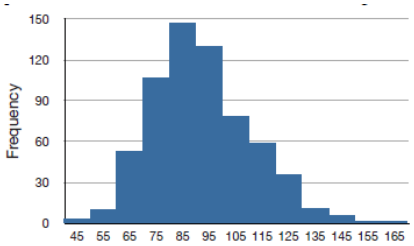
12. Match each description to one of the graphs.
    a. Normal distribution
    b. Positive or right skewed
    c. Negative or left skewed
    d. Bimodal



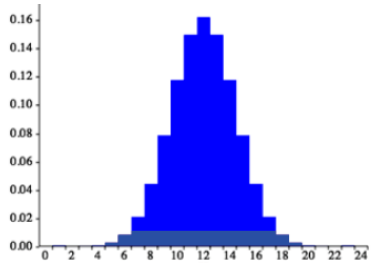
The frequency of times between eruptions of the Old Faithful geyser.



Scores on a 20-point statistics quiz.



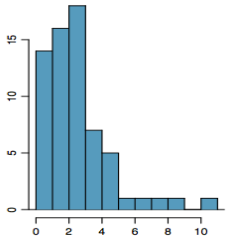
Distribution of scores on a psychology test.



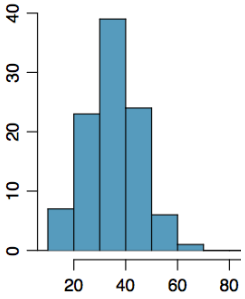
The number of heads in 24 sets of 100 coin flips.

13. Write a sentence or two to describe each distribution in terms of modality, symmetry, skewness and outliers.

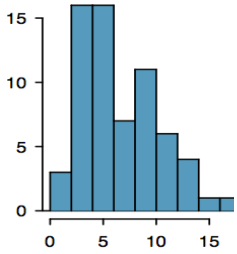
a.



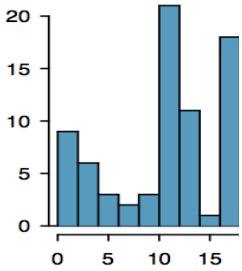
b.



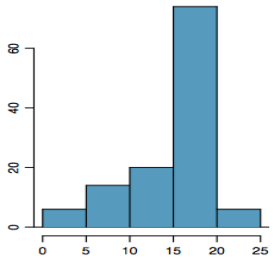
c.



d.



e.



Section 3.3 Summary Statistics: Measures of Center

Calculating Summary Statistics

In addition to graphical and verbal descriptions, we can use numbers to summarize quantitative distributions. We want to know what an "average" value is (where the data is centered), and how spread out the values are. Together, the center and spread provide important information which can be used estimate our population parameters. In this section we will talk about the measures of center and in the next section we will talk about the measures of spread.

Measures of Center

There are a few different types of "averages" that measure the center, and the one we use will depend on the shape of the distribution. We will mention the mode but focus mainly on the two most common "averages": the **mean** and the **median**.

Mode

In the previous section, we saw that the **modes** are related to the peaks where similar values are grouped. A mode is the value where a peak occurs. One way to calculate the mode(s) is to take the midpoint of each peak in the histogram.

Mean

The **mean**, or more formally the arithmetic mean, is what probably comes to mind when you hear the word average. The calculation of the mean uses every data value in the distribution and is therefore strongly affected by skew and outliers.

To calculate the mean of a distribution, we divide the sum of the data values by the number of data values we have. The sample mean is usually represented by $\bar{x}$, a lower-case x with a bar over it, read x-bar. The lower-case letter n is used to represent the number of data values or **sample size**.

Mean

$$\bar{x} = \frac{\text{sum of data values}}{n}$$

Example 1: Mirabel's exam scores for her last math class were: 79, 86, 82, 94. What is her mean test score?

To find the mean test score we need to find the sum of her test scores, then divide the sum by the number of test scores ($n = 4$). The mean is:

$$\bar{x} = \frac{79 + 86 + 82 + 94}{4}$$
$$= 85.25 \text{ points}$$

We will round the sample mean to one more decimal place than the original data. In this case, we would round 85.25 to 85.3 points. Also notice that the mean has the same units as the data and it is important to label it.

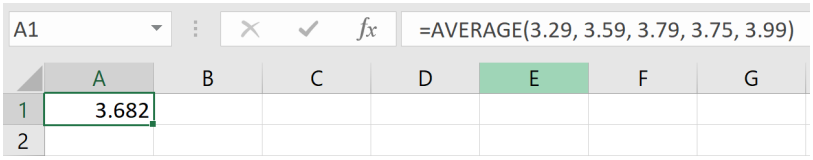
It is reasonable to calculate the mean by hand when the data set is small, but if the data set is large, or if you will be finding additional statistics, then technology is the way to go. We can find the mean of a data set using the spreadsheet formula =AVERAGE.

Example 2: The price of peanut butter at 5 stores was $3.29, $3.59, $3.79, $3.75, and $3.99. Find the mean price using a spreadsheet.

There are two ways to use the =AVERAGE formula. If your data set is not too large, you can enter each value directly into the formula. Using this method, we write

=AVERAGE(3.29 3.59, 3.79, 3.75, 3.99)

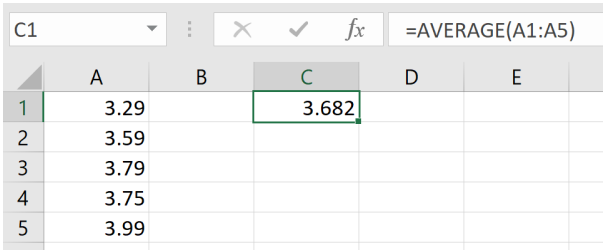
=$3.68



The other method is to enter the data values into a single column (or row) of the spreadsheet and reference the column (or row) range in the formula. We can enter the range by highlighting the data values. As illustrated below, if we enter the data into column A, the formula is

=AVERAGE(A1:A5)

=$3.68



Sometimes when there is a lot of data with repeated values we are given a frequency table.

Example 3: One hundred families from a particular neighborhood are randomly selected and asked to give their annual household income rounded to the nearest $5,000. The results are shown in the frequency table below.

Income (thousands of dollars)	Frequency
$15	6
$20	8
$25	11
$30	17
$35	19
$40	20
$45	12
$50	7

Calculating the mean by hand could get tedious if we try to type in all 100 values:

$$\bar{x} = \frac{\overbrace{15+\cdots+15}^{6 \text{ terms}} + \overbrace{20+\cdots+20}^{8 \text{ terms}} + \overbrace{25+\cdots+25}^{11 \text{ terms}} + \cdots}{100}$$

We could calculate this more easily by noticing that adding 15 to itself six times is the same as $(15)(6) = 90$. Using this simplification, we get

$$\bar{x} = \frac{(15)(6)+(20)(8)+(25)(11)+(30)(17)+(35)(19)+(40)(20)+(45)(12)+(50)(7)}{100}$$

$$= \frac{3390}{100}$$

$$= 33.9$$

The mean household income of our sample is 33.9 thousand dollars, or $33,900.

We could also use =AVERAGE to find the mean for this example, but it would require entering each repeated value individually. If the mean is all we need, then taking advantage of multiplication as repeated addition is the more straightforward way to go. We could also enter the frequency table and the calculation above in a spreadsheet.

Example 4: Extending the last example, suppose a new family moves into the neighborhood and has a household income of $5 million ($5000 thousand). Adding this to our sample, our mean becomes:

$$\bar{x} = \frac{(15)(6)+(20)(8)+(25)(11)+(30)(17)+(35)(19)+(40)(20)+(45)(12)+(50)(7)+(5000)(1)}{101}$$

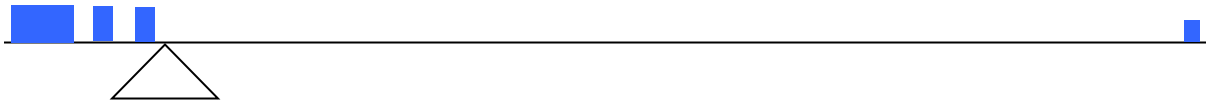
$$= \frac{8390}{101}$$

$$= 83.069$$

While 83.1 thousand dollars, or $83,100 is the correct mean household income for the new sample, it is no longer representative of the neighborhood – in fact, it is greater than every income in the sample aside from the new one we added!

Imagine the data values on a see-saw or balance scale. The mean is the value at the tip of the triangle that keeps the data in balance, like in the picture below.

If we graph our household data, the $5 million value is so far out to the right that the mean has to adjust to keep things in balance.



For this reason, when working with data that is skewed or has outliers, it is common to use a different measure of center, the median.

Median

The **median** of a data set is the "middle" value, when the data are listed in order from smallest to largest. We can also think of the median as the value that has 50% of the data below it and 50% of data above it. As we will discover later, this also makes the median what we call the **50th percentile**.

Median

If the number of data values is odd, then the median is the middle data value
If the number of data values is even, then the median is the mean of the middle pair

Example 5 (odd number of values): Find the median of these quiz scores: 5, 10, 8, 6, 4, 8, 2, 5, 7, 7, 6

We must start by listing the data in order: 2, 4, 5, 5, 6, 6, 7, 7, 8, 8, 10.

It is helpful to mark or cross off the numbers as you list them to make sure you don't miss any. Also, be sure to count the number of data values in your ordered list to make sure it matches the number of data values in the original list.

In this example there are 11 quiz scores. When the distribution contains an odd number of data values there will be a single number in the middle and that is the median. For small data sets, we can "walk" one value at a time from the ends of the ordered list towards the center to find the median

$$\overbrace{2,4,5,5,6}^{\text{Lower Half}}\ \overbrace{6}^{\text{Median}}\ \overbrace{7,7,8,8,10}^{\text{Upper Half}}$$

The median test score is 6 points.

Example 6 (even number of values): Now suppose we add another quiz score to the list. Suppose someone in the class got a perfect score of 20 on this very difficult quiz.

Then the ordered list would be: 2, 4, 5, 5, 6, 6, 7, 7, 8, 8, 10, 20.

There are now 12 quiz scores in our sample. When the distribution contains an even number of data values there will be a pair of values in the middle rather than a single value. Then we take the average of the middle two values.

$$\overbrace{2,4,5,5,6}^{\text{Lower Half}}\ \overbrace{6,7}^{\text{Middle Pair}}\ \overbrace{7,8,8,10,20}^{\text{Upper Half}}$$

$$\text{Median} = \frac{6+7}{2}$$
$$= 6.5 \text{ points}$$

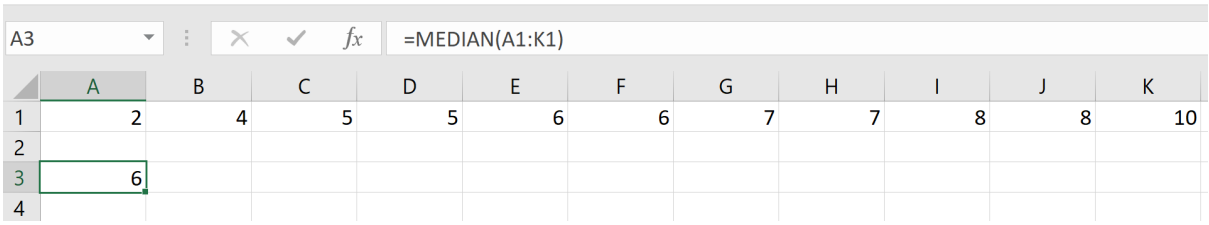
What is important to notice is that despite adding an outlier to our data set, the median is largely unaffected. The median quiz score for the new distribution is 6.5 points.

We can also find the median using the spreadsheet formula =MEDIAN. Just like the spreadsheet function =AVERAGE, we can either list the individual data values in the formula, or we can enter the data values into a row (or column) and use the row range (or column range) in the formula.

Using the data values of the original distribution, we can write function as
=MEDIAN(2, 4, 5, 6, 6, 7, 7, 8, 8, 10) or

=MEDIAN(A1:AK)

=6 points



Example 7: Let's continue with our peanut butter example and find the median both by hand and with a spreadsheet. The price of peanut butter at 5 stores was $3.29, $3.59, $3.79, $3.75, and $3.99.

To find the median by hand, we must first list the prices in order. This give us: $3.29, $3.59, $3.75, $3.79, $3.99

Since there are an odd number of data values in the sample ($n = 5$), we know that the median will be the single data value in the middle of the ordered list.

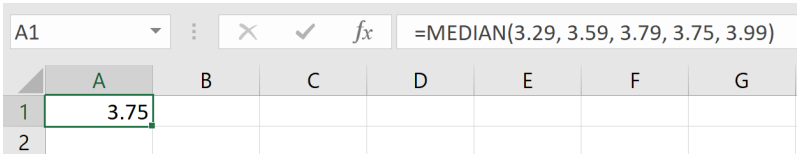
Lower Half: 3.29, 3.59 | Median: 3.75 | Upper Half: 3.79, 3.99

The median price of peanut butter at these five stores is $3.75.

Using a spreadsheet, we write

=MEDIAN(3.29, 3.59, 3.79, 3.75, 3.99)

=$3.75



It is worth noting that when you use a spreadsheet to find the median you do not have to order the data first. You can enter the data values in the order they are given to you.

The Relationship Between the Mean and the Median

If a distribution is skewed, the mean is pulled in the direction of the skew, as we saw in the see-saw diagram. In a right skewed distribution, the mean is greater than the median, while in a left skewed distribution, the mean is less than the median. If the distribution is symmetric, the mean and the median will be approximately equal.

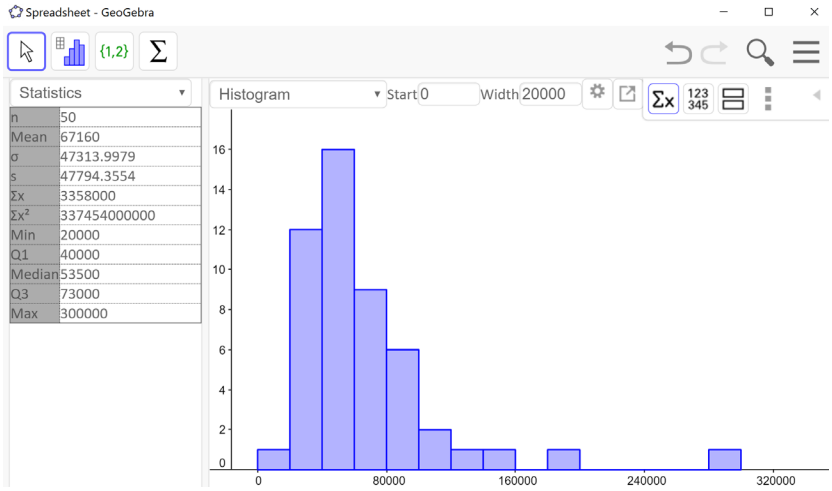
To demonstrate this, we have entered some data in GeoGebra, as previously explained, and made histograms. To see the statistics that GeoGebra calculates, we click on the summation symbol $\left(\sum x\right)$ on the right-hand menu bar.

Example 8: Fifty people from the Portland Metro area who are employed full time were sampled and their annual salaries were recorded (to the nearest thousand dollars). The histogram and summary statistics from GeoGebra are shown below.

From the histogram we can see that the shape of the distribution is unimodal and skewed to the right. We can see from the statistics output on the left that the mean is greater than the median. This is because the few people with higher incomes bring the average up.

$$\text{Mean} = \$67{,}160 \qquad \text{Median} = \$53{,}500$$

$$\text{Mean} > \text{Median}$$

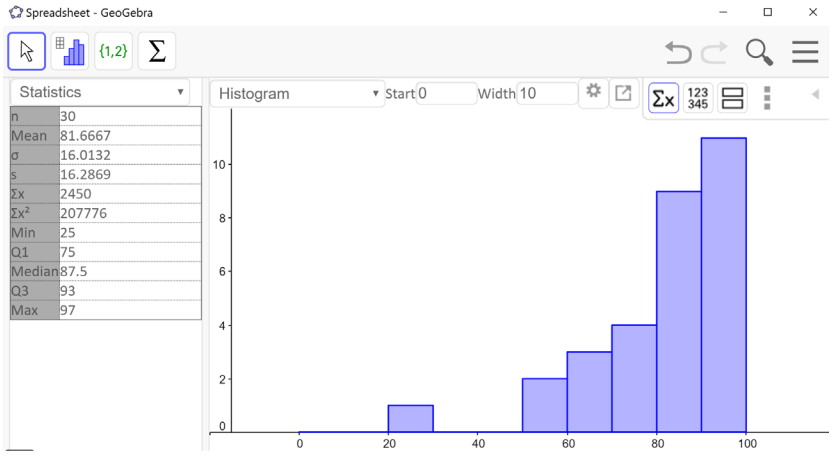


Example 9: A random selection of 30 math 105 exams at PCC were sampled and their scores were recorded. The histogram of the resulting distribution is shown below.

The shape of the distribution is unimodal and skewed to the left. There also appears to be an outlier between 20 and 30. We can see from the statistics output that the mean is less than the median. This is because the low test score brought the average down.

$$\text{Mean} = 81.7 \text{ points} \qquad \text{Median} = 87.5 \text{ points}$$

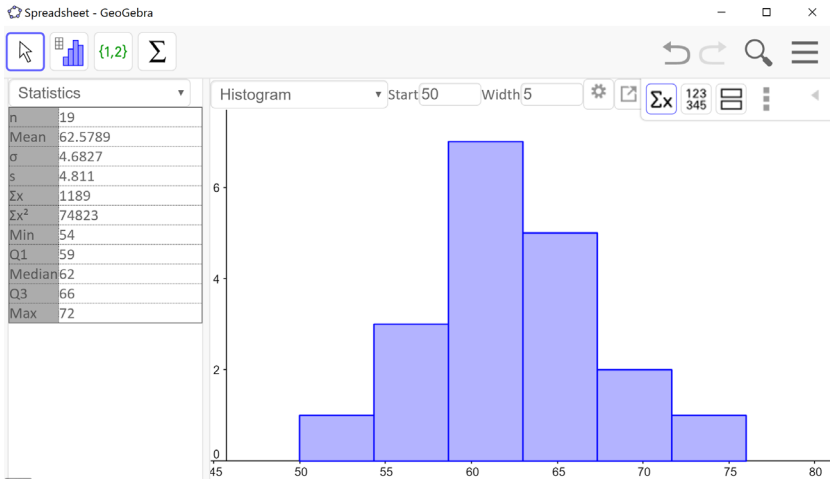
$$\text{Mean} < \text{Median}$$



Example 10: Nineteen people identifying as female were sampled and their heights (in inches) were recorded. The histogram of the resulting distribution is shown below.

The shape of the distribution is unimodal and roughly symmetric. We can also see from the statistics output that the mean and the median are approximately equal.

$$\text{Mean} = 62.6 \text{ inches} \qquad \text{Median} = 62 \text{ inches}$$
$$\text{Mean} \approx \text{Median}$$



We can use these observations in reverse as well. If we know the mean is greater than the median, then we can expect the distribution to be skewed to the right. If the mean is less than the median, then we can expect the distribution to be skewed to the left. When the mean and the median are approximately equal, the distribution is likely to be symmetric.

Example 11: Recent college graduates were asked how much student loan debt they have. The data has a mean of $46,265 and a median of $33,652. Just based on this information, do you expect the distribution to be symmetric, skewed to the left, or skewed to the right?

Since the mean is greater than the median, we can expect the distribution to be skewed to the right.

Exercises 3.3

1. A group of diners were asked how much they would pay for a meal. Their responses were: $7.50, $25.00, $10.00, $10.00, $7.50, $8.25, $9.00, $5.00, $15.00, $8.00, $7.25, $7.50, $8.00, $7.00. $12.00.
    a. Find the mean, including units.
    b. Find the median, including units.
    c. Based on the mean and the median, would you expect the distribution to be symmetric, skewed left, or skewed right? Explain.

2. You recorded the time in seconds it took for 8 participants to solve a puzzle. The times were: 15.2, 18.8, 19.3, 19.7, 20.2, 21.8, 22.1, 29.4.

a. Find the mean, including units.
b. Find the median, including units.
c. Based on the mean and the median, would you expect the distribution to be symmetric, skewed left, or skewed right? Explain.

3. Use the following table is the cost of purchasing a car at a local dealership. Some of the cars sold were new and some were used.
   a. Calculate find the mean, including units.
   b. Can you figure out how to find the median using the frequency table? See if you can do it without listing out all the data values.
   c. Based on the mean and the median, would you expect the distribution to be symmetric skewed left or skewed right? Explain.

Cost (Thousands of dollars)	Frequency
15	3
20	7
25	10
30	15
35	13
40	11
45	9
50	7

4. Studies are often done by pharmaceutical companies to determine the effectiveness of a treatment. Suppose that a new cancer drug is currently under study. Of interest is the average length of time in months patients live once starting the treatment. Two researchers each follow a different set of 40 cancer patients throughout their treatment. The following data (in months) are collected.
   a. Find the mean and median of each group.
   b. Compare and contrast the two groups.

   Researcher 1: 3, 4, 11, 15, 16, 17, 22, 44, 37, 16, 14, 24, 25, 15, 26, 27, 33, 29, 35, 44, 13, 21, 22, 10, 12, 8, 40, 32, 26, 27, 31, 34, 29, 17, 8, 24, 18, 47, 33, 34

   Researcher 2: 3, 14, 11, 5, 16, 17, 28, 41, 31, 18, 14, 14, 26, 25, 21, 22, 31, 2, 35, 44, 23, 21, 21, 16, 12, 18, 41, 22, 16, 25, 33, 34, 29, 13, 18, 24, 23, 42, 33, 29

5. An experiment compared the ability of three groups of participants to remember briefly-presented chess positions. The data are shown below. The numbers represent the average number of pieces correctly remembered from three chess positions.
   a. Make a histogram for each group.
   b. Find the mean of each group.

c. Find the median of each group.
d. Compare the shapes of the distributions as well as the centers of the three groups.

Non-players	Beginners	Tournament Players
22.1	32.5	40.1
22.3	37.1	45.6
26.2	39.1	51.2
29.6	40.5	56.4
31.7	45.5	58.1
33.5	51.3	71.1
38.9	52.6	74.9
39.7	55.7	75.9
43.2	55.9	80.3
43.2	57.7	85.3

6. There is evidence that smiling can attenuate judgments of possible wrongdoing. This phenomenon termed the "smile-leniency effect" was the focus of a study by Marianne LaFrance & Marvin Hecht in 1995[6]. The following data are measurements of how lenient the sentences were for three different types of smiles and one neutral control. A higher number indicates greater leniency. The same subject was used for all of the conditions so that may affect the results.
    a. Make a histogram for each smile type and the neutral control.
    b. Find the mean for each type of smile and the neutral control.
    c. Find the median for each type of smile and the neutral control.
    d. Compare the shapes of the distributions as well as the centers for each type of smile and control.

    (The data continues on the next page)

False Smile	Felt Smile	Miserable Smile	Neutral Control
2.5	7	5.5	2
5.5	3	4	4
6.5	6	4	4
3.5	4.5	5	3
3	3.5	6	6
3.5	4	3.5	4.5
6	3	3.5	2
5	3	3.5	6
4	3.5	4	3
4.5	4.5	5.5	3

[6] LaFrance, M., & Hecht, M. A. (1995) Why smiles generate leniency. Personality and Social Psychology Bulletin, 21, 207-214. Adapted from www.onlinestatbook.com, by David M. Lane, et al, used under CC-BY-SA 3.0.

5	7	5.5	4.5
5.5	5	4.5	8
3.5	5	2.5	4
6	7.5	5.5	5
6.5	2.5	4.5	3.5
3	5	3	4.5
8	5.5	3.5	6.5
6.5	5.5	8	3.5
8	5	5	4.5
6	4	7.5	4.5
6	5	8	2.5
3	6.5	4	2.5
7	6.5	5.5	4.5
8	7	6.5	2.5
4	3.5	5	6
3	5	4	6
2.5	3.5	3	2
8	9	5	4
4.5	2.5	4	5.5
5.5	8.5	4	4
7.5	3.5	6	2.5
6	4.5	8	2.5
9	3.5	4.5	3
6.5	4.5	5.5	6.5

7. Make up three data sets with 5 values each that have:
    a. The same mean but different medians.
    b. The same median but different means.

8. The frequency table below shows the number of women's shoes that were sold in an hour at a local shoe store.
    a. Would you treat this data as categorical or quantitative?
    b. How would the bar graph be different from the histogram?
    c. Treat the data as quantitative and find the mean and the median. Are these useful statistics?

Non-players	Frequency
5	4
6	4
7	6
8	6
9	5

Section 3.4 Summary Statistics: Measures of Variation

Measures of Variation

Consider these three sets of quiz scores for a 10-point quiz:

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 data sets have a mean of 5 points and median of 5 points, 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. Section C was not as consistent as section A, but not as widely varied as section B.

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

There are several ways to measure the variation of a distribution. In this section we will look at the **standard deviation**, **range** and the **interquartile range (IQR)**.

Standard Deviation

The **sample standard deviation, s,** is a measure of variation that tells us how far, on average, the data values deviate, or are different from, the mean. The mean and standard deviation are paired to provide a measure of center and spread for symmetric distributions.

Sample Standard Deviation

$$s = \sqrt{\frac{\text{Sum of the squared deviations from the mean}}{n-1}}$$

where n is the sample size, or the number of data values

We will go through the whole process for calculating the standard deviation. Let's say there is another section of quiz scores:

Section D: 0, 5, 5, 5, 5, 5, 5, 5, 5, 10

The mean quiz score, like Sections A, B and C, is 5 points.

The first step in finding the standard deviation is to find the deviation, or difference, of each data value from the mean. We will do this in a table. You could also use a spreadsheet to do these calculations.

Data Value	Deviation: (Data Value – Mean)
0	$0 - 5 = -5$
5	$5 - 5 = 0$
5	$5 - 5 = 0$
5	$5 - 5 = 0$
5	$5 - 5 = 0$
5	$5 - 5 = 0$
5	$5 - 5 = 0$
5	$5 - 5 = 0$
5	$5 - 5 = 0$
10	$10 - 5 = 5$

We would like to get an idea of the "average" deviation from the mean, but if we find the average of the values in the second column, the negative and positive values cancel each other out (this will always happen), so to prevent this we square the deviations.

Data Value	Deviation: (Data Value –Mean)	Deviation Squared
0	$0 - 5 = -5$	$(-5)^2 = 25$
5	$5 - 5 = 0$	$0^2 = 0$
5	$5 - 5 = 0$	$0^2 = 0$
5	$5 - 5 = 0$	$0^2 = 0$
5	$5 - 5 = 0$	$0^2 = 0$
5	$5 - 5 = 0$	$0^2 = 0$
5	$5 - 5 = 0$	$0^2 = 0$
5	$5 - 5 = 0$	$0^2 = 0$
5	$5 - 5 = 0$	$0^2 = 0$
10	$10 - 5 = 5$	$5^2 = 25$

Next, we add the squared deviations and we get

$25 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 25 = 50$.

Ordinarily, we would then divide by the number of scores, *n*, (in this case, 10) to find the mean of the deviations, but the division by *n* is only done if the data set represents a population. When the data set represents a sample (as it almost always does), we instead divide by $n - 1$ (in this case, $10 - 1 = 9$).

We assume Section D represents a sample, so we will divide by 9. Note that our units are now points-squared since we squared all of the deviations. It is much more meaningful to use the units we started with, so to convert back to points we take the square root.

The sample standard deviation is

$$s = \sqrt{\frac{50}{9}}$$
$$\approx 2.2 \text{ points}$$

As with the mean, we will round answers to one more decimal place than the original data. This tells us that on average, scores were 2.2 points away from the mean of 5 points. In summary, here are the steps to calculate the standard deviation by hand.

Calculating the Sample Standard Deviation

1. Find the deviations by subtracting the mean from each data value
2. Square each deviation
3. Add the squared deviations
4. Compute the square root of the sum divided by $n-1$:

$$s = \sqrt{\frac{\text{Sum of the squared deviations from the mean}}{n-1}}$$

There are a few important characteristics we want to keep in mind when finding and interpreting the standard deviation.

- The standard deviation is never negative. It will be zero if all the data values are equal and get larger as the data spreads out.
- The standard deviation has the same units as the original data and it is important to label it.
- The standard deviation, like the mean, can be highly influenced by outliers.

Example 1: To continue our peanut butter example, we will find the standard deviation of this sample: \$3.29, \$3.59, \$3.79, \$3.75, and \$3.99.

The first thing we need to find is the sample mean, and we know it is \$3.68 from our previous work.

Next, we need to find the deviation from the mean for each data value and square it.

Data Value	Deviation	Deviation Squared
\$3.29	3.29 – 3.68 = -0.39	$(-0.39)^2 = 0.1521$
\$3.59	3.59 – 3.68 = -0.09	$(-0.09)^2 = 0.0081$
\$3.79	3.79 – 3.68 = 0.11	$(0.11)^2 = 0.0121$
\$3.75	3.75 – 3.68 = 0.07	$(0.07)^2 = 0.0049$
\$3.99	3.99 – 0.68 = 0.31	$(0.31)^2 = 0.0961$

The sum of the deviations squared is

$0.1521 + 0.0081 + 0.0121 + 0.0049 + 0.0961 = 0.2733$ dollars-squared.

The sample standard deviation is

$$s = \sqrt{\frac{0.2733}{4}}$$
$$\approx \$0.2613$$

Since the units are dollars, we will round to two decimal places rather than one more than the data. This gives us a standard deviation of $0.26. Together with the mean this tells us that on average, the cost of a jar of peanut butter is $0.26 away from the mean of $3.68.

Calculating the standard deviation by hand can be quite a nuisance when we are dealing with a large data set, so we can also use technology. We use the spreadsheet function =STDEV.S to find the *sample* standard deviation. Notice that this is different from the population standard deviation, which uses the function =STDEV.P.

Just like the spreadsheet functions =AVERAGE and =MEDIAN, we can either list the individual data values in the formula, or we can enter the data values into a row or column and use the row or column range in the formula.

Example 2: The total cost of textbooks for the term was collected from 36 students. Use a spreadsheet to find the mean, median, and standard deviation of the sample.

$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

Since we are finding more than one statistic for this data set, it is much more efficient to enter the data values into a row or column and reference the range in each of the formulas. Entering the data in column A, the formulas are:

Mean: =AVERAGE(A1:A36)

= $299.58

Median: =MEDIAN(A1:A36)

= $307.50

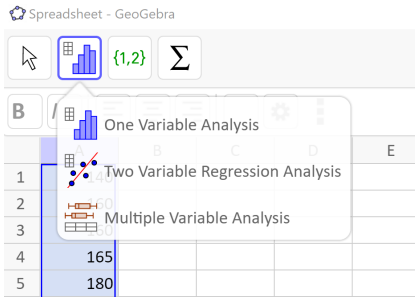
Standard Deviation: =STDEV.S(A1:A36)

= $78.68

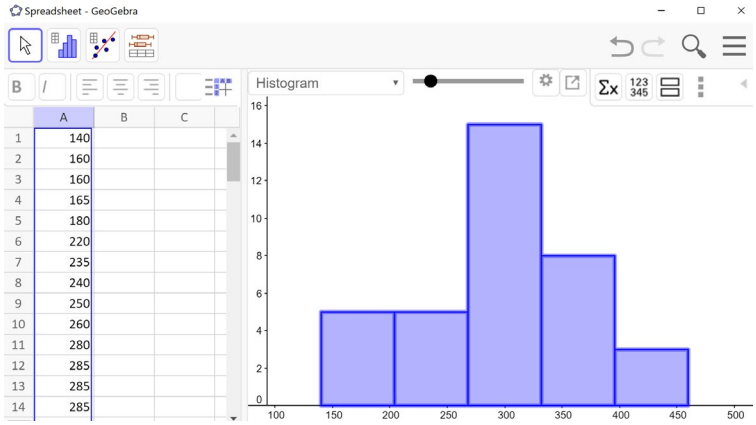
The mean and the median are relatively close to each other, so we can expect the distribution to be approximately symmetric with maybe a slight skew to the left, since the mean is smaller. The mean and standard deviation together tell us that the average cost of textbooks for a term is about $299.58, give or take $78.68.

In addition to a spreadsheet, we will continue our use of GeoGebra. Let's take a look at how to use GeoGeogebra to find the mean, median and standard deviation for the last example. We begin just like we did for making a histogram.

Example 2 Continued: We enter the textbook data into column A of the spreadsheet in GeoGebra. (Main Menu →View →Spreadsheet). Next, select the column title of your data, click on the histogram in the menu bar on the left, and select One Variable Analysis.

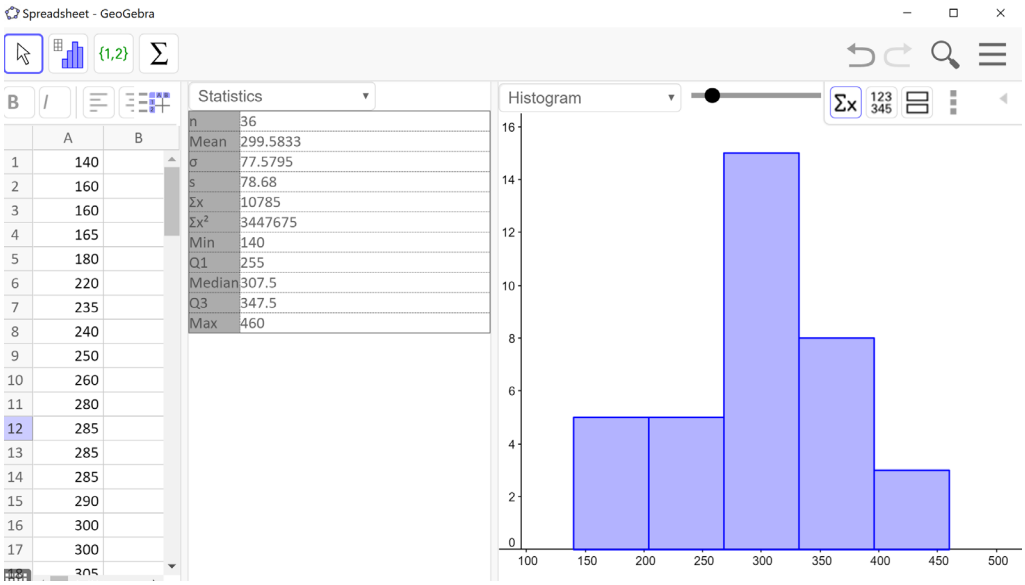


Then you will see the histogram. It is always a good idea to check the shape of your distribution before calculating anything. Notice our histogram matches what we found about the distribution from the mean and median. It is approximately symmetric or slightly skewed to the left.



Next, we click the summation symbol $\left(\sum x\right)$ in the menu bar on the right. The list of summary statistics will pop up as you can see in the image below. We see that the mean is \$299.58, the median is \$307.50 and the sample standard deviation is \$78.68 – just like we found using the spreadsheet formulas.

The statistics we will use are the sample size (n), the mean, median, and the sample standard deviation (s). The last five entries in the table – min, Q1, median, Q3, and max – together make up the 5-number summary which we will learn about shortly!



The standard deviation is the measure of variation that we pair with the mean for approximately symmetric distributions. This pairing should make sense because the standard deviation uses the mean in its calculation. But what about the median? What measure of variation do we pair with it?

Range

One candidate is the **range**. The range tells us the spread or width of the entire data set. We calculate the range as the difference between the maximum and minimum value.

Range

Range = Max – Min

However, the range is not a very good measure of variation since it is very strongly affected by skew and outliers. Consider, for example, the distribution of full time salaries in the United States. Many people earn a minimum wage salary, while others like Jeff Bezos (Amazon) and Bill Gates (Microsoft) earn millions (if not billions!). A range this large does very little to help us get a sense of the spread where most of the data values lie.

Quartiles and the Interquartile Range

Instead, the measure of variation that we pair with the median is the **interquartile range (IQR)**. The IQR tells us the width of the middle 50% of data values. By cutting off the lower and upper 25% of data values, we are able to ignore extreme values and provide a more accurate sense of how spread out the distribution is.

The IQR is calculated as the difference between the third quartile (Q_3) and the first quartile (Q_1). Before we can calculate the interquartile range, though, we need to learn how to find the first and third quartiles.

Interquartile Range (IQR)

$$\text{IQR} = Q_3 - Q_1$$

As the name implies, **quartiles** are values that divide the data into quarters. The **first quartile (Q_1)** is the value that 25% of the data lie below. The **third quartile (Q_3)** is the value that 75% of the data lie below. As you might have guessed, the second quartile is the same as the median since 50% of the data values lie below it.

We have seen that the data is split in half by the median so to split it into quarters, we find the median of each half of the data.

Quartiles

Q_1 the median of the lower half of the data
Q_2 is the median of the whole data set
Q_3 is the median of the upper half of the data

If there is an odd number of data values, we don't use the median in either half

Example 3 (even number of data values): Suppose we have measured the height, in inches, of 12 people who identify as female. The data values are listed below. Find the interquartile range.

59 60 69 64 70 72 66 64 67 66 63 61

Just like when finding the median, we must first order the data.

59 60 61 63 64 64 66 66 67 69 70 72

Then we divide the data into two halves. In the case of an even sample size, we split the distribution down the middle. The first 6 data values are the lower half and the next 6 data values are the upper half. Then we find the median of each. The median of the lower half is Q_1 and the median of the upper half is Q_3.

$$\overbrace{59, 60, \underbrace{61, 63}_{Q_1=\frac{61+63}{2}=62}, 64, 64}^{\text{Lower Half}} \overbrace{66, 66, \underbrace{67, 69}_{Q_3=\frac{67+69}{2}=68}, 70, 72}^{\text{Upper Half}}$$

In this data, $Q_1 = 62$ inches and $Q_3 = 68$ inches. Then we subtract to find the IQR.

$$\begin{aligned}\text{IQR} &= Q_3 - Q_1 \\ &= 68 - 62 \\ &= 6 \text{ inches}\end{aligned}$$

This tells us the that the middle 50% of the women's heights lie within an interval of 6 inches.

Example 4 (odd number of data values): Suppose we added one more height (68 inches) to the data set from the previous example. We will again find the interquartile range of the heights.

59 60 61 63 64 64 66 66 67 68 69 70 72

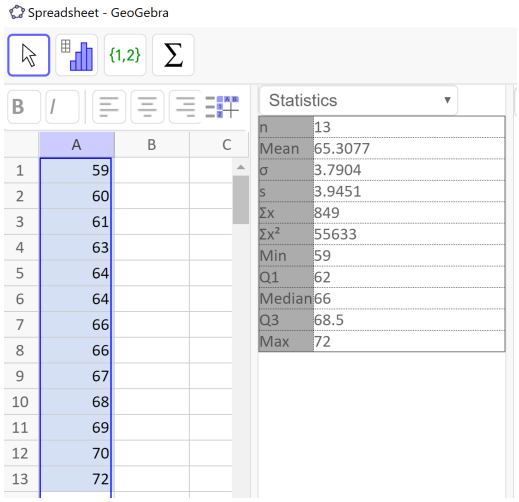
The data are already in order, but we have an odd number of values. To deal with this we do not use the median in the upper or lower halves. The lower half will include the values strictly below the median, and the upper half will include the values strictly above the median.

Lower Half: $59, 60, \underbrace{61, 63}_{Q_1=\frac{61+63}{2}=62}, 64, 64$ Median (Ignore): 66 Upper Half: $66, 67, \underbrace{68, 69}_{Q_3=\frac{68+69}{2}=68.5}, 70, 72$

Then the interquartile range is:

$$\begin{aligned} \text{IQR} &= Q_3 - Q_1 \\ &= 68.5 - 62 \\ &= 6.5 \text{ inches} \end{aligned}$$

If the data set is small, we can find the first and third quartiles by hand, but we have also seen that they are part of the output from GeoGebra. Here is the output for this data.



From the list of summary statistics we can see that $Q_1 = 62$ and $Q_3 = 68.5$ inches. Now we can calculate the interquartile range.

$$\begin{aligned} \text{IQR} &= 68.5 - 62 \\ &= 6.5 \text{ inches} \end{aligned}$$

This is the same value we found by hand.

It is important to note that we are not using spreadsheets for the five-number summary because they do not calculate the quartiles in the same way, so they will not give the same results. Now that we have learned how to find the quartiles we can make a five-number summary and boxplot.

The Five-Number Summary and Boxplots

The **five-number summary** is made up of the minimum, Q_1, median, Q_3, and the maximum. These five values divide the data into quarters. A **boxplot**, also called a **box-and-whisker plot**, is a graphical representation of the five-number summary. Each region of the boxplot contains approximately the same number of data values, so we can see the spread for each region. We can find the five-number summary and draw a boxplot by hand or by using GeoGebra. In our last example the five-number summary from GeoGebra is: 59, 62, 66, 68.5, 72 inches.

Five-Number Summary

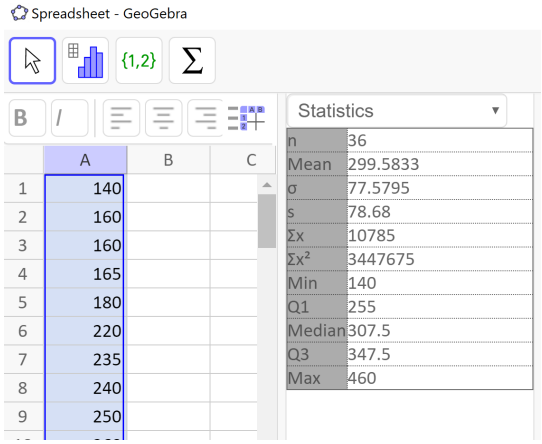
Minimum, Q_1, Median, Q_3, Maximum

We will use GeoGebra to find the five-number summary for the next example and then explain how to draw a boxplot.

Example 5: Let's continue with the cost of textbook data from Example 2. Use GeoGebra to find the five-number summary for this sample and draw a boxplot by hand.

$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

As we found before, here is the GeoGebra output. The last five entries of the summary statistics are the five-number summary. Remember to label all of your statistics with units.

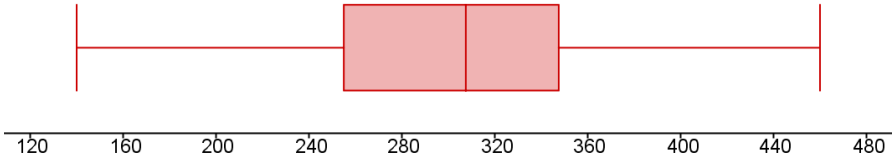


The five-number summary is

Min	Q_1	Median	Q_3	Max
\$140	\$255	\$307.50	\$347.50	\$460

To draw the boxplot, we will first draw a number line that extends a little beyond the minimum and maximum values, and choose a scale. We decided to draw our number line from \$120 to \$480, in increments of \$40. Then we add a meaningful title and units.

Next, make vertical lines at the first quartile, median and third quartile and connect them to form a box. This is the middle 50% of the data and you might notice that the width of the box is the IQR. Then, extend the "whiskers" out to the minimum and maximum values. Note that a boxplot does not have a vertical scale and the height of the box does not matter. Our boxplot looks like this:



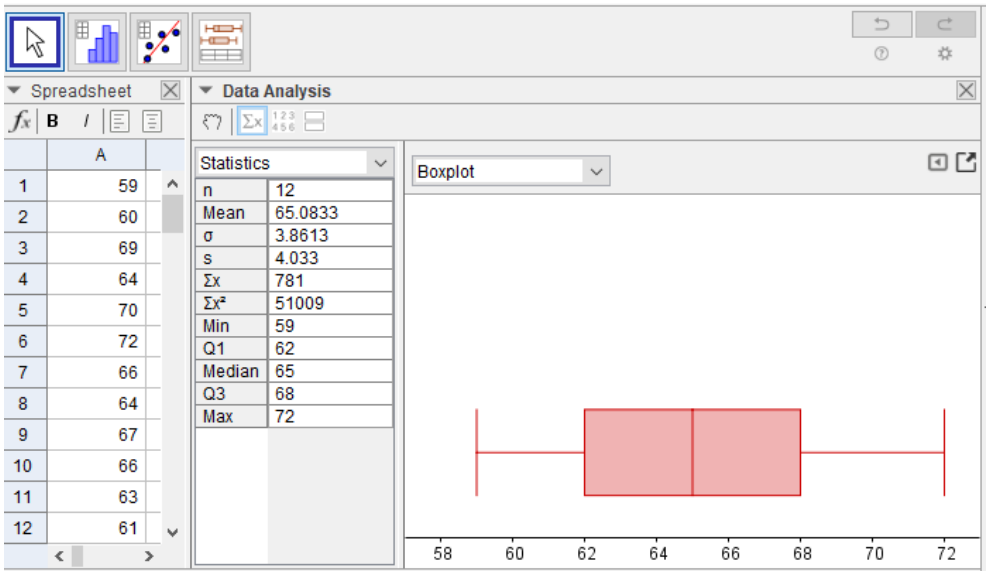
Cost of Textbooks for One Term in Dollars, for a Sample of 36 Students

GeoGebra will also draw boxplots for us. We enter and select the data values like we have done before and select One Variable Statistics. This brings up the graphics window with a histogram by default. Use the drop-down menu to select the boxplot. We also click on $\sum x$ to show the summary statistics.

Example 6: We will continue with our height data from the 12 people who identify as women. Find the five-number summary and create a boxplot using GeoGebra.

59 60 69 64 70 72 66 64 67 66 63 61

Following the steps above we have the following GeoGebra output. The last five entries in the statistics table are the five-number summary and we have the boxplot on the right.

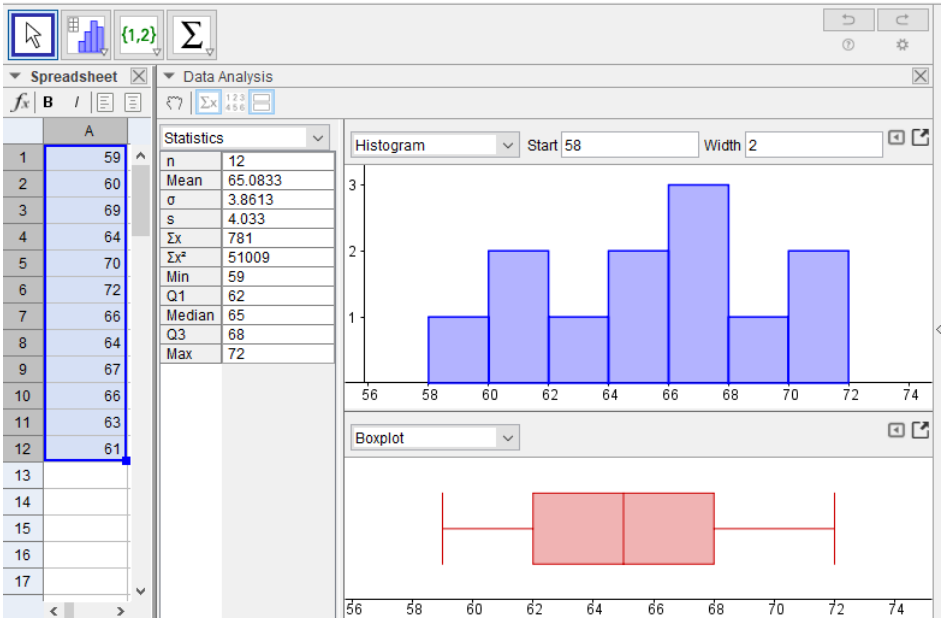


Here is the five-number summary:

Min	Q_1	Median	Q_3	Max
59 in	62 in	65 in	68 in	72 in

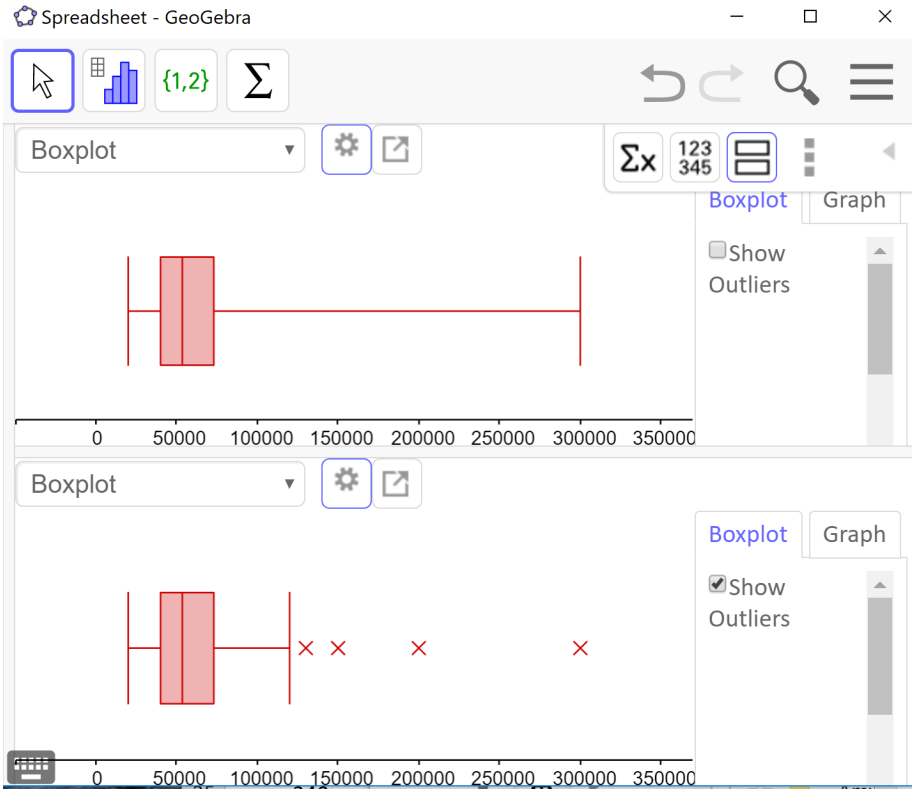
For this data, the two sides of the box and the two whiskers are approximately the same width. This suggests that the distribution is symmetric. We can verify this by noticing that the mean is approximately equal to the median.

The boxplot can tell us the shape of the distribution, but we cannot tell how many peaks the data has. For that we need a histogram. We can see the histogram and boxplot together by selecting the icon that looks like two rectangles stacked, or an = sign.



Now we have a full picture of this data.

The default boxplot in GeoGebra is called a **modified box plot**, which shows the data values that are outliers with an X but requires a few more steps to make by hand. To change from the modified box plot to a regular box plot, click on the left pointing arrow in the boxplot window (downloaded version) or the settings wheel (online version) for options, and uncheck "show outliers." The output window below shows two side-by-side boxplots (the regular boxplot on top and the modified boxplot on the bottom) illustrating the distribution of the annual salaries for 50 randomly selected full-time workers in the Portland Metro area.



From the upper boxplot we can see that this distribution is skewed to the right and the upper quarter of the data is very spread out. It is natural to think of the data values as being evenly spread out in each region, but that is quite often not the case. From the lower boxplot we can see that there are 4 data values that are considered outliers and how far away the last data value is from the others. This is why it is useful to show outliers on a boxplot.

Modified Boxplot (Optional)

If you are curious to know how a modified boxplot is made, we will explain it briefly. There is a rule called the **1.5*IQR** rule to determine which points are considered outliers. An outlier is a point that is more than 1.5 IQRs away from the middle 50% of the data (the box in the boxplot). We know how to calculate the IQR and then we multiply that by 1.5. We subtract that from Q_1 to find the **lower fence** and add that value to Q_3 to find the **upper fence**. Values beyond the fences are considered outliers and are

drawn with an X or a star. Then we draw the whiskers of a modified box plot to the furthest data value inside the fence on each side.

Percentiles

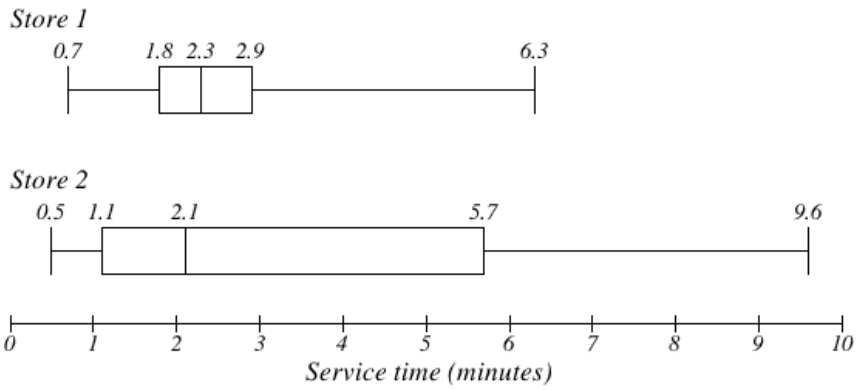
Back when we were finding the median, we mentioned that the median is also called the 50th percentile, because 50% of the data values lie below it. We can define any **percentile** as the data value with that percentage of values below it. Since we have found the quartiles, we can also identify the 25th and 75th percentiles for our data. Q_1 is the 25th percentile because 25% of the data values lie below it and Q_3 is the 75th percentile because 75% of the data values lie below it.

Percentiles are used when comparing the growth of children to the population and in the results of standardized tests, such as the SAT test. If a person scored in the 83rd percentile, that means they scored higher than 83% of the people who took the test.

Comparing Distributions

Box plots and percentiles are particularly useful for comparing data from two populations.

Example 7: The box plots of service times for two fast-food restaurants are shown below. Compare the length of time to get served at the two restaurants. Which one should you go to if you are in a hurry?



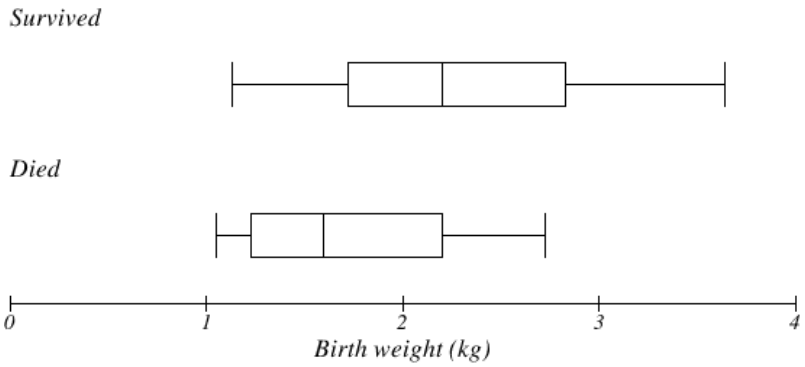
Store 2 has a slightly shorter median service time (2.1 minutes vs. 2.3 minutes), but the service times are less consistent, with a wider spread of the data.

The 75th percentiles are 2.9 and 5.7 minutes. That means at store 1, 75% of customers were served within 2.9 minutes, while at store 2, 75% of customers were served within 5.7 minutes.

Which store should you go to in a hurry? That depends upon your opinion about luck – 25% of customers at store 2 had to wait between 5.7 and 9.6 minutes.

Example 8: The boxplots below show the 5-number summaries of the birth weights, in kilograms, of infants with severe idiopathic respiratory distress

syndrome (SIRDS)[7]. The boxplots are separated to show the birth weights of infants who survived and those that did not. What can we conclude from this data?



Comparing the two groups, the boxplot reveals that the birth weights of the infants that died appear to be, overall, smaller than the weights of infants that survived. In fact, we can see that the median birth weight of infants that survived is about the same as the third quartile of the infants that died.

Similarly, we can see that the 25th percentile of the survivors is larger than the 50th percentile of those that died, meaning that over 75% of the survivors had a birth weight larger than the median birth weight of those that died.

Looking at the maximum value for those that died and the third quartile of the survivors, we can see that over 25% of the survivors had birth weights higher than the heaviest infant that died.

The box plots give us a quick, though informal, way to determine that birth weight is quite likely linked to the survival of infants with SIRDS.

Z-Scores

Have you ever heard the saying that you can't compare apples and oranges? It turns out that you can - provided we standardize their measures first!

We will be using the standard score called a Z-score, which is a method commonly used with unimodal and symmetric distributions (called **normal** or **nearly normal distributions**). Z-scores may be used with any data, but if the distribution is skewed, then the distribution of Z-scores will also be skewed.

To calculate the Z-score for a data value, we find out how far away from the mean it is by subtracting. Then we divide by the standard deviation to see how many standard deviations that is. Thus, the **Z-score** of a data value is the number of standard deviations it is away from the mean.

[7] van Vliet, P.K. and Gupta, J.M. (1973) Sodium bicarbonate in idiopathic respiratory distress syndrome. *Arch. Disease in Childhood*, 48, 249–255. As quoted on http://openlearn.open.ac.uk/mod/oucontent/view.php?id=398296&section=1.1.3

Z-score

$$Z = \frac{\text{data value - mean}}{\text{standard deviation}}$$

Be sure to calculate the difference first, then divide

If a data value is above the mean, its Z-score will be positive. If a data value is below the mean, its Z-score will be negative. Therefore, if a data value is one standard deviation above the mean, its Z-score is +1. If it is 2.5 standard deviations below the mean, its Z-score is -2.5. Note that the units of Z-scores are standard deviations, not the units of the data values.

We can use Z-scores to determine the relative unusualness of a data value with respect to its own distribution. That is what allows us to compare two unlike items. The convention in statistics is to say that a data value is **unusual** if it is more than 2 standard deviations from the mean, or in other words, if its Z-score is less than -2 or greater than +2.

Example 9: The oranges at a local grocery store have a mean diameter of 5.8 inches and a standard deviation of 1.2 inches. The apples, on the other hand, have a mean diameter of 4.2 inches and a standard deviation of 0.6 inches.

Ali closes their eyes and selects and apple and an orange. When they look at both pieces of fruit, they seem small. If the orange has a diameter of 4.2 inches and the apple has a diameter of 3.5 inches, which is smaller relative to their respective piles of fruit?

To determine which fruit is relatively smaller, Ali can find each of their Z-scores.

$$Z_{\text{Orange}} = \frac{4.2 - 5.8}{1.2} \qquad Z_{\text{Apple}} = \frac{3.5 - 4.2}{0.6}$$
$$= -1.33 \text{ standard deviations} \qquad = -1.17 \text{ standard deviations}$$

By convention, Z-scores are rounded to two decimal places, so we see that the orange is 1.33 standard deviations below its mean and the apple is 1.17 standard deviation below its mean. The orange is therefore smaller relative to its distribution since its Z-score is less than the apple's Z-score.

We can also see from the Z-scores that neither fruit has an unusually small diameter since each piece of fruit is less than 2 standard deviations from its mean.

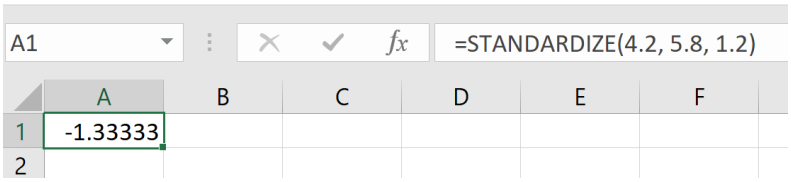
We can also find Z-scores using a spreadsheet with this formula:

=STANDARDIZE(data value, mean, standard deviation)

To verify our apple and orange Z-scores, we would write:

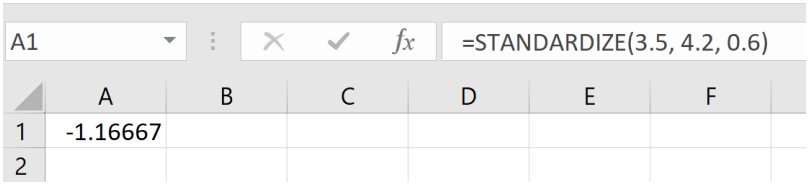
Apple: =STANDARDIZE(4.2, 5.8, 1.2)

= -1.33 standard deviations



Orange: =STANDARDIZE (3.5, 4.2, 0.6)

= -1.17 standard deviations



Example 10: The mean weight of men over the age of 20 is 195.7 pounds[8] with a standard deviation of 29.8 pounds. The mean weight of domestic cats is 8.6 pounds with a standard deviation of 1.2 pounds. (The standard deviation for men's weights is estimated. The cat's mean weight is based on ideal cat weight and the standard deviation is approximate).

At his peak, Andre the Giant, the 7-foot-4-inch French professional wrestler and actor, weighed 520 pounds. When Georgie the cat was at his peak he weighed 24 pounds. Who was more giant – Andre the Giant or Georgie the cat?

Since the weights of cats and men cannot be compared directly, we will need to calculate the Z-scores.

$$Z_{\text{Andre}} = \frac{520 - 195.7}{29.8} = 10.88 \text{ standard deviations}$$

$$Z_{\text{Georgie}} = \frac{24 - 8.6}{1.2} = 12.83 \text{ standard deviations}$$

Using the standardize function, we would write:

Andre: =STANDARDIZE (520, 197.5, 29.8)

= 10.88 standard deviations

Georgie: =STANDARDIZE (24, 8.6, 1.2)

= 12.83 standard deviations

[8] 2016 CDC Report. https://www.cdc.gov/nchs/data/series/sr_03/sr03_039.pdf. The study included all ethnicities but the report does not say whether transgendered men were included.

Since both Z-scores are greater than 2 standard deviations, both weights are extremely unusual. However, since the Z-score for Georgie's weight is larger, he is even more giant than Andre the Giant.

Exercises 3.4

Many of the datasets from Exercises 3.3 are repeated here so you can use your previous work to help you.

1. A group of diners were asked how much they would pay for a meal. Their responses were: $7.50, $25.00, $10.00, $10.00, $7.50, $8.25, $9.00, $5.00, $15.00, $8.00, $7.25, $7.50, $8.00, $7.00. $12.00.
    a. Using your mean from section 3.3, find the standard deviation of this data. Explain what the mean and standard deviation tell you about how much the group of diners would pay for a meal.
    b. Calculate the five-number summary for this data.
    c. Calculate the range and IQR for this data.
    d. Create a boxplot for the data.

2. You recorded the time in seconds it took for 8 participants to solve a puzzle. The times were: 15.2, 18.8, 19.3, 19.7, 20.2, 21.8, 22.1, 29.4.
    a. Using your mean from section 3.3, find the standard deviation of this data. Explain what the mean and standard deviation tell you about how much the group of diners would pay for a meal.
    b. Calculate the five-number summary for this data.
    c. Calculate the range and IQR for this data.
    d. Create a boxplot for the data.

3. Use the following table is the cost of purchasing a car at a local dealership. Some of the cars sold were new and some were used.
    a. Find the standard deviation of this data. Explain what the mean and standard deviation tell you about how much the cars are selling for.
    b. Calculate the five-number summary for this data.
    c. Calculate the range and IQR.
    d. Create a boxplot for the data.

Cost (Thousands of dollars)	Frequency
15	3
20	7
25	10
30	15
35	13
40	11
45	9
50	7

4. Studies are often done by pharmaceutical companies to determine the effectiveness of a treatment. Suppose that a new cancer drug is currently under study. Of interest is the average length of time in months patients live once starting the treatment. Two researchers each follow a different set of 40 cancer patients throughout their treatment. The following data (in months) are collected.
    a. Find the standard deviation of each group.
    b. Calculate the 5-number summary for each group.
    c. Calculate the range and IQR for each group.
    d. Create side-by-side boxplots and compare and contrast the two groups.

    Researcher 1: 3, 4, 11, 15, 16, 17, 22, 44, 37, 16, 14, 24, 25, 15, 26, 27, 33, 29, 35, 44, 13, 21, 22, 10, 12, 8, 40, 32, 26, 27, 31, 34, 29, 17, 8, 24, 18, 47, 33, 34

    Researcher 2: 3, 14, 11, 5, 16, 17, 28, 41, 31, 18, 14, 14, 26, 25, 21, 22, 31, 2, 35, 44, 23, 21, 21, 16, 12, 18, 41, 22, 16, 25, 33, 34, 29, 13, 18, 24, 23, 42, 33, 29

5. An experiment compared the ability of three groups of participants to remember briefly-presented chess positions. The data are shown below. The numbers represent the average number of pieces correctly remembered from three chess positions.
    a. Find the standard deviation of each group.
    b. Calculate the 5-number summary for each group.
    c. Calculate the range and IQR for each group.
    d. Create side-by-side boxplots and compare and contrast the two groups.

Non-players	Beginners	Tournament Players
22.1	32.5	40.1
22.3	37.1	45.6
26.2	39.1	51.2
29.6	40.5	56.4
31.7	45.5	58.1
33.5	51.3	71.1
38.9	52.6	74.9
39.7	55.7	75.9
43.2	55.9	80.3
43.2	57.7	85.3

6. There is evidence that smiling can attenuate judgments of possible wrongdoing. This phenomenon termed the "smile-leniency effect" was the focus of a study by Marianne LaFrance & Marvin Hecht in 1995[9]. The following data are measurements of how lenient the sentences were for three different types of smiles and one neutral control.

[9] LaFrance, M., & Hecht, M. A. (1995) Why smiles generate leniency. Personality and Social Psychology Bulletin, 21, 207-214. Adapted from www.onlinestatbook.com, by David M. Lane, et al, used under CC-BY-SA 3.0.

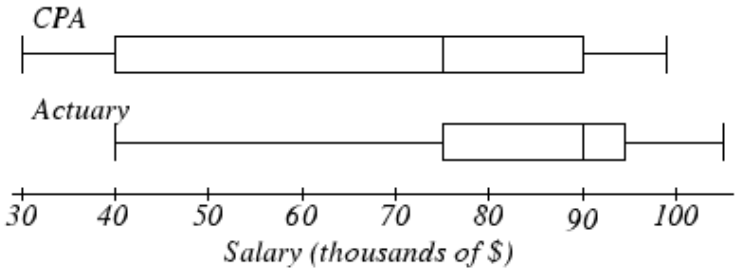
The same subject was used for all of the conditions so that may affect the results. The second column is a continuation of the first column.

a. Find the standard deviation for each type of smile and the neutral control.
b. Calculate the 5-number summary for type of smile and the neutral control.
c. Calculate the range and IQR for each type of smile and the neutral control.
d. Create side-by-side boxplots and compare and contrast the four groups.

False Smile	Felt Smile	Miserable Smile	Neutral Control
2.5	7	5.5	2
5.5	3	4	4
6.5	6	4	4
3.5	4.5	5	3
3	3.5	6	6
3.5	4	3.5	4.5
6	3	3.5	2
5	3	3.5	6
4	3.5	4	3
4.5	4.5	5.5	3
5	7	5.5	4.5
5.5	5	4.5	8
3.5	5	2.5	4
6	7.5	5.5	5
6.5	2.5	4.5	3.5
3	5	3	4.5
8	5.5	3.5	6.5
6.5	5.5	8	3.5
8	5	5	4.5
6	4	7.5	4.5
6	5	8	2.5
3	6.5	4	2.5
7	6.5	5.5	4.5
8	7	6.5	2.5
4	3.5	5	6
3	5	4	6
2.5	3.5	3	2
8	9	5	4
4.5	2.5	4	5.5
5.5	8.5	4	4
7.5	3.5	6	2.5
6	4.5	8	2.5
9	3.5	4.5	3
6.5	4.5	5.5	6.5

7. Make up two data sets with 5 numbers each that have:
    a. The same mean but different standard deviations.
    b. The same standard deviation but different means.

8. The side-by-side boxplots show salaries for actuaries and CPAs.
    a. Estimate the 25th, 50th and 75th percentiles for CPA and actuary salaries.
    b. Deshawn makes the median salary for an actuary. Kelsey makes the first quartile salary for a CPA. Who makes more money? How much more?
    c. What percentage of actuaries make more than the median salary of a CPA?
    d. What percentage of CPAs earn less than all actuaries?



9. Suppose you buy a new car whose advertised gas mileage is 25 mpg (miles per gallon). After driving the car for several months, you find that you are getting only 21.4 mpg. You phone the manufacturer and learn that the standard deviation of gas mileage for cars of that model is 1.15 mpg.
    a. Find the Z-score for the gas mileage of your car.
    b. Does it appear that your car is getting unusually low gas mileage? Explain your answer using your Z-score.

10. This data is a sample of the average number of hours per year that a driver is delayed by road congestion in 11 cities: 56, 53, 53, 50, 46, 45, 44, 43, 42,40, 36
    a. Find the mean and the standard deviation, including units.
    b. What is the Z-score for the city with an average delay time of 42 hours per year?

11. You scored an 89 on a math test where the class mean and standard deviation are 75 points and 7 points respectively. You scored a 65 on an English test where the mean and standard deviation are 53 points and 4 points, respectively. In which class did you do better? Explain your answer using Z-scores.

12. Poe, the Clydesdale horse has a world record breaking height of 20.2 hands. All Clydesdale horses have a mean height of 16.5 hands and a standard deviation of 1.85 hands. The last Great Dane to hold the world record for dog height was Gibson who was 107 cm tall. Great Danes have a mean height of 81 cm and a standard deviation of 13 cm. Which animal is taller compared to their respective breed? Explain your answer using Z-scores.