§ 1.1 Introduction and Terms

How tall are you? How old are you? What's the average lifespan of Americans today? What was the average lifespan of Americans in 1960? What percentage of people are going to vote for a certain presidential candidate?

All of the above questions are ones to which we don't have *exact* answers. In the case of the first two questions, you likely could answer immediately with a number that isn't *quite* exact, but is close enough - at least it's close enough in the sense that we all agree upon some level of precision when measuring age and height. For the last three questions, it's seemingly a 5-second Internet search away from having the "answer." But how do we know how precise these values are?

These kinds of questions - how we measure things precisely, and what is considered precise enough to be "accurate" - fall under the umbrella of *statistics*. Statistics is the study of variability, among other things: the idea that things aren't *the same*.

Statistics comes with a good chunk of vocabulary, but it's primarily front-loaded. Let's identify some key terms here.

Definitions

The overall group one would like to discuss is called the **population of interest.** The number - a mean or proportion, for instance - one would like to try to determine for the population is called a **parameter**. While a **census** would collect data from every individual in the population and directly compute the parameter, a subset of the population, or a **sample**, is usually obtained. From this sample, the same kind of number (mean, proportion, etc.) is computed. This is called a **statistic**.

Example 1.1.1
A school system would like to determine the percentage of all parents of elementary school children in the school system that would support a new proposal that would start school 30 minutes later each day. The school system conducts a survey by reaching out to all 3,255 parents of elementary school children and asking about the proposal, but only 1,340 of the parents respond. 755 of these parents do not support the proposal.

Identify the population of interest, parameter, sample, and statistic.

Solution
The population is all 3,255 parents of elementary school children in this school system. The parameter is the number the school system wants to determine, which is the proportion of all parents that support the proposal. The sample is the 1,340 parents who respond. The statistic is the *sample* proportion who support the proposal, which is $\frac{1340-755}{1340} \approx 0.437$, or 43.7%.

Exercise 1.1.2 A tackle shop has received a shipment of 1,000 worms, but a delay in shipping means there may be an unusually high number of dead worms. The shop wants to estimate how many living worms are in the shipment, so it randomly selects 40 worms from the shipment and finds that 6 are dead.

(a) Identify the population, parameter, sample, and statistic.

(b) Estimate how many dead worms are in the shipment.

In the previous example, you may be wondering: how did the school system reach out? Did they mail a survey, ask students to take a form home with them, or, most likely, post a form on Facebook? Regardless of the method, what can we say about the 43.7% statistic the school system obtained? Does this number *accurately* reflect the parents who did not respond?

When the school system collected a sample and computed the sample statistic, the goal was hopefully to be able to apply this to the larger population of all parents.

Definition

Using the results of a survey, observational study, or experiment to make a conclusion about a larger population is called inference, or inferring to the population.

Exercise 1.1.3 Discuss whether you think it is reasonable for the school system to infer the 43.7% to all parents of elementary school children in the system.

In theory, it seems foolish to try to infer from a sample to a larger population. After all, how can you know what the rest of the population looks like or how it would respond? Realistically, though, we *have* to be able to make inferences, as censuses are too costly or time-consuming in practice. How can we make a reasonable inference?

Exercise 1.1.4 Suppose you wanted to take a sample of 30 students at your school. If you wanted to be able to take any conclusions you reached from this sample and infer it to the population of all students at your school, what would you want the sample to look like?

In practice, the best we can hope is for our sample to "look like" our population as a whole, or to be **representative** of it. But how do we obtain a representative sample?

Definition

A **random sample** is obtained by using a chance process to select individuals.

Exercise 1.1.5 Discuss how randomly selecting a sample from a population *should* create a sample that is representative of the population.

There are many ways to randomly sample, including:

- Putting names in a hat
- Numbering individuals and using a random number generator
- Dialing randomly generated telephone numbers

Exercise 1.1.6 Discuss how the school system from the earlier example could have tried to obtain a random sample of parents in the school system.

Oftentimes, samples are *not* obtained randomly, which can obviously present serious problems. Most commonly, samples that are not obtained randomly end up being non-representative of the population. When this occurs, it is called bias.

Definition

Bias occurs in a sampling method if the method would consistently or systematically result in a non-representative sample of the population. This would in turn produce statistics that consistently overestimate or consistently underestimate the population.

Bias can take many forms and there are many kinds, including, but not limited to:

- Selection bias: when a subset of the population that may differ from the rest of the population can't be selected or is underrepresented
- Non-response bias: when a high percentage of individuals don't respond to a survey, it's likely that those who did respond feel differently than those who didn't.
- Response bias: when individuals' answers will be consciously or subconsciously directed towards a certain answer.

Exercise 1.1.7 How might the school system's sampling method have resulted in bias? Explain what effect this may have had on the estimate of the actual percentage of all parents who support the proposal.

Example 1.1.8
Identify how each of the following sampling methods could result in bias.

(a) A fast food manager wants to estimate the average time that employees showed up on time to work over the course of a week relative to when their shift started. To make his life simpler, he just looks at the time stamps for when employees arrived on Monday.

(b) Someone is interested in whether people like their new haircut, so they post a poll on their Instagram asking people to answer yes or no.

(c) A local polling organization is conducting a poll to estimate the percentage of city residents who plan on voting to reelect the current city mayor. They use a random sample of 250 landlines to call people to respond to the poll.

(d) A college student is doing a research paper on people's dating lives, trying to estimate, among other things, the proportion of the college's students that are in a relationship. They decide to randomly select students on campus and ask them if they are in a relationship or not.

Solution

(a) People may be more likely to show up earlier on Mondays than on, say, Fridays, so the manager's estimate may end up being lower than the actual average time employees showed up relative to when their shifts started over the course of the week.

(b) Generally, younger people are far less likely to own landlines than older people, and younger and older people may feel differently about the city mayor. This could lead to an overestimate if older residents are more in favor of the mayor than younger residents (or vice-versa).

(c) Because saying no to the question might hurt the person's feelings, people who respond to the poll are most likely to say yes. Further, the poll can only be answered by people who follow the person who posted the poll, which means they are more than likely the person's friends or acquaintances, which again makes them more likely to say yes.

(d) Individuals' dating lives are a sensitive subject, so asking face-to-face may lead people to lie out of privacy or to make themselves look better.

Exercise 1.1.9 Discuss how each of the sampling methods from the previous exercise could be improved to reduce the impact of bias. Come up with one concrete improvement for each method.

§1.1 Homework

1. At a rice mill, trucks enter every day with 40,000 to 50,000 pounds of rice. Before the rice is accepted to be milled, the rice must be checked for the amount of aflatoxin, in parts per billion (ppb), which is poisonous for some animals. For every truck that arrives, a mill worker takes a device that gathers small amounts of rice from 5 different depths (the rice is nearly 6 feet deep!) and uses this device in 3 different locations. Suppose a truck arrives with 45,000 pounds of rice, and the mill worker goes and obtains 2 pounds of rice. The 2-pound sample shows 8 parts per billion of aflatoxin.

   (a) Identify the population, the parameter, the sample, and the statistic.

   (b) What is the purpose of the worker obtaining rice from 5 different depths?

   (c) The worker only obtains a 2-pound sample out of the 45,000 pounds. Why do you think this is? Do you think this is reasonable?

   (d) At a different mill, workers obtain samples by scooping 5-pounds off the top of every truck that comes in. Explain how this could result in bias.

2. A Boston University study that found that 99% of the donated brains of deceased former college and National Football League football players had chronic traumatic encephalopathy (CTE), a degenerative brain disease. After news stoies about the study came out, the study was noted to have substantial selection bias: the brains were primarily donated from families who noticed cognitive decline in the former players.

   (a) Theoretically, what is the parameter of interest in this study?

   (b) Explain how this selection bias could affect the estimate of the parameter.

   (c) A follow-up study was conducted to address the issue of selection bias. Read a short article about it at the link below.

   https://www.sciencedaily.com/releases/2022/04/220428103958.htm

   Discuss the findings of the follow-up study.

3. An online survey was conducted of 308 individuals (243 women and 65 men) in Canada. In the survey, individuals filled out the "Partner Phubbing Scale," which measured how much the participants perceived their significant others ignored them by being on their phone, and the "Perceived Romantic Relationship Quality Scale," which is self-explanatory. In it, researchers found that partners "phubbing" was negatively related to the quality of relationships.[1]

   (a) What do you think the population of interest in this study is?

   (b) Do you think it's reasonable to infer the findings of this study to the population of interest?

   (c) What do you think "negatively affected" means?

4. Below is a quote from an article on PBS.org from September 2022.

   > "There's a dirty little secret that we pollsters need to own up to," wrote polling expert David Hill, president of Hill Research Consultants and a 2020 fellow at the University of Southern California's Dornsife Center for the Political Future, in The Washington Post in 2020. "People don't talk to us anymore, and it's making polling less reliable."[2]

   Explain what David Hill might be talking about. How might people "not talking to" pollsters making polling less reliable?

5. Read the short article "Coffee drinking is associated with increased longevity" at the link below:

   https://tinyurl.com/2p9by4xb

   Then, answer the following.

   (a) Identify the sample, the sample size, and where the sample came from.

   (b) It is implied that the study did not include a random sample of individuals. Do you think this is a problem or not? Explain.

   (c) What were some of the parameters being estimated in this study?

   (d) Do you think it's reasonable to assume that drinking two to three cups of coffee a day will lead to a longer lifespan?

[1] https://journals.sagepub.com/doi/10.1177/00332941221144611
[2] https://www.pbs.org/publiceditor/blogs/pbs-public-editor/the-problem-with-polls/

§ 1.2 Observational Studies & Experiments

In the previous lesson, we discussed sample surveys, which are one major kind of statistical study. The other two are defined below.

Definition

An **observational study** is a study in which individuals are not assigned to treatment groups. An **experiment**, on the other hand, is a study in which individuals are assigned to treatment groups.

Exercise 1.2.1 A study in Brazil[3] was conducted to investigate the effects of smoking during pregnancy on the birth weight of newborn babies. Data was collected from women who gave birth at either of two maternity hospitals in January 2012. The women were classified as having either not smoked or smoked 1 to 5, 6 to 10, or 11 or more cigarettes per day during pregnancy. Infant data was then collected to investigate the effects of smoking on infant birth weight.

(a) Do you think this was an experiment or an observational study?

(b) Why do you think researchers chose the study design you identified in (a)?

(c) Data showed that newborn weight was, on average, 320 grams lower for infants whose mothers who had smoked 6 or more cigarettes per day than for infants with mothers who had not smoked during pregnancy. What do you think the purpose of this study was?

In the case of the Brazilian study, it's clear that doing an experiment would have been highly unethical and even illegal: people cannot be forced to smoke during pregnancy, especially if it's predicted that this will be unhealthy for infants. Ethics are indeed a *major* concern when it comes to many studies - not just when trying to determine whether it's ethical to conduct a randomized experiment.

Exercise 1.2.2 In the Brazilian study, participants all had to give informed written consent prior to their participation in the study. Why do you think this is?

Exercise 1.2.3 Nearly every experiment, in fact, must rely upon volunteers to participate in the experiment. Why do you think this is? What are some limitations this may cause?

Even once a study design has been chosen that will be completely ethical, many problems may abound. In the case of the Brazilian mothers, the data revealed that infants of women that smoked 6 or more cigarettes a day had a lower mean birth weight than infants whose mothers didn't smoke during pregnancy: was it reasonable to assume that the smoking *caused* this reduction in mean birth weight, though? Here, it's worth defining four terms that will enable us to better answer this question.

Definition

A **response variable** is a variable you measure at the end of a study. An **explanatory variable** is a variable that may affect the response variable.

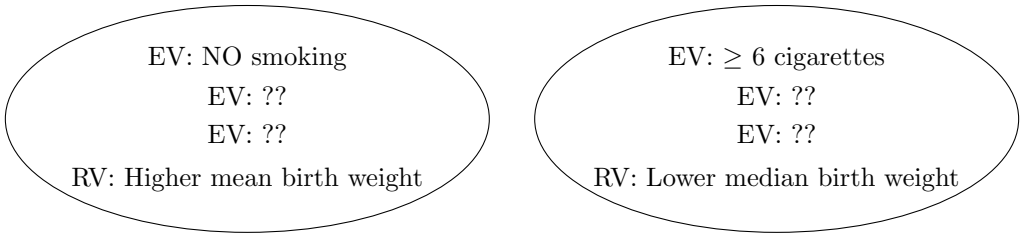
Two variables are **associated** if changes in one variable can *predict* changes in another variable. On the other hand, **causation** occurs when the changes in a response variable can be directly attributed to *only one* explanatory variable.

For the Brazilian study, we can identify the following.

- The explanatory variable of interest was amount of cigarettes smoked per day during pregnancy.
- The response variable was birth weight of newborns.
- The study showed an association between smoking and birth weight because infants whose mothers had smoked on average had lower birth weight.

[3] *Reproductive Health, 2018*

We return to the question, though: did the study show that smoking caused the reduction in mean birth weight? Let's look at a diagram of what the data showed.



For the study to show causation, we must be able to say that the amount of cigarettes smoked per day was the *only* variable that led to the reduction in birth weight. The issue with an observational study, though, is that the two groups may differ in some other way that also affects the response variable.

Exercise 1.2.4 Can you think of *another* way in which the two groups shown above may have differed? Could this also be tied to a change in birth weight of infants?

When two (or more) associated variables can *both* predict changes in a response variable, confounding may occur.

Definition

Confounding (i.e. "confusing") occurs in an observational study when two or more associated explanatory variables both predict changes in the response variable. Then, it is impossible to tell which of the explanatory variables *caused* the changes in the response variable.

In the study of smoking and infant birth weight, researchers had to consider the presence of numerous confounding variables, including

- Age
- Education
- Whether the mother had a partner
- Whether it was the mother's first pregnancy or not
- Pre-pregnancy weight of the mother
- Whether the mother participated in a prenatal education group
- Previous advice regarding warning signs in pregnancy

and numerous others. In this particular study, confounding was accounted for using statistical models that are well beyond the scope of this course, but statistically significant results were demonstrated even after accounting for the above variables. In short, it was indeed reasonable to conclude that this study, especially when considered alongside hundreds of other similar studies, *did* show that smoking caused a reduction in mean birth weight.

Many studies may not go to such lengths to consider confounding, though; even worse, even researchers that do consider confounding may not find these considerations included in news stories covering their studies.

Example 1.2.5
Here's an article title from RealSimple.com: "7 Science-Backed Health Benefits of Drinking Red Wine."[4] Among the benefits listed is "It improves heart health," where heart health is measured by the amount of HDL, or "good" cholesterol.

(a) In the study recording HDL, identify the explanatory and response variable.

(b) Do you think this study was an experiment or observational study?

(c) Identify another variable that may be associated with drinking red wine that could also be predictive of higher HDL. Explain how this could lead to confounding.

[4] `https://www.realsimple.com/health/nutrition-diet/red-wine-health-benefits`

Solution

(a) The explanatory variable is amount of red wine (or whether an individual drinks red wine) and the response variable is amount of HDL.

(b) No study can ethically force people to drink or not drink red wine, so this must have been an observational study.

(c) Individuals that drink red wine may be wealthier and therefore have better access to healthcare. This could also lead to higher HDL. This could make it impossible to determine if drinking red wine or having more money and better access to healthcare caused the higher HDL.

Unfortunately, headlines like the Real Simple one are all too common. Such articles rarely address any of the deeper statistics involved in these research studies, regardless of how well researchers tried to address confounding. (Note: In the referenced HDL study, confounding was actually *not* addressed!)

Exercise 1.2.6 A university wants to investigating how incorporating history into its teaching of Calculus I would affect student scores on the final exam. It decides to conduct a study where one of its Calculus I teachers will teach a Calculus I class the way always have (without history) and another Calculus I teacher will teach a Calculus I class with history incorporated. At the end of the semester, the university will look at the average final exam score of the two Calculus I classes.

(a) Identify the explanatory variable and the response variable.

(b) Is this an experiment or observational study?

(c) Identify 2 other explanatory variables that could also affect students' final exam scores.

(d) Explain how one of the two variables you identified in (c) could lead to confounding.

(e) How could the university address the potential for confounding variables you identified?

---

§1.2 Homework

1. A study was conducted in Germany[5] in which 10th-12th grade students were given the option of starting school at either 8 A.M. or 8:50 A.M. each day. For each student, the amount of hours slept was recorded at regular intervals over the course of 4 years, as was quarterly grades as provided by the school.

   (a) Identify the explanatory variable.

   (b) Identify the response variables listed.

   (c) Was this study an observational study or experiment? Explain.

   (d) The researchers had to obtain informed consent from all of the participants. Why was this an important step, even if students were the ones choosing when they started school?

2. A 2017 study[6] showed a positive association between consumption of food within 2 hours of bedtime and increased body fat. Explain what "positive association" means in this context.

3. Numerous studies have shown that children who eat dinner with their family multiple times per week earn better grades in school.

   (a) Identify 2 potential confounding variables associated with eating with one's family multiple times per week.

   (b) Explain how the variables listed in (a) could "confound" the results of these studies.

[5] https://www.nature.com/articles/s41598-022-06804-5
[6] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5657289/

4. Suppose a history teacher wants to conduct a study in which they investigate whether students taking notes on a laptop is more effective than taking notes on paper. The teacher plans on conducting the study in their first period class.

   (a) Describe how the teacher could use an observational study design to conduct this study with their first period.

   (b) Describe how the teacher could use an experimental design to conduct this study with their first period.

   (c) Describe an advantage of each of the study designs in (a) and (b).

   (d) Describe a disadvantage of each of the study designs in (a) and (b).

5. A university obtained 252,300 death certificates from a national health organization in 2008 to determine the leading causes of death that year.

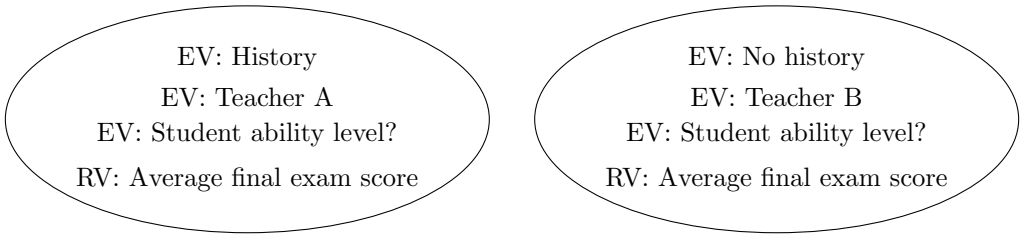
   (a) What is the sample, and what is the population?

   (b) The sample was not randomly selected. Discuss one reason this might be. Do you think this is an issue?

§ 1.3 Scope of Inference

Okay, so bias and confounding can be major problems when taking a sample or conducting an observational study. Sophisticated statistical techniques can be used to deal with confounding in certain cases, but is there something more elementary that can be done to reduce the possibility of confounding? It turns out, there is.

Let's go the example from Exercise 1.2.6, where a university wanted to investigate the effects of incorporating history into teaching Calculus I. The primary flaw in the university's design was that it would create two treatment groups that looked like those pictured below



The two groups differ in *more than one way* that might affect final exam score. Perhaps Teacher A is a more effective teacher than teacher B; perhaps the no history class has a stronger group of calculus students; perhaps the history class is at 8:00 A.M., when students are more likely to skip or sleep in, and the no history class is at 10:00 A.M. The experiment as currently constructed is loaded with potential confounding variables! This will make it impossible to determine whether any differences in average final exam scores is due to the history, the teacher, the student ability level, etc. What can the university do?

Definition

To **control** for a variable is to keep it the same for all treatment groups.

Control is used to try to make treatment groups look as similar as possible. This way, any differences in the response variable can theoretically be attributed *only* to the treatment.

Exercise 1.3.1 Discuss how the university could control for the teacher of the class/es and the time.

Controlling the teacher and class time, though, does not deal with the issue of *student ability level*. How can the classes be split up so that the average ability level is reasonably similar? On the one hand, students could be placed into classes according to their high school GPAs... but this doesn't account specifically for math. What about their math grades in their senior year of high school or their previous year of college? This seems time consuming and assuredly invasive of students' privacy. What can be done?

The solution, as it frequently is, is to assign students *randomly*.

Definition

Random assignment is when individuals are placed into treatment groups using a chance process.

Exercise 1.3.2 If the students are randomly placed into one of the two Calculus I classes, how *should* the average ability levels of the two classes compare?

The beauty of random assignment is that it shouldn't just "even out" the students according to ability level: it should create groups that look similar in *every* way that can't be directly controlled for.

Exercise 1.3.3 A pharmaceutical company is conducting randomized trials to test the effectiveness of its new cholesterol-lowering drug on 40-50 year old males with high cholesterol. It recruits 500 such males to participate in the study and randomly assigns 250 to take the company's current cholesterol drug and the remaining 250 to take the new drug.

(a) What is the purpose of randomly assigning the men to the two groups instead of, say, letting them choose?

(b) What are some other variables the company should try to directly control?

(c) Assume the company directly controls for as many variables as it can. If the randomly assigned men who took the new cholesterol drug show a greater average reduction in cholesterol than the men who took the current drug, is it reasonable to conclude the new drug is more effective than the current drug?

This brings us to a vital discussion: when can you infer what?

Definition

The **scope of inference** of a study is what the study can reasonably infer. This usually boils down to whether the study can **infer to a larger population** or if it can **infer causation**.

We already discussed in the first lesson that results can reasonably be inferred to a larger population only when the sample is representative of that population; this can primarily be assured through random sampling. To infer causation, however, a study's design must include control and random assignment to reduce the impact of any confounding variables; only in this way can it reasonably rule out any other explanatory variables from having impacted the response variable.

The table below provides a breakdown of when it's reasonable to infer to a larger population or causation.

		Random Assignment & Control	
		Yes	**No**
Random selection	**Yes**	You can infer to the population. You can infer causation.	You can infer causation. You cannot infer causation.
	No	You can't infer to the population. You can infer causation.	You can't infer to the population. You can't infer causation.

Exercise 1.3.4 Of the four possibilities in the table above, two are by far the most common. Which two do you think it is, and why?

Example 1.3.5
A study[7] in Finland in 2016 investigated the gambling habits of people living in one of three Finnish regions. Twenty thousand individuals (the "population sample") were randomly selected to complete a survey anonymously, of which 7,186 (about 36%) responded. Additionally, 119 individuals who had sought treatment for gambling problems (the "clinical sample") were selected to participate in the study. Among the results were the following:

- 83% of the population sample stated they had gambled on at least one game type during 2016
- A much higher proportion of the clinical sample had played EGMs (electronic gaming machines) frequently than the proportion of those in the population sample.

(a) Why was it important that the survey was anonymous?

(b) Is it reasonable to infer that roughly 83% of *all* Finnish people gambled on at least one game in 2016? If not, to what population *can* it be inferred?

(c) Based on this study, is it reasonable to infer that EGMs are more responsible for gambling problems than other game types?

[7] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7434154/

Solution

(a) Without anonymity, people may not respond truthfully, as gambling is a sensitive topic relating to personal finances and potentially morality.

(b) No, as the sample was not randomly selected from *all* Finnish people. The population sample was randomly selected from all people in just three Finnish regions, so the 83% can only reasonably be inferred to all people in these three regions.

(c) No. A randomized experiment was not conducted. Simply identifying that people who sought treatment for gambling were more likely to have played EGMs is not in and of itself proof that EGMs lead to gambling problems.

Exercise 1.3.6 In 2019, a study[8]was conducted to determine if the Nike Vaporfly 4% shoe would reduce the amount of oxygen needed to run. Below is the abstract of the study:

"Nineteen subjects performed two 5-minute trials at 4.44m/s wearing the Adidas Adios Boost (AB), Nike Zoom Streak (ZS), and Nike Vaporfly 4% (VP) in random order. Oxygen uptake was recorded during minutes 3–5 and averaged across both shoe trials. On a second day, subjects wore reflective markers, and performed a 3-minute trial in each shoe. Motion and force data were collected over the final 30 seconds of each trial. VP oxygen uptake was 2.8% and 1.9% lower than the AB and ZS. Stride length, plantar flexion velocity, and center of mass vertical oscillation were significantly different in the VP. The percent benefit of the VP over AB shoe was predicted by subject ground time. These results indicate that use of the VP shoe results in improved running economy, partially due to differences in running mechanics. Subject variation in running economy improvement is only partially explained by variation in ground time."

Answer the following questions related to this study.

(a) Why was the order that the runners ran in the three shoes randomized?

(b) Is it reasonable to infer that the Nike Vaporfly 4% caused the reduced oxygen intake?

(c) To what population, if any, is it reasonable to infer the results of this study to?

Understanding scope of inference is vital whether you end up going into scientific research and conducting studies of your own or simply reading the news as an ordinary citizen. Statistically literacy is more than just scope of inference, but understanding scope of inference is a huge start.

§1.3 Homework

1. Suppose Pepsi Co. wants to do a study to see if people prefer their new Pepsi Zero Sugar to Coke Zero. They want to do a taste test where they have people drink a cup of each soda and then determine which they like better. What is one major variable the researchers need to control for in such a study?

2. A 20-story apartment complex in a city has a working, but slow elevator. The landlord of the complex is considering getting a new elevator that would be faster, but it would take 1 month to install. The landlord wants to randomly survey 20 people who live in the complex and see if they think it would be worth improving the elevator even if it would take a month.

   (a) Why should the manager randomly sample instead of sampling, say, the people who live on the first floor?

   (b) The landlord is considering randomly selecting one person from *each floor* (this is called stratified random sample). Explain one benefit and one drawback of such a sampling method.

   (c) Will the landlord reasonably be able to assume that a stratified random sample of 20 people will be representative of all the residents of the complex?

[8]Hunter, Iain (2019). Running economy, mechanics, and marathon shoes. *Journal of Sports Science*, 2367-2373.

3. A 2020 study[9] at the University of Texas at Austin was conducted to investigate the effects of social media "likes" on adolescents' perceptions of themselves. Below is a quote from an article about the study.

> "Study participants were told they were helping test drive a new social media program that allowed them to create a profile and interact with same-age peers by viewing and liking other people's profiles. The 'likes' received were tallied, and a ranking of the various profiles displayed them in order of most to least liked. In reality, these 'likes' were assigned by computer scripts.
>
> Participants were randomly assigned to receive either few 'likes' or many 'likes' relative to the other displayed profiles. In a post-task questionnaire, students in the fewer-"likes" group reported more feelings of rejections and other negative emotions than those who received more 'likes."'

   (a) Was this study an experiment or an observational study?

   (b) Do you think the study randomly selected individuals? Why or why not?

   (c) *Why* did the computer *randomly* assignment the numbers of "likes"?

   (d) The article continues: "This study is an important scientific advance because it uses an experiment, and it shows that not getting enough 'likes' actually causes adolescents to reduce their feelings of self-worth." Explain how this experiment shows causation.

   (e) Is it reasonable to infer the results of this study to all adolescents?

   (f) Do you feel the study was ethical? (Note: Study participants were notified after the study that the "likes" they had received were random.)

4. Many scientific studies use animals like mice or rats, as opposed to humans, to conduct scientific experiments.

   (a) Why do you think this is? Name 3 reasons.

   (b) Beyond the potentially dubious morals of animal testing, discuss one major limitation of the scope of inference of these kinds of studies.

5. Explain why random selection allows for inference to a larger population.

6. Explain the purpose of random assignment and direct control in an experiment.

[9] `https://www.researchgate.net/publication/344195460_Getting_Fewer_Likes_Than_Others_on_Social_Media_Elicits_Emotional_Distress_Among_Victimized_Adolescents`

§ 1.4 Describing Distributions of Data

Once a study has been designed, approvals can be obtained, and data has been collected, it's time to assess and analyze the data. What are some key characteristics we look for when analyzing data? It would behoove us to first identify the main kinds of variables.

Definition

A **categorical variable** is a variable in which the values are categories. These categories could be numbers, but they are numbers that only serve as labels.

A **quantitative variable** is a variable in which the values are numerical measurements of some kind.

Categorical variables include things like sex, race, hair color, political party, etc. It could even include something numerical like *age group*, where someone may fit in a certain category like 15-29 or 30-44.

Quantitative variables are any measurements, which include time, height, weight, length, and so on. Age could be a quantitative variable if a person's particular age factored in, as opposed to simply an age *range*.

Unfortunately, the data you may work with can take *many* forms - one quantitative variable, two or more quantitative variables, one categorical variable, two or more categorical variables, or some combination of each - and there is no perfect playbook that teaches one how to analyze any and all data. That said, there are a few things to look out for when it comes to each kind of variable.

For categorical variables, here are a few major considerations:

- Which categories have the highest proportion/number of individuals in them? Which categories have the least?
- When analyzing two or more groups, look for similarities and differences!
- If comparing two or more groups, you must compare *proportions*, as the group sizes may not be the same.

Example 1.4.1
The table below displays a summary of the Coordinate Algebra End-of-Course test scores for two teachers' students.

	0-69	70-79	80-89	90-100
# of students (Teacher A)	2	12	15	22
# of students (Teacher B)	1	32	18	24

Teacher A and Teacher B don't know each other's students' scores, but they meet at a conference and are talking about how their students did. Teacher A says, "Yes, I had 37 students get a B or higher," to which Teacher B responds, "Yes, about the same for me - I think I had 42 students get at least a B."

(a) Teacher A feels slightly bad that Teacher B had more students get at least a B. Should they?

(b) Compare the two distributions of test scores.

Solution

(a) No! Teacher A failed to consider that Teacher B might also just *have more students*. In fact, $\frac{37}{51} \approx 72.5\%$ of Teacher A's students got at least a B, which is much higher than the $\frac{42}{75} = 56\%$ of Teacher B's students that got at least a B.

(b) Both teachers had a high majority of students pass the exam (assuming 70 is passing). Teacher A, however, had a much higher proportion get an A or a B, while Teacher B had a higher proportion of students get a C.

For quantitative variables, it's more frequent that we discuss some relevant subset of the following characteristics: the shape, center, variability, and unusual features.

Definition

The **shape** of a distribution may include

- if the distribution is roughly symmetric, right skewed, or left skewed

  A distribution is right skewed if it has more lower values and fewer higher values (making a longer right "tail"). A distribution is left skewed if ti has more higher values and fewer lower values (making a longer left "'tail").

- if the distribution has one peak (unimodal), two peaks (bimodal), or more

Definition

The **central tendency** of a data set is an estimate of where the "center" or "middle" of the data is.

Most often, this is computed as either the **mean** or **median**. The **mean** is the average value of the data set, while the **median** is the middle value (or average of the two middle values).

Definition

The **variability** of a data set is one of multiple measures of how varied the data is. This could include the

- Range = Maximum − Minimum
- Interquartile Range = Third Quartile − First Quartile
- Standard Deviation = Typical distance between the values in the data set and the *mean*

Definition

The **unusual features** (this isn't quite a technical definition) is anything striking about a distribution, such as

- if any values might be potential outliers
- if there are any "gaps" in the data
- if there are any unique clusters

All put together, the following provide a general set of guidelines for features of a quantitative distribution that might be worth discussing. Given any particular distribution, perhaps one or more of these are not especially relevant (there may be no unusual features, for instance).

Example 1.4.2
Shown below are the AP exam pass rates (as percentages) for the 15 AP courses offered at a certain Georgia high school in 2020-21.

57.1, 68.8, 83.3, 35.3, 60, 58.6, 56.3, 52, 33.3, 40, 91.7, 10, 63.4, 59.4, 60

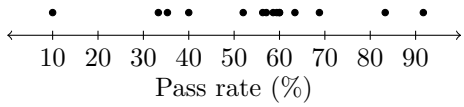
(a) Create a dotplot to display this data.

(b) Create a histogram to display this data. What advantages does this have over the dotplot for this kind of data?

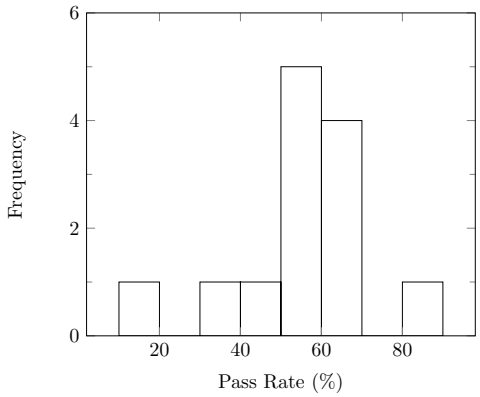
(c) Describe the shape and any unusual features of the distribution.

Solution

(a) The dotplot is below.



(b) For a histogram, it makes sense to make the intervals have width 10 - this way there will be 8 different intervals (not too many, not too few).



The histogram groups close values together, so we are better able to see the peak around 50-60%.

(c) The distribution of pass rates is unimodal and slightly left skewed. There were no classes with pass rates between 20-30% or 70-80%, and the classes with pass rates of 10% and 91.7% may be outliers.

Now, it's worth discussing how we can go about computing some summary statistics to measure central tendency and variability. Let's talk about two ways you can use technology to do it: with a TI-84 calculator and with Microsoft Excel.

TI-84 Skills

1. To type data into a list, press `stat`, then `1:Edit`.
2. To compute summary statistics, press `2nd mode` (`quit`) to go to the home screen. Then press `stat`, go to `CALC`, and press `1:1-VarStats`.
3. In the `1:1-VarStats` menu, type in your list using `2nd` then the number of your list (so `2nd` then `1` for L_1). Leave `FreqList` blank (or type the number `1`). Press `enter`.
4. In the resulting summary statistics, $\bar{x}$ is the sample mean and `Sx` is the sample standard deviation (σ`x` is the population standard deviation if your list was an entire population). Scroll down to see the median and quartiles.

Exercise 1.4.3 Compute the mean, median, standard deviation, IQR, and range of the AP exam pass rate data using a TI-84.

For Microsoft Excel, you need only to be comfortable with how to type ranges of cells. If you have data in cells A1, A2, A3, and so on all the way up to A50, then communicating this range of cells to Excel is done with a colon: so these cells would be A1:A50. You can always drag and highlight to select a range of cells as well.

Excel Skills

1. Input your data into a column.
2. Go to an empty cell and type = to start a new formula.
3. Excel knows many formulas, most of which are intuitive to figure out (and which it will help autofill!). For instance
   - AVERAGE will give the mean of values in a range of cells.
   - STDEV gives the standard deviation of a sample, and STDEV.P gives the standard deviation of a population.
   - There is no command for range, but you can use the MAX and MIN functions for this.
   - QUARTILE.INC will give the quartiles if you add the additional argument of *which quartile*. For instance, =QUARTILE.INC(A1:A10,3) will give the 3rd quartile (the 75th percentile) of the values in cells A1 through A10.

Exercise 1.4.4 Use Excel to compute the mean, median, standard deviation, IQR, and range of the AP exam pass rate data. Did you get the same values as you did in Exercise 1.3.9?

Being able to compute summary statistics and describe what you see is an invaluable skill for any future researcher. Even though we're addressing relatively simple data sets here, the skills and considerations made can generalize to more complex and messy data sets.

Example 1.4.5
With the summary statistics now included, describe the AP pass rates at this school.

Solution
Pass rates for AP exams at this school were unimodal and slightly left skewed with a mean of 55.3% and median of 58.6%. The standard deviation was 19.3%, so most subjects had a pass rates between about 36% and 74%. However, no subjects had pass rates between 20% and 30% or 70% and 80%, and two subjects were potential outliers with pass rates of 10% and 91.7%, respectively.

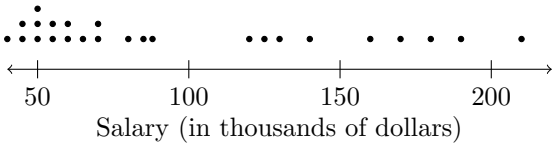
---

§1.4 Homework

1. A teacher was reporting back test scores to a class and stated, "The mean was 74 and the standard deviation was 12." Explain in everyday language what the standard deviation means in this context.
2. Shown below are the AP exam pass rates for another Georgia high school in 2020-21.

   76.2, 52.2, 58.3, 54.1, 63.6, 60, 57.1, 60.8, 30.2, 20, 48, 33.3, 70.3

   (a) Use a TI-84 or Excel to compute the mean, median, standard deviation, and IQR.
   (b) Make graph of this data and describe its shape and any unusual features.
3. A law school is trying to discuss with its students what kinds of salaries they might be able to earn after graduation. It obtains a random sample of 25 of its previous year's graduates and asks them to anonymously fill out a survey in which, among other things, they provide their salary. Shown below is a dotplot of the reported salaries.



The law school then reports to students that, for the random sample, the average starting salary was a whopping $97,625.

(a) Explain why this mean of $97,625 is not a good measure of central tendency for this data set.

(b) Compute the median salary for the 25 graduates. Is this a better measure of central tendency?

(c) What about this data set made the median and mean so different? Explain.

4. In Major League Baseball (MLB), every team has a payroll (the total amount it pays its players). Some players make more than others, and the table below is a list of the *percentage of the payroll* that each team's highest paid player will earn in 2023. For instance, Matt Olson is the highest paid Atlanta Braves player at $21,000,000 per year, which is 11.0% of the team's payroll.

Team	% Payroll	Team	% Payroll
Chicago White Sox	10.9	Oakland Athletics	17.1
Atlanta Braves	11	St. Louis Cardinals	17.6
San Francisco Giants	11.4	Colorado Rockies	17.9
Toronto Blue Jays	11.6	Miami Marlins	18.5
Philadelphia Phillies	11.9	Los Angeles Angles	19.6
Boston Red Sox	12	Baltimore Orioles	19.7
New York Mets	12.9	Texas Rangers	19.8
Los Angeles Dodgers	12.9	Cleveland Guardians	22
San Diego Padres	13.5	Arizona Diamondbacks	23.9
Chicago Cubs	14.8	Minnesota Twins	24.1
New York Yankees	14.9	Milwaukee Brewers	24.8
Seattle Mariners	16.4	Kansas City Royals	25.9
Pittsburgh Pirates	16.5	Detroit Tigers	30.3
Houston Astros	16.6	Washington Nationals	31.5
Tampa Bay Rays	16.9	Cincinnati Reds	35.4

(a) Construct a histogram for this data.

(b) Compute summary statistics and provide a 2-3 sentence description of the distribution of percentages of payrolls dedicated to highest players for the MLB teams in 2023.

§ 1.5 Working with Messy Data

Unfortunately, most data sets are not as clean or small as those we work with in classroom settings. Classroom examples must be made to fit in a class period or have nice round numbers to not obscure the statistical concepts. In real life, though, data sets are often much, much bigger... and even messier. To give even the slightest idea of such a data set, we'll use the American Statistical Association's *Census at School Random Sampler*. The Census at School is a database of responses to online surveys given all around the country, and the Random Sampler allows you to select random samples of responses from any state.

Suppose we want to get a random sample of the responses of 100 Georgia high school students who completed the survey in 2022.

Exercise 1.5.1

(a) Go to `https://ww2.amstat.org/censusatschool/` and click on Random Sampler.

(b) Select the appropriate options to get our desired sample. Then, hit "Submit."

(c) Click "Download Data."

(d) Open the downloaded CSV file with Microsoft Excel. (If your device tries to open it with a text editor, then *right click it* and choose "Open With.")

You should now have a random sample of 100 real Georgia students who completed this survey in 2022. You may notice that the Excel file is... a lot. The students' responses generated data for *59* different variables (not including country, state, and year)! Can you imagine dealing with this by hand?

Luckily, Excel can happily chug away and compute summary statistics for 1,000 cells as easily as it can for 5 cells. Unluckily, though, Excel will do so *mindlessly*. Excel will pay absolutely no mind to the joker who decided to say they were 5,000 centimeters (∼164 feet) tall or the student who wrote that they speak "1 1/2" languages.

Values like this may be funny, the result of an honest mistake, or even meaningful - but regardless, they are *going* to exist. The question is, do they need to be dealt with, and if so, how?

Exercise 1.5.2

(a) Suppose the actual average height of students in the sample of 100, ignoring the student who wrote 5,000 cm, is 170 cm. What is the *sum* of the heights of these 99 students then?

(b) If we *don't* remove the 5,000 cm from the data set, what is the mean of the 100 heights now? How much did it change?

(c) Do you think it's okay if we keep the 5,000 cm in the data set?

Exercise 1.5.3 What is your opinion on the "1 1/2" languages data point? Should it be kept or removed?

Exercise 1.5.4

(a) Look at your own sample of data and find at least 4 other instances of data values that are missing or errant in some way. Find values in at least 4 *different* columns.

(b) Discuss how you would handle each of the types of missing/errant values you identified in (a).

The process of fixing or removing errant values, replacing values, or making any other necessary adjustments is called cleaning data. Some softwares come with packages that will automatically clean data! Oftentimes, though it is necessary to clean data manually.

In the case of the Census at School data, a new website was actually formed to house the data. It is:

`https://new.censusatschool.org.nz/random-sampler/?database=Cas_US`

Exercise 1.5.5

(a) Go to the new Census at School site.

(b) Use the options provided to generate a new sample of 100 students from Georgia. Choose "All Variables" for Select Variables and "Random Sample" for Sample Type.

(c) Press Download sample.

(d) Open the CSV file for the sample. Look for any messiness (missing or errant values). Do you see anything?

The new Census at School site enables a few different features that give it advantages over the original site. Beyond simply cleaning data, it also enables generating data for just specific variables, which will make data far more presentable. Depending on where you find data, it may already be cleaned, or you may have to do it yourself!

Exercise 1.5.6 Consider a data set for all home sales in Connecticut in 2020. Among the values listed are the assessed value of the home, the actual sale price of the home, and the sales ratio (or the actual price divided by the assessed value). Upon looking at the data set, one of the first homes you see is the one listed below:

List Year	Town	Address	Assessed value	Sale Amount	Sales Ratio
2020	Groton	3 WATER ST UNIT 304	0	600000	0

(a) What's the issue with this data point?

(b) What would you do this with data point?

One other feature of the new Census at School site is worth talking about briefly: it allows for *stratified* random samples.

Definition

A **stratified random sample** is a sampling method in which a population is divided into subgroups, or **strata** (singular: **stratum**), that are expected to *differ* in some way related to what is being measured. Then, a random sample of each stratum is selected, and the subsequent random samples are combined to make the overall stratified sample.

Exercise 1.5.7 Suppose you were looking at your sample of 100 Georgia students and you noticed that 75 of the 100 students selected were all seniors. Would this be problematic?

Exercise 1.5.8 If sampling too many seniors - or too many students from any one grade level - were particularly problematic, how could a stratified random sample eliminate these problems from occurring?

Exercise 1.5.9 A student gets a data set of movie sales per weekend for 200 movies over the course of 26 weekends. They find that the numbers are in the millions of dollars, and they would like to not have so many 0's clogging up the data and any potential tables or graphs they make.

(a) How could they use an Excel formula to take care of any column that has values in the millions to no longer be in the millions?

(b) If the student goes through with converting the values to smaller (one- and two-digit) numbers, what is important that they keep in mind when making any tables or graphs?

Stratified random sampling is remarkably common when it comes to polls of all kinds: researchers want to get people of different ages, genders, races, ethnicities, etc. Stratified sampling allows this to

happen. Even experiments can make sure of stratified sampling - even though people have to volunteer for experiments, those in charge of the experiment can ensure they don't begin the study until they have people from various strata.

Exercise 1.5.10 During the Super Bowl in 2023, TurboTax ran an ad where a man was dancing by a fountain for 30-60 seconds. Suppose TurboTax initially considered whether a man or a woman doing the dancing would be more enjoyable for people watching the ad. They film one of each ad - with a man dancing and with a woman dancing - and decide to get a group of volunteers who will randomly assigned to view one of the two ads.

(a) What are some different "strata" that TurboTax might want to consider getting some individuals from each of?

(b) Suppose that, among the volunteers, 60 identify as male and 60 identify as female. How might TurboTax want to go about randomly assigning the ads in order to avoid confounding?

---

§1.5 Homework

1. Shown below is an example of a data set of sales from a fictional company.[10]

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1		Ship Mode	First Class			Same Day			Second Class			Standard Class		
2		Segment	Consumer	Corporate	Home Office	Consumer	Corporate	Home Office	Consumer	Corporate	Home Office	Consumer	Corporate	Home Office
3	Order ID	Order Date												
4	CA-2011-100293	14-Mar-13												91.056
5	CA-2011-100706	16-Dec-13							129.44					
6	CA-2011-100895	2-Jun-13										605.47		
7	CA-2011-100916	21-Oct-13											788.86	
8	CA-2011-101266	27-Aug-13							13.36					
9	CA-2011-101560	28-Nov-13								542.34				
10	CA-2011-101770	31-Mar-13											1.869	
11	CA-2011-102274	21-Nov-13											865.5	
12	CA-2011-102673	1-Nov-13											1044.44	
13	CA-2011-102988	5-Apr-13								4251.92				
14	CA-2011-103317	5-Jul-13		242.546										
15	CA-2011-103366	15-Jan-13	149.95											
16	CA-2011-103807	2-Dec-13											21.19	
17	CA-2011-103989	19-Mar-13		590.762										
18	CA-2011-104283	27-Jun-13										616.14		
19	CA-2011-106054	6-Jan-13		12.78										
20	CA-2011-106810	14-May-13											310.88	
21	CA-2011-107573	12-Dec-13										23.472		
22	CA-2011-107811	29-Apr-13											661.504	
23	CA-2011-108707	24-Oct-13												10.368
24	CA-2011-109043	15-Aug-13	243.6											
25	CA-2011-109232	13-Jan-13							545.94					
26	CA-2011-110100	25-Apr-13										302.376		
27	CA-2011-110408	18-Oct-13								2216.8				
28	CA-2011-110422	21-Jan-13							25.248					
29	CA-2011-111500	17-Aug-13											484.79	
30	CA-2011-111934	5-May-13		47.32										
31	CA-2011-112718	16-Dec-13											1.167	

   (a) Explain why all the empty white space in the spreadsheet is less than ideal.

   (b) Propose a method for reorganizing the data to handle the issue of empty white space.

2. Go to `https://tinyurl.com/y3n8sr2t`. You should see a spreadsheet of "Ask A Manager Salary Survey." Find at least 3 different ways or areas in which this data could be cleaned.

3. Suppose you want to construct some kind of poll or survey that you will administer online. One of the questions you are considering asking is, "How many years of higher education do you have?"

   (a) Explain an advantage of having the answer box for this question be free text (i.e. the respondent types in their answer).

   (b) Explain a disadvantage of having the answer box for this question be free text.

   (c) Would you choose to go free text or not?

4. A law firm is conducting an investigation into the workplace culture of a professional sports organization. It is going to conduct a random sample of employees and anonymously ask them about experiences within the organization.

[10] Source: `https://foresightbi.com.ng/microsoft-power-bi/dirty-data-samples-to-practice-on/`

(a) Explain why the firm might want to conduct a stratified sample as opposed to a random sample.

(b) Identify two different stratifying variables (i.e. "age" or "hair color") by which the firm might want to stratify employees. Explain your choices of variables.

5. While the idea of getting a stratified sample is simple - break the population into strata and then randomly sample from each stratum - it is often much more difficult in practice. One novel approach was discussed in the abstract of a 2020 study[11] below.

> "Most studies among Hispanics have focused on individual risk factors of obesity, with less attention on interpersonal, community and environmental determinants. Conducting community based surveys to study these determinants must ensure representativeness of disparate populations. We describe the use of a novel Geographic Information System (GIS)-based population based sampling to minimize selection bias in a rural community based study."

(a) Explain what the first sentence might mean.

(b) Explain what the second sentence means.

(c) How do you think this "area-based" random sampling works?

(d) What is meant by "minimize selection bias"?

[11] https://bmcpublichealth.biomedcentral.com/articles/10.1186/s12889-020-09793-0

§ 1.6 Standardized Scores & Normal Distributions

It happens all the time: two friends are arguing over who did better. A conversation might go something like this:

> Student A: "I got an 1240 on the SAT and you only got an 1180. See, I *told* you I'm just smarter than you."
>
> Student B: "Yeah, but I got a 27 on the ACT, and you only got a 25!"
>
> Student A: "Wow, so you beat me by TWO POINTS. I beat you by 60 points!"

Though neither friend should be belittling the other one for standardized test scores - or thinking that those measure intelligence all that well - it's worth digging into who is "right."

The issue at hand is that the students are trying to compare different *units*. While Student A scored higher than Student B by 100 points, that is 100 SAT points. Student B only outscored Student A by two points, but this was 2 ACT points. Are points equally valued on both tests?

This kind of issue happens all the time when you are trying to compare values with different units. At first glance comparing values with different units is impossible. This, then, provides the need for some kind of *unitless* measure of location within a distribution with *any* kind of units.

Definition

A **standardized score**, or **z-score**, is the number of standard deviations above or below the mean a value is. The formula for a standardized score is

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

Example 1.6.1
Suppose that SAT scores have a mean of 1060 and a standard deviation of 200 and that ACT scores have a mean of 20.8 and a standard deviation of 5.8. Compute and interpret each of Student A and Student B's standardized scores on both the SAT and ACT.

Solution
Let z_{AS}, z_{AA}, z_{BS} and z_{BA} be the z-scores of Student A on the SAT and ACT and Student B on the SAT and ACT, respectively. We have

$$z_{AS} = \frac{1240 - 1060}{200} = 0.9 \qquad z_{BS} = \frac{1180 - 1060}{200} = 0.6$$
$$z_{AA} = \frac{25 - 20.8}{5.8} \approx 0.724 \qquad z_{BA} = \frac{27 - 20.8}{5.8} \approx 1.069$$

Student A scored 0.9 standard deviation higher than the mean SAT score and 0.724 standard deviation higher than the mean ACT score. Student B scored 0.6 standard deviation higher than the mean SAT score and 1.069 standard deviations higher than the mean ACT score.

Standardized scores are an incredibly powerful tool because they are unitless. They enable us to make comparisons even across distributions with wildly different centers and amounts of variability. In the case of the two students, we determine that, by the slightest margin, Student A performed better overall. They performed $1.069 - 0.724 = 0.345$ standard deviation better than Student B on the ACT, while Student B only performed $0.9 - 0.6 = 0.3$ standard deviation better than Student A on the SAT.

It is of course worth stepping back for a moment to again recognize that their conversation was silly in the first place.

Exercise 1.6.2 A statistics teacher has decided that, to better facilitate his students' understanding of z-scores, he will not put their actual test grades on their test papers when he hands them back. Instead, he will provide them with only a z-score, followed by providing the class with the class average and standard deviation. Suppose that on a certain test, the class average was 82 and the standard deviation was 7.4.

(a) Joel received a score of $z = 0$ and initially exclaimed, "I thought I did alright; how did I get a 0?!?" What did Joel actually score?

(b) Marina's test came back with $z = 2.43$ and "GREAT JOB!" written on it. What did Marina get?

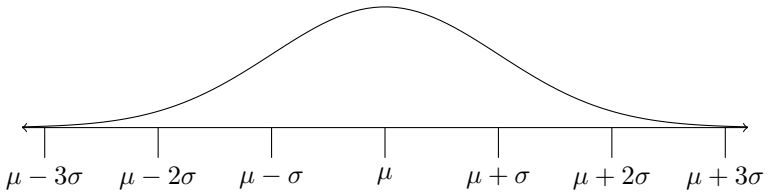
Exercise 1.6.3 A National Football League (NFL) franchise is looking at which of two players, Player 1 and Player 2, who play the same position, to draft. The team values size and speed equally. Nearly all of Player 1 and Player 2's measurements were identical except for two: arm length and 40-yard dash time. Player 1 had an arm length of 32 inches and a 40-yard dash time of 4.52 seconds; Player 2 had an arm length of 31 inches and a 40-yard dash time of 4.43 seconds. Suppose that the distribution of arm lengths for all players at this position has a mean of 30.8 inches and a standard deviation of 0.8 inch and the distribution of 40-yard dash times for all players at this position has a mean of 4.5 seconds and a standard deviation of 0.06 second.

Based on the above information, which player should the team draft?

Standardized scores can be computed in *any* distribution, but they are of particular convenience for one very special (and inaptly named) kind of distribution.

Definition

A **Normal distribution** is a symmetric, unimodal distribution with a shape known as a bell curve. A Normal distribution with mean μ (Greek *mu*) and standard deviation σ (Greek *sigma*) is shown below.



The Normal distribution is actually a mathematical function[12]that obeys the **Empirical Rule**, which states that, for any normal distribution,

- Approximately 68% of observations lie within ± 1 standard deviation of the mean
- Approximately 95% of observations lie within ± 2 standard deviations of the mean
- Approximately 99.7% of observations lie within ± 3 standard deviations of the mean

There's a useful fact for drawing Normal distributions by hand: the values one standard deviation away on each side of the mean occur at the *inflection points* - where the curve changes from curved up to curved down or vice-versa.

Additionally, we should make a note about Normal distributions. These are known as density curves - it's like they are histograms, but with *infinitely many infinitesimally small intervals*. To compute the probability of any particular value/values occurring in a histogram, you just add up the heights of the bars: in a Normal distribution, though, you are adding up infinitesimally thin bars, and you're adding infinitely many of them. This is tantamount to computing *area*.

So, for instance, the area under the curve is about *0.68* between $\mu - \sigma$ and $\mu + \sigma$. This is equivalent to stating that $0.68 \to 68\%$ of observations are between $\mu - \sigma$ and $\mu + \sigma$.

Exercise 1.6.4 What do you think the area under an entire Normal curve is?

Example 1.6.5
Many biological variables are normally distributed in populations in nature. For instance, the weights of animals in any sizable population will almost inevitably be a normal distribution.

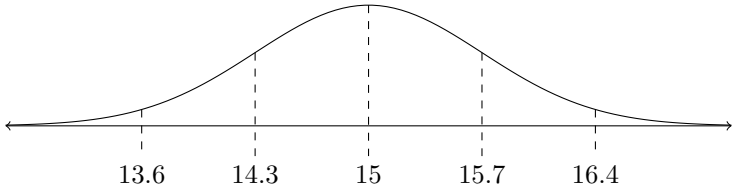
Suppose that the lengths of all adult female green anacondas are normally distributed with a mean of 15 feet and a standard deviation of 0.7 feet.

(a) Draw this distribution and label the values one and two standard deviations from the mean.

(b) Anna (sorry, couldn't resist) is an adult female green anaconda that is 16.4 feet long. Compute and interpret Anna's z-score.

(c) Approximately what percentage of adult female green anacondas are longer than Anna?

Solution (a) Our diagram is pictured below.



(b) $z = \frac{16.4-15}{0.7} = 2$. Anna is 2 standard deviations longer than the average adult female green anaconda.

(c) We need the *area* under the Normal curve to the *right* of 16.4. Since we know that 16.4 is 2 standard deviations above the mean, we can use the Empirical Rule and symmetry. We know that 95% of adult female green anacondas should have lengths between 13.6 and 16.4, which means that 5% should be shorter than 13.6 or longer than 16.4. The two leftover regions should have the same area, as the Normal distribution is symmetric. Therefore, the desired probability is $\frac{1}{2}(5\%) = 2.5\%$, or 0.025.

Anna the anaconda was one long snake! While the numbers provided thus far have been quite clean - nice whole numbers of standard deviations - getting an intuitive for the Normal distribution is a useful skill when it comes to interpreting computations that will later be made using technology.

Exercise 1.6.6 We defined standard deviation in an earlier lesson as " typical" deviation from the mean. Explain how this can be interpreted quantitatively in the context of a Normal distribution.

Exercise 1.6.7

(a) In any Normal distribution, what percentage of observations are greater than the mean?

(b) In any Normal distribution, what percentage of observations are greater than the mean, but less than 1 standard deviation above the mean?

(c) In any Normal distribution, what percentage of observations are more than one standard deviation below the mean?

Exercise 1.6.8 Recall that SAT scores are normally distributed with a mean of 1060 and a standard deviation of 200. What is the probability that a randomly selected student who's taken the SAT scored between 960 and 1260?

[12] The probability density function for a normal distribution with mean of μ and a standard deviation of σ was first discovered and formalized in the 18th and 19th centuries and is given by $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$.

§1.6 Homework

1. Explain how the process of computing a standardized z-score removes units.

2. Over the course of your life, you've likely taken many, many standardized tests (sadly). What about these tests do you think was "standardized," and what does this have in common with standardized z-scores?

3. Victor took a quiz and scored a 40 out of 45. The mean score for the quiz was 37 out of 45 with a standard deviation of 4.8. Calculate and interpret Victor's standardized score.

4. Two runners are comparing how much they won their recent races by. The first runner, who runs the 200 meter, says he won by 2 seconds. The second runner, who runs the 800 meter, says he also won by two seconds. One could argue that both runners did equally well, both winning by 2 of the same unit (seconds). Make a counterargument for why one runner's race was more impressive. Which runner?

5. Two of the lesser known breeds of salmon are chinook salmon, commonly found in Alaska and Oregon, and coho salmon, found along all of the Pacific Ocean. The distribution of weights of chinook salmon has a mean of 10 pounds and a standard deviation of 7 pounds, while the distribution of weights of coho salmon has a mean of 30 pounds and a standard deviation of 14 pounds. A fisherman catches a chinook salmon that weighs 27.5 pounds. What would the weight of a coho salmon, in pounds, need to be in order to have the same standardized z-score as the chinook salmon the fisherman caught?

6. At a certain store, the weights of oranges are approximately normally distributed with a mean of 140 grams and a standard deviation of 2 grams. Approximately what proportion of oranges in this store have a weight...

   (a) greater than 140 grams?

   (b) greater than 142 grams?

   (c) less than 144 grams?

   (d) between 138 and 146 grams?

7. Isaac and Alexa are arguing about the following problem.

   Home prices in a certain town have a mean of \$200,000 and a standard deviation of \$80,000. Approximately what percentage of homes cost less than \$280,000?

   Isaac says the answer is ~ 0.84, and Alexa says it's impossible to answer without more information.

   (a) Where did Isaac get his answer from?

   (b) Where did Alexa get her (lack of an) answer from?

   (c) Who is correct?

§ 1.7 More Normal Computations

We now know how to compute areas under normal curves when we have nice, clean standardized z-scores like $0, 1, 2$, etc. What if we can't rely on the Empirical Rule, though?

In comes math (and technology)! To find the area underneath a curve is a procedure called integration, which you will learn about in calculus. Not all integrals are easily computable by hand, but calculators and computers can compute integrals of even nasty-looking functions like $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$, the normal probability density (parent) function, in a second.

TI-84 Skills

To compute the area under a normal curve, press `2nd stat` (`vars`) and select `2:normalcdf` (the cdf stands for *cumulative density function*). You will then input four arguments:

`lower:` Lower value (if none, just type really negative number)

`upper:` Upper value (if none, just type really big number)

μ: Mean

σ: Standard Deviation

Example 1.7.1
Do part (b) of #7 from the last section's homework using a TI-84.

Solution
We want the area underneath the normal curve for values *above* 142, so we put a lower bound of 142 and an upper bound of 10,000.

`normalcdf(140,10000,140,2)` ≈ 0.159

Exercise 1.7.2 Do (a), (c), and (d) from #7 of the last section's homework using a TI-84.

We can also compute probabilities in normal distributions utilizing Excel.

Excel Skills

In Excel, the functions NORM.DIST and NORM.S.DIST will compute areas underneath *any* normal distribution and the *standard* normal distribution, respectively. The inputs are x, mean, standard deviation, where x is the value of interest. However, unlike `normalcdf`, it only computes areas under the curve to the left (below) the given x value. So, in essence, you can only input an upper bound.

Furthermore, there is a fourth argument: whether to set Cumulative to True or False. For cumulative areas under the curve, you should set Cumulative to True.

Utilizing Excel to compute normal probabilities requires precision and a bit of creativity at times. First off, you must know whether your distribution is standardized or not. If it is already standardized, or you standardize a value yourself, you'll use NORM.S.DIST; If not, you'll use NORM.DIST. For the creativity part, it's worth trying your hand at it.

Exercise 1.7.3 Consider a normal distribution with a mean of 20 and a standard deviation of 4. How would you compute...

(a) the proportion of values in the distribution *above* 15?

(b) the proportion of values in the distribution *between* 15 and 22?

Example 1.7.4
Suppose that the number of AP courses taken by seniors at a certain high school is approximately normally distributed with mean $\mu = 8$ and standard deviation $\sigma = 3.6$. Use Excel to answer the following questions. Assume all students being discussed are seniors at this high school.

(a) Approximately what percentage of seniors have taken 10 or fewer AP courses?

(b) Laura has taken 12 AP courses. What percentage of students have taken more?

(c) Approximately what percentage of seniors have taken between 3 and 5 AP courses?

Solution

(a) In this case, 10 is our upper bound, so we can simply type =NORM.DIST(10,8,3.6,true) into Excel. We obtain a proportion of 0.711, so our answer is that approximately 71.1% of seniors have taken 10 or fewer AP courses.

(b) We want the area under the normal curve *above* 12, so we must compute a cumulative probability *below* a value and subtract from 1. Depending on your perspective, we could use an upper bound of 12 or 13 for our computation; we choose to use 12 here. We type =NORM.DIST(12,8,3.6,true and obtain ∼0.867; in a new cell, we can type =1- and then whatever cell the previous value is in. We get a value of 0.133, so approximately 13.3% of seniors at this school have taken more than 12 AP courses. (Note: It would be faster to realize what we need to do and just type =1-NORM.DIST(12,8,3.6,true) right from the start - it saves a cell, too!)

(c) To get the area under the curve *between* two values, we will find the area under the curve to the left of both 5 and 3 and will subtract the values we get. Therefore, we type the most concise command possible, =NORM.DIST(5,8,3.6,true) - NORM.DIST(3,8,3.6,true), and get ∼0.12. Therefore, approximately 12% of seniors at this school have taken between 3 and 5 AP courses.

Exercise 1.7.5 For each part of Example 1.7.4, compute the standardized values and then use NORM.S.DIST to find the corresponding percentages.

When research studies - observational studies or experiments - are conducted, sample data is often standardized to obtain what is called a test statistic.

Definition

A **test statistic** is a standardized score that measures how many standard deviations away a sample statistic is from the assumed value of the parameter.

Generally, the more *unusual* or *unlikely* a test statistic is, the better evidence the researcher has for whatever hypothesis they are looking to demonstrate. Think of it like a criminal case against someone who, say, is accused of having stolen art from a museum. In the U.S. justice system, this person is presumed innocent - therefore, the "assumed parameter" is not guilty. The "researcher," or the jury, is looking to determine whether a different hypothesis is true: that the person is guilty. Evidence is gathered and presented. If the evidence is not unusual - say the accused person is the same height as the person who stole the artwork - this isn't considered enough to have demonstrated the person's guilt. However, if the evidence is very unusual - say the accused person's fingerprints were left on the wall next to the stolen art - then it is a much stronger demonstration of their guilt.

Exercise 1.7.6
Which of the following test statistics do you think are particularly unusual or unlikely?

(a) $z = 0$ (b) $z = 1.8$ (c) $z = 2.5$ (d) $z = -4.1$

Which one would give the *best* evidence?

Historically, a sample statistic that is less than 5% likely to occur is considered too unlikely to have occurred by chance given whatever assumed value of the parameter. When such a statistic is obtained, it's called statistically significant.

Definition

When a sample statistic is highly unlikely to have occurred by chance, it is called **statistically significant**.

Exercise 1.7.7 Based on the historical 5% rule, how many standard deviations away from the assumed parameter must a statistic be to be considered statistically significant?

As we move on, we'll see these ideas and terms pop up repeatedly. The questions of "What's unusual?" and "Is this unusual enough?" are surprisingly vital to just about any scientific research.

§1.7 Homework

1. At a fast food restaurant, the soda machine has a button that can be pressed to automatically dispense liquid for a 16-ounce drink. Suppose that the amount of liquid distributed by the machine is approximately Normally distributed with a mean of 15.9 and a standard deviation of 0.06. Approximately what percentage of the time does the machine overfill a drink?

2. Without typing it in, determine the approximate value of what the function

   `=NORM.S.DIST(1,TRUE) - NORM.S.DIST(-1,TRUE)`

   would output in Excel.

3. Find the area under a standard Normal curve between $z = 0.5$ and $z = 0.8$.

4. Referring back to Exercise 1.6.2 (p.24), a statistics teacher gives each of his students a z-score instead of their actual score on a test. The test average was 82 and the standard deviation of 7.4. Christina scored a 90 and, after computing `=NORM.DIST(90,82,7.4,TRUE)` = `0.86`, says that she beat 86% of the class. Do you agree?

5. Up to now, we have only computed areas under Normal curves given values in a Normal distribution; what if we wanted to go the other way?

   Recall that SAT scores are approximately Normally distributed with a mean of 1060 and a standard deviation of 200.

   (a) Use your calculator or Excel to estimate what SAT score would be required in order to score at the 90th percentile. What was your process?

   (b) Use your calculator or Excel to estimate the IQR of the distribution of SAT scores. What was your process?

§ 1.8 Activity: Smelling Parkinson's

As reported by the Washington Post[13], Joy Milne of Perth, United Kingdom, smelled a "subtle musky odor" on her husband Les that she had never smelled before. At first, Joy thought maybe it was just from the sweat after long hours of work. But when Les was diagnosed with Parkinson's disease 6 years later, Joy suspected the odor might be a result of the disease.

Scientists were intrigued by Joy's claim and designed an experiment to test her ability to "smell Parkinson's." Joy was presented with 12 different shirts, each worn by a different person, some of whom had Parkinson's and some of whom did not. The shirts were given to Joy in a random order and she had to decide whether each shirt was worn by a Parkinson's patient or not.

Exercise 1.8.1 Why should we care whether someone can smell Parkinson's disease?

Exercise 1.8.2 How many correct decisions (out of 12) would you expect Joy make if she *couldn't* really smell Parkinson's disease and was just guessing?

Exercise 1.8.3 How many correct decisions (out of 12) would it take to *convince* you that Joy wasn't guessing?

Although the researchers wanted to believe Joy, there was a chance that she may not really be able to tell Parkinson's by smell. Before they were willing to commit to a larger and more expensive investigation, they needed to be convinced Joy wasn't just guessing.

Definition

When researchers have a claim they suspect is true or are looking for evidence of, it's called the **alternative hypothesis** H_a. To test such a claim, researchers must assume that what they suspect is *not* true: this assumption is called the **null hypothesis** H_0.

In this case, the parameter of interest is a population proportion, which we might call p. Here, $p =$ true proportion of times Joy can correctly identify whether a shirt belongs to someone with Parkinson's or not.

Exercise 1.8.4 What is the alternative hypothesis H_a in terms of p?

Exercise 1.8.5 What is the null hypothesis H_0 in terms of p?

After the researchers showed Joy the shirts in a randomized order and determined whether she was correct for each shirt, it was found that Joy got... *11* of her 12 guesses correct.

Exercise 1.8.6 Does this give evidence for the alternative hypothesis or not?

Exercise 1.8.7 There are two possible explanations for how Joy could have identified so many shirts correctly. What are they?

When researchers are faced with evidence like Joy's getting 11 of 12 correctly, they must determine how *likely* it is that they are wrong - that the null hypothesis is *true* - and that the sample results just occurred by chance. The resulting value is called a P-value.

Definition

Assuming that the null hypothesis is true, the **P-value** of a study is the probability that evidence as strong as that observed in the sample/experiment occurred by chance.

The question is, then: what is the value of P for this study? To estimate it, we will conduct a simulation.

Definition

A **simulation** is a process which models a chance process with matching probabilities.

[13] Credit for this activity goes to Doug Tyson. An online version of the simulation is available at `https://stapplet.com/parkinsons.html`

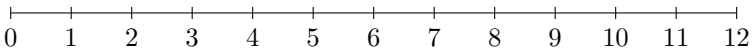
Why a simulation? Well, we don't have 6 Parkinson's patients' shirts on us! Furthermore, we don't actually need the shirts to model the probabilities involved. The probability that Joy is correct *if she is guessing* is 0.5 and the probability she is incorrect is 0.5.

Exercise 1.8.8 What could we use to simulate "sniffing 12 shirts"?

There are many, many ways we could indeed simulate such a simple random process, but we'll go with arguably the easiest: flipping a coin.

Exercise 1.8.9 Let each coin flip represent sniffing a shirt. If the coin shows heads, consider this a correct guess; if the coin shows tails, consider this an incorrect guess. Flip your coin 12 times and record the number of correct guesses.

In order to get a sense of just how unlikely it would've been for Joy to get 11 or more correct by guessing, we'll make a class dotplot.



Exercise 1.8.10 Based on this small-scale simulation, what proportion of the simulations resulted in 11 or more shirts correctly identified, assuming the person was guessing? In other words, what is the simulated P-value?

Exercise 1.8.11 Is the P-value low enough for you to be convinced that Joy *wasn't* guessing?

Usually, simulations aren't conducted by flipping a coin - they're done with a computer - and they are conducted thousands and thousands of times, rather than only 15 or 20.

Exercise 1.8.12 As a class, we will go to `https://stapplet.com/parkinsons.html` and conduct 1,000 simulations. What proportion of these resulted in 11 or more correct guesses?

When the likelihood that an observed statistic would occur if the null hypothesis was true is incredibly unlikely, it means that we have incredibly strong evidence for the alternative hypothesis. In this case, we would say we reject the null hypothesis - there is indeed statistically significant evidence that Joy really could "smell Parkinson's."

In fact, the one shirt Joy got wrong was from a patient who did not have Parkinson's (and Joy guessed they did)... but this patient did eventually get diagnosed with Parkinson's. Joy therefore actually guessed *all 12 shirts correctly.* Today, new research is being done to utilize skin-based biomarkers to detect Parkinson's years before symptoms appear.

---

§1.8 Homework

1. A tire company is testing out how long, on average, their tires last (in miles). They want to advertise that their tires last more than 60,000 miles on average, so they conduct testing on a set of 50 tires. The sample mean is 60,300 miles. Let the parameter of interest be μ = true mean number of miles all of these tires last. (μ is Greek "mu.")

   (a) What are the null and alternative hypotheses for this study?

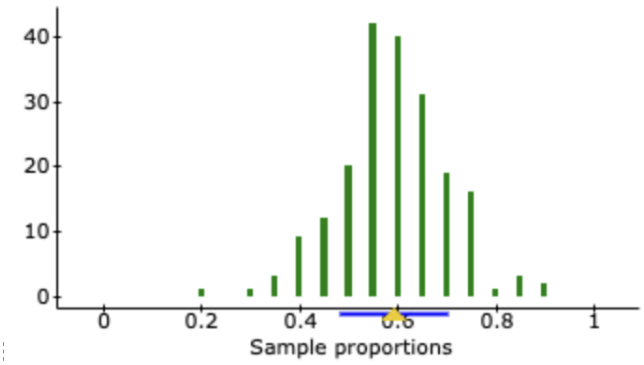
   (b) The study resulted in a P-value of 0.082. What does this mean?

   (c) Do you think the study produced convincing evidence that the tires last on average more than 60,000 miles?

2. A basketball card company sells boxes that, the claim, give a 40% chance of getting an autograph in. An excited young fan buys 10 boxes and doesn't get a single autograph. He thinks the company is lying about the 40%. Describe a simulation the young fan could conduct to estimate the likelihood of not getting a single autograph in 10 boxes if the company is telling the truth.

3. In Exercise 1.8.2, you identified a certain number of correct guesses (out of 12) that it would take to convince you that someone could smell Parkinson's. Based on the dotplot made in class, would you revise your answer? If so, to what?

4. Erika plays basketball, but her friends tease her about her poor free throw shooting. Last season, Erika shot 60% on her free throws. Over the summer, Erika says she worked on it a lot, so her friends decide to test her out. They have her shoot 20 free throws, and she makes 16 of them. A simulation was conducted to assess possible outcomes in 20 trials of a chance process with a 60% chance of success. Shown below is a graph of the outcome of 200 such simulations.



Based on the simulations, is there statistically significant evidence that Erika did in fact improve?

§ 1.9 Sampling Distributions

In the Smelling Parkinson's activity, we tested the hypothesis that Joy Milne was able to smell Parkinson's by comparing her result - 11 of 12 correct - with *many possible* results of the same chance process. This is a very common idea in statistics: you get sample data, but you don't know if it's ordinary or extreme. In order to determine this, you must compare it with all *possible* sample data.

Definition

Suppose you compute a statistic x from a random sample of size n. Then the **sampling distribution of** x is the distribution of x for all possible random samples of size n.

The study of sampling distributions is vital to statistics because statistics *vary* from sample to sample. Only from the study of all possible values of the statistics can one glean information about a particular value of the statistic.

Example 1.9.1
The heights (in inches) of the 5 starting players for a basketball team are given below.

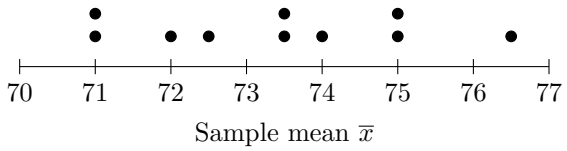
Player	1	2	3	4	5
Height (in)	70	72	72	75	78

Suppose we randomly sample 2 players and compute the sample mean height $\overline{x}$. Draw a dotplot of the sampling distribution of $\overline{x}$.

Solution
We need to look at every possible $\overline{x}$ we could get. There are 10 possible samples of 2 players, each of which will have its own sample mean $\overline{x}$.

Sample Players	$\overline{x}$	Sample Players	$\overline{x}$
1, 2	71	2, 4	73.5
1, 3	71	2, 5	75
1, 4	72.5	3, 4	73.5
1, 5	74	3, 5	75
2, 3	72	4, 5	76.5

The sampling distribution is shown in the dotplot below.



Ordinarily, sampling distributions involve populations much larger than 5 individuals, and therefore the possible numbers of samples are typically in the millions, billions, and beyond. Technology - and math - are helpful in at least estimating what these distributions look like.

Exercise 1.9.2 A midwest college is interested in estimating the average cumulative GPA μ of all 2,919 of its students. Rather than go compute the average of all 2,919 values, registrars decide to just take a post a flyer asking 20 students to anonymously fill out an online survey with their GPA. The registrars obtain a sample with a mean GPA of $\overline{x} = 3.5$.

(a) Why do you think the registrars might have decided to have students fill out a survey rather than conduct a random sample?

(b) Is it reasonable to infer the sample mean GPA $\overline{x} = 3.5$ to the population of all students at this college? Why or why not?

For the college GPA data, we can actually simulate what the sampling distribution looks like.

Exercise 1.9.3 Go to `rossmanchance.com/applets` and select Sampling Words. Click the drop-down menu that says "Gettysburg" and choose "College Midwest." Then, check the box to the right that says Show Sampling Options. Put 1 for Number of Samples and 20 for Sample Size. Press Draw Samples.

(a) What sample mean $\overline{x}$ did you obtain? Is this the same or different than the one obtained by the registrars?

(b) Press Draw Samples repeatedly until you've taken 50 samples. Describe what you see. What does each blue (then gray) box at the bottom right represent?

(c) Now, input 950 in Number of Samples. Press Draw Samples. You should now have 1,000 different sample means. Is the graph you see the exact sampling distribution of $\overline{x}$?

(d) The registrars claim that the students who filled out the survey are a reasonably random, and therefore representative, sample of all students at the college. Use the Count Samples: Greater than $\geq$ to input a value to assess this claim.

Sampling distributions are a very complex topic because of how abstract they are, but simulation can help us to understand exactly what they represent and how they can be used to assess the likelihood that a certain kind of statistic could occur by chance. In future statistics courses, you may learn about the math that enables us to *precisely* describe many sampling distributions without having to use simulations.

Exercise 1.9.4 As a class, choose something quantitative to measure for each student in the class (height, shoe size, hours spent on homework last night, etc.).

(a) Measure the value for each student in the class.

(b) Type all of the class data into a column in Excel.

(c) Go back to the Rossman Chance website and hit Clear below the Midwest College data in the upper left.

(d) Highlight all of the class data in Excel and copy it (Ctrl-C will copy). Then, paste it into the box on the Rossman Chance site (Ctrl-V will paste).

(e) Take 500 samples of size 2. What are the shape, mean, and standard deviation of the simulated sampling distribution?

(f) Now, take 500 samples of size 5. What are the shape, mean, and standard deviation of the simulated sampling distribution?

(g) Predict what the simulated sampling distribution will look like if you take samples of size 10.

(h) Test your predictions from (g).

(i) Go back to Excel and compute the actual average for all students in the class. What do you notice about this value?

(j) Complete the sentence: *The mean of the sampling distribution of any sample statistic is equal to* _____.

---

§1.9 Homework

1. Describe in your own words why sampling distributions are useful to study.

2. Suppose a population distribution is right skewed with a mean of 8.08.

   (a) What do you think the shape and mean of the sampling distribution for random samples of size 2 from this population would be?

(b) What do you think the shape and mean of the sampling distribution for random samples of size 5 from this population would be?

(c) What do you think the shape and mean of the sampling distribution for random samples of size 20 from this population would be?

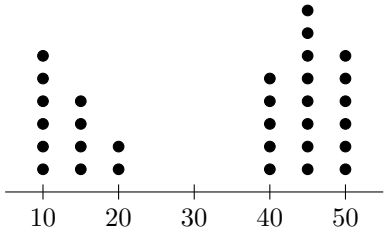
(d) Go to `https://onlinestatbook.com/stat_sim/sampling_dist/`, hit Begin, and then select "Skewed" from the dropdown menu to the right of the topmost graph.

(i) On the third graph, select N=2. Press Animated a few times to get a feel for the applet. Then, press 10,000. Were your answers to (a) correct?

(ii) Now, on the third graph, select N=5. Press Animated a few times again. Then, press 10,000. Were your answers to (b) correct?

(iii) On the third graph, select N=20. Press Animated a few times again. Then, press 10,000. Were your answers to (c) correct?

3. Shown below are quiz scores from a recent periodic table quiz that a chemistry teacher gave. Scores are out of 50.



(a) Describe the shape, central, and variability of the distribution of quiz scores.

(b) Describe loosely what the sampling distribution of the sample mean for random samples of size 2 from the population of quiz scores would look like.

4. Go back to the Rossman Chance site and choose the "Gettysburg" dataset. This is a list of 268 words from the Gettysburg Address. A student was interested in the average word length in this portion of the Gettysburg Address, so they decided to sample 5 words from it. They obtained an average word length of 6.6 letters.

Use the simulated sampling distribution of the sample mean for $n = 5$ to assess whether this student's estimate is reasonable. Do you think they selected a random sample? Why or why not?

§ 1.10 Confidence Intervals, part 1

In the previous lesson, we discussed sampling distributions, which describe every *possible* value of a statistic from a random sample of a certain size. Beyond simply being able to determine whether a statistic will be plausible or not, we can actually use the theory behind sampling distributions to get really reliable estimates of population parameters *without* having to simulate.

The reason we are able to do this is because of one simple fact: most sampling distributions are approximately normal. This remarkable fact takes many forms, but one is the Central Limit Theorem, which you will learn about in a future statistics course. For now, it will suffice to know that, for sufficiently large sample sizes, sampling distributions will be approximately normal.

Let's do an extensive example to try to answer one question: what was the average birth weight, in pounds, of babies born in the U.S. in 2021?

Data was obtained from the National Vital Statistics System of the National Center for Statistics[14] on all births in the United States in 2021. However, the data would not *fit* in an Excel file, as Excel only contains up to 1,048,576 rows. Rather than compute the average weight of all million plus babies in the sample, we could instead obtain a random sample of 50 babies. Shown below are the weights, in grams, of 50 randomly selected babies.

3600	3775	2863	2650	3170	3255	3912	2920	3204	3400
3005	3130	3590	3910	1820	2977	3374	3909	2925	3350
3450	3090	3070	3033	3138	3030	3799	3550	3480	3585
2971	2675	4890	2625	3015	3485	2870	3002	3402	3605
2720	2948	3475	2910	2410	2340	3599	3390	3580	2637

Exercise 1.10.1 Use a TI-84 or Excel to compute the sample mean weight of the 50 babies in pounds (1 pound = 454.5 grams).

Exercise 1.10.2 Do you think the sample mean from (a) is exactly equal to the mean weight of all babies born in the U.S. in 2021? Do you think it's close?

With any statistic comes the idea that a *different* sample would've resulted in a *different* statistic. The likelihood our sample mean is *exactly* correct is incredibly low - rather, we hope to be close. This leads to the concept of margin of error.

Definition

The margin of error is a value that is added and subtracted to a sample statistic to create a range of plausible[15] values for the parameter.

Margin of error is often presented as a plus or minus value. For instance, if we selected a margin of error of 1 pound, our range of plausible values would be 7.06 ± 1, or 6.06 to 8.06 pounds.

Exercise 1.10.3

(a) Describe an advantage of having a *small* margin of error.

(b) Describe an advantage of having a *large* margin of error.

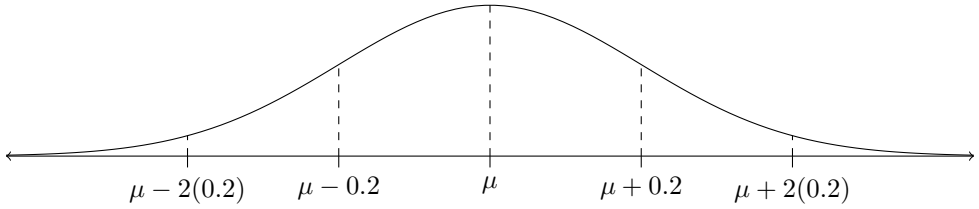
Exercise 1.10.4 How large would you make the margin of error?

The size of the margin of error depends on how *confident* we want to be that our range of values will contain the parameter. If we use a huge margin of error, then we will be very confident - 100% even, maybe - that our interval will capture the parameter. This comes at the cost of precision, though: how would you feel if a teacher told you they were 100% confident that you got between a 0 and a 100 on a test? While accurate, would you feel you had a good idea of what you scored?

[14] `https://www.nber.org/research/data/vital-statistics-natality-birth-data`

[15] *Plausible* means reasonable - not necessarily likely, but also not so unlikely as to be unreasonable.

To determine how confident we can be, we return to our old friend: sampling distributions. Through some computations that are, admittedly, beyond the scope of this course, we can determine that the sampling distribution of all *possible* sample means from 50 babies looks like that pictured below.



From a previous homework question, we know that the mean of all these possible sample means should be *exactly* the population mean: μ is therefore the mean weight of *all* babies born in the U.S. in 2021. The question is, where in this distribution is *our* sample mean of 7.06? Ideally, we'd like to be *confident* that our sample mean is within a certain distance of the population mean.

This is where Normality comes in!

Exercise 1.10.5 Shade a region under the curve that you are *95%* confident that our sample mean is in. How do you know?

Because of the Empirical Rule, we can always be 95% confident that our sample statistic will be within roughly 2 standard deviations of the mean. This means that a margin of error of roughly 2 standard deviations will give us a 95% chance of capturing the parameter! This can enable us to construct our range of plausible values with 95% confidence.

Definition

A **C% confidence interval** is a range of plausible values for the parameter. The C% means that, if we took many of these samples and constructed a C% interval for each one, approximately C% of the intervals would indeed capture the parameter.

For our data, we can therefore construct a 95% confidence interval by adding and subtracting two standard deviations of the sampling distributions.

$$7.06 \pm 2(0.2) \Rightarrow 7.06 \pm 0.4 \Rightarrow (6.66, 7.46)$$

How do we interpret this? We can simply state

We are 95% confident that the mean weight of all babies born in the U.S. in 2021 is between 6.66 and 7.06 pounds.

Exercise 1.10.6 To get a feel for what exactly the 95% means in "95% confidence", go to the applet at `https://www.geogebra.org/m/m6kjsr5z`.

(a) Press Generate new sample mean. Did the new interval capture the parameter? Record Y if yes and N if no.

(b) Press Generate new sample mean 49 more times. For each generated interval, record a Y if the interval captured the parameter and a N if it didn't. Feel free to use the table below.

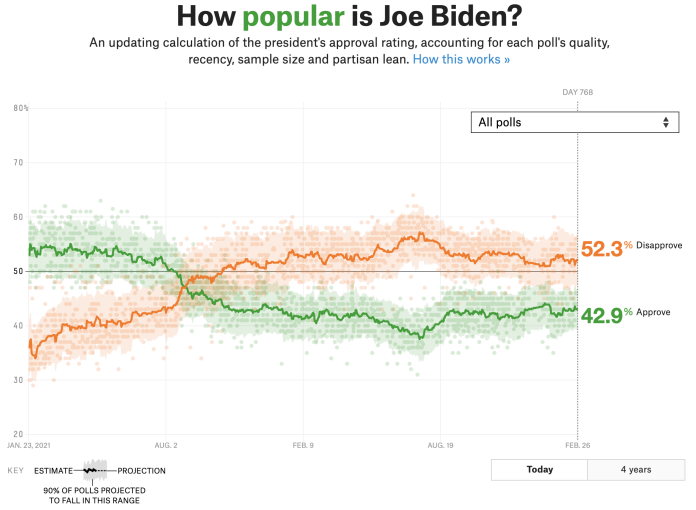
(c) Out of your 50 samples and corresponding intervals, how many captured the parameter? What percentage is this?

(d) If you generated 10,000 samples and each corresponding intervals, what percentage do you think would capture the parameter?

Confidence intervals are incredibly powerful estimation tools: they allow us to account for the sample-to-sample variability by adding and subtracting a margin of error. The normality of sampling distributions means that we can add precisely the margin of error that will give us a certain level of confidence.

Let's look at an example in context.

Example 1.10.7
The popular polling and all-things-statistics website FiveThirtyEight.com[16]maintains a "How Popular/Unpopular" is Joe Biden site, which keeps track of a number of presidential approval rating polls. Shown below is the graphic on the site on February 26, 2023 (when this was written - feel free to check it out today!).



(a) What does the dark orange squiggly line represent?

(b) What does the the shaded orange area represent?

(c) What do you think the margin of error of these aggregated polls is?

(d) Based on the graphic, is it plausible that less than 50% of people disapprove of Joe Biden?

(e) Based on the graphic, is it plausible that more than 50% of people approve of Joe Biden?

(f) The percentages for approval and disapproval don't add up to 100%. Why do you think this is?

Solution

(a) Based on the graphic in the bottom left, the orange squiggly line represents the estimate for what percentage of people disapprove of Joe Biden.

(b) Based on the graphic in the bottom left, the shaded orange area represents the "projection," or the margin of error, of these polls.

(c) The height from the orange squiggly line to the peak of the orange shaded area on any given day seems to hover around 5-7%, so this is the margin of error.

(d) Yes. 50% is within the orange shaded area.

(e) No. 50% is not within the green shaded area.

(f) Well, we'd need to read more about these polls! One reasonable guess is that some individuals stated (or were given the option to state) that they neither approve nor disapprove.

[16] `https://projects.fivethirtyeight.com/biden-approval-rating/`

§1.10 Homework

1. A poll was conducted to determine what percentage of likely voters plan on voting for Candidate A. The poll results were: "52% of likely voters plan on voting for Candidate A. The margin of error was 2.5 percentage points at 95% confidence."

   (a) Construct the actual confidence interval from these results.

   (b) Interpret the interval.

   (c) Is it plausible that less than half of likely voters will vote for Candidate A? Explain.

   (d) Someone reads the poll results on election morning and says, "Oh, there's a 95% chance they are going to win." Explain why this is an incorrect interpretation of the interval and what the actual meaning of the 95% is.

2. Go to the following link for a recent Marist poll: `https://tinyurl.com/4rjtc6ua`

   (a) What question was being asked?

   (b) Identify a parameter (or parameters) of interest.

   (c) What was the sample size?

   (d) Identify the margin of error of the poll.

   (e) Construct a confidence interval for the percentage of Americans that are more excited about the Super Bowl than Valentine's Day.

   (f) In a romantic dispute about why one partner is more excited about the Super Bowl than taking the other partner out for Valentine's Day, the partner says, "Oh, come on, like half the country cares more about the Super Bowl!" Does your interval from (d) back up this claim?

3. A student is conducting a research project in which they measure the growth of plants, in inches, after 6 months of being in soil with a particular fertilizer. They find that the mean amount of growth for their 40 plants was 9.6 inches. Assume that the standard deviation of the sampling distribution of all possible means of 40 plants is 0.44 inch.

   (a) Construct a 95% confidence interval for the mean amount of growth of plants after 6 months in soil with this fertilizer.

   (b) Interpret the interval from (a).

   (c) If the student instead were to compute a 99% confidence interval, do you think the interval would be wider or narrower than the interval in (a)? (Note: You don't have to (and don't know how to) compute the new interval to answer this.)

   (d) If the student had instead used 100 plants and then constructed a 95% confidence interval, do you think the interval would be wider or narrower than the interval in (a)?

§ 1.11 Confidence Intervals, part 2

To begin this lesson, you're going to investigate the effects that sample size and the confidence level have on the margin of error of an interval. As our setting, suppose someone is going to conduct a survey estimating the proportion of all people in Georgia who are going to vote for a Republican candidate for governor in 2024. Additionally, for the purposes of the applet, suppose the true proportion is 0.5 (or 50% of people).

Keep in mind with this applet that this is measuring *proportion* of voters, so 0.03 is 3%, 0.1 is 10%, and so on.

Go to the link at: `https://www.geogebra.org/m/sgbvvg3f`.

Exercise 1.11.1

(a) The applet starts with a sample size of $n = 5$ and a confidence level of $C = 90\%$. Approximately what is the margin of error? Do you think this is good or bad?

(b) Play around with the confidence level C using the slider. What effect does increasing or decreasing C have on the margin of error? How does this make sense?

(c) Play around with the sample size n using the slider. What effect does increasing or decreasing sample size have on the margin of error? How does this make sense? margin

(d) You likely observed in (d) that increasing sample size decreases the margin of error. Describe the rate that this occurs at as n increases.

(e) Suppose you wanted to make a poll that would have a margin of error of no more than 5% at 95% confidence. What sample sizes will achieve this?

(f) Obviously, smaller margins of error are better - the smaller margin of error, the more precise the estimate will be. However, nearly every political poll uses sample sizes of anywhere from 500 to 2,000. Based on the applet and your intuition, why do you think this is?

(g) Suppose the polling institution wants a margin of error of no more than 2.5%, and they want a 99% confidence interval. You're tasked with getting them the ideal sample size. What would you recommend?

Now, let's look at some various examples of confidence intervals in context.

Example 1.11.2

When people take any kind of assessment - IQ test, SAT, ACT, etc. - they are given a score. The question may remain, though - what if the person had a particularly good (or bad) day? Would they score the same on the test if they took it again the next day, or the day after that? The measure of how "repeatable" as test is, in terms of returning nearly the same score on repeated tests, is called *reliability*, and is calculated as a number r between 0 and 1, with $r = 0.7$ being adequate reliability and $r = 0.9$ being excellent reliability.

Reliability is used to compute what is called the *standard error of measurement* SE_m, which is given by the formula

$$SE_m = SD\sqrt{1 - r}$$

where SD is the standard deviation of the test scores for a suitably large sample size.

Estimates of the SAT have given a standard deviation of approximately 200 points and a reliability of $r \approx 0.91$. One particular student scores a 1200 on the SAT.

(a) Compute and interpret a 68% confidence interval for this student's score.

(b) Compute and interpret a 95% confidence interval for this student's score.

Solution
We first compute $SE_m = 200\sqrt{1-0.91} = 60$.

(a) A 68% interval requires adding and subtracting one SE_m from the statistic, so in this case, our interval is $1200 \pm 60 \to (1140, 1260)$. We are 68% confident that this student's "true" SAT score lies somewhere between 1140 and 1260.

(b) A 95% interval requires adding and subtracting two SE_m from the statistic, so in this case, our interval is $1200 \pm 2(60) \to (1080, 1320)$. We are 95% confident that this student's "true" SAT score lies somewhere between 1080 and 1320.

Exercise 1.11.3 If you were in charge of reporting SAT scores to students, would you choose to report a 68% or a 95% confidence interval, and why?

Exercise 1.11.4 Suppose the student is a junior who is determined to go to a university where 75% of students have at least a 1360 on the SAT. If you were this student, would you encouraged or discouraged by the intervals? How would you internalize these intervals?

Example 1.11.5
As of 1998, the World Health Organization (WHO) suggested that at-home systolic blood pressure readings of 140 milligrams (mg) or higher indicated hypertension or other cardiovascular disease. A study[17]conducted in 1998 involving 25 patients with "hypertensive small vessel disease" investigated the difference between at-home blood pressure readings versus readings calculated by physicians. Each patient conduct their own at-home systolic blood pressure test and then had a physician compute a systolic blood pressure test in a clinic. The 95% confidence interval for the at-home readings was 143.4 ± 13.6 mg, while a 95% confidence interval for the clinical readings was 169.2 ± 16.5 mg.

(a) What do you notice about the two confidence intervals?

(b) One cutoff used for determining whether someone is at high risk for developing cardiovascular disease is a systolic blood pressure of 160 mg.

(i) Based on the at-home confidence interval, is it plausible that the 25 patients were, on average, at high risk for cardiovascular disease?

(ii) Based on the clinical confidence interval, is it plausible that the 25 patients were, on average, at high risk for cardiovascular disease?

(iii) All 25 patients in this study did indeed have cardiovascular disease. What does that make you think about at-home readings when considering (i) and (ii)?

Exercise 1.11.6 As a result of the study from the previous example, it was recommended that the WHO change their definition of hypertension to 135 mg for patients using at-home monitors. As of 2023, some websites, like Heart.org[19], actually indicate that a systolic blood pressure of 130-139 indicates hypertension.

(a) Additionally, Heart.org recommends that people take multiple blood pressure readings at any given time.

(b) If you took your own blood pressure two times and saw that the systolic readings were off by 5 mg, would you be surprised or shocked? How about 10 mg?

Hopefully, you've gotten a feel for how ubiquitous confidence intervals are in the world at large - confidence intervals enable us to compute *ranges* of values that help us deal with the fact that measurements of all kinds are constantly changing. In future statistics courses, you'll learn more about different kinds of intervals, the origins of some of the formulas used, and just how such intervals can be computed with more complicated data.

[19]https://academic.oup.com/ajh/article/11/7/813/285964
[19]https://www.heart.org/en/health-topics/high-blood-pressure/understanding-blood-pressure-readings/monitoring-your-blood-pressure-at-home

§1.11 Homework

1. A company that makes deodorant wants to use focus groups to determine how the public would react to a new advertising campaign for the deodorant. Ava is in the marketing department and has to submit a proposal discussing how the focus groups will be obtained, the cost and time required to do so, and other relevant details.

   (a) Ava would like to obtain a random sample, but thinks this is impractical. Instead, she is going to create online postings where people can volunteer for the study. What are some concerns this might raise about the focus group, and how could this be addressed?

   (b) Ava is going to record, among other things, what percentage of the focus group says they would purchase the deodorant after watching the new ad. She will construct a confidence interval to estimate the percentage of *all* people who would do so.

      (i) What confidence level would you recommend for Ava?

      (ii) Ava is debating between getting 50 or 75 people for the focus group. Describe an advantage and a disadvantage of each sample size.

2. A 2005 Behavioral Risk Factor Survey in Wisconsin[20] resulted in a 95% confidence interval of 20.7 ± 1.1 for the percentage of all Wisconsin adults who smoke cigarettes.

   (a) Explain why the interpretations of the 95% below are *incorrect*.

      (i) There is a 95% chance that any future sample would result in the sample interval.

      (ii) There's a 95% chance that 20.7% of adults in Wisconsin in 2005 smoke cigarettes

   (b) Provide a correct interpretation of the 95%.

   (c) The same survey resulted in two separate 95% confidence intervals for the percentage of Wisconsin men and women who smoke, respectively. The intervals were 21.9 ± 1.8 and 19.4 ± 1.5 for men and women, respectively. Do you think there is a statistically significant difference between the percentages of Wisconsin men and women that smoke?

3. The fast food restaurant Wendy's has been advertising for a while now that people prefer their fries to McDonald's fries "2 to 1," essentially stating that 2/3 of individuals' prefer their fries to McDonald's fries.

   (a) What is the parameter of interest here? (There could be more than one correct answer.)

   (b) What process should Wendy's go through to back up a claim like this?

   (c) Do some research online and find the actual study. Report your findings in the next class period.

[20] `https://dhs.wisconsin.gov/wish/brfs/confidence-intervals.htm`

§ 1.12 Review

Overview Questions

NOTE: Even an AP Statistics or first-year college statistics student would have a hard time perfectly answering all of these questions. Yes, you want to try to communicate using precise statistical vocabulary, but it's okay if you at least understand the big-picture ideas.

1. What is the purpose of random sampling when estimating a parameter?
2. Why don't most experiments randomly sample?
3. Why do experiments random assign treatments and control for other variables?
4. What is bias, and what are some ways it can be reduced?
5. What's the difference between association (correlation) and causation?
6. What kinds of studies can infer causation, and why?
7. What is confounding? Give an example.
8. When is it reasonable to infer a statistic to a larger population?
9. What are some major characteristics to consider when describing the distribution of a quantitative variable?
10. Why are standardized z-scores so useful?
11. Do z-scores only apply to Normal distributions? If no, why are they so commonly tied to Normal distributions?
12. What is the sampling distribution of a statistic?
13. What is the role of sampling distributions when evaluating or analyzing a statistic?
14. What are null and alternative hypotheses, and how do we go about assessing evidence of an alternative hypothesis?
15. What is a P-value?
16. What is margin of error, and why does it exist?
17. What's the relationship between sample size and margin of error? How do researchers decide on how big of a sample size to use?
18. What is a $C\%$ confidence interval, and what does the C mean?

Skill Questions

19. Organizers of a business conference are curious how pleased conference attendees were with the keynote speaker. They give each attendee a paper survey that asks them to rate the speaker on a scale of 1 to 5, where 1 is "strongly dissatisfied" and 5 is "strongly satisfied." They will then compute the average satisfaction score. Of the 1540 attendees, 350 complete the survey, and the average satisfaction score of these surveys was 3.6.
    (a) Identify the population, parameter, sample, and statistic.
    (b) Explain how the sampling method could result in bias. Indicate whether you think the statistic might be an over- or under-estimate of the parameter.
20. Twenty students in a class were asked how many siblings they had. Results are shown below.

3	2	5	1	0	0	3	4	2	1
1	1	2	0	8	1	1	2	1	3

    (a) Find the mean, median, standard deviation, minimum, maximum, Q1, Q3, and range.

(b) Construct a dotplot of this data.

(c) Construct a histogram of this data.

(d) Describe the histogram you made in (c) in two to three sentences (in context).

21. The weights of Granny Smith apples at a certain orchard are approximately normally distributed with a mean of 155 grams and standard deviation of 15 grams.

(a) (No calculator) Approximately what percentage of Granny Smith apples at this orchard weigh between 140 and 170 grams?

(b) (No calculator) Approximately what percentage of Granny Smith apples at this orchard weigh more than 185 grams?

(c) (Calculator or Excel) A Granny Smith apple is randomly selected from this orchard. What's the probability it weighs less than 150 grams?

(d) (Calculator or Excel) A Granny Smith apple is randomly selected from this orchard. What's the probability it weighs more than 145 grams?

(e) (Calculator or Excel) A certain store will only accept Granny Smith apples that weigh between 150 and 165 grams. Approximately what percentage of this orchard's apples will be accepted by the store?

Application Questions

22. Suppose you want to conduct a research study investigating whether the college that a person attended has an influence on how likely they are to get a job working at a college. You decide to write 100 resumés that are completely identical, except 50 of the resumés state that the applicant attended College A and 50 state they attended College B.

(a) Your next job (no pun intended) is to determine which colleges to send the resumés to. Discuss how you might choose the 100 colleges. Why would you choose them this way?

(b) Next, you need to actually send out the 100 resumés. Discuss how you would decide which colleges received which resumés.

(c) What parameter/s would you be measuring after resumés are sent out?

23. Numerous studies have found that individuals that drive red cars are more likely to be involved in automobile accidents that individuals that drive non-red cars.

(a) Do you think these studies are observational studies or experiments? Explain.

(b) In response to some of these studies, it has been commented that the personality of the drivers may be a confounding variable. Explain what this means in context.

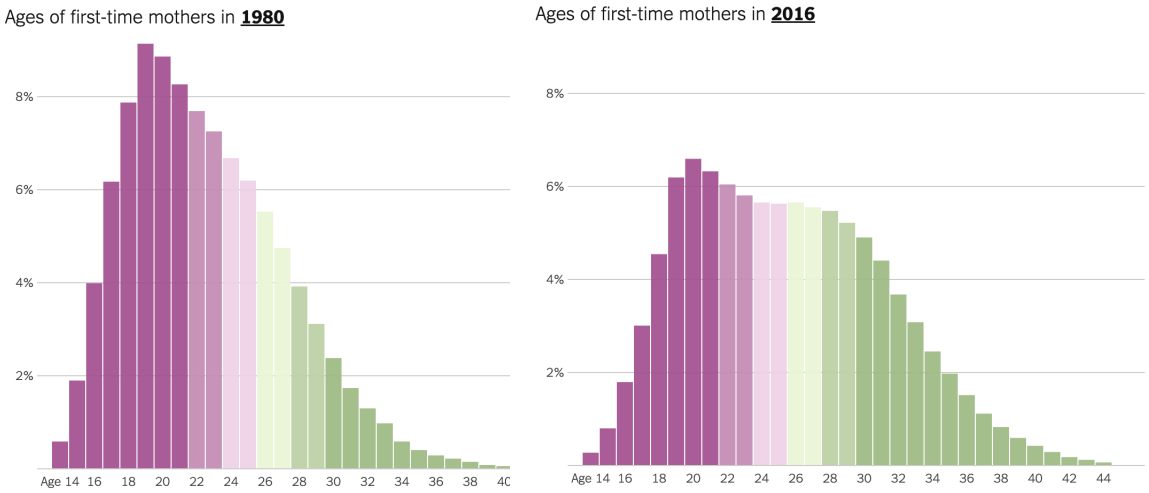
24. An SAT prep teacher has read some research that indicates that Khan Academy may produce more improvement in SAT scores than a traditional SAT prep book. They decide to investigate this by conducting an experiment using their 4 SAT prep classes.

(a) Describe the design of an experiment the teacher could use to test their hypotheses with the 4 classes.

(b) Describe some potential ethical concerns with the study from (a).

(c) The teacher is considering letting the students choose the prep program they prefer - online or book. While ultimately "fair," discuss how this might make it more difficult to determine which program is more effective at increasing SAT scores.

25. Shown below are histograms[21] of the average ages of first-time mothers in 1980 and 2016.



(a) Compare the distributions in context. (What's similar? What's different?)

(b) Do you think the mean age of first-time mothers is higher, lower, or about the same as the median age in 1980? How about 2016?

26. Suppose that the distribution of average spin rate, in revolutions per minute (rpm), for a certain Major League Baseball pitcher is approximately normally distributed with a mean of 2400 rpm and a standard deviation of 85 rpm.

(a) One game, the pitcher's average spin rate is 2500. Calculate and interpret the z-score of this value.

(b) In approximately what percentage of games will the pitcher throw a higher average spin rate than 2500?

(c) The pitcher suffers an injury, and after a certain amount of time off, returns to the mound. He pitches a game and his average spin rate is only 2250. The pitcher claims they are completely healthy, and that it just wasn't a great game, but commentators believe the pitcher is still nagged by injuries. Make a normal distribution calculation to assess the commentators' claim.

27. Every day in a math class, a teacher randomly selects 1 of 15 students to clean all of the whiteboards after class. One week, a student, Amy, is chosen two times in the span of 5 days. She feels that there's no way this should've happened if the teacher was truly picking randomly.

(a) Describe a simulation Amy could conduct to assess the likelihood that her being selected 2 or more times could occur just by chance.

(b) Carry out 100 trials of your simulation.

(c) Based on your simulation, does Amy have a legitimate complaint?

28. A law school boasts that 90% of its graduates secure a job by the time they graduate. One student at this law school thinks this is an overly optimistic claim. They consult the registrar and get a random sample of 50 students in their final year at the law school with the plans of calculating the sample percentage of students that have secured a job.

(a) What are some considerations the student should make before conducting their survey?

(b) What does "the sampling distribution of the sample percentage for 50 students" mean?

(c) Suppose the student obtains a sample percentage of 80, and that the standard deviation of the sampling distribution is approximately 5.5%. Compute and interpret a 95% confidence interval for the percentage of students at this law school that have secured a job.

(d) Based on the interval, is there statistically significant evidence that the law school's claim is overly optimistic?

[21] `https://www.nytimes.com/interactive/2018/08/04/upshot/up-birth-age-gap.html`

Homework Answers

§1.1

1. (a) Population: all 45,000 pounds of rice, parameter: true amount of aflatoxin in ppb, sample: 2-pounds of rice, statistic: 8 parts per billion
   (b) This should get a more representative sample of all 45,000 pounds than just sampling from one depth.
   (c) Convenience! What did you say about how reasonable it is?
   (d) Farmers may put the best rice on top, as this is what's visible.
2. (a) The actual percentage of all former college and NFL players' brains that have CTE.
   (b) The families that donated had noted cognitive decline, so the sample percentage of brains with CTE may be higher than that of the overall population.
3. (a) Theoretically, it would be all people (or all adults).
   (b) No, as the sample consisted only of people in Canada.
   (c) Partners who experienced more phubbing generally experienced less quality relationships.
4. What'd you say?
5. (a) The sample is the nearly 450,000 people who filled out the questionnaire.
   (b) Likely not. The sample size is enormous, and people aren't likely to lie about their coffee intake. Additionally, there's no obvious selection bias in the sampling method.
   (c) Percentage of participants who developed cardiovascular disease is one.
   (d) No. This was an observational study, not an experiment. We cannot conclude causation.

§1.2

1. (a) Starting time of school each day
   (b) Amount of hours slept, quarterly grades
   (c) Observational study
   (d) What'd you say?
2. People who consume more food within 2 hours of bedtime tend to have increased body fat.
3. What variables did you come up with?
4. (a) The teacher could give a pre-test, then let students use their device of choice (laptop, paper). Afterward, they can give a post-test and compare the differences in average improvements in test scores between the two groups.
   (b) The teacher can do the same thing as in (a), but randomly assign who uses laptop and who uses paper.
   (c) The design in (a) won't account for confounding variables that can't be directly controlled.

(d) Some students may really prefer one method and be assigned the other. Additionally, if the school does not provide laptops, equity may be a concern.

5. (a) The sample is the 252,300 people, and the population is theoretically all people (or all people in that nation).
   (b) What'd you say?

§1.3

1. Researchers should blind participants so that they don't know which beverage they are drinking, otherwise participants' expectations could produce confounding.
2. (a) The people on the first floor are much less likely to care about whether an elevator is installed; therefore, they may not be representative of all residents of the complex.
   (b) The benefit is that it guarantees people from each floor are selected, which is important since people on different floors may have different opinions about an elevator. The drawback is that it will take longer to obtain this sample.
   (c) Yes, as they are randomly selected from all residents of the complex.
3. (a) Experiment
   (b) No
   (c) So that the more and less likes groups would be as similar as possible with regards to variables that couldn't be directly controlled
   (d) What'd you say?
   (e) No, as the adolescents weren't randomly selected (though it's reasonable to infer to individuals "similar" to those in the study).
   (f) This is debatable!
4. These studies may not be inferable to humans!
5. The sample should be representative of the population.
6. Random assignment and direct control makes treatment groups as similar as possible so that any differences between the groups can theoretically be attributed to the differences in treatments.

§1.4

1. Students typically scored between 62 and 86.
2. A dotplot or histogram would be acceptable.
3. (a) The mean is higher than a large majority of the salaries.
   (b) Roughly $75,000; yes
   (c) The very high salaries dragged the mean upwards, but not the median.
4. What'd you write?

§1.5

1. (a) It's just a tremendous waste of space!
   (b) What was your suggestion? There are numerous acceptable answers.
2. What did you notice?
3. (a) People can write details that can't be captured by just a number.

   (b) Any spreadsheet result of the data will be harder to classify if there is free text (which could be a few words or a rambling paragraph!).

   (c) What do you think?

4. There are many good variables to stratify by - department, years at the company, gender, etc. - what'd you choose, and why?

5. This was all about what you think and for you and the teacher to discuss.

§1.6

1. How'd you describe it?

2. The tests are the same, or very close to it, for everyone.

3. $z = 0.625$; Victor scored 0.625 standard deviation above the mean

4. The first runner.

5. 65 pounds

6. (a) 0.5 (b) 0.16 (c) 0.975 (d) 0.8385

7. (a) Isaac assumed the distribution was approximately Normal.

   (b) Alexa realized the distribution wasn't necessarily Normal.

   (c) Alexa.

§1.7

1. 4.8%

2. ~ 0.68

3. 0.097

4. Not necessarily - we don't know if the distribution was Normal.

5. (a) The 90th percentile is 1389. (b) The IQR is about 270.

§1.8

1. (a) $H_0 : \mu = 60,000, H_a : \mu > 60,000$

   (b) Assuming the mean is actually 60,000, the probability of getting a sample mean as high or higher than 60,300 just by chance is about 0.082.

   (c) It depends on what you consider highly unlikely!

2. What'd you come up with?

3. A good argument could be made for 9 or 10 correct guesses.

4. Yes.

§1.9

1. What'd you say?
2. (a) Still somewhat right skewed
   (b) Maybe still right skewed, but less so
   (c) Starting to look approximately normal!
   (d) How'd you do?
3. The sampling distribution would be unimodal with a peak around 25-30.
4. It is very highly unlikely that the student selected a random sample.

§1.10

1. (a) $(49.5\%, 54.5\%)$
   (b) We are 95% confident that the true percentage of likely voters who will vote for Candidate A is between 49.5% and 54.5%.
   (c) Yes, as 50% is in the interval.
   (d) The 95% means that, if *many* samples were taken (and an interval constructed for each one), about 95% of these intervals would contain the actual percentage.
2. What were you able to come up with?
3. (a) $(8.72, 10.44)$
   (b) Stick to the script!
   (c) Wider
   (d) Narrower

§1.11

1. (a) One concern is that this sample may not be representative of the general public.
   (b) At least 90%.
2. (a) (i) This is absurd. (ii) No, 95% is about the success rate of the interval if many intervals were made.
   (b) See previous answers asking the same question.
   (c) Yes, as the intervals do not overlap.
3. (a) Proportion of all people who prefer Wendy's fries to McDonald's fries
   (b) Probably a well-designed experiment or survey!
   (c) I couldn't find it ANYWHERE. Just vague references to "research firms."

§1.12

1. Should result in representative sample and therefore allow for inference to population
2. Because people must be willing to participate, among other reasons
3. To theoretically ensure that any differences between the groups are due to the treatments and no other variables
4. When a method will consistently over-estimate or consistently over-estimate the parameter; careful sampling process, vetting questions, etc.

5. Association means it is unknown whether one variable is the sole cause of the other variable, whereas causation is when that is known.

6. Well-designed experiments; see #3

7. When another explanatory variable, associated with the explanatory variable of interest, *also* leads to similar changes in the response variable

8. When the sample was randomly selected from the population

9. Shape, central tendency, variability, unusual features

10. They are unitless, enabling comparisons between distributions with different units or amounts of variability

11. No, they apply to any distributions. However, knowing z-scores allows for easy computations when the distribution is approximately Normal.

12. The distribution of all possible values of that statistic (for samples of the same size)

13. We use sampling distributions to get an idea of how that statistic behaves for all possible samples; this can help us gauge whether a particular statistic could've reasonably occurred by chance or not.

14. The alternative hypothesis is what we are looking for evidence of, and the null hypothesis is what we assume to be true (typically, that the alternative is *not* true). We then determine how likely it is that evidence could've occurred by chance *if* the null hypothesis is true.

15. The P-value is the likelihood of getting as good of evidence as was observed if the null hypothesis was true.

16. It's what we add and subtract from a statistic to get a confidence interval. It exists because, typically, it is very, very unlikely that a statistic will exactly equal the parameter.

17. Larger sample sizes lead to smaller margins of error. However, this comes at a cost, so researchers typically determine the smallest possible sample size that will guarantee the margin of error will be no larger than a certain number.

18. It is an interval constructed using a process that has a $C\%$ chance of successfully containing the parameter.

19. (a) Population: All 1540 attendees, Parameter: mean satisfaction for all 1540 attendees, Sample: 350 responders, Statistic: sample mean of 3.6

    (b) The people willing to respond may have felt more positive about the speaker, so this 3.6 may be an overestimate of the actual mean.

20. (a) Mean: 2.05, median: 1.5, SD: 1.883, min: 0, max: 8, Q1: 1, Q3: 3, range: 7

    (b) You can construct this!

    (c) You can construct this!

    (d) The distribution of number of siblings for this class is unimodal and heavily right skewed with a median between 1 and 2.

21. (a) 68% (b) 2.5% (c) 36.9% (d) 0.748 (e) 37.8%

22. (a) is subjective. (b) should involve random assignment. (c) is also subjective, though it's likely the proportion (percentage) of universities that offer the job.

23. (a) Certainly observational studies.

    (b) People who buy red cars may be riskier or more aggressive, which also may lead to a higher likelihood of getting in accidents. It's impossible to know if it's truly the color or peoples' personalities that are leading to the increased likelihood of accidents.

24. (a) The teacher could randomly assign 2 classes to use Khan Academy and 2 classes to use the prep book. They would first have everyone take the SAT to develop a baseline, and, after a fixed amount of time with each program, have students take the SAT again. They could then compare the mean changes in score between the two groups.

    (b) If the teacher believes Khan Academy is more effective, then it's questionable whether it's ethical to force two classes to use an inferior product.

    (c) There are many possible answers here, but the general problem is that the two groups of students may differ in some *other* way (besides just the program) that affects SAT scores, making it impossible the cause of any differences in effectiveness.

25. (a) Both distributions are unimodal and right skewed, but the median age of first time mothers in 2016 is a few years higher than it was in 1980. Additionally, there is more variability in ages of first-time mothers in 2016.

    (b) The mean will be higher than the median in both 1980 and 2016.

26. (a) $z = 1.176$. The pitcher's average spin rate this game was 1.176 standard deviations higher than his mean average spin rate.

    (b) ~ 0.12

    (c) It depends on your threshold for being convinced, but only about 4% of this pitcher's games have that low of an average spin rate assuming he's healthy.

27. Amy may have a legitimate complaint, but it depends again on your threshold for being convinced.

28. (a) Among others: they likely want to ask anonymously.

    (b) All possible sample percentages based on random samples of 50 students from the law school.

    (c) $(69, 91)$. They can be 95% confident that the true percentage of this year's graduates from this law school that have secured a job by graduation is between 69% and 91%.

    (d) No. 90% is in the interval, so it's plausible.

