INTRODUCTORY STATISTICS



Hannah Seidler-Wright
Chaffey College



Introductory Statistics

Hannah Seidler-Wright

Chaffey College

This text is disseminated via the Open Education Resource (OER) LibreTexts Project (https://LibreTexts.org) and like the thousands of other texts available within this powerful platform, it is freely available for reading, printing, and "consuming."

The LibreTexts mission is to bring together students, faculty, and scholars in a collaborative effort to provide an accessible, and comprehensive platform that empowers our community to develop, curate, adapt, and adopt openly licensed resources and technologies; through these efforts we can reduce the financial burden born from traditional educational resource costs, ensuring education is more accessible for students and communities worldwide.

Most, but not all, pages in the library have licenses that may allow individuals to make changes, save, and print this book. Carefully consult the applicable license(s) before pursuing such effects. Instructors can adopt existing LibreTexts texts or Remix them to quickly build course-specific resources to meet the needs of their students. Unlike traditional textbooks, LibreTexts' web based origins allow powerful integration of advanced features and new technologies to support learning.



LibreTexts is the adaptable, user-friendly non-profit open education resource platform that educators trust for creating, customizing, and sharing accessible, interactive textbooks, adaptive homework, and ancillary materials. We collaborate with individuals and organizations to champion open education initiatives, support institutional publishing programs, drive curriculum development projects, and more.

The LibreTexts libraries are Powered by NICE CXone Expert and was supported by the Department of Education Open Textbook Pilot Project, the California Education Learning Lab, the UC Davis Office of the Provost, the UC Davis Library, the California State University Affordable Learning Solutions Program, and Merlot. This material is based upon work supported by the National Science Foundation under Grant No. 1246120, 1525057, and 1413739.

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation nor the US Department of Education.

Have questions or comments? For information about adoptions or adaptions contact info@LibreTexts.org or visit our main website at https://LibreTexts.org.

This text was compiled on 06/17/2026

TABLE OF CONTENTS

Licensing

1: Designs of Statistical Studies

- 1.1: Welcome to Statistics
  - 1.1.1: Exercises
- 1.2: The Statistical Analysis Process
  - 1.2.1: Exercises
- 1.3: Research Questions, Types of Statistical Studies, and Stating Reasonable Conclusions
  - 1.3.1: Exercises
- 1.4: Random Sampling and Bias
  - 1.4.1: Exercises
- 1.5: Experiments and Random Assignment
  - 1.5.1: Exercises

2: Descriptive Statistics

- 2.1: Descriptive Statistics - Dotplots and Histograms
  - 2.1.1: Exercises
- 2.2: Quantifying the Center of a Distribution
  - 2.2.1: Exercises
- 2.3: Quantifying Variability Relative to the Median
  - 2.3.1: Exercises
- 2.4: Quantifying Variability Relative to the Mean
  - 2.4.1: Exercises

3: Probability

- 3.1: Introduction to Probability
  - 3.1.1: Exercises
- 3.2: Marginal, Joint, and Conditional Probability
  - 3.2.1: Exercises
- 3.3: The Addition and Complement Rules
  - 3.3.1: Exercises

4: Discrete Probability Distributions

- 4.1: Discrete Random Variables
  - 4.1.1: Exercises
- 4.2: The Geometric Distribution
  - 4.2.1: Exercises
- 4.3: The Binomial Distribution
  - 4.3.1: Exercises

5: Continuous Probability Distributions and The Normal Distribution

- 5.1: Probability Distributions of Continuous Random Variables
  - 5.1.1: Exercises
- 5.2: Characteristics of the Normal Distribution and The Empirical Rule
  - 5.2.1: Exercises
- 5.3: The Standard Normal Distribution
  - 5.3.1: Exercises
- 5.4: Finding Critical Values from the Normal Distribution
  - 5.4.1: Exercises

6: Inference Involving a Single Population Proportion

- 6.1: The Sampling Distribution of Sample Proportions
  - 6.1.1: Exercises
- 6.2: Estimating a Population Proportion
  - 6.2.1: Exercises
- 6.3: Introduction to Hypothesis Testing
  - 6.3.1: Exercises
- 6.4: Hypothesis Tests for a Single Population Proportion
  - 6.4.1: Exercises
- 6.5: Conclusions (1)
  - 6.5.1: Exercises

7: Inference Involving a Single Population Mean

- 7.1: The Sampling Distribution of Sample Means
  - 7.1.1: Exercises
- 7.2: The Student's T-Distribution
  - 7.2.1: Exercises
- 7.3: Estimating a Population Mean
  - 7.3.1: Exercises
- 7.4: Hypothesis Tests for a Single Population Mean
  - 7.4.1: Exercises
- 7.5: Conclusions (2)
  - 7.5.1: Exercises

8: Inference Involving Two Population Parameters

- 8.1: Paired Samples
  - 8.1.1: Exercises
- 8.2: Distributions of Differences
  - 8.2.1: Exercises
- 8.3: Inference for a Difference in Two Population Means
  - 8.3.1: Exercises
- 8.4: Inference for a Difference in Two Population Proportions
  - 8.4.1: Exercises

9: Linear Regression

- 9.1: Scatterplots
  - 9.1.1: Exercises
- 9.2: Quantifying Direction and Strength
  - 9.2.1: Exercises
- 9.3: The Line of Best Fit
  - 9.3.1: Exercises

10: Inference Involving More Than Two Parameters

- 10.1: The Chi-Square Distribution
  - 10.1.1: Exercises
- 10.2: Goodness-of-Fit
  - 10.2.1: Exercises
- 10.3: Testing for Independence
  - 10.3.1: Exercises
- 10.4: ANOVA
  - 10.4.1: Exercises

Index

Glossary

Detailed Licensing

Licensing

*A detailed breakdown of this resource's licensing can be found in **Back Matter/Detailed Licensing**.*

CHAPTER OVERVIEW

1: Designs of Statistical Studies

1.1: Welcome to Statistics

1.1.1: Exercises

1.2: The Statistical Analysis Process

1.2.1: Exercises

1.3: Research Questions, Types of Statistical Studies, and Stating Reasonable Conclusions

1.3.1: Exercises

1.4: Random Sampling and Bias

1.4.1: Exercises

1.5: Experiments and Random Assignment

1.5.1: Exercises

Thumbnail: https://courses.lumenlearning.com/at...ntroduction-8/

This page titled 1: Designs of Statistical Studies is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Hannah Seidler-Wright.

1.1: Welcome to Statistics

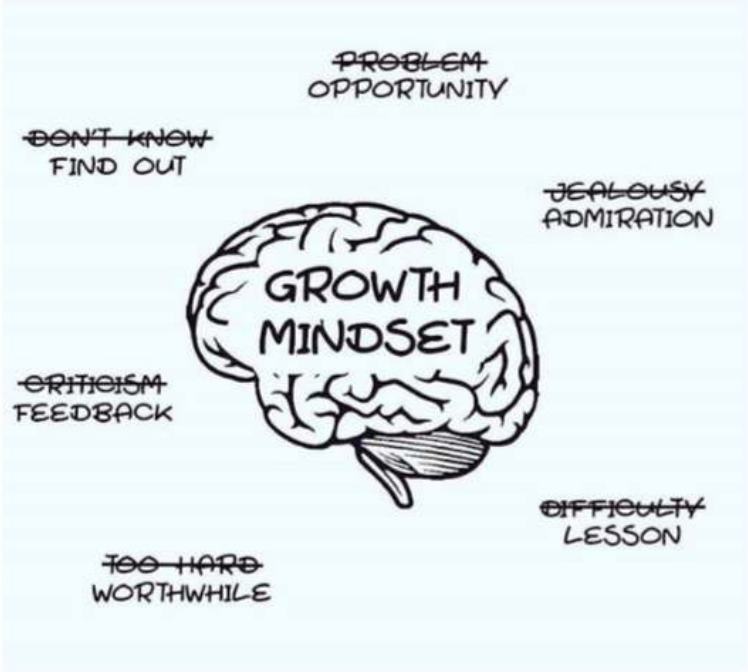
Not your *average* stats class!

This class will be different in many ways from other math classes you might have taken in the past. You are going to interact with your classmates and your instructor quite a bit. You will talk to each other about statistics. You will be in groups with your classmates, and I expect that you will contribute to the discussion of the concepts you are learning. In this class, everyone will have a chance to share their viewpoint and we respond to whatever anyone says with respect. I expect that students in this class will support each other in working together.

In survey after survey of employers, teamwork skills (along with communication skills) are at the top of the list of attributes they would like to see more of in their new hires. We will develop these skills, among others, in this class.

Growth mindset and neuroplasticity

We will adopt "hope theory" and "mindset thinking" to improve ourselves. As stated in this model, "hope can be thought of as the 'will power' to move toward action as well as the 'way power' or the pathway that will lead to goal achievement." Many of you may be able to improve your success in mathematics by adopting a high hope/growth mindset in your own life, to take on challenges that might have previously seemed unattainable. During our class, we will learn about hope, growth mindset, and neuroplasticity, and we will practice applying it. We will also try to take an evidenced-based approach to learn about our learning. We will develop practical strategies for improving our growth mindset, dealing with stress and frustration, and goal setting.



"A growth mindset is needed for #4IR Shift the negative to positive ⭐💡 #entrepreneur #DigitalTransformation #startup #innovation #fintech #insurtech #thinkbigsundaywithmarsha Cc @Clagett @SpirosMargaris @kimgarst @psb_dc @leimer @Paula_Piccard @" by Paula Piccard is marked with CC0 1.0. To view the terms, visit https://creativecommons.org/publicdo...?ref=openverse.

Materials

You will need a few materials to get set up for the class:

- A computer equipped with a microphone, a camera, and Google Chrome.
- Some mechanism for making a PDF document.
  - If you don’t have a scanner, there are various apps you can use on your phone to scan documents and convert them into PDFs including CamScanner, and the Microsoft Outlook app (where you can add an image and select the document icon).
- I will provide you with text resources we will be using. It is recommended that you have a binder to store your printed/lecture notes, as well as a place to keep solved exercises.

Getting set up with Amplify

Desmos is a wonderful free calculator that has statistical capabilities. We can also use the Amplify platform to engage in activities that help develop deep understanding of challenging mathematical concepts that have the desmos calculator built into it.

In order to utilize Amplify to its fullest, you will need to create an account. This way, any progress you make in an activity will be saved. You can also access feedback that I leave for you when you have an account.

To create an account:

1. Go to student.amplify.com and enter the code _____
2. Click “Sign up”
3. Enter your first AND last name (as it appears on the roster), your email, and create a password. Make your password easy to remember so that you can easily sign in for activities in our class.
4. Click the create account button

That's it! Now you are ready to complete any Amplify activities.

This page titled 1.1: Welcome to Statistics is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Hannah Seidler-Wright.

1.1.1: Exercises

1. Define a growth mindset and a fixed mindset. Describe one way you will apply your mindset to this class.

2. Describe three expectations of students in our class.

3. Go to https://www.desmos.com/calculator. Use the desmos calculator to convert $\frac{3}{8}$ to a decimal. Round the decimal to two decimal places.

This page titled 1.1.1: Exercises is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Hannah Seidler-Wright.

1.2: The Statistical Analysis Process

Statistics is the science of collecting, organizing, summarizing, and analyzing information to draw conclusions or answer questions. Additionally, statistics is about providing a measure of confidence in any conclusions. Statistics play a big role in our day to day decision making. Data can help us answer many questions:

- Students can use data to help pick a college that is a good fit for them
- Teachers use data to improve their teaching methods
- Medical researchers use data to learn if treatments actually work sufficiently
- Scientists use data to measure the effect humans have on the environment
- Car companies use data to determine how safe a vehicle is

Statistical analysis is the process of looking at a sample of data to learn something about a larger population that may be difficult to understand because of its size. It allows us to make generalizations about populations based on sample data. There are four steps:

1. Ask a question that can be answered by collecting data.
2. Decide what to measure and collect the data.
3. Summarize the data and analyze the data.
4. Draw a conclusion and communicate the results to your audience.

Step 1: Ask a question that can be answered by collecting data

Many people believe that personality, habits, likes and dislikes are affected by the time of year you were born. One such theory involves chronotypes. The term chronotype refers to a person's natural tendency to be most active and alert at certain times of day. People who are most active early in the day are labeled morning people, early birds, or larks. Those at their best in the evening are called night people or owls.

This chronotype theory states that there are three chronotypes and that each corresponds to four birth months.

Chronotype	**Birth Months**
Morning	May, June, July, August
Evening	January, February, November, December
None	March, April, September, October

We are going to use the four step statistical analysis process to try to answer this question: Is there a significant correlation between someone's birth month and their chronotype?

Step 2: Decide what to measure and then collect data

Instead of collecting data from all **individuals** of an entire group for the study (called the **population**), we could instead select a **sample** of the population. In order to investigate this question, we need to collect data from a random group of participants. We would need to know the _____ _____ and _____ of each individual. We could have participants take an assessment to determine their chronotype. We would see what proportion of the participants had matching birth month and chronotype according to the chronotype theory (found in the table above).

Step 3: Summarize and analyze the data

In a group of 30 randomly selected adults, it was found that 11 had matching chronotype and birth month according to the provided chronotype theory. The proportion (percentage) of matches was _____. Is this proportion high enough to convince us that this theory applies to the entire population? In other words, can we infer the results to some group bigger than those in the sample?

Suppose this chronotype theory is false. Is it possible that a participant in the study could still select a chronotype that matched their birth month according to the theory? What fraction or proportion of the participants in the study do we expect to have matching birth month and chronotype according to the theory?

In order to analyze the data, we need to use probability to obtain strong enough evidence to support or reject a claim. If this chronotype theory is false, the birth month does not match the corresponding chronotype, there is a _____ chance that a student would select the chronotype predicted. How far above _____ would the proportion need to be in order to convince us that this chronotype theory is reasonable?

Chance variation is the type of differences we would naturally expect to see between many different samples. We will see what proportions in a sample are likely to occur just by chance by rolling a die. In a 6-sided die, each of the six outcomes are equally likely to occur. The event of rolling a 1 or a 2 should occur around 33% or exactly $\frac{1}{3}$ of the time. If you'd like to try this yourself, type in "roll dice" into google. Click the "roll" button and record your outcomes.

Die value	3	2	5	4	2	5	5	1	1	3	4	4	1	1	3	1	3	4	4	2
Resulted in 1 or 2	N	Y	N	N	Y	N	N	Y	Y	N	N	N	Y	Y	N	Y	N	N	N	Y

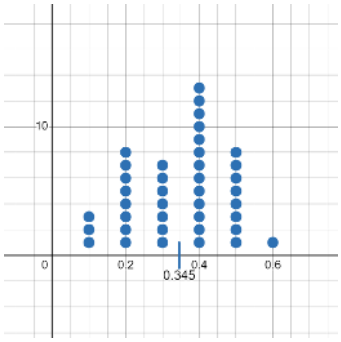
Proportion of die rolls resulting in 1 or 2: $\frac{8}{20} = 0.4 = \_\_\%$

Die value	6	4	5	2	6	1	1	5	5	1	2	5	2	2	1	4	1	1	4	5
Resulted in 1 or 2	N	N	N	Y	N	Y	Y	N	N	Y	Y	N	Y	Y	Y	N	Y	Y	N	N

Proportion of die rolls resulting in 1 or 2: ____=_____=_____

Note: A **proportion** is a number between 0 and 1. It represents a fraction or portion of a total. We usually write proportions as decimals or percents. To calculate the decimal, use a calculator to divide the numerator (top of the fraction) by the denominator (bottom of the fraction). For example, if you roll the die 25 times and 7 of those rolls resulted in a 1 or a 2, then the proportion as a fraction would be written as $\frac{7}{25}$ and you would enter $7 \div 25$ in your calculator to get 0.28. To change this to a percent, we multiply the decimal by 100 or move the decimal twice to the right and write a percent symbol. For example, $0.28 \cdot 100\% = 28\%$.

You can see from above that we did not get the same proportion in both times we tried this experiment, and the proportions we found were both close to $\frac{1}{3}$. We could repeat this experiment many times to see what outcomes are unusual. Below is a desmos graph of a dotplot of proportions from 40 repetitions of this experiment.



Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

This is an example of a **distribution** of data. We can now use it to make a decision about chronotype theory.

Step 4: Draw a conclusion and communicate the results

In the study, we found the proportion of matching chronotypes to birth month was around _____%. Since this proportion is likely to occur based on the distribution of data above, we can't generalize this result to the population because the proportion isn't high enough (or unusual enough) to convince us that the chronotype theory is true.

Identifying important information:

A group of researchers wondered if fewer than half of the women who visit a particular fertility clinic would want to choose the sex of their future child if it was possible. A total of 561 women responded to the survey. 229 of them said they wanted to choose the sex of their future child. The researchers analyzed the data and concluded that their was convincing evidence that fewer than half of the women who visit the clinic would choose the sex of a future child because, if, in reality, at least half of women who visit a fertility clinic would like to choose the sex of a future child, it would be very unusual to observe a percentage as low as 41% in a sample of 561 women.

1. What is the question being asked?

2. What is the way the data was collected and measured? What is the population? What is the sample?

3. Summarize and analyze the data:

4. What is the conclusion of the study?

This page titled 1.2: The Statistical Analysis Process is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Hannah Seidler-Wright.

1.2.1: Exercises

The statistical analysis process enables us to use data to make decisions about situations when we only have a limited amount of information. The statistical investigation we explored in this section involved using data from students in a class to assess whether a person's chronotype can be predicted by their birth month. Using data from the class we made an inference about the relationship between birth month and chronotype for all people. Statistical investigations allow us to use data from small samples to make generalizations about much larger populations.

1. A student flips a coin ten times and finds that the coin landed "Heads" on eight of the ten flips.
   a. If a coin is fair (i.e. the probability that the coin lands "Heads" is equal to the probability that the coin lands "Tails"), what is the probability that the coin lands "Heads" on a single coin flip?

   b. If a coin is fair and is flipped 20 times, approximately what fraction of the coin flips do you expect will land heads? Choose the best answer below.

      i. $\frac{0}{20}$

      ii. $\frac{5}{20}$

      iii. $\frac{10}{20}$

      iv. $\frac{20}{20}$

   c. If a coin is fair and is flipped 20 times, approximately what proportion of the coin flips do you expect will land heads? Choose an incorrect answer, and explain why it is wrong.

      i. -0.40

      ii. 0.50

      iii. 0.50%

      iv. 9.75

   d. Do you think the coin is fair based on this observation (8 out of 10 flips landing on heads)? Explain.

2. Read the following study description and answer the following questions: Researchers wanted to know if people think a task will be hard to accomplish when the instructions are difficult to read.[1] To answer this question, researchers randomly divided

twenty student volunteers into two groups of 10 students each. Researchers gave instructions to each group of students using different fonts (see below). Instructions for one group were written in a large upright font. The other group was given the same instructions but in a font that used hard-to-read italics. Researchers asked students to read the directions and say how many minutes they thought the task would take. Researchers did this in order to figure out if the fonts used for the instructions made a difference.

This is the easy-to-read upright font that was used in the study.

This is the hard-to-read italic font that was used in the study.

The first group of students, those that read the instructions printed in the easy font, had an average time estimate of 8.23 minutes. The other group, the group that read the instructions in the hard-to-read italic font, had an average time estimate of 15.1 minutes.

Researchers concluded that such a large difference between the averages was not likely to have occurred by chance. There was evidence that people think a task will be harder when instructions are difficult to read.

a. Which question below is a reasonable research question for this investigation? Explain how you made your decision.

   i. Do people like reading in different fonts?

   ii. Do people prefer reading one font to another font?

   iii. Do people think a task will be harder if the instructions for the task are harder to read?

   iv. Do people think that some instructions are easier to follow than other instructions?

b. What variables are used to answer the research question?

   i. Type of font & Amount of time a person thinks a task will take to complete.

   ii. Type of font & Amount of time a task takes to be completed.

   iii. Preferred font & Amount of time a person thinks a task will take to complete.

   iv. Preferred font & Amount of time a task takes to be completed.

c. How are the data summarized?

i. Researchers compared the total amount of time that the two groups took to complete the tasks.

ii. Researchers compared the average amount of time that the two groups took to complete the tasks.

iii. Researchers compared the total amount of time that the two groups estimated that it would take to complete the tasks.

iv. Researchers compared the average amount of time that the two groups estimated that it would take to complete the tasks.

d. What did the researchers conclude?

i. The 10 students in the sample who read the instructions in the hard-to-read font took longer to complete the task than the 10 students who read the instructions in the easy-to-read font.

ii. There is evidence that people will take a longer amount of time to complete a task when the instructions are harder to read.

iii. The 10 students in the sample who read the instructions in the hard-to-read font think the task will be more difficult than the 10 students who read the instructions in the easy-to-read font.

iv. There is evidence that people think a task will be harder when the instructions are harder to read.

3. Read the following study description and answer the following questions: The United States Government recommends that to stay physically fit, middle-aged adults (ages 40 to 60) need to burn 150 to 400 calories per day doing exercise. Researchers at Minnesota State University, Mankato, wanted to learn whether middle-aged adults who used the Wii Fit video game exercised enough to meet the government's fitness recommendations.[2] The Wii Fit is a video game that includes exercises. The researchers taught 20 middle-aged adult volunteers how to use the Wii Fit video game. On the day after they were trained, the adults exercised for 20 minutes with the Wii Fit. Researchers measured the total amount of energy each of the adults in the study used in calories. They found that the average energy used was 116 calories for the 20 minute session. Based on the results of the study, the researchers concluded the Wii Fit video game could be a helpful form of exercise for middle aged adults. But, for exercise with Wii Fit to meet the government's recommendation, the researchers stated that the length of the exercise session should be increased from 20 minutes to 30 minutes.

a. Which question below is a reasonable research question for this investigation? Explain how you came to this answer.

i. Do people think that playing the Wii Fit video game burns calories?

ii. Does the Wii Fit video game burn enough calories to be considered suitable exercise?

iii. Does the Wii Fit video game burn more calories than traditional exercise?

iv. What is the average amount of time that middle-aged adults spend playing Wii Fit video games?

b. What data did the researchers collect to answer the research question?

i. The amount of time that the adults exercised.

ii. The name of the adults.

iii. The type of exercise the adults completed.

iv. The total amount of calories the adults burned through exercising.

c. How are the data summarized?

i. Researchers found the proportion of adults who exercise using the Wii Fit video game.

ii. Researchers found the proportion of adults who prefer exercising with the Wii Fit video game over traditional exercises.

iii. Researchers found the average amount of calories that adults consumed through exercising using the Wii Fit video game.

iv. Researchers found the average amount of time that adults spent exercising using the Wii Fit video game.

d. What did the researchers conclude?

i. The Wii Fit video game is a preferred exercise for some middle-aged adults.

ii. The Wii Fit video game does not appear to burn enough calories in a 20–minute session, but a 30–minute session would possibly be enough.

iii. The Wii Fit video game does appear to burn enough calories in a 20–minute session, but a 30–minute session would be even better.

iv. The sample size is too small to make any reasonable conclusions about all middle-aged adults.

Reference

[1] Hyunjin Song, “The Effects of Processing Fluency on Judgment and Processing Style: Three Essays on Effort Prediction, Risk Perception, and Distortion Detection” (PhD diss., The University of Michigan, 2009).

[2] http://www.ncbi.nlm.nih.gov/pubmed/21178930

This page titled 1.2.1: Exercises is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Hannah Seidler-Wright.

1.3: Research Questions, Types of Statistical Studies, and Stating Reasonable Conclusions

In the last section, we discussed the statistical analysis process. We begin the process by asking a question that can be answered by collecting data. Understanding the type of research question that is being asked helps us to know how we collect data in the next step.

Research Questions and Types of Statistical Studies

In a **statistical study,** a **population** is a set of all people or objects that share certain characteristics. A **sample** is a subset of the population used in the study. **Subjects** are the individuals or objects in the sample. Subjects are often people, but could be animals, plants, or things. **Variables** are the characteristics of the subjects we study. For example, a variable could be hair color, age, salary, etc. In a previous lesson, we examined the relationship between a person's birth month and chronotype. The population was all adults. The sample was the 30 randomly selected adults. The variables were sets of personality traits and birth date groups.

- A research question about a population.
- A research question about the causal relationship between two variables.

---

1. Throughout this course, (a) we will learn how to make estimates about a population, (b) we will test claims about a population, (c) we will compare two populations, and (d) investigate a relationship between two variables (using means, proportions, or standard deviations) by asking **research questions about a population**. State which scenario (a-d) the following questions connect with.

- Is there a relationship between the number of hours a full-time student works at a job and their GPA?
- What is the average amount of hours spent studying community college students complete per week?
- What proportion of community college students work full-time?
- Do the majority (more than 50%) of community college students work full-time?
- Do community college students who study more than 36 hours per week have a higher average GPA than those who do not?
- Does the average amount of hours spent studying by community college students exceed 36 hours per week?

In all of the questions above, the researchers only *observe* subjects in a sample to learn about a population's characteristics. They do not control any of the variables. We call this study an **observational study**.

2. We will also examine **research questions about a causal relationship between variables**. For each of the following example questions, circle the word or words that suggest causation.
    a. Does studying more hours for a college class improve test grades?
    b. Does caffeine reduce the number of migraines (long-lasting headaches) for women?
    c. Do violent video games increase crime in the US?

To answer these questions, we investigate how one variable responds as another variable is manipulated or changed. An **explanatory variable** is the (input) variable being manipulated. A **response variable** is the (output) variable used to measure the impact from manipulation of the explanatory variable. An **experiment** involves a manipulation or change to the explanatory variable.

3. For the following questions, determine if researchers should conduct an observational study, or an experiment. Justify your choice.
    a. What is the average time it takes to recover from heart surgery?

    b. Do vehicle emissions cause climate change?

    c. What is the approval rating for the governor of California?

    d. Does more regular attendance in high school improve college success?

    e. Is race associated with the maternal mortality rate?

4. Read the following statistical study and answer the questions that follow:

   We are interested in learning whether getting more sleep improves one's emotional state (emotional score out of 10 points determined by a professional assessment). We want to see if there is a difference between the emotional score of adults that sleep for 4 hours every night for a week and the emotional score of adults that sleep for 8 hours every night for a week. To investigate this question, we use 100 adult volunteers. The emotional score of each subject will be measured at the beginning of the study. 50 of the volunteers will participate in a sleep program where they are limited to 4 hours of sleep every night for a week. The other 50 participants can sleep for 8 hours every night for a week. At the end of a week, the emotional score will be measured again.

   a. What is the research question?

   b. Is the question about a population or a causal relationship between two variables?

   c. Is this an observational study or an experiment?

      1. If this is an observational study, what is the population?

      2. If this is an experiment, what are the explanatory and response variables?

   d. We need to divide the group of 100 volunteers into two groups of 50 so that there is a fair comparison between the 4 hour and 8 hour sleep groups. What would be a way to create two groups that have similar volunteers?

Drawing Reasonable Conclusions

The last step in the statistical analysis process is drawing a conclusion. In this step, researchers must communicate their results clearly. They will extend what they learned from the data to explain what they learned from the study.

There are two types of reasonable conclusions that can be drawn from a study.

We may generalize our results from a sample to the population. In order to do this, we must create a sample that is representative of the larger population. The best way to create a sample that is representative of the population is to use random sampling.

We may conclude that there is a causal relationship between two variables. This conclusion arises from an experiment when a significant change in the response variable was caused by the manipulation of the explanatory variable. In order to conclude causation, we must make sure that we create experimental groups that are similar. The best way to achieve this is through random assignment.

If we create a sample using random sampling and we use random assignment to create similar experimental groups, then we can reasonably draw both of the conclusions above. If random sampling is not used to create a sample and random assignment is not used to create similar groups in an experiment, no conclusions can be reliably drawn.

5. Read the following statistical study and answer the questions that follow:

   The SAT exam is used in admissions decisions by many four-year colleges and universities. In the past, The College Board carried out a study of 5,342 SAT essays that were selected at random from more than a million SAT exams. For this sample of essays, 15% were written in cursive and 85% were not. The results showed that the average score for essays written in cursive was higher than for essays that were not written in cursive.

   a. Is this an observational study or an experiment? Justify your answer.

   b. Is it reasonable to conclude that writing the essay in cursive was the cause of the higher scores? Explain your answer.

   c. Is it reasonable to generalize the conclusion to the population? In other words, is it reasonable to conclude that the score for essays written in cursive was higher than for essays that were not written in cursive, on average? Explain your answer.

6. Read the following statistical study and answer the questions that follow:

   Imagine that a psychologist is interested in finding out if listening to classical music has an effect on one's ability to recall material that has been read. The psychologist recruits volunteer students who say they like to study while listening to music. She randomly assigns them into two groups. Each group is told to read a famous poem. One group reads the poem in silence, and the other group reads the poem while they listen to classical music. After reading the poem, they take a brief assessment that asks the students to recall information about the poem. The psychologist concludes that students who listen to classical music while they read score lower than students who read in silence, on average.

   a. Is this an observational study or an experiment? Justify your answer.

   b. What is one possible reason for why the students who listened to classical music scored lower than those who did not?

   c. The psychologist found that the difference was so large that it was unlikely due to chance variation alone. Is it reasonable to conclude that listening to classical music caused students in the sample to score lower on the assessment? Explain.

   d. Is it reasonable to generalize this conclusion to the population? Why or why not?

This page titled 1.3: Research Questions, Types of Statistical Studies, and Stating Reasonable Conclusions is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Hannah Seidler-Wright.

1.3.1: Exercises

1. Give an example of a research question that involves estimating a characteristic about the population of Registered Nurses in California.

2. Improve this poorly stated research question: Do registered nurses work a lot?

3. Create a cause-and-effect research question.

"Alkaline water: the secret to glowing skin" is the headline of an article that appeared in *Scratch Magazine* (February 10, 2021). The article claims that consuming alkaline water instead of tap water improves the hydration of skin and therefore, improves skin appearance. Consider the following hypothetical study designs. For each study, answer the questions that follow.

4. Study design 1: Two hundred students were selected at random from those enrolled at a large college in California. Each student in the sample was asked whether they drank alkaline water more than once in a typical week. A skin specialist rated skin health for each student on a scale of 1 to 10. It was concluded that skin health was significantly better on average for the group that reported drinking alkaline water more than once a week than it was for the group that did not.

   a. Explain why this is an observational study.

   b. Was random selection used to create the sample? Explain.

   c. Did the study use random assignment to experimental groups? If so, explain what method was used to randomly assign students.

   d. Is the conclusion "drinking alkaline water leads to healthier skin" reasonable given the study description? Explain your answer.

   e. Is it reasonable to generalize conclusions from this study to some larger population? Justify your answer. If so, what population?

5. Study design 2: One hundred people volunteered to participate in a statistical study. For each volunteer, a coin was tossed in order to place them into a group. If the coin landed head up, the volunteer was assigned to group 1. If the coin landed tail up, the volunteer was assigned to group 2. Those in group 1 were asked to drink one cup of alkaline water daily for three months. Those in group 2 were asked to drink one cup of tap water daily for three months. At the end of the three months, a skin specialist rated skin health on a scale of 1 to 10 for each of the volunteers. It was concluded that skin health was significantly better on average for those in group 1 than for those in group 2.
   a. Explain why this is an experiment.

   b. Was random selection used to create the sample? Explain.

   c. Did the study use random assignment to experimental groups? If so, explain what method was used to randomly assign students.

   d. Is the conclusion “drinking alkaline water leads to healthier skin” reasonable given the study description? Explain your answer.

   e. Is it reasonable to generalize conclusions from this study to some larger population? Justify your answer. If so, what population?

6. Study design 3: One hundred students were selected at random from those enrolled at a large college. Each of the selected students was asked to participate in a study and all agreed. For each student, a coin was tossed in order to place them into one of two groups. If the coin landed head up, the student was assigned to group 1. If the coin landed tail up, the student was assigned to group 2. Those in group 1 were asked to drink one cup of alkaline water daily for three months. Those in group 2 were asked to drink one cup of tap water daily for three months. At the end of the three months, a skin specialist rated skin health on a scale of 1 to 10 for each of the volunteers. It was concluded that skin health was significantly better on average for those in group 1 than for those in group 2.
   a. Is this an observational study or an experiment? Justify your answer.

   b. Was random selection used to create the sample? Explain.

   c. Did the study use random assignment to experimental groups? If so, explain what method was used to randomly assign students.

   d. Is the conclusion “drinking alkaline water leads to healthier skin” reasonable given the study description? Explain your answer.

   e. Is it reasonable to generalize conclusions from this study to some larger population? Justify your answer. If so, what population?

7. Study design 4: One hundred people who live in Miami volunteered to participate in a statistical study. The volunteers were divided into two experimental groups based on sex, with females in group 1 and males in group 2. Those in group 1 were asked to drink one cup of alkaline water daily for three months. Those in group 2 were asked to drink one cup of tap water daily for three months. At the end of the three months, a skin specialist rated skin health on a scale of 1 to 10 for each of the volunteers. It was concluded that skin health was significantly better on average for those in group 1 than for those in group 2.

   a. Is this an observational study or an experiment? Justify your answer.

   b. Was random selection used to create the sample? Explain.

   c. Did the study use random assignment to experimental groups? If so, explain what method was used to randomly assign students.

   d. Is the conclusion “drinking alkaline water leads to healthier skin” reasonable given the study description? Explain your answer.

   e. Is it reasonable to generalize conclusions from this study to some larger population? Justify your answer. If so, what population?

8. Study design 5: Two hundred cosmetology students enrolled at a large college in California were chosen to participate in the statistical study. Each student in the sample was asked whether they drank alkaline water more than once in a typical week. A skin specialist rated skin health for each student on a scale of 1 to 10. It was concluded that skin health was significantly better on average for the group that reported drinking alkaline water more than once a week than it was for the group that did not.
   a. Is this an observational study or an experiment? Justify your answer.

   b. Was random selection used to create the sample? Explain.

   c. Did the study use random assignment to experimental groups? If so, explain what method was used to randomly assign students.

   d. Is the conclusion “drinking alkaline water leads to healthier skin” reasonable given the study description? Explain your answer.

   e. Is it reasonable to generalize conclusions from this study to some larger population? Justify your answer. If so, what population?

9. What are the four steps in the statistical analysis process?

This page titled 1.3.1: Exercises is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Hannah Seidler-Wright.

1.4: Random Sampling and Bias

Random Sampling

1. Explain what it means for a sample to be representative of a population.

2. Suppose that Chaffey is thinking of ways to raise money. The administration is considering offering reserved parking spots for a fee of $80. The college wants to know the percentage of students who would support this fee. One way to do this is to conduct a campus wide survey called a _____. Is this reasonable? Why or why not? Recall that the student population at Chaffey is almost 20,000.

3. Read each of the following ways to sample students at Chaffey and decide whether the sample produced would be representative of the population from 2. Explain why or why not.
   a. Choose three 7:30 am classes at random and survey all of the students in each class.

   b. Put a poll on the front page of the college website. A poll is an opinion survey. Use the students who answer the poll as the sample.

   c. Talk to students as they enter the library.

When we sample, our goal is for every member of the population to have the same chance of being selected. We would like to avoid selection bias, which is when a sample differs from the population in some way so that some individuals are more likely to be selected than others. To avoid selection bias, we could use a _____ _____ _____ in which all samples of a given size have the same chance of being chosen.

4. A **voluntary response sample** is one in which the participants are self-selected. Each member chooses to participate. Another biased sample is a **convenience sample** in which sampling does not use random selection but instead uses an available or convenient group to form the sample. Categorize a, b, and c from question 3. into one of these samples or state neither.

The college has a list of all registered students. In order to produce a **simple random sample**, a number can be assigned to each individual on the list. One can produce a random integer list that has the desired number of data values. The individuals corresponding to those numbers could be emailed a survey or contacted by the college.

What are three other types of random sampling?

1.

2.

3.

Bias

If the results of a sample are not representative of a population, then the sample has bias. There are three sources of bias in sampling:

1. Sampling bias - _____ _____ means that the technique used to obtain the samples individuals tends to favor one part of the population over another. Any convenience sample has sampling bias because the individuals are not chosen through random sampling.

2. Nonresponse bias - _____ _____ exists when individuals selected to be in the sample who do not respond to the survey have different opinions from those who do. It can occur when individuals selected for the sample do not want to respond or can not be contacted.

    a. Write two reasons why a recipient of a survey might not want to respond

        i. _____

        ii. _____

3. Response bias - _____ _____ exists when the answers on a survey do not reflect the true feelings of the respondent.
    a. For example, an interviewer who does not elicit trust with an individual may get responses that are not honest.

    b. Some survey questions result in responses that are not complete truths. People can over or underestimate their own abilities.

    c. The way a question is worded can lead to unintended responses. The phrasing may influence the response of an individual.

    d. The arrangement of words or questions in a survey can influence the response of an individual.

This page titled 1.4: Random Sampling and Bias is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Hannah Seidler-Wright.

1.4.1: Exercises

1. Describe the main difference between an observational study and an experiment.

2. Imagine that you want to learn about the average number of hours, per day, that students at your college spend on their mobile devices. You want to select a simple random sample of 75 students from the full-time students at your college. You have a list of all full-time students, whose names are arranged in alphabetical order. How would you select a simple random sample of 75 students from this population? Describe your process.

3. Imagine you want to know how many hours per week students at your school spend studying, on average. Determine if the following sampling methods will reasonably produce representative samples. Justify your answers.
   a. Select 40 students randomly using a list of all student IDs and a random number generator.

   b. Select 80 students as they enter the library.

   c. Select 150 students randomly using a list of all student IDs and a random number generator.

   d. Select the 200 students enrolled in calculus III at the college this semester.

   e. Which of the methods above would produce the most representative sample and the best estimate? Explain.

4. Examining the benefit of random sampling:
   a. Take 10 samples of size 5 using this simulation (by clicking on 5 circles, clicking reset after recording the average diameter, and repeating 10 times) and record the provided average diameters below:

      You can use the QR code below to access the applet.

   b. The average diameter for this population of 60 circles is 19.3. For the samples you selected, how many had an average diameter greater than 19.3? How many had an average diameter less than 19.3?

   c. Generate 10 random samples using this simulation (by clicking the "generate sample" button, recording the average diameter, and repeating 10 times) and record the provided average diameters below:

      You can use the QR code below to access the applet.

d. The average diameter for this population of 60 circles is 19.3. For the random samples, how many had an average diameter greater than 19.3? How many had an average diameter less than 19.3?

e. Which gave better estimates of the true population mean: your samples or the randomly generated samples? Why do you think this?

5. Give an example of a voluntary response sample. Explain how the sample does not represent the population.

6. Determine the type of sampling used (simple random, stratified, systematic, or convenience)

a. A market researcher polls every tenth person who walks into a store.

b. The first 50 people who walk into a sporting event are polled on their television preferences.

c. A computer generates 100 random numbers, and 100 people whose names correspond with the numbers on the list are chosen.

d. Which of the above sampling techniques will prevent the sample from being representative of the population? Explain.

7. A professor is curious what motivates students to cheat. They want to know if cheating is more pervasive in STEM classes compared to other disciplines. They randomly select 50 students from various STEM classes and 50 students from various non-STEM classes. They ask participants if they have cheated in the class and compare the proportion of students who say yes in each group. What type of bias should the professor be concerned about? Explain.

8. A market research company wants to gauge interest in a bingo facility in a small city. The researchers send out a text containing a link to the survey to randomly selected phone numbers in the city. What type of bias should the researchers be concerned about? Explain.

This page titled 1.4.1: Exercises is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Hannah Seidler-Wright.

1.5: Experiments and Random Assignment

Experiments

An _____ is a controlled study conducted to determine the effect varying one or more explanatory variables or **factors** has on a response variable. Any combination of the values of the factors is called a _____ . In an experiment, the **experimental unit (or _____)** is a person, object, or some other well-defined item upon which a treatment is applied.

The goal in an experiment is to determine the effect various treatments have on the response variable. When participation in a study prompts a physical response from a participant, it is difficult to isolate the effects of the explanatory variable. To counter the power of suggestion, researchers set aside one treatment group as a **control group**. This group is given a **placebo** treatment–a treatment that cannot influence the response variable. The control group helps researchers balance the effects of being in an experiment with the effects of the active treatments.

Of course, if you are participating in a study and you know that you are receiving a pill which contains no actual medication, then the power of suggestion is no longer a factor. **Blinding** in a randomized experiment preserves the power of suggestion. When a person involved in a research study is blinded, they do not know who is receiving the active treatment(s) and who is receiving the placebo treatment. A **double-blind experiment** is one in which both the subjects and the researchers involved with the subjects are blinded.

Random Assignment

In previous lessons, we stated that random assignment helps to make experimental groups similar. In this exercise we will see how well random assignment actually works.

An article in the journal Pediatrics reported on the results of an experiment that compared recovery times for two types of hernia surgery for children.

- Method 1: laparoscopic repair (a surgery that uses three small incisions)
- Method 2: open repair (a surgery that uses one large incision)

To compare the two treatments (hernia surgery methods), the researchers needed to create two groups of children that were similar with respect to any variables that might affect the response variable (recovery time).

Imagine that a new group of researchers thought that another variable - a child's age - might also affect his or her recovery time. The researchers wanted to control for age, so they wanted the two treatment groups to contain children who were similar in age. This would prevent the age variable from influencing the response variable in the experiment.

One way to do this is to randomly assign children to one of the two groups. This might be done by flipping a coin to assign each child to a group. If the coin lands heads, the researchers assign the child to the Method 1 group. If the coin lands tails, the researchers assign the child to the Method 2 group.

Let's investigate whether this method of random assignment creates similar groups. Suppose there are 30 children with hernias who volunteered to participate in the experiment. The identification numbers and ages of these 30 children are given in the following table.

We will randomly assign children to one of the two groups by using a **random number generator**. Then we will look at the results and see if random assignment actually works. Below is a list of 30 children with their ages included.

Child ID #	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Age	12	11	9	8	11	10	11	10	7	6	12	10	10	9	10

Child ID #	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
Age	7	7	8	6	9	7	9	8	11	9	12	12	11	12	12

Use a random number generator (access through the link or through the QR code below) to determine which 15 children will have the surgery method 1. The remaining children should be placed into the method 2 table.

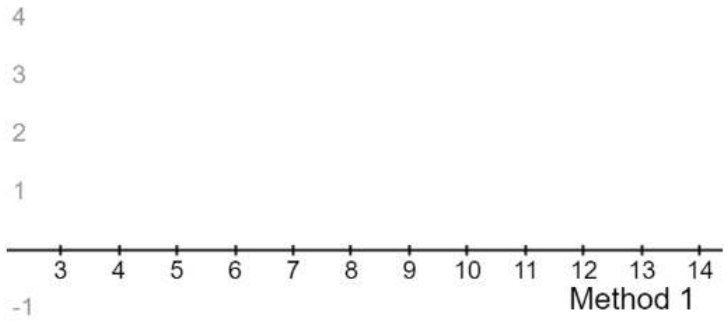
Method 1

Child ID #															
Age															

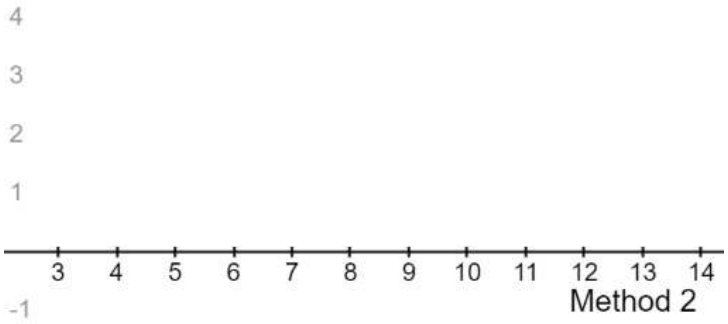
Method 2

Child ID #															
Age															

Below are two labeled number lines. Record your data on them and calculate the average age for each method.



Average age for method 1:



Average age for method 2:

Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

1. Think about the dotplots and their averages. Was the method of random assignment successful in creating groups with similar ages? Explain your answer.

2. Other variables that might affect recovery time are weight and fitness level. Do you think that our random assignment to experimental treatments (method 1 group and method 2 group) would create groups of similar weight and fitness level? Why do you think so?

This page titled 1.5: Experiments and Random Assignment is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Hannah Seidler-Wright.

1.5.1: Exercises

1. Compare and contrast random sampling and random assignment.

2. When random sampling is used to create a sample, what types of conclusions can reasonably be made?

3. When random assignment is used to create similar groups from a sample, what types of conclusions can reasonably be made?

4. One hundred students were selected at random from those enrolled at a large college. Each of the selected students was asked to participate in a study and all agreed. For each student, a coin was tossed in order to place them into one of two groups. If the coin landed head up, the student was assigned to group 1. If the coin landed tail up, the student was assigned to group 2. Those in group 1 were asked to drink one cup of alkaline water daily for three months. Those in group 2 were asked to drink one cup of tap water daily for three months. At the end of the three months, a skin specialist rated skin health on a scale of 1 to 10 for each of the participants. It was concluded that skin health was significantly better on average for those in group 1 than for those in group 2. In this experiment, what is the explanatory variable, and what are its values (the individual treatments)? What is the response variable in this experiment?

5. With so much advancement in technology, people have been given access to doing many tasks at once. How effectively can people multitask? Imagine researchers want to perform an experiment to answer this question. These researchers divided volunteers into two groups. Each subject was given a literature passage to analyze. One group had to check email and respond to messages while they were analyzing the passage for 30 minutes. The other group analyzed the passage without any distractions for 30 minutes. All subjects were then given a short 10 point assessment directly after. Researchers found that the distracted group's average assessment score was 4 points lower than the average assessment score for the group that was not distracted.
   a. Identify the explanatory variable and the individual treatments. Then identify the response variable.

   b. Explain why it would be good for the researchers to use random assignment to put each volunteer in one of the experimental groups. Why should the researchers do this rather than letting the volunteers decide which group they wanted to be in.

   c. Identify the control group in this experiment.

   d. Is it possible for the subjects of this study to be blinded? Explain your answer.

6. Compare and contrast voluntary response samples and nonresponse bias.

7. In order to answer a question about a population, what type of study should we conduct?

8. In order to answer a cause-and-effect type question, what type of study should we conduct?

This page titled 1.5.1: Exercises is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Hannah Seidler-Wright.

CHAPTER OVERVIEW

2: Descriptive Statistics

2.1: Descriptive Statistics - Dotplots and Histograms

2.1.1: Exercises

2.2: Quantifying the Center of a Distribution

2.2.1: Exercises

2.3: Quantifying Variability Relative to the Median

2.3.1: Exercises

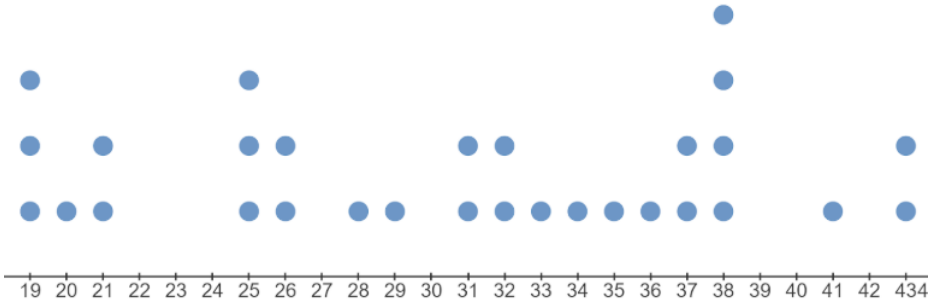
2.4: Quantifying Variability Relative to the Mean

2.4.1: Exercises

This page titled 2: Descriptive Statistics is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Hannah Seidler-Wright.

2.1.1: Exercises

1. Given is a dot plot of MPG rating for 30 randomly selected cars.



Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

a. What does a dot on the dot plot represent?

b. What is the most frequent MPG?

c. How many vehicles have an MPG less than 25?

d. What is the proportion of vehicles that have a MPG that is more than 35? Round to three decimal places.

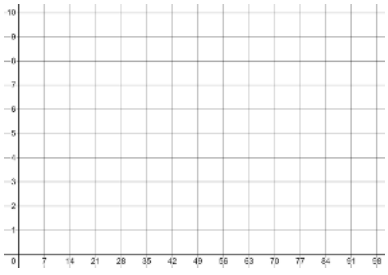
2. Given is the average life expectancy at birth of women from 20 randomly selected countries:

$$W = [78, 79, 49, 77, 51, 55, 71, 56, 62, 69, 70, 76, 77, 80, 90, 81, 82, 83, 56, 84]$$

a. Complete the table below.

Bin	Tally	Frequency	Relative Frequency		
			Fraction	Decimal	Percent
42-48					
49-55					
56-62					
63-69					
70-76					
77-83					
84-90					
Total					

b. Create a frequency histogram for the average life expectancy at birth of women. Label the axes.



Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

c. In how many countries is the average life expectancy at birth for women at least 77?

d. Explain in words the meaning of the height of the third bar in the histogram.

e. Describe the center, shape, and spread of the histogram.

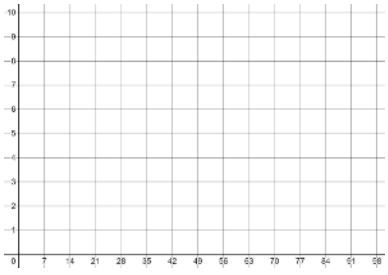
3. Given is the average life expectancy at birth of men from 20 randomly selected countries:

$$M = [89, 56, 76, 49, 77, 49, 55, 62, 62, 63, 76, 78, 79, 90, 80, 83, 81, 84, 90, 82]$$

a. Complete the table below.

Bin	Tally	Frequency	Relative Frequency		
			Fraction	Decimal	Percent
42-48					
49-55					
56-62					
63-69					
70-76					
77-83					
84-90					
Total					

b. Create a frequency histogram for the average life expectancy at birth of men. Label the axes.



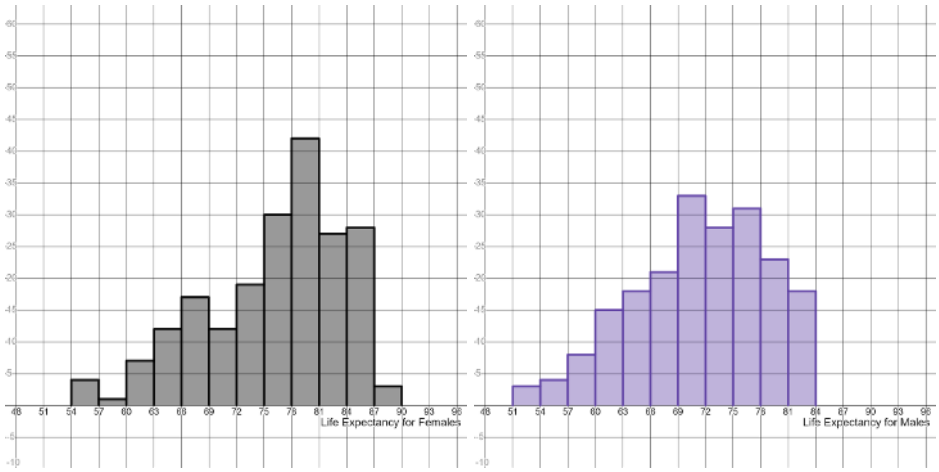
Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

c. In how many countries is the average life expectancy at birth for men at least 77?

d. Explain in words the meaning of the height of the fifth bar in the histogram.

e. Describe the center, shape, and spread of the histogram.

4. Given below are histograms for life expectancy for 202 countries for women and men. Compare the center, shape, and spread of the histograms.



Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

2.1.1: Exercises is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by LibreTexts.

- **Current page** is licensed CC BY-NC-SA 4.0.
- **1.2: The Statistical Analysis Process** by Hannah Seidler-Wright is licensed CC BY-NC-SA 4.0.

2.2: Quantifying the Center of a Distribution

A stimulant is a type of drug that is often found in weight loss medications. We will examine the effect of a stimulant on the weight gains of a treatment group of rats. These are compared to a control group of rats who receive no stimulant treatment.

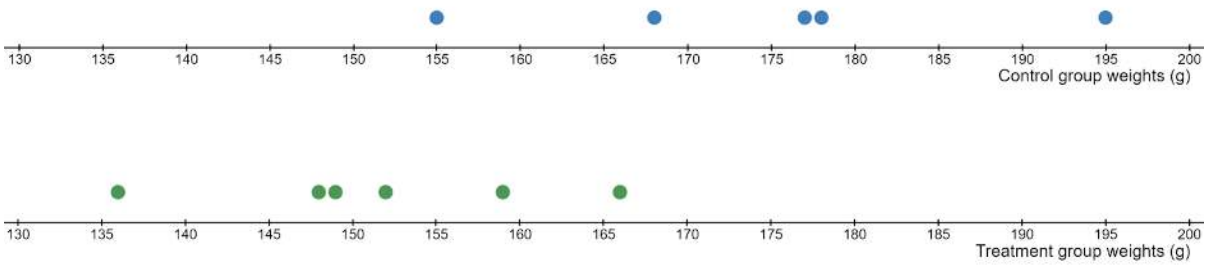
What is a typical value in the data set?

Suppose we observe the following weight gains (in grams) for twelve adolescent lab rats over a one-month period. The weight gain for the rats in the treatment and control groups are given below:

Control group weights in grams (no stimulant)	168	155	178	203	195	177
Treatment group weights in grams (stimulant)	136	159	152	149	166	148

To determine whether there might be an effect on weight gain due to the stimulant, we will determine representative (or central) values of the two groups, namely, the **sample mean** and the **sample median**.

Below are dotplots for the control and treatment groups of rats:



Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

Sample Means

1. Imagine the dotplot as a scale that can tip left or right or stay balanced. Where do you think the control group's dotplot balances? That is, on the number line, where would you set a balance point so that the distribution does not tip to the left or right?

This value is an estimation of the **mean** or average. A mean is one way we could describe a typical data value in a set. To calculate the exact mean, we add all data values to find a sum and divide the sum by the number of data values in the set.

2. Compute the average weight gain for the rats in the control group.

This value is the exact **sample mean** since it is the mean of a sample of six rats in the control group. Luckily, this set is very small and therefore, the computation is not too difficult to do by hand. For most data sets, we will use technology to compute the sample mean for a set. The mathematical symbol we use to denote a sample mean is $\bar{x}$ (pronounced "x-bar"). Formulaically, we say

$$\bar{x} = \frac{\sum x_i}{n} = \frac{\text{sum of all values in the set}}{\text{number of values in the set}}$$

For the control group,

$$\bar{x} = \frac{168 + 155 + 178 + 203 + 195 + 177}{6} = 179.\overline{3} \approx 179.3 \text{ grams}$$

3. Compute the mean weight gain for the rats in the treatment group and call this y ("y-bar"). Round to one decimal place.

$$\bar{y} = \text{\_\_\_\_\_} = \text{\_\_\_\_} \approx$$

4. Compare $\bar{x}$ and $\bar{y}$. Which sample mean is larger? Is the difference between the sample means large enough to make you believe that the stimulant has an effect on weight gain in adolescent rats? Why or why not?

Sample Medians

A **sample median** is the middle number of a sorted list of data values. Here is the process for computing a median applied to the control group values.

- First we sort the data values from smallest to largest:

unsorted	168	155	178	203	195	177
sorted	155	168	177	178	195	203

- Notice that the middle number in this ordered set is between the 3rd and 4th values. This will always be the case when we have an even number of values in our set. To find the location of the median for a set that has an even number of elements, we can divide the sample size by 2 and the median will be between this quotient and the next number on the list. (For example, let's say we have a set of 8 data values. The median will be exactly between the 4th and 5th values in the sorted list). For our control group, the median falls between 177 and 178. This means that the sample median is halfway between these two values or in other words, the median is the average of these two middle numbers:

$$\text{median } = \frac{177 + 178}{2} = 177.5 \text{ grams}$$

- If there are an odd number of values in the set, the median is the data value exactly in the middle of the sorted list, and there will be an equal number of data values on each side of the median. For example, consider the set $A = [1, 1, 12, 15, 17, 21, 22, 25, 40]$ which has 9 values in it (which is odd). The middle number or median of the sorted set is 17. There are 4 data values to the left of 17 and 4 data values to the right of 17. When the sample size is odd, we can divide by 2 as we did before, however, this will result in a number that is not whole. 9 divided by 2 is 4.5. To find the location of the median for an odd sized set, we can divide by 2 and round up to the nearest whole number. 4.5 rounds to 5, so the 5th value in this set is the median.

5. Compute the sample median for the treatment group.

The sample median is another way to describe a central/representative/typical value in a set of data.

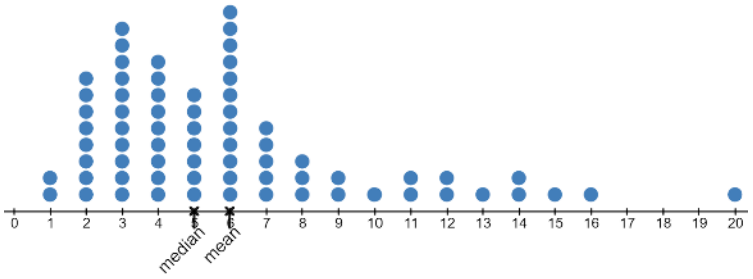
6. Compare the median weight gains of the two groups. Which sample median is larger? Is the difference between the sample medians large enough to make you believe that the stimulant has an effect on weight gain in adolescent rats? Why or why not?

7. Suppose we made an error when we recorded the largest weight gain in the treatment group. Instead of writing 166, we wrote 616. Recalculate the sample mean and sample median for the treatment group with this new value. In what way(s) did this error impact the mean? In what way(s) did this error impact the median?

Resistant Measures of Center

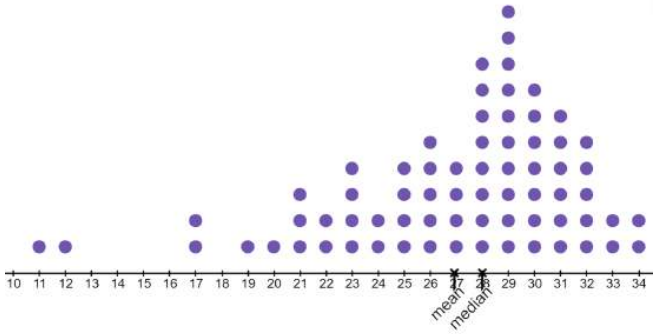
When should we avoid choosing a mean as a representative value for a data set? We call the mean and median measures of center. Measures of center are single values that represent a typical value for a given set of data. You've just seen that the mean is strongly affected by extreme values (values that are far away from most of the other data). We say that the mean is *not* a **resistant** measure of center. You also saw in the last example that the median was unaffected by the existence of an extreme value. We say that the median *is* a **resistant** measure of center. Extreme values are often present when a distribution is **skewed**. A distribution is skewed when it is not symmetric and one side has a long tail of values. When a distribution is skewed, the mean is pulled in the direction of the tail.

The following dotplot is an example of a distribution that is **skewed to the right**. When the distribution is right-skewed, the mean tends to be greater than the median.



Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

The following dotplot is an example of a distribution that is **skewed to the left**. When the distribution is left-skewed, the mean tends to be less than the median.

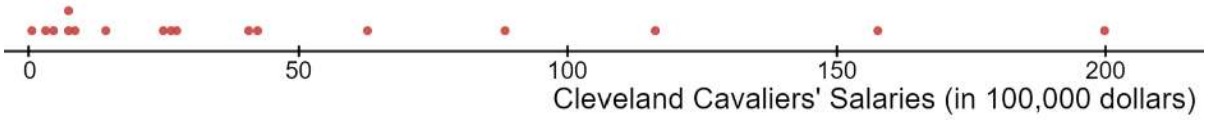


Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

8. Below are the salaries of the seventeen players on the Cleveland Cavaliers basketball team during the 2009-2010 season (access this website using QR code below). The 2009-2010 season was LeBron James' last season playing for the Cavaliers (until he later returned to the team). Lebron James was not the highest paid player on the team that season, Shaquille O'Neal was.

1	Shaquille O'Neal	20000000	10	Sebastian Telfair	2500000
2	LeBron James	15779912	11	J.J. Hickson	1429200
3	Antawn Jamison	11641095	12	Leon Powe	855189
4	Mo Williams	8860000	13	Darnell Jackson	736420
5	Anderson Varejao	6300000	14	Jawad Williams	736420
6	Delonte West	4254250	15	Danny Green	457588
7	Daniel Gibson	4088500	16	Coby Karl	311896
8	Jamario Moon	2750000	17	Cedric Jackson	53834
9	Anthony Parker	2644230			

A dotplot of the salaries is given below:



Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

a. Calculate the mean salary for the Cavaliers during the 2009-2010 season.

b. Calculate the median salary for the Cavaliers during the 2009-2010 season.

c. Would the mean or the median be most representative of the Cleveland Cavalier players' salaries in the 2009-2010 season? Justify your answer.

d. How does Shaquille O'Neal's salary impact the mean? You can examine this question by computing the mean without Shaq's salary included and compare.

This page titled 2.2: Quantifying the Center of a Distribution is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Hannah Seidler-Wright.

- **Current page** by Hannah Seidler-Wright is licensed CC BY-NC-SA 4.0.
- **1.2: The Statistical Analysis Process** by Hannah Seidler-Wright is licensed CC BY-NC-SA 4.0.

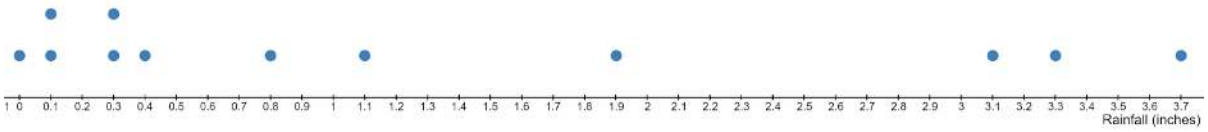
2.2.1: Exercises

1. The table below lists the average monthly rainfall in California and Utah.

Month	CA (inches)	UT (inches)
Jan	3.3	1.4
Feb	3.7	1.3
Mar	3.1	1.9
Apr	0.8	2
May	0.3	2.1
Jun	0.1	0.8
Jul	0	0.7
Aug	0.1	0.8
Sept	0.3	1.3
Oct	0.4	1.6
Nov	1.1	1.4
Dec	1.9	1.2

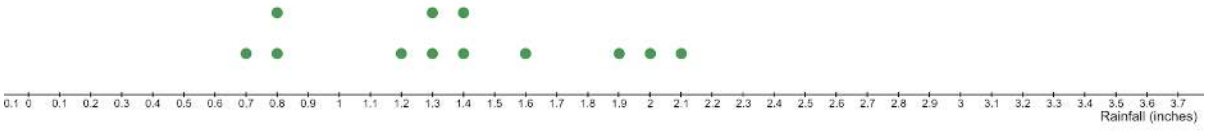
These data are summarized in dotplots below.

California Rainfall



Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

Utah Rainfall



Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

a. Compute the sample mean of the typical monthly rainfall for both California and Utah. Use the symbols $\bar{x}$ and $\bar{y}$ for these sample means.

b. Compute the sample medians of the monthly rainfall for both California and Utah.

c. Compare and contrast the means and medians for the two states.

2. Students in a literature class took a test where the highest score was 100. Most of the students did quite well, with more than half of the students getting in the 80s and 90s on the test. Most of the rest of the students earned 70s, and only a few students earned scores in the 60s. Only two students got less than a 60, and their scores were very low in the 20s.
   a. If a grade of A is earned by getting a score in the 90s, B in the 80s, C in the 70s, D in the 60s, and F for scores below 60, what is the shape of the distribution of scores? Draw a sketch of the graph to show your thinking.

   b. Is the mean or median likely to be larger? Explain your answer.

   c. Students often want to know the average test score. Which of the values would be more representative of the typical grade on the test, the mean or median? Explain your reasoning.

3. Given below is a frequency distribution for the age of students in a statistics class. The frequency tells you how many students take on the given value of the variable (age). For example, $f_1 = 5$ and $x_1 = 18$, so there are 5 students in the class who are 18 years old.

Age (x_i)	Frequency (f_i)
18	5
19	6
20	4
21	1
22	3
23	2
27	1
32	1
36	1
40	1

Rather than finding the mean by adding every individual value and dividing by the total number of values $\left(\frac{18+18+18+18+18+19+...+40}{25}\right)$, we can instead find the mean of this *grouped data* by using the frequencies listed in the table. We can multiply each value of x by its associated frequency, find the sum, and divide by the total frequency $\left(\frac{(18\cdot 5)+...+(40\cdot 1)}{5+...+1}\right)$. The formula for the sample mean can be written as

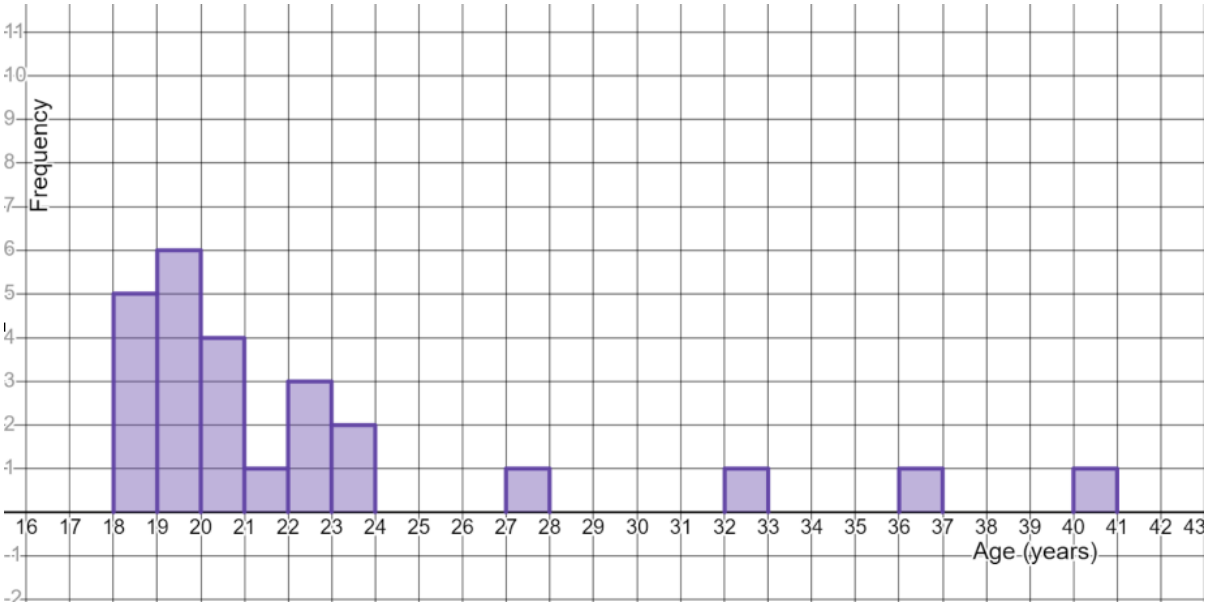
$$\bar{x}=\frac{\sum_i f_i\cdot x_i}{\sum_i f_i}$$

a. Fill in the missing values in the table below (including the totals)

Age (x_i)	Frequency (f_i)	$f_i\cdot x_i$
18	5	$f_1\cdot x_1=5\cdot 18=90$
19	6	
20	4	
21	1	$f_4\cdot x_4=1\cdot 21=21$
22	3	$f_5\cdot x_5=3\cdot 22=66$
23	2	
27	1	
32	1	$f_8\cdot x_8=1\cdot 32=32$
36	1	
40	1	$f_{10}\cdot x_{10}=1\cdot 40=40$
Totals	$\sum_i f_i=$ _____	$\sum_i f_i\cdot x_i=$ _____

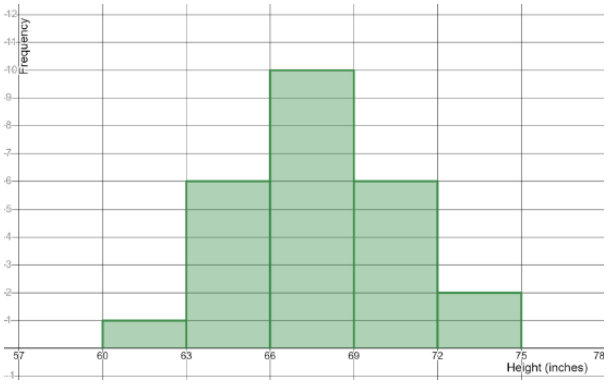
b. Compute the mean by dividing the appropriate values you computed in the totals row above.

c. Below is a frequency histogram. Is the mean likely to be greater than or less than the median? Explain your answer.



Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

4. Below is a frequency histogram for the heights students in a statistics class. Estimate the center, shape, and spread.



Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

5. Below is Samantha's calculation of the median for the heights of students in her statistics class. Explain any errors she made and write a sentence or two explaining to Samantha how to fix her mistake.

   Samantha's work and solution:

   $$H = [67, 67, 67, 67, 67, 62, 67, 64, 64, 64, 65, 65, 66, 68, 68, 69, 69, 69, 71, 71, 71, 73, 74, 67, 64]$$

   Sort the list from smallest to greatest:

   $$[62, 64, 64, 64, 64, 65, 65, 66, 67, 67, 67, 67, 67, 67, 67, 68, 68, 69, 69, 69, 71, 71, 71, 73, 74]$$

   There are 25 students in the class, so to find the median, we divide by 2 and get 12.5 so the median is in between the 12th and 13th numbers on the list. Therefore, $\text{median } = \frac{67+67}{2} = 67 \text{ inches}$

2.2.1: Exercises is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by LibreTexts.

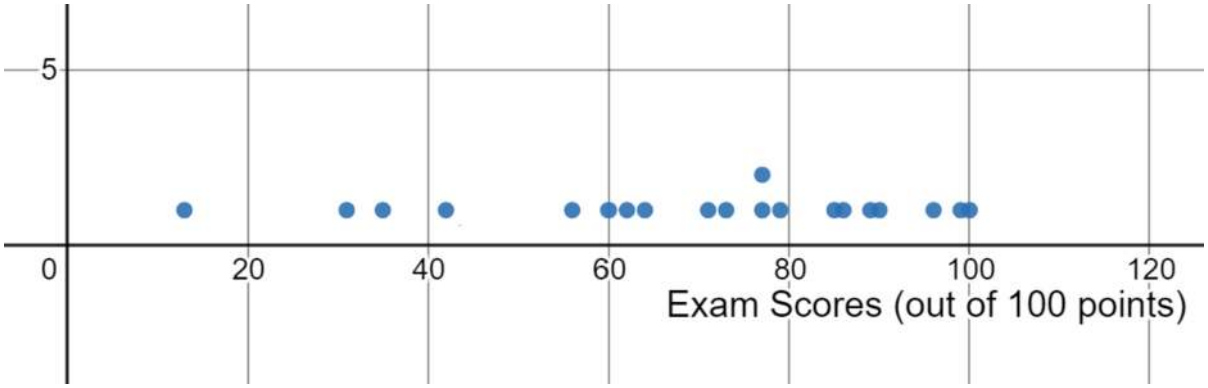
- **Current page** is licensed CC BY-NC-SA 4.0.
- **1.2: The Statistical Analysis Process** by Hannah Seidler-Wright is licensed CC BY-NC-SA 4.0.

2.3: Quantifying Variability Relative to the Median

In a previous lesson, we estimated the center, shape, and spread of distributions. We quantified the center by computing the mean and median for a data set. In this section, we will examine the spread or variability of a distribution.

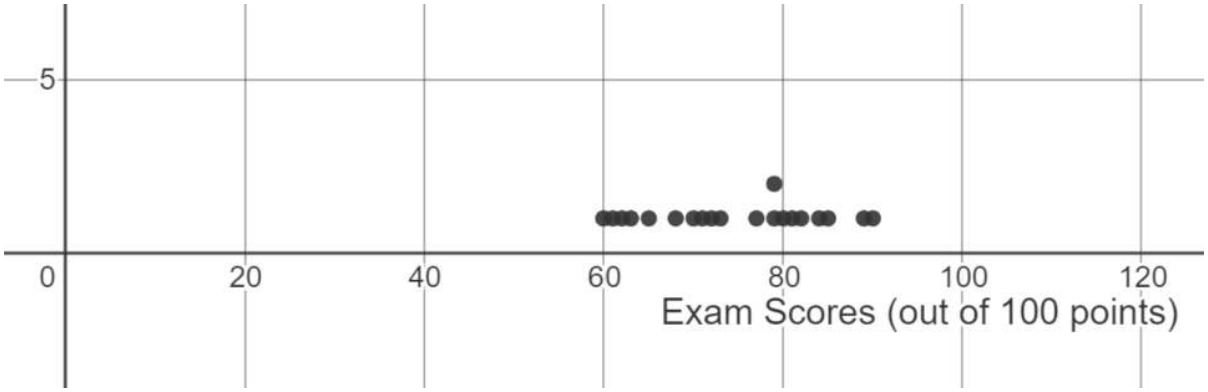
1. Given below are the dotplots for two samples of quiz scores.

   Scores for class A: [13,31,35,42,56,60,62,64,71,73,77,77,79,85,86,89,90,96,99,100]



Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

   Scores for class B: [60,61,62,63,65,68,70,71,72,73,77,79,79,80,81,82,84,85,89,90]



Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

   Estimate the center, shape, and spread of the distributions. Then compare the two distributions. What do you notice is similar about the distributions? What do you notice is different about the distributions?

2. In a previous section, we summarized the center of a distribution by using the sample mean and median. This is an important aspect of reading a graph, but it is not enough. The center of the distributions above are similar, but one distribution has more variability than the other distribution. Choose between the following two numbers to represent the variability for the class A distribution. Then explain your choice: 85 or 30

Range

3. One way to represent the variability in data is with the **range**. The range is the difference between the maximum and minimum data values.

$$\text{range} = \text{maximum} - \text{minimum}$$

Compute the range for the two classes. Which class has the larger range?

One issue with the range is that it is influenced by outliers. In other words, the range is **not a resistant** measure of spread. This is because the calculation of the range always uses the two most extreme data values in the set.

Five Number Summary

Another way to describe variability is with **quartiles**. When the data is sorted from lowest to greatest, the quartiles are numbers that divide the data into four equal parts. Recall that the median is the number that divides the data into two equal parts. 50% of the data is below the median, and 50% of the data is above the median.

Let's compute the quartiles for the class A scores. To find the quartiles, we should first compute the median.

First, we make sure that the data are in order from smallest to largest.

$$[13, 31, 35, 42, 56, 60, 62, 64, 71, 73, 77, 77, 79, 85, 86, 89, 90, 96, 99, 100]$$

There are 20 values in this set, so we divide 20 by 2 to find the location of the median. 20 divided by 2 is 10 so the median is between the 10th and 11th data values in the ordered list.

$$[13, 31, 35, 42, 56, 60, 62, 64, 71, 73] \text{ Median } [77, 77, 79, 85, 86, 89, 90, 96, 99, 100]$$

$$\text{Median } = \frac{73 + 77}{2} = 75$$

Notice that the median is *not* an existing value in the set. We see that the median divides the set into two *new* sets. To find the quartiles, we find the median of the lower half and upper half of the data set. There are 10 values to the left of the median. To find the location of the **first quartile Q1**, we divide 10 by 2 to get 5 and therefore, there will be 5 data values on each side of Q1. Similarly, to find the location of the **third quartile Q3**, there will be 5 data values on each side of Q3.

$$[13, 31, 35, 42, 56] \text{ Q1 } [60, 62, 64, 71, 73] \text{ Median } [77, 77, 79, 85, 86] \text{ Q3 } [89, 90, 96, 99, 100]$$

$$\text{Q1} = \frac{56 + 60}{2} = 58 \qquad \text{Q3} = \frac{86 + 89}{2} = 87.5$$

$$[13, 31, 35, 42, 56]58[60, 62, 64, 71, 73]75[77, 77, 79, 85, 86]87.5[89, 90, 96, 99, 100]$$

The median, first quartile Q1, and third quartile Q3, together with the minimum and maximum values in the set, make up **the five number summary**. The five number summary is given in the table below:

Class A Five Number Summary

Minimum	13
Q1	58
Median = Q2	75
Q3	87.5
Maximum	100

4. Compute the five number summary from Class B's scores.

Class B Five Number Summary

Minimum	
Q1	
Median	
Q3	
Maximum	

We have learned that quartiles are medians! When a set contains an odd number of values, the middle value is not included in the lower or upper half of the given data set to compute the quartiles.

5. Another way to describe variability in a data set is to find the distance between the first and third quartiles (Q1 and Q3). This distance is called the **interquartile range**, abbreviated as **IQR**.

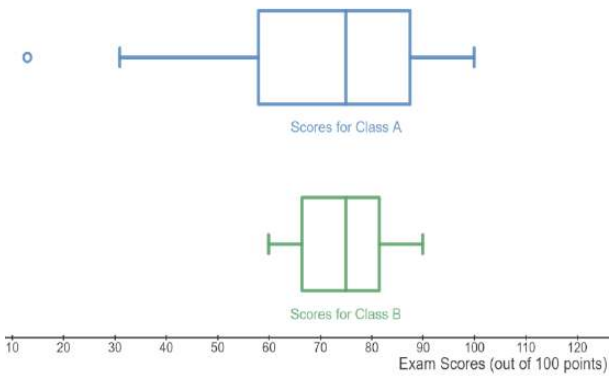
   $$\text{IQR} = \text{Q3} - \text{Q1}$$

   The IQR gives the range of the middle 50% of the data. For class A, the IQR is $\text{Q3} - \text{Q1} = 87.5 - 58 = 29.5$ points. Calculate the IQR for the scores from class B.

6. Compare the IQRs for scores for class A and class B. Does this comparison agree with your intuition about the spread of the distributions? Do you think the IQR is highly affected by extreme values?

Boxplots (Box and Whisker Plots)

The values in the five-number summary can be represented in a graph called a **boxplot** (sometimes referred to as a box and whiskers plot). Given below are the boxplots for the scores for each of the classes.



Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

7. Mark the minimum, Q1, median, Q3, and maximum on each of the boxplots above.

8. Are the boxplots consistent with your interpretation of the spread of the distributions?

9. What do you think the open circle on the boxplot for scores for class A indicates?

Recall, an *outlier* is a value that is extremely far away from the rest of the values in the set of data. When we are working with boxplots, values that deviate further than 1.5 IQRs from the quartiles are considered outliers. Outliers are outside of the range that extends from $\mathrm{Q1} - (1.5 \cdot \mathrm{IQR})$ to $\mathrm{Q3} + (1.5 \cdot \mathrm{IQR})$.

These boundaries are called the **fences**. Any value that is below the lower fence, $\mathrm{LF} = \mathrm{Q1} - (1.5 \cdot \mathrm{IQR})$, or above the upper fence, $\mathrm{UF} = \mathrm{Q3} + (1.5 \cdot \mathrm{IQR})$, is considered an outlier. On the graph we represent the outlier with an open circle, or an asterisk. The attached whisker is connected to the next non-outlier in the data set.

10. Explain why the score 13 from the class A data is an outlier. Compute the fences in your answer.

11. You try! The ages of 15 Academy Award winners for best actress are given below:

$$[48, 47, 54, 35, 45, 65, 77, 36, 33, 37, 36, 44, 21, 37, 38]$$

a. Sort the ages from smallest to largest and compute the five number summary.

Ages of Academy Award winners - Five Number Summary

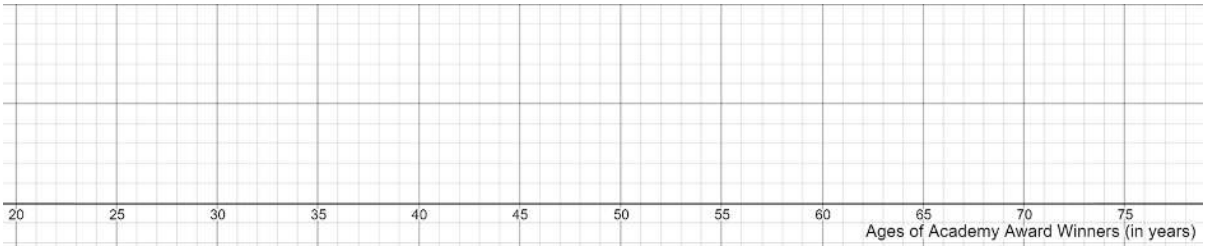
Minimum	
Q1	
Median	
Q3	
Maximum	

b. Compute the range of the ages of academy award winners.

c. Compute the IQR of the ages of academy award winners.

d. Compute the fences and use them to identify any outliers.

e. Graph the boxplot for the age of academy award winners.



Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

f. Go to https://www.desmos.com/calculator and complete the following steps:

1. Enter $A = [48, 47, 54, 35, 45, 65, 77, 36, 33, 37, 36, 44, 21, 37, 38]$ into the first line.
2. Tap Enter on your keyboard.
3. Type boxplot(A). Alternatively, you can open the keypad, click on the functions button, scroll to visualizations, and select boxplot. Entered the name of the set, A, in the empty parentheses.
4. Click the Zoom Fit icon and confirm the box next to "Exclude Outliers" is checked.

g. Use this tool to check your work.

This page titled 2.3: Quantifying Variability Relative to the Median is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Hannah Seidler-Wright.

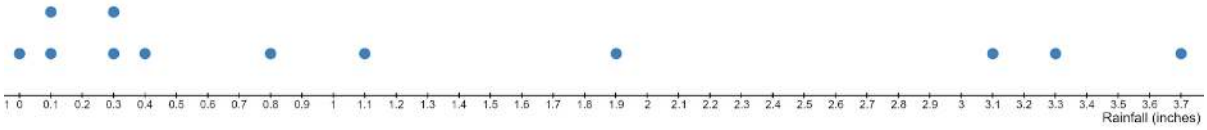
- **Current page** by Hannah Seidler-Wright is licensed CC BY-NC-SA 4.0.

- **1.2: The Statistical Analysis Process** by Hannah Seidler-Wright is licensed CC BY-NC-SA 4.0.

2.3.1: Exercises

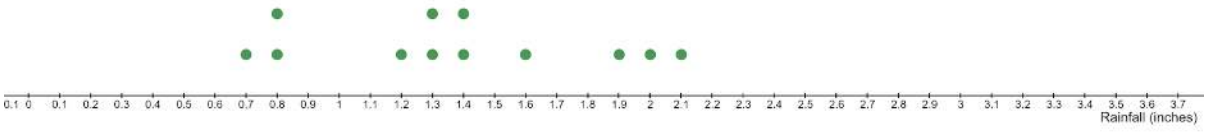
1. The dotplots below summarize the average monthly rainfall in California and Utah.

California Rainfall



Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

Utah Rainfall



Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

a. Compare the center, shape, and spread of the two distributions.

b. In how many months was the average rainfall less than 1 inch in California? In what proportion of a year was average rainfall less than 1 inch in California?

c. Decide which measure of spread (range or IQR) would be best to use to compare the variation in the rainfall of CA and UT. Explain.

d. Select the value that best represents the IQR for California rainfall: 1 inch or 2 inches? Justify your answer.

2. Given below are the heights (in feet) of 30 randomly selected pro football and basketball players.

Heights of 30 Pro Football Players

6.5	6.3	6.8	6.1	5.8
6.4	6.3	6.7	6.5	5.8
6.6	6.3	6	5.8	6.6
6.3	5.9	6	6	6.3
5.8	6.5	6.3	6.2	5.8
5.8	6.5	6.2	6.2	6.1

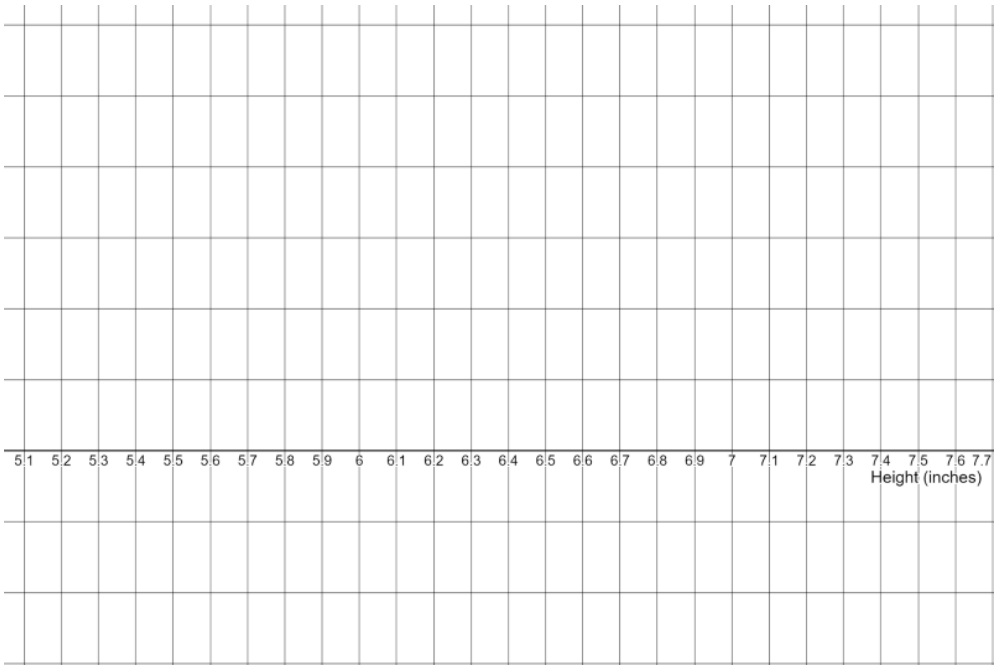
Heights of 30 Pro Basketball Players

6.1	6	6.6	6.5	6.6
7	6.5	6.6	6.5	6.6
6.3	6.8	6.4	6.7	6.3
6.9	6.3	6.6	6.1	6.4
6.3	6.9	6.8	6.5	6.6
5.8	6.3	6.9	6.7	6.8

a. Calculate the five-number summary for the data values from the Football and Basketball samples by hand.

Five Number Summary	Football Sample	Basketball Sample
Minimum		
Q1 (first quartile)		
Median		
Q3 (third quartile)		
Maximum		

b. Graph the boxplots (stacked) for the heights of football players and the heights of basketball players by hand. Use the graphs to compare the distributions.



Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

c. Calculate the range of heights for the football and basketball samples. Write the calculations. How is the range represented graphically in the boxplot?

d. Calculate the IQR for the heights of football and basketball players. Write the calculations.How is the IQR represented graphically in the boxplot?

3. Below is Nate's calculation of the five number summary and graph of the corresponding boxplot. Explain any errors they made and write a sentence or two explaining to Nate how to fix their mistakes.

Nate's work and solution:

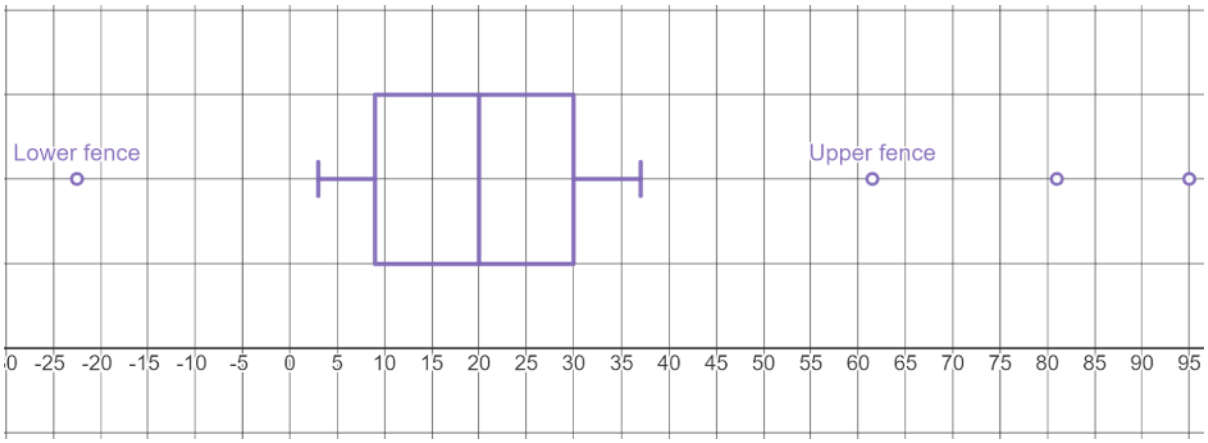
The following data represent the observed number of native plant species from random samples of study plots on different islands in the Galapagos Island chain:

$$P = [23, 26, 33, 21, 35, 30, 16, 3, 17, 9, 8, 9, 19, 12, 11, 7, 23, 95, 81, 4, 37, 28]$$

Sort the data from smallest to largest:

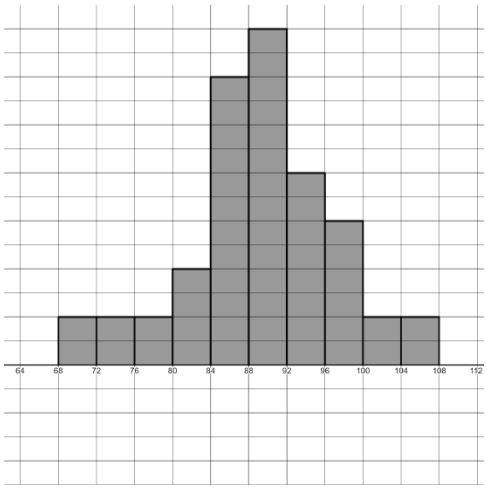
$$[3, 4, 7, 8, 9, 11, 12, 16, 17, 19, 21, 23, 23, 26, 28, 30, 33, 35, 37, 81, 95]$$

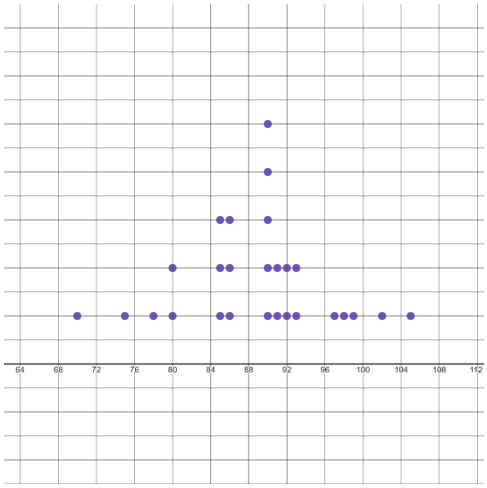
There are 22 values in the data set, so the median will be between the 11th and 12th value in the set. Therefore $\text{median} = (19 + 21)/2 = 20$. The median separates the set into two sets each of size 11. Half of 11 is 5.5, so Q1 and Q3 will be the 6th value in their respective sets. $Q1 = 11$, and $Q3 = 30$. The minimum is 3 and the maximum is 95. The lower fence is $9 - (1.5 \cdot 21) = -22.5$ and the upper fence is $30 + (1.5 \cdot 21) = 61.5$ so the outliers are -22.5, 61.5, 81, 95.

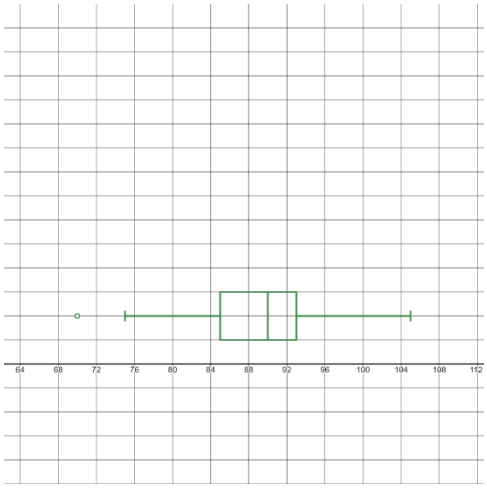


Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

4. The following graphs represent glucose levels (mg/100ml) in the blood for a random sample of 27 non-obese adults. Use the graphs to estimate the center, shape, and spread of the distribution. Identify some benefits/challenges with using each of the graphs to describe this distribution.







Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

5. Below is Danny's calculation of the five number summary and graph of the corresponding boxplot. The following data represent glucose levels (mg/100ml) in the blood for a random sample of 27 non-obese adults.

   Danny's work and solution:

$$G = [80, 85, 75, 90, 70, 97, 91, 85, 90, 85, 105, 86, 78, 92, 93, 90, 80, 102, 90, 90, 99, 93, 91, 86, 98, 86, 92]$$

   Sort the list from smallest to greatest:

$$[70, 75, 78, 80, 80, 85, 85, 85, 86, 86, 86, 90, 90, 90, 90, 90, 91, 91, 92, 92, 93, 93, 97, 98, 99, 102, 105]$$

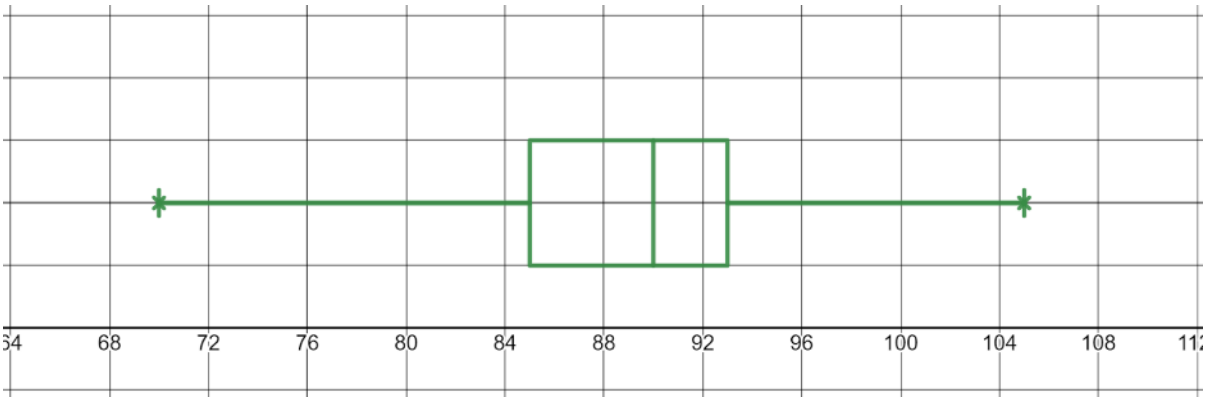
   The median is the 14th number on the list which is 90. Then we get two sets each with 14 numbers in them:

   Lower $= [70, 75, 78, 80, 80, 85, 85, 85, 86, 86, 86, 90, 90, 90]$ so $Q1 = 85 + 85/2 = 85$.

   Upper $= [90, 90, 90, 91, 91, 92, 92, 93, 93, 97, 98, 99, 102, 105]$ so $Q3 = 92 + 93/2 = 92.5$.

   Minimum is 70, Q1 is 85, Median is 90, Q3 is 92.5, and maximum is 105. $\text{IQR} = 92.5 - 85 = 7.5$.

   Lower fence $= 85 - (1.5 \cdot 7.5) = 73.75$ and upper fence $= 92.5 + (1.5 \cdot 7.5) = 103.75$.



Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

This page titled 2.3.1: Exercises is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Hannah Seidler-Wright.

- **Current page** by Hannah Seidler-Wright is licensed CC BY-NC-SA 4.0.
- **1.2: The Statistical Analysis Process** by Hannah Seidler-Wright is licensed CC BY-NC-SA 4.0.

2.4: Quantifying Variability Relative to the Mean

In the previous lesson, we examined variability with respect to the median. In this lesson, we will develop a measure of variability that depends on the mean. Recall that the mean is the average of all values in a given set. We use all data values in its computation. The sample mean is denoted by $\bar{x}$ and is found by adding all the values in the sample and dividing by the sample size. It's formula is $\bar{x} = \frac{\Sigma x_i}{n}$. When referring to a population mean, we use the symbol μ ("mu"). It is computed in the same way; we add all values in the population, and divide by the population size.

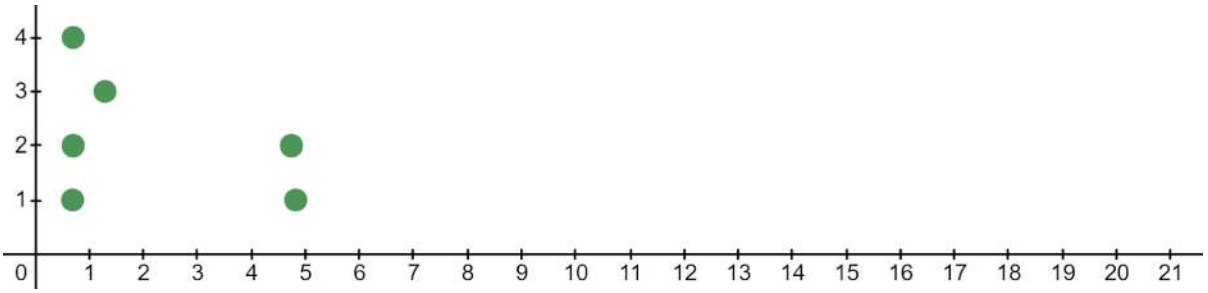
We have seen a few measures of variability so far. The range is the difference between the maximum and minimum values in a set of data. The range is not resistant to outliers because its calculation involves only the two most extreme values in the set of data. The interquartile range or IQR uses only two values as well, but it *is resistant* because it is the difference between the third and first quartiles. Quartiles are found using the resistant measure of center, the median. Therefore, the IQR depends on the median. We will now develop a measure of variability that depends on the mean and uses all values in a given set of data.

The table and dotplots below show the 2022 salaries (in millions of dollars) of a sample of players from the Kansas City Royals and Los Angeles Dodgers.

2022 Salaries (in millions of dollars)

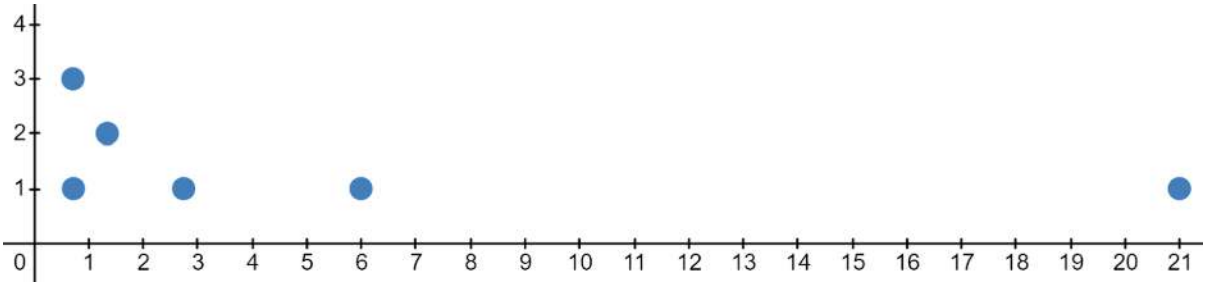
Kansas City Royals	4.83	0.7	4.75	0.71	1.3	0.71
Los Angeles Dodgers	6	0.73	2.75	1.35	0.72	21

Kansas City Royals



Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

Los Angeles Dodgers



Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

1. Let’s start by computing the mean salary for each sample. How do they compare?

A measure of variability should tell us how spread out data is. Since we are using the mean as our center, it would be useful to know the distances between each value and the mean. These distances are called **deviations**. Data values that are above the mean have positive deviations. Data values that are below the mean have negative deviations. Formulaically, we say

$$\text{deviation } = (\text{ data value } - \text{ mean }) = x_i - \bar{x}$$

2. Calculate the deviations for each of the samples and enter them in the tables below.

Kansas City Royals

Value	Deviation
4.83	
0.7	
4.75	
0.71	
1.3	
0.71	

Los Angeles Dodgers

Value	Deviation
6	
0.73	
2.75	
1.35	
0.72	
21	

3. Based on the deviations, which sample is more spread out?

Data values with large deviations (that are farther away from the mean) contribute more to the variability in the data set. Values with small deviations do not contribute as much to the overall variability of a data set. To measure the total amount of variability, we need to combine the deviations into a single number.

4. One way we might do this is by finding the average deviation from the mean. Let's try it! Add all the deviations for the Kansas City Royals. Interpret the result.

Standard Deviation

The standard deviation is a measure of variability that describes the typical deviation from the mean for all values in a set of data. To compute the **standard deviation of a sample**, we complete the following steps:

Step 1: Find the mean and find all deviations from the mean.

Step 2: Square each deviation.

Step 3: Add all squared deviations and divide by one less than the sample size.

Step 4: Take the square root of the result from step 3.

---

5. Complete the table with the square deviations for each value in the sample. Find the total square deviations.

   Kansas City Royals

Value	Deviation	$(\text{Deviation})^2$
4.83		
0.7		
4.75		
0.71		
1.3		
0.71		

   Total square deviations:

6. The sum of squared deviations is one way to represent the variability in a distribution. But it is not a commonly used measure. A more commonly used measure is the **sample variance** which is found by dividing the sum of squared deviations by one less than the sample size. In other words, the sample variance is an average of the squared deviations. Formulaically, we say

$$s^2 = \frac{\text{sum of squared deviations}}{\text{sample size minus 1}} = \frac{\Sigma\,(x_i - \bar{x})^2}{n-1}$$

   Compute the sample variance for the Kansas City Royals.

7. The **sample standard deviation**, denoted by s, is the square root of the sample variance. Calculate the sample standard deviation for the Kansas City Royals. Include units in your answer.

8. Can the standard deviation be negative? Explain.

9. Can the standard deviation be zero? Explain.

The process of calculating the standard deviation by hand has many steps and is time consuming, even for small sets of data. Therefore, we will usually use technology to compute the mean and standard deviation for us.

10. Use the following instructions to use the desmos graphing calculator to compute the sample standard deviation for the Los Angeles Dodgers.
    a. Go to https://www.desmos.com/calculator.
    b. Copy the values from the data set into the first line. Type $D = [6, 0.73, 2.75, 1.35, 0.72, 21]$. Hit enter on your keyboard to go to the next line.
    c. Type mean(D) to compute the sample mean. Hit enter on your keyboard to go to the next line. Alternatively, you can open the keypad, click on the functions button, scroll to statistics, and select mean. Enter the name of the set, D, in the empty parentheses. Hit enter on your keyboard to go to the next line. $\bar{x} =$ _____
    d. Type stdev(D) to compute the sample standard deviation. Alternatively, you can open the keypad, click on the functions button, scroll to statistics, and select stdev. Enter the name of the set, D, in the empty parentheses. $s =$_____

11. Compare the sample standard deviations found in 7 and 8. Which standard deviation is higher? Which distribution is more spread out? Do these values adequately represent the spread/variability in each sample of salaries?

12. In the sample of salaries from the Los Angeles Dodgers, which value impacts the standard deviation the most?

13. Let's say that 21 was incorrectly recorded as 12. Use desmos to find the standard deviation of the set with the error in it: $D = [6, 0.73, 2.75, 1.35, 0.72, 12]$. How does this error affect the sample standard deviation?

    Old standard deviation:

    New standard deviation (from set with error):

Outliers and skewing have a large effect on the standard deviation. Therefore, we say that the standard deviation is not a resistant measure of variability.

Summary

We have examined in detail two measures of center (mean and median) and three measures of variability or spread (range, interquartile range, and standard deviation). We calculated the IQR using the median. We calculated the standard deviation using the mean. In deciding which measures of center and spread to use, we need to remember two things:

- Mean and standard deviation go together. Median and IQR go together.
- Both the mean and standard deviation are influenced by outliers and skew.

When the data are skewed or contain outliers, we usually use the median and IQR to summarize the data. When the data are reasonably symmetric we use the mean and standard deviation. In addition, these summary values are never enough. We should always look at a graph as well. This can be a dotplot, histogram, or boxplot.

2.4: Quantifying Variability Relative to the Mean is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by LibreTexts.

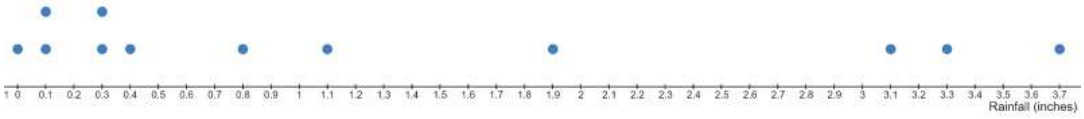
- **Current page** is licensed CC BY-NC-SA 4.0.

- **1.2: The Statistical Analysis Process** by Hannah Seidler-Wright is licensed CC BY-NC-SA 4.0.

2.4.1: Exercises

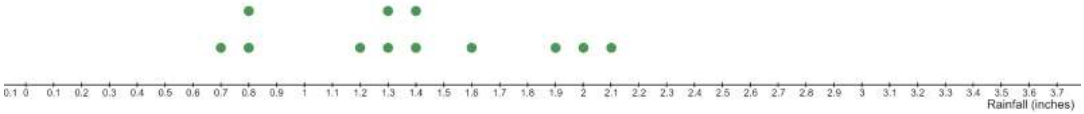
1. The dotplots below summarize the average monthly rainfall in California and Utah.

California Rainfall



Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

Utah Rainfall



Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

a. Which of the distributions have the most variation? Justify your answer.

b. Decide which measure of spread (standard deviation or IQR) would be best to use to compare the variation in the rainfall of CA and UT. Explain.

c. Select the value that best represents the standard deviation for Utah rainfall: 1 inch or 2 inches? Justify your answer.

2. Below is Bruno's calculation of the sample standard deviation for a sample of five pro basketball players by hand. Explain any errors he made and write a sentence or two explaining to Bruno how to fix his mistakes.

Bruno's work and solution:

Heights of 5 Pro Basketball Players

6.1	6.3	6.3	6.6	6.5

$$\text{The mean is } \frac{6.1+6.3+6.3+6.6+6.5}{5} = 6.36 \text{ feet.}$$

Height	6.1	6.3	6.3	6.6	6.5
Deviation	$6.1-6.36=-2.6$	$6.3-6.36=-0.06$	$6.3-6.36=-0.06$	$6.6-6.36=0.24$	$6.5-6.36=0.14$
Squared Deviation	$-2.6^2=-0.0676$	$-0.06^2=-0.0036$	$-0.06^2=-0.0036$	$0.24^2=0.0576$	$0.14^2=0.0196$

The sum of squared deviations is 0.0024. We divide by the sample size to get $\frac{0.0024}{5} = 0.00048$.

Then we square to get $\sqrt{0.00048} \approx 0.022$ feet.

3. Given below are the heights (in feet) of 5 randomly selected pro football and basketball players.

Heights of 5 Pro Football Players

6.7	6.3	6.2	6.5	5.9

Heights of 5 Pro Basketball Players

6.1	6.3	6.3	6.6	6.5

a. Compute the mean for each of the samples.

b. Compute the sample standard deviation for each set by hand. Round to three decimal places.

c. Compare the variation in the samples using the calculation from b.

d. Suppose an error was made when inputting the data for the sample of football players and 6.7 got recorded as 7.6. *Without doing any calculations*, decide which measures of center and spread will change. Explain how each measure is affected and why.

4. The following data represents miles per gallon gasoline consumption (highway) for a random sample of 55 makes and models of passenger cars.

30	25	33	18	30	20	29	25	13	35	28
27	24	52	20	24	25	27	24	13	32	28
22	15	49	23	24	27	24	28	21	33	25
25	35	10	24	24	24	27	33	28	29	29
24	35	27	25	18	32	26	30	37	31	31

a. Use desmos to compute the mean and sample standard deviation for the miles per gallon. Round the mean to one decimal place and the sample standard deviation to two decimal places.

$\bar{x} =$ _____

$s =$ _____

b. Interpret the standard deviation in context.

This page titled 2.4.1: Exercises is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Hannah Seidler-Wright.

- **Current page** by Hannah Seidler-Wright is licensed CC BY-NC-SA 4.0.
- **1.2: The Statistical Analysis Process** by Hannah Seidler-Wright is licensed CC BY-NC-SA 4.0.

CHAPTER OVERVIEW

3: Probability

3.1: Introduction to Probability

3.1.1: Exercises

3.2: Marginal, Joint, and Conditional Probability

3.2.1: Exercises

3.3: The Addition and Complement Rules

3.3.1: Exercises

This page titled 3: Probability is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Hannah Seidler-Wright.

3.1: Introduction to Probability

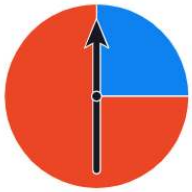
In this section, we will explore probability through the context of experiments. An experiment is a process whose outcomes are unknown. Imagine we are playing a game with a spinner.

Probability from a Spinner

Here's how it works. You'll spin the spinner. How do you win? That's up to you. Two options:

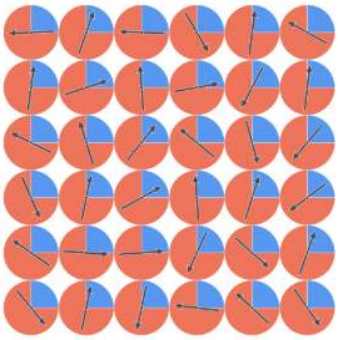
- Spinner lands on RED, you win.
- Spinner lands on BLUE, you win.

Which option do you want? Explain why.



Images were created with Polypad by Amplify, polypad.amplify.com.

If you were to play the game many times (say, 36 times in a row), which option would you choose? Explain why.



Images were created with Polypad by Amplify, polypad.amplify.com.

Damian plays the game 36 times. The number of times the spinner landed on red and on blue are given in the table below. Complete the third column in the table with the relative frequencies written as fractions and decimals rounded to three decimal places.

	Number of Spins	**Relative Frequency**
Landed on Red	26	
Landed on Blue	10	
Total	36	

The table shows the results for an entire class. Write the relative frequencies (**the empirical probability**) as fractions and decimals rounded to three decimal places into the table. Which event appears more likely? Explain.

	Number of Spins	**Relative Frequency**
Landed on Red	935	
Landed on Blue	325	
Total	1260	

Probability is the measure of the likelihood of a random event or chance behavior occurring. When we use outcomes from previous repetitions or trials of an experiment to calculate the probability of the next trial, we are calculating **Empirical Probability**. **Theoretical Probability** is calculated by dividing the number of outcomes in an event by the total number of all possible outcomes.

- What is the (theoretical) probability that the spinner lands on red?

- What is the (theoretical) probability that the spinner lands on blue?

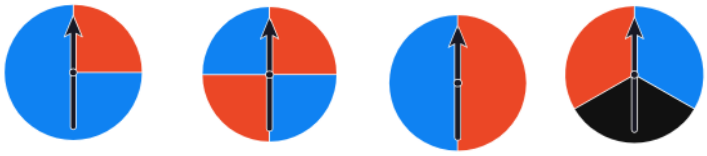
In this class, the experiment has been repeated _____ times (the spinner was spun _____ times). The empirical probability (or proportion, or relative frequency) for the event "Landed on Red" is trending toward the theoretical probability, $\frac{3}{4} = 0.75 = 75\%$. Similarly, the empirical probability (or proportion or relative frequency) for the event "Landed on Blue" is trending toward the theoretical probability, $\frac{1}{4} = 0.25 = 25\%$. This phenomenon is referred to as **the law of large numbers**. As the number of repetitions of a probability experiment increases, the proportion with which a certain outcome is observed gets closer to the probability of the outcome. For this spinner, it is more likely to land on red, and less likely to land on blue.

Can probability be more than 1? Explain your reasoning.

Can probability be negative? Explain your reasoning.

Malik and Mia played 36 rounds with a different spinner. Their results are shown in the table. Select the spinner(s) that they most likely used. Explain your reasoning.

	Number of Spins
Landed on Red	17
Landed on Blue	19
Total	36



Images were created with Polypad by Amplify, polypad.amplify.com.

Here's a spinner Calon created where landing on red is almost impossible, and landing on blue is almost certain.



Images were created with Polypad by Amplify, polypad.amplify.com.

Enter the relative frequencies into the table. Write the number of spins you might expect to see that landed on red and on blue in the table. Then calculate the relative frequencies as decimals rounded to three decimal places.

	Number of Spins	Relative Frequency
Landed on Red		
Landed on Blue		
Total	36	$\frac{36}{36} = 1 = 100\%$

- What number should probabilities always add to?

Malik claims he can create a spinner where red and blue are both unlikely. Mia says that's not possible. Who do you think is correct? Justify your answer.

Summary

The Law of Large Numbers states that empirical probability approaches theoretical probability as the number of trials of a probability experiment are repeated.

The **sample space**, S, of a probability experiment is the collection of all possible outcomes. For example, the sample space for our spinner experiment is $S = \{\text{lands on blue, lands on red}\}$. An **event**, sometimes denoted E, is any collection of one or more outcomes from a probability experiment. For example, one event from our spinner experiment could be $E = \{\text{lands on blue}\}$.

Basic Rules of Probability:

- The probability of any event E, is between 0 and 1 (inclusive). If the probability of an event is 0, then the event is impossible. If the probability of an event is 1, then the event is certain. Probability is never negative (less than 0) and is never more than 1.
- The sum of the probabilities of all outcomes must equal 1.

Probability from a Pair of Dice

Now, consider a pair of fair dice (a fair die is a die in which every outcome is equally likely to be rolled). Write all possible outcomes from rolling *one* die.

There are 6 possible outcomes from the first die, and 6 possible outcomes from the second die. Therefore, the sample space is composed of all combinations of these outcomes. There are 36 elements in the sample space. Fill in the table with the remaining outcomes.

Roll a…	**1 on the 2nd Die**	**2 on the 2nd Die**	**3 on the 2nd Die**	**4 on the 2nd Die**	**5 on the 2nd Die**	**6 on the 2nd Die**
1 on the 1st Die	1, 1	1, 2	1, 3	1, 4	1, 5	1, 6
2 on the 1st Die	2, 1	2, 2	2, 3	2, 4	2, 5	
3 on the 1st Die	3, 1	3, 2	3, 3	3, 4		
4 on the 1st Die	4, 1	4, 2	4, 3			
5 on the 1st Die	5, 1	5, 2				
6 on the 1st Die	6, 1					

There are 36 elements in the sample space. We will consider the total number of dots on the dice as an outcome in the sample space. Fill in the table with the remaining sums.

Roll a…	1 on the 2nd Die	2 on the 2nd Die	3 on the 2nd Die	4 on the 2nd Die	5 on the 2nd Die	6 on the 2nd Die
1 on the 1st Die	2	3	4	5	6	7
2 on the 1st Die	3	4	5	6		
3 on the 1st Die	4	5	6			
4 on the 1st Die	5	6				
5 on the 1st Die	6					
6 on the 1st Die	7					

The sample space that contains all possible outcomes from rolling two fair dice is

$$[2, 3, 4, 5, 6, 7, 3, 4, 5, 6, 7, 8, 4, 5, 6, 7, 8, 9, 5, 6, 7, 8, 9, 10, 6, 7, 8, 9, 10, 11, 7, 8, 9, 10, 11, 12]$$

The probability of rolling a 1, (which can be written as $P(1)$, read "P of one") is $\frac{0}{36} = 0$. It is impossible to roll a 1. The probability of rolling a 2, $P(2) = \frac{1}{36}$ which is about 0.028. It is very unlikely to roll a 2. The probability of rolling a 3 is $P(3) = \frac{2}{36}$ which is about 0.056. Notice that there are two ways to roll a 3:

- We roll 1 on the first die and 2 on the second die
- We roll 2 on the first die and 1 on the second die

$$\text{In general, } P(E) = \frac{\text{the number of ways to roll } E}{\text{the total number of outcomes}}.$$

$P(4) = \frac{3}{36} \approx$

$P(5) = \frac{4}{36} \approx$

$P(6) = \frac{5}{36} \approx$

Compute the following probabilities. Enter your answers as decimals rounded to three decimal places:

$P(7) =$ _____$\approx$_____

$P(8) =$ _____≈_____

$P(9) =$ _____≈_____

$P(10) =$ _____≈_____

$P(11) =$ _____≈_____

$P(12) =$ _____≈_____

Pick a number. If it is rolled from two dice, you win!

- Which number should you pick? (Think about probabilities). Explain how you made your choice.

- Try it! Go to https://www.random.org/dice/ and click roll dice. Did you win?

This page titled 3.1: Introduction to Probability is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Hannah Seidler-Wright.

3.1.1: Exercises

1. You are playing a game that involves picking different colored blocks from a bag. The bag contains 5 blue blocks and 3 green blocks. The game is won by picking a green block from the bag.
   a. What is the probability that the game is won?

   b. What is the probability that the game is not won?

2. You are creating a game for your classmates to play. You will fill a bag with a mixture of blue blocks and green blocks. You want the probability of winning the game by selecting a green block to be 35%. How many blue and green blocks should you fill the bag with?

3. John's bag of blocks contains 10 blocks, some of which are blue and some of which are green.
   a. You play John's game 20 times. Below is a table containing the frequencies for which you picked a blue block and a green block.

	Frequency
Blue	5
Green	15
Total	20

   What is the experimental probability of picking a blue block from the bag? What is the experimental probability of picking a green block from the bag?

   b. Based on these observations, how many blocks do you think are blue?

   c. Based on these observations, what do you think the theoretical probability of selecting a blue block is?

d. You play John's game 200 times. Below is a table containing the frequencies for which you picked a blue block and a green block. Complete the table by filling in relative frequencies. Estimate how many blue and green blocks are in John's bag of 10 blocks.

	Frequency	Relative Frequency
Blue	61	
Green	139	
Total	200	

4. Tatiana's game has 2 red blocks and 6 blue blocks. If you play her game 100 times, about how many times do you expect to pick a red block?

5. Pablo is playing a game that involves flipping a coin and rolling a 8-sided die.
   a. Construct the sample space. How many possible outcomes are in the sample space?

   b. What is the probability of rolling an even number and heads?

   c. What is the probability of rolling a 1 or 8 and tails?

6. Imagine that a pharmaceutical company has developed a new drug to treat anxiety. The company thoroughly tested the new drug in many clinical trials. Clinical trials, by law, must involve a very large number of patients. The results of the clinical trials for the new drug show that it lowers anxiety in 75% of patients. A local doctor gave the new drug to her patients with anxiety. After taking the drug, 90% of those patients experienced less anxiety. The doctor was so excited that she now claims the drug is a medical breakthrough.

   a. Which is more likely to be closer to the true percentage of patients who will have their anxiety reduced by the drug? 75% or 90%? Explain your answer.

   b. The drug's success rate in reducing patients' anxiety in the clinical trials was different from the drug's success rate for the doctor's patients. Why might that be? In explaining your reasoning, think about the Law of Large Numbers.

3.1.1: Exercises is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by LibreTexts.

3.2: Marginal, Joint, and Conditional Probability

We have learned how probabilities can be estimated using relative frequencies from a table that separates data into groups according to the values of some variable. Similar methods can be used to investigate relationships between variables in a two-way table. We can compute various types of probabilities from a two-way table. In this section, we will explore **marginal, joint, and conditional probabilities**. We will do this using the data presented in the following two-way table.

The United States Census Bureau collects large amounts of data about the population of the United States.[3] The two-way table below presents marital status and whether an individual is male or female for the population of the United States. Numbers in this table are in millions of people. For example, 32.8 represents 32,800,000. Each value has been rounded to the nearest 0.1 million which is equivalent to 100,000.

Marital Status	Married	Widowed	Divorced	Separated	Never Married	Total
Male	64.5	3.4	12.2	2.1	47.5	129.7
Female	63.4	11.8	16.5	2.9	41.5	136.1
Total	127.9	15.2	28.7	5	89	265.8

Marginal Probability

If a probability is computed using only totals in the margins from the table (the far right column, or the bottom row in the above table), it is called a **marginal probability**. Let A represent a row or column heading in the table above. Then the marginal probability, $P(A)$, is given by

$$P(A) = \frac{\text{number of ways } A \text{ can happen}}{\text{total number of outcomes}}$$

In this fraction, the numerator (top of the fraction) is a row or column total from the margins of the table, and the denominator (bottom of the fraction) is the grand total (found in the lower right cell in the table above).

Use the two-way table above to find some probabilities:

1. What is the probability that a randomly selected person from the adult population in the US is male? In probability notation, you are asked to find $P(\text{male}) = \frac{\text{number of males}}{\text{grand total}}$. Round your answer to three decimal places.

2. What is the probability that a randomly selected person from the adult population in the US is divorced? Include probability notation in your answer. Round your answer to three decimal places.

   $P(\_\_\_\_\_) = \_\_\_\_\_$

Joint Probability

Frequently, we need to find a probability that involves two outcomes, both A and B. In a two-way table, probability of both events occurring, $P(AandB) = P(A \cap B)$, is found by taking the number of ways A and B can happen, and dividing by the grand total.

$$P(A \text{ and } B) = P(A \cap B) = \frac{\text{number of ways } A \text{ and } B \text{ can happen}}{\text{total number of outcomes}}$$

3. What is the probability that a randomly chosen adult is male and divorced? Round your answer to three decimal places.

   $P(\text{male} \cap \text{divorced}) =$ _____ $=$

4. What is the proportion of adults that are female and never married? Include probability notation in your answer. Round your answer to three decimal places.

Multiplication Rule for Independent Events

When frequencies are given for values of two categorical variables in a two-way table, the joint probability can be computed without using any special rules. Without a two-way table, joint probability can sometimes be computed using the multiplication rule. If the events are **independent**, where the probability of one event does not depend on the occurrence of the other, then the probability can be computed using multiplication. If A and B are independent, formulaically, we can say

$$P(A \cap B) = P(A) \cdot P(B)$$

This is called the **multiplication rule for independent events.**

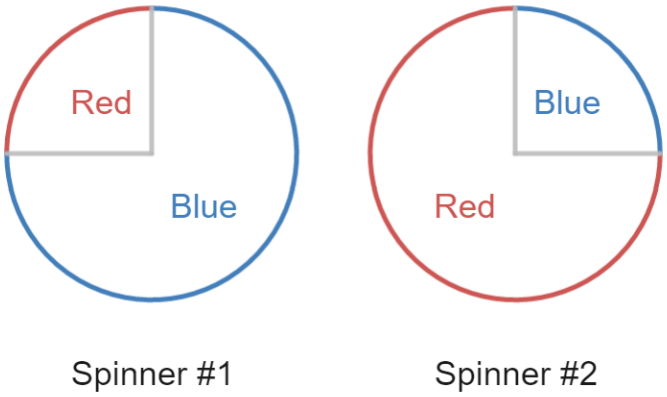
Consider a standard 52 deck of cards. What is the probability of selecting a queen and then an ace? The probability of selecting a queen is $\frac{4}{54}$ since there are 4 queens and 52 cards. This event changes the number of cards remaining in the deck. The probability of selecting an ace depends on these new frequencies. The probability of selecting an ace is now $\frac{4}{51}$ since there are 4 aces and 51 cards remaining in the deck. In this scenario, the events are dependent because **we do not return the chosen card(s) to the deck**. The events occurred **without replacement**.

$$P(\text{ queen } \cap \text{ ace }) = \frac{4}{52} \cdot \frac{4}{51} = \frac{4}{663} \approx 0.0060$$

Alternatively, if we return the chosen card(s) to the deck, the events are independent. In this scenario, the events occurred **with replacement**.

$$P(\text{ queen } \cap \text{ ace }) = \frac{4}{52} \cdot \frac{4}{52} \approx 0.0059$$

5. Consider a new spinner game:



Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

a. You win if both spinners land on red. What is the probability that you win?

b. You win if both spinners land on blue. What is the probability that you win?

c. You win if the first spinner lands on red and the second spinner lands on blue. What is the probability that you win?

Conditional Probability

When two variables are **dependent**, the probability of one variable's outcome is influenced by the outcome of the other variable. If A and B are dependent, then the probability of B depends on a condition: whether A happens or not. We will calculate the probability of B given that A will occur. This is called **conditional probability**. The conditional probability of B, given that A has occurred is denoted $P(B \mid A)$ (read as "the probability of B given A").

6. Let's return to our census data from 2020. Let A represent widowed adults in the US, let B represent male adults in the US, and let C represent female adults in the US.

Marital Status	Married	A = Widowed	Divorced	Separated	Never Married	Total
B = Male	64.5	3.4	12.2	2.1	47.5	129.7
C = Female	63.4	11.8	16.5	2.9	41.5	136.1
Total	127.9	15.2	28.7	5	89	265.8

   a. Find the proportion of widowed individuals who are male. In other words, what is the probability that a randomly selected *widowed individual* is male? Now, we are not selecting from the entire population, but rather, from the subset that is defined by a condition. Therefore, the denominator of the fraction will be the total *widowed individuals* and the numerator will be the number of people who are *male and widowed*. Round your answer to three decimal places.

$$P(\text{male} \mid \text{widowed}) = P(B \mid A) = \frac{\text{number of male and widowed}}{\text{total number of widowed}} = \frac{P(A \cap B)}{P(A)} = \qquad \approx$$

   b. Find the proportion of widowed individuals who are female. Include probability notation in your answer and round to three decimal places.

   c. The number of widowed individuals who are male is close to the number of separated individuals who are male. Does this mean that the probabilities are the same? Explain.

The multiplication rule is altered for dependent events. If A and B are dependent (where the probability of one depends on the occurrence of the other), then $P(A \cap B) = P(A) \cdot P(B \mid A)$. Applying this rule, we get

$$P(A \cap B) = P(A) \cdot P(B \mid A) = P(\text{ widowed }) \cdot P(\text{ male } \mid \text{ widowed }) = \frac{15.2}{265.8} \cdot \frac{3.4}{15.2} = \frac{3.4}{265.8}$$
$$= \frac{\text{ways } A \text{ and } B \text{ can both happen}}{\text{total number of outcomes}}$$

This is the **multiplication rule for dependent events**. Remember that A and B is a single event in a two-way table, and when this is the case, the multiplication rule is unnecessary.

Reference

[3] https://data.census.gov/cedsci/table?q=Marital%20Status%20and%20Marital%20History&tid=ACSST5Y2020.S1201&moe=false

This page titled 3.2: Marginal, Joint, and Conditional Probability is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Hannah Seidler-Wright.

3.2.1: Exercises

1. Imagine a survey of students, faculty, and administrators is conducted to gauge the importance of using social media to interact with peers and colleagues. The data is summarized in the two-way table below.

	Very important	Somewhat important	Not important	Total
Students	65	82	52	199
Faculty	15	31	36	82
Administrators	9	15	15	39
Total	89	128	103	320

   a. Name one way in which students, faculty, and administrators differ in their attitudes regarding social media.

   b. What is the probability that a subject felt social media was very important? Use probability notation in your answer. Write your answer as a fraction and decimal rounded to three decimal places.

   c. What is the probability that a subject felt social media was somewhat important? Use probability notation in your answer. Write your answer as a fraction and decimal rounded to three decimal places.

   d. What is the probability that a subject felt social media was not important? Use probability notation in your answer. Write your answer as a fraction and decimal rounded to three decimal places.

   e. What is the probability that a subject is a student? Use probability notation in your answer. Write your answer as a fraction and decimal rounded to three decimal places.

f. What is the probability that a subject is a faculty? Use probability notation in your answer. Write your answer as a fraction and decimal rounded to three decimal places.

g. What is the probability that a subject is an administrator? Use probability notation in your answer. Write your answer as a fraction and decimal rounded to three decimal places.

h. What is the probability that a subject is a student and felt that social media was very important? Use probability notation in your answer. Write your answer as a fraction and decimal rounded to three decimal places.

i. What is the probability that a subject is a faculty and felt that social media was very important? Use probability notation in your answer. Write your answer as a fraction and decimal rounded to three decimal places.

j. What proportion of students felt that social media was very important? Use probability notation in your answer. Write your answer as a fraction and decimal rounded to three decimal places.

k. What proportion of those that felt that social media was somewhat important were faculty? Use probability notation in your answer. Write your answer as a fraction and decimal rounded to three decimal places.

l. Which is more likely: Finding a student who thinks social media is not important, or finding a faculty member who thinks social media is not important. Defend your response.

2. The forecast for an upcoming weekend shows a 30% chance of rain on Saturday and a 20% chance of rain on Sunday. Assume these events are independent. What is the probability of it raining on both days?

This page titled 3.2.1: Exercises is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Hannah Seidler-Wright.

3.3: The Addition and Complement Rules

The Addition Rule

Two events are considered **mutually exclusive** if they cannot both occur. If events A and B are mutually exclusive, the probability that event A or event B will occur is the sum of the individual probabilities. Formulaically speaking, If events A and B are mutually exclusive, $P(A \text{ or } B) = P(A \cup B) = P(A) + P(B)$.

If events are **not mutually exclusive**, they *can* both occur. Here we add the probability of each event, and subtract the overlap (A and B) because it was counted twice. Formulaically speaking, If events A and B are not mutually exclusive, $P(A \text{ or } B) = P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

These make up the **addition rule**.

1. Let's return to the census data on sex and marital status.

Marital Status	Married	Widowed	Divorced	Separated	Never Married	Total
Male	64.5	3.4	12.2	2.1	47.5	129.7
Female	63.4	11.8	16.5	2.9	41.5	136.1
Total	127.9	15.2	28.7	5	89	265.8

   a. Are the events "divorced" and "separated" mutually exclusive? Explain.

   b. What is the probability that a randomly selected adult in the US is divorced or separated? Round your answer to three decimal places. Show your thinking using the table.

Marital Status	Married	Widowed	Divorced	Separated	Never Married	Total
Male	64.5	3.4	12.2	2.1	47.5	129.7
Female	63.4	11.8	16.5	2.9	41.5	136.1
Total	127.9	15.2	28.7	5	89	265.8

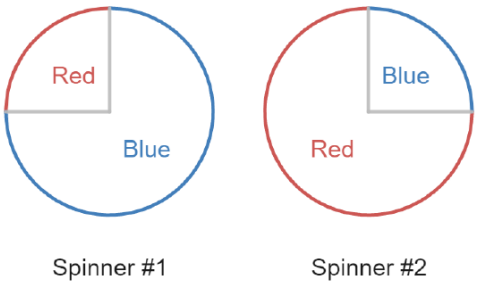
$$P(\text{ divorced } \cup \text{ separated }) = \_\_\_\_\_ + \_\_\_\_\_ \approx$$

   c. Are the events "separated" and "female" mutually exclusive? Explain.

d. What is the probability that a randomly selected adult in the US is separated or female? Use probability notation in your answer and round your answer to three decimal places. Show your thinking using the table.

Marital Status	Married	Widowed	Divorced	Separated	Never Married	Total
Male	64.5	3.4	12.2	2.1	47.5	129.7
Female	63.4	11.8	16.5	2.9	41.5	136.1
Total	127.9	15.2	28.7	5	89	265.8

2. Here's a new game with two spinners. For this game, we say the spinners "match" if they land on the same color (e.g., both red, or both blue).



Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

a. What is the probability that both spinners land on red?

b. What is the probability that both spinners land on blue?

c. What is the probability that the spinners match (both red or both blue)? Use probability notation in your answer.

The Complement Rule

The **complement** of an event A is "not A" and is denoted A^c. Because one of these events must occur, their probabilities must add to one. Therefore, we can compute the probability of event A using the rule of complements.

$$P(A) = 1 - P(\text{ complement of } A) = 1 - P(A^c) \text{ and } P(\text{ complement of } A) = P(A^c) = 1 - P(A)$$

This rule is useful when the computation of the probability of event A is complicated, but the probability of the complement of A is simpler.

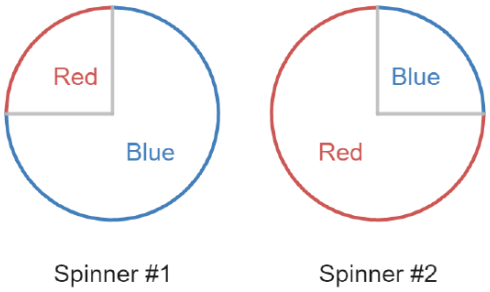
3. Let's return to the census data on sex and marital status.

Marital Status	Married	Widowed	Divorced	Separated	Never Married	Total
Male	64.5	3.4	12.2	2.1	47.5	129.7
Female	63.4	11.8	16.5	2.9	41.5	136.1
Total	127.9	15.2	28.7	5	89	265.8

a. Define event A to be "widowed, divorced, separated, or never married". What is the complement of A in words?

b. Compute $P(A^c)$.

4. Here's a new game with two spinners. For this game, we say the spinners "do not match" if they don't land on the same color.



Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

Use your answer to 2c. to find the probability that the spinners don't match. Use probability notation in your answer.

5. Marjorie forgets to prepare for a multiple choice quiz so she randomly guesses on all 3 questions. Each question has 5 answer options.

   a. What is the probability that Marjorie gets a question correct on the quiz?

   b. What is the probability that Marjorie gets a question incorrect on the quiz?

   c. What is the probability that Marjorie gets all 3 questions correct?

   d. What is the probability that Marjorie gets at least 1 question incorrect?

This page titled 3.3: The Addition and Complement Rules is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Hannah Seidler-Wright.

3.3.1: Exercises

1. Imagine a survey of students, faculty, and administrators is conducted to gauge the importance of using social media to interact with peers and colleagues. The data is summarized in the two-way table below.

	Very important	Somewhat important	Not important	Total
Students	65	82	52	199
Faculty	15	31	36	82
Administrators	9	15	15	39
Total	89	128	103	320

   a. What is the probability that a subject is a student or an administrator? Use probability notation in your answer. Write your answer as a fraction and decimal rounded to three decimal places.

   b. What is the probability that a subject feels social media is very important or somewhat important? Use probability notation in your answer. Write your answer as a fraction and decimal rounded to three decimal places.

   c. What is the probability that a subject is an administrator or feels social media is somewhat important? Use probability notation in your answer. Write your answer as a fraction and decimal rounded to three decimal places.

   d. What is the probability that a subject is not an administrator? Use probability notation in your answer. Write your answer as a fraction and decimal rounded to three decimal places.

   e. What is the probability that a subject does not feel social media is very or somewhat important? Use probability notation in your answer. Write your answer as a fraction and decimal rounded to three decimal places.

2. In a national pet owners survey, 39% of US households own at least one dog and 34% of households own a cat. Assume that 60% of US households own a dog or a cat.
   a. What is the probability that a randomly selected household owns neither a cat nor a dog?

   b. What is the probability that a randomly selected US household owns both a cat and a dog?

3. Your friend Ainsley asks you to play a game. They win if the roll of two fair 4-sided dice results in different values. You win if you roll matching dice. You will play three rounds with Ainsley.
   a. Compute the probability that you win all three rounds.

   b. Compute the probability that you lose at least one round.

4. The forecast for an upcoming weekend shows a 40% chance of rain on Saturday and a 65% chance of rain on Sunday. Assume these events are independent. What is the probability that it will rain at least once on the weekend?

5. Rafael is planning a 3-day vacation. The forecast for Friday, Saturday, and Sunday shows a 25% chance of rain on Friday, a 25% chance of rain on Saturday, and a 60% chance of rain on Sunday. Assume these events are independent. What is the probability that it will rain at least once on the weekend?

6. The director of a team of software engineers is tasked with selecting three of the engineers to work on a special short-term project. There are three female and five male engineers. All eight of the engineers are equally qualified, so random selection is used to form the group. What is the probability that the director randomly selects three females to make up the team? (Note that each individual can only be selected *once*).

This page titled 3.3.1: Exercises is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Hannah Seidler-Wright.

CHAPTER OVERVIEW

4: Discrete Probability Distributions

4.1: Discrete Random Variables

4.1.1: Exercises

4.2: The Geometric Distribution

4.2.1: Exercises

4.3: The Binomial Distribution

4.3.1: Exercises

This page titled 4: Discrete Probability Distributions is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Hannah Seidler-Wright.

4.1: Discrete Random Variables

Wordle is a popular puzzle game that was created by the New York Times in October of 2021. To play the game, a player tries to guess the five letter word of the day. Each player has a total of six tries to guess the word. After each guess, the colors of tiles will change to indicate how close the guess was to the word. If a tile turns green, the letter is in the correct spot. If a tile turns yellow, the letter is in the word but not in the correct spot. If a letter turns gray, it is not a letter in the word. According to twitter, on April 28, 2022, 98,967 individuals played Wordle.



Adapted from "Wordle Emoji Screenshot" by Levi OP is licensed under CC BY-SA 4.0.

The following table sorts the number of people who played by the number of guesses taken to solve the puzzle.

1. Complete the table by computing the relative frequencies (rounded to three decimal places) for each value of x. Recall that n is the total number of people in the table.

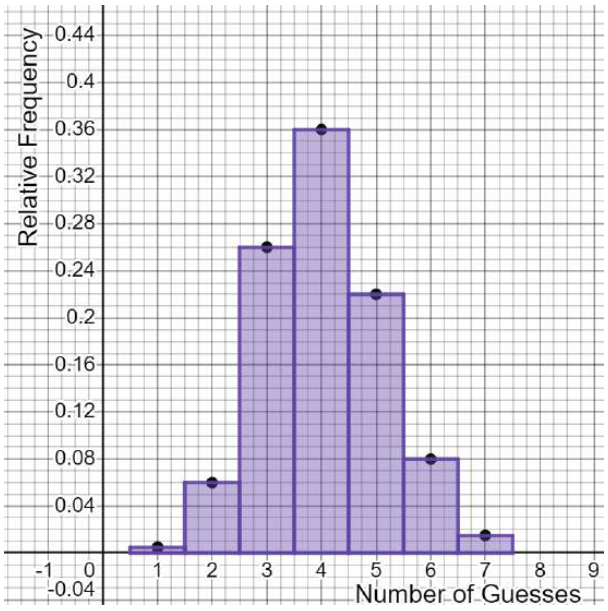
Number of Guesses, x_i	**Number of people, f_i**	**Relative Frequency, f_i/n**
1	494	
2	5938	
3	25731	
4	35628	
5	21772	
6	7917	
More than 6	1487	

The number of guesses is an example of a **discrete random variable**. We can list the values of a discrete random variable in order. In other words, we can *count* the values of the variable. In this example, the variable x can take on the values {1, 2, 3, 4, 5, 6, 7} (if an individual takes 7 guesses, they have lost the game because they have run out of tries). If we randomly select a game from this population, the probability that it was won in a certain number of tries is equal to the relative frequency of the chosen number of guesses.

A **discrete probability distribution** assigns a probability to all possible values of a discrete random variable. Each probability is between 0 and 1, and the sum of all probabilities is 1.

Finding Probability from a Discrete Probability Distribution

The discrete probability distribution in the Wordle example above can be represented with a histogram.



Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

2. What is the shape of the histogram? What does the shape suggest to you about the number of guesses taken to win Wordle?

3. Imagine that we randomly select a Wordle game. Use the probability distribution above to answer the following questions.

   a. Find the probability that the game is won in more than 5 guesses. Using probability notation, we would denote this probability as $P(x > 5)$. (Hint: we are asking what the probability is that the word was guessed on the sixth try OR the word was not guessed).

   b. Find the probability that the game is won in at most 4 guesses. Use probability notation in your answer.

c. Find the probability that the game is won between 3 and 6 guesses. Use probability notation in your answer.

d. Find the probability that the game is won (in other words, the number of guesses is *not* 7). Use probability notation in your answer.

e. Three friends are competing with each other through Wordle. What is the probability that all three individuals guess in less than 3 tries? Assume guesses are independent. Use probability notation in your answer. Round your answer to five decimal places.

f. You want to know how you compare to other Wordle players. You want to know the average number of guesses it takes to win Wordle. Use the histogram to estimate the mean number of guesses.

Finding the Mean of a Discrete Probability Distribution

The mean of a probability distribution is the mean of a distribution of relative frequencies. If the probabilities are exact, the mean is a **population mean**. We have seen that the symbol used to denote a population mean is μ (read as "mew") and the symbol used to denote a sample mean is $\bar{x}$ (read as "x bar"). In a previous exercise, you were tasked with computing a mean from a frequency distribution table. The formula used was

$$\text{mean } = \frac{\sum x_i \cdot f_i}{n}$$

where $n = \sum f_i$. We can use algebra to write an equivalent equation,

$$\text{mean } = \sum \frac{f_i}{n} \cdot x_i$$

Noting that f_i/n is a relative frequency, and knowing that relative frequencies are regarded as probabilities in a discrete probability distribution, we can say,

$$\text{population mean } = \mu = \sum x_i \cdot P(x_i)$$

In summary, the **mean** or **expected value** (denoted $E(x)$) **of a discrete probability distribution** is found using

$$\mu = E(x) = \sum x_i \cdot P(x_i)\,.$$

The mean of a probability distribution represents both the average value of the random variable and the center of the distribution.

4. Compute the expected number of guesses it takes to win Wordle. Compare this to the estimation you made in 3f.

Finding the Standard Deviation of a Discrete Probability Distribution

The population standard deviation of a probability distribution is denoted by the Greek letter σ (read "sigma"). It is computed by adding all square deviations, dividing by the total frequencies, and square rooting. Using algebra, we can manipulate the inside of the square root:

$$\text{population standard deviation } = \sqrt{\frac{\Sigma (x_i - \mu)^2}{n}}$$

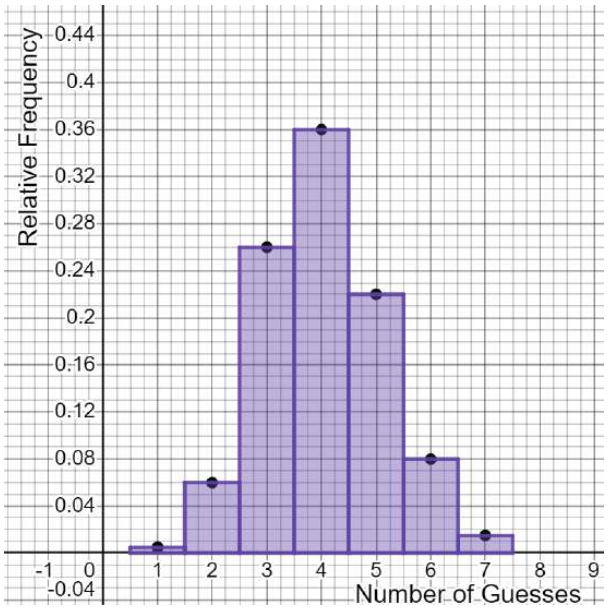
$$= \sqrt{\frac{f_1 (x_1 - \mu)^2 + f_2 (x_2 - \mu)^2 + \ldots + f_n (x_n - \mu)^2}{n}} = \sqrt{\sum \frac{f_i}{n} \cdot (x_i - \mu)^2}$$

where $n = \sum f_i$. Notice that f_i/n is the same as the probability. Therefore, we can find the standard deviation of a probability distribution using the following formula:

$$\text{population standard deviation } = \sigma = \sqrt{\sum P(x_i) \cdot (x_i - \mu)^2}$$

5. For the number of guesses in the game of Wordle, the population standard deviation is $\sigma = 1.1$ guesses. Interpret the standard deviation in context.

Area and Probability



Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

6. In the Wordle histogram, the width of each bar is 1. Recall, the area of a rectangle is found using the width times the length.
   a. What is the area of the bar centered at 1?

   b. What is the area of the bar centered at 2?

   c. What is the relationship between a value's area and its probability?

   d. What is the total area of all the bars?

Apply it!

7. Finn is purchasing a new car from Ford. Ford offers buyers a maintenance plan which will cost an additional $35 per month. It will cover oil changes, routine maintenance, and any major repairs. The maintenance plan provides coverage for the first 6 years of ownership. On average, Finn drives 10,000 miles per year. He will get his car serviced every 5,000 miles. The average cost of service is $150.

   a. What is the cost of services for the first six years if Finn does not purchase the maintenance plan?

   b. Below is a probability distribution where you can see the potential cost of routine maintenance and possible major repairs in the first six years of car ownership. Standard services are not included in the costs. What is the expected cost of routine maintenance and possible major repairs for the first 6 years?

Annual Maintenance and Repair Costs for the First Six Years, \(x\)	0	$500	$1,000	$1,500	$2,000	$2,500
Probability, \(P(x)\)	0.02	0.52	0.21	0.14	0.08	0.03

   c. Compute the total expected cost for services, routine maintenance, and repairs for the first six years.

   d. What would Finn pay if he purchases the maintenance plan?

   e. Should Finn purchase the maintenance plan? Support your answer.

This page titled 4.1: Discrete Random Variables is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Hannah Seidler-Wright.

- **Current page** by Hannah Seidler-Wright is licensed CC BY-NC-SA 4.0.
- **1.2: The Statistical Analysis Process** by Hannah Seidler-Wright is licensed CC BY-NC-SA 4.0.

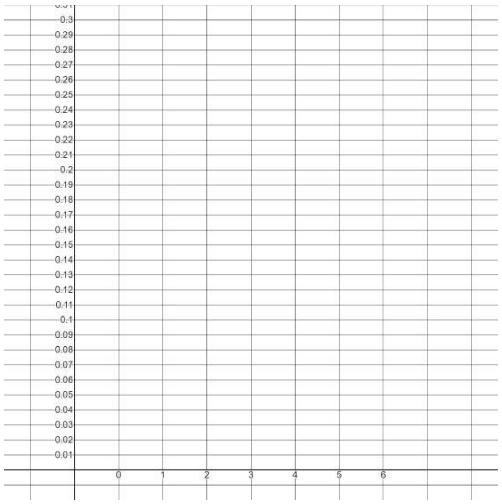
4.1.1: Exercises

1. A program at a local community college wants to evaluate its attrition rate, this is the number of semesters a student remains in the program. Over the years, they have established the following probability distribution.

   a. Using what you know about probability distributions, find $P(4)$ and enter it in the table below.

x = the number of semesters a student will remain in the program	0	1	2	3	4	5	6
$P(x)$	0.12	0.18	0.30	0.15		0.10	0.05

   b. Graph the probability histogram. Be sure to label the axes.



Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

   c. Find the probability that a student remains in the program for more than 3 semesters. Use probability notation in your answer.

   d. Find the proportion of students who remain in the program for at most 3 semesters. Use probability notation in your answer.

   e. On average, how long do you expect a student to remain in the program? Include units in your answer.

2. You set up a booth at a local fund-raising event. The game consists of rolling two six-sided dice. The dice are fair, so each individual roll of one die has a probability of 1/6. Players pay $5 per roll. A player who rolls a 2 or a 3 wins a prize that costs you $3. Players who roll an 11 or 12 win a prize that costs you $8. Players who roll other numbers win nothing. An average of 30 guests play your game each hour, and the event will go on for 8 hours. How much money do you expect to raise during the event? Use the table below to help guide your thinking.

Roll	2, 3	4, 5, 6, 7, 8, 9, 10	11, 12
How much money you make or lose			
Probability from roll			

3. If A and B are mutually exclusive, then find the probability of A and B.

4. The proportion of tweets made by adults in the US that are political is 33%.

   a. Compute the proportion of tweets made by adults in the US that are not political.

   b. You randomly read 5 tweets made by adults in the US. What is the probability that all 5 tweets are political? Assume the events are independent. Round to four decimal places.

   c. You randomly read 5 tweets made by adults in the US. What is the probability that all 5 tweets are not political? Assume the events are independent. Round to four decimal places.

   d. You randomly read 5 tweets made by adults in the US. What is the probability that at least one of the 5 tweets are political? Round to four decimal places.

5. Write a learning strategy that you haven't tried yet that you are interested in.

This page titled 4.1.1: Exercises is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Hannah Seidler-Wright.

- **Current page** by Hannah Seidler-Wright is licensed CC BY-NC-SA 4.0.
- **1.2: The Statistical Analysis Process** by Hannah Seidler-Wright is licensed CC BY-NC-SA 4.0.

4.2: The Geometric Distribution

There are many probability experiments where a trial has only two outcomes. For example, asking a group of individuals if they vote yes on a proposition, or randomly guessing on a multiple choice test. When we conduct a sequence of independent trials with only two outcomes per trial, we are conducting a binomial experiment.

1. Which of the following only have two possible outcomes:

   a. Rolling a 4 on a 6-sided die

   b. Examining the global temperature change over time.

   c. Measuring the height of adult in California

   d. Meeting a person that is infected with Covid-19

Characteristics of a Geometric Experiment

A geometric experiment is a probability experiment with the following characteristics:

- Each trial has exactly two possible outcomes which are labeled success and failure.
- The probability of success is the same for each trial. We denote the probability of success as p and the probability of failure as $q = 1 - p$.
- We look for when the first and only success occurs. There must be at least one trial, and in theory, we could repeat trials forever.

2. Go to https://www.random.org/dice/ and roll 1 die. Roll the die counting the number of trials it took to roll a 5. Keep track of your rolls in the table below.

Tally	On what attempt did you succeed in rolling a 5?

3. Assume we will roll a fair six-sided die.
   a. What is the probability of rolling a 5? We define rolling a 5 as success, and therefore, we are computing the probability of success.

   b. What is the probability that we will not roll a 5? Use the complement rule to compute the probability of failure.

4. Suppose we are rolling a fair six-sided die.
   a. What is the probability that we will roll a 5 (succeed) on the first attempt?

   b. What is the probability that we will roll a 5 (succeed) on the second attempt? In this case, we fail on the first try and succeed on the second try. Use the multiplication rule for independent events.

5. Suppose we are rolling a fair six-sided die.
   a. What is the probability that we will roll a die and succeed (roll a 5) on the third attempt? In this case, we fail on the first and second tries and succeed on the third try. Use the multiplication rule for independent events.

   b. What is the probability that we will roll a die and succeed (roll a 5) on the fourth attempt? In this case, we fail on the first and second and third tries and succeed on the fourth try. Use the multiplication rule for independent events.

6. What is the formula for computing geometric probability? So far, this is what we have come up with:

$$\begin{aligned} P(1) &= \frac{1}{6} \\ P(2) &= \frac{5}{6} \cdot \frac{1}{6} \\ P(3) &= \frac{5}{6} \cdot \frac{5}{6} \cdot \frac{1}{6} = \left(\frac{5}{6}\right)^2 \cdot \frac{1}{6} \\ P(4) &= \frac{5}{6} \cdot \frac{5}{6} \cdot \frac{5}{6} \cdot \frac{1}{6} = \left(\frac{5}{6}\right)^3 \cdot \frac{1}{6} \end{aligned} \tag{4.2.1}$$

   What patterns do you notice?

7. What is P(5)?

Geometric Probability

In general, the probability of succeeding only once on the xth attempt is

$$P(x) = q^{x-1}p$$

where p is the probability of success and $q = 1 - p$ is the probability of failure.

8. **You try!** You enter a darts tournament. The probability of hitting the bullseye is 17%. What is the probability that you hit the bullseye on the 7th attempt? You can upload an image to show your thinking.

This page titled 4.2: The Geometric Distribution is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Hannah Seidler-Wright.

4.2.1: Exercises

1. Determine which of the following scenarios only have two outcomes:

    a. Flipping a coin
    b. Attempting to score on a penalty kick in soccer
    c. Distance to the nearest landfill
    d. Finding a political tweet

2. If you are flipping a biased coin that lands on tails 72% of the time, find the probability that you will get your first “tails” on the fourth flip. Round your answer to four decimal places.

3. You are rolling a fair die. What is the probability that you will roll your first 6 on the eighth try? Round your answer to four decimal places.

4. A soccer player is practicing their penalty kicks. They have a 0.4 success rate. What is the probability that they make a goal for the first time on the fifth attempt. Round your answer to four decimal places.

5. In the board game Monopoly, one way to get out of jail is to roll doubles. Find the probability that it takes three rolls to get out of jail. Round your answer to four decimal places.

This page titled 4.2.1: Exercises is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Hannah Seidler-Wright.

4.3: The Binomial Distribution

Recall, there are many probability experiments where a trial has only two outcomes. For example, asking a group of individuals if they vote yes on a proposition, or randomly guessing on a multiple choice test. When we conduct a sequence of independent trials with only two outcomes per trial, we are conducting a binomial experiment. In the last lesson, you learned how to calculate the probability of success achieved on the xth trial when there are only two outcomes per each trial, and independent trials.

Now, we will turn to computing the probability of x successes in n independent trials, where there are only two possible outcomes per trial. We will examine this distribution by conducting an experiment. We will look at the possible outcomes from flipping four fair coins. Tossing a coin is an example of a trial with only two outcomes which are heads and tails.

1. Go to this website or use the QR code below to simulate flipping coins (or if you want, use four actual coins for the experiment). We will define landing on heads as success, and landing on tails as failure. We define the random variable x to be the number of successes (heads) out of four trials (4 coin tosses).

   a. There are many possible outcomes when flipping four coins. For example, the first coin could land on heads, the second on tails, the third on heads, and the fourth on tails. Let's denote this outcome as HTHT. In this outcome, there are 2 successes and 2 failures. Write two other possible outcomes from flipping four coins and the corresponding number of successes and failures.

   b. What is the maximum number of successes possible (this is the largest value of x)?

   c. What is the minimum number of successes possible (this is the smallest value of x)?

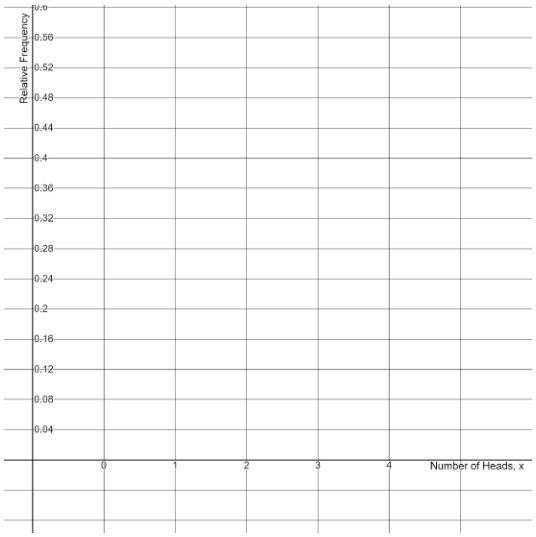
d. Time to flip the coins! Conduct the experiment 20 times. Each time you toss the four coins, use tally marks to keep track of the number of successes (heads) in the table below. When you have made 20 tally marks, you have finished.

Number of Heads, x	Tally
0	
1	
2	
3	
4	

e. Here are the results from the experiment repeated 200 times. Complete the table with the relative frequency.

Number of Heads, x	Frequency	Relative Frequency
0	14	
1	43	
2	85	
3	42	
4	16	
Total	200	

f. Use the data from e. to create a relative frequency histogram. The relative frequencies are approximations of the probabilities of each outcome, $P(x)$.



Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

g. Describe the center, shape, and spread of the relative frequency histogram.

h. Which value of the random variable occurred most often?

The Binomial Distribution

A **binomial experiment** is a probability experiment with the following characteristics:

- The experiment consists of n independent trials.
- Each trial has exactly two possible outcomes which are labeled **success** and **failure**.
- The probability of success is the same for each trial. We denote the probability of success as p and the probability of failure as $q = 1 - p$.

For the random variable x representing the number of successes in a fixed number of trials (n trials), the distribution of probabilities is called the **binomial distribution**.

2. In the binomial experiment above, how many trials were there?

$n =$ _____

3. In the binomial experiment above, what is the probability of success? What is the probability of failure?

$p =$ _____

$q = 1 - p =$ _____

4. Next, we will compute the actual probabilities from a binomial distribution. We will apply our knowledge of probability laws to help with the computation. We will lower the number of trials to 2 unfair coins in this computation. An unfair coin is a coin that is unbalanced and therefore lands on one side more or less than the other. Let's assume we have a collection of biased coins where the probability of getting heads is 60%.

   a. Recall, if events A and B are independent, then $P(A \cap B) =$ _____. This is the multiplication rule for independent events.

   b. If events A and B are mutually exclusive, then $P(A \cup B) =$ _____. This is the addition rule for mutually exclusive events.

   c. If you flip 2 coins, are the 2 coin flips independent? Explain.

   d. If 2 coins are tossed, there are 4 possible outcomes. One possible outcome is the first coin lands on heads, and the second coin lands on heads (H and H). In this outcome, there are 2 successes so the random variable takes on the value 2. List the other 3 possible outcomes from tossing 2 coins.

   e. Compute the probability of both coins landing on heads.

   $P(H \cap H) = P(2) =$ _____

   f. Compute the probability of both coins landing on tails. $P(T \cap T) = P(0) =$ _____

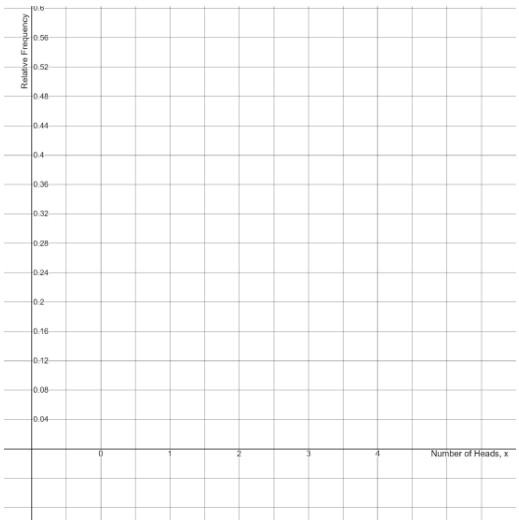
   g. Compute the probability of one coin landing on heads. Careful, there is more than one way this can happen!

   $P((H \text{ and } T) OR (T \text{ and } H)) = P(1) =$ _____

h. Summarize the distribution in the table:

Heads, x	P(x)
0	
1	
2	
Total	

i. Graph the binomial probability distribution as a histogram.



Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

j. Estimate the mean of the distribution.

Next, we construct the binomial distribution for three biased coins where the probability of a coin landing on heads is 60%.

$P(0) = P(T \text{ and } T \text{ and } T) = P(T) \cdot P(T) \cdot P(T) = (0.4)(0.4)(0.4) = 1 \cdot (0.4)^3$
$P(T \text{ and } T \text{ and } H) = P(T) \cdot P(T) \cdot P(H) = (0.4)(0.4)(0.6) = (0.4)^2(0.6)^1 = 0.096$
$P(T \text{ and } H \text{ and } T) = P(T) \cdot P(H) \cdot P(T) = (0.4)(0.4)(0.6) = (0.4)^2(0.6)^1 = 0.096$
$P(H \text{ and } T \text{ and } T) = P(H) \cdot P(T) \cdot P(T) = (0.4)(0.4)(0.6) = (0.4)^2(0.6)^1 = 0.096$
$P(1) = P\left(TTH \text{ or } THT \text{ or HTT } \right. = 3 \cdot (0.4)^2(0.6)^1 = 3(0.096) = 0.288$
$P(T \text{ and } H \text{ and } H) = P(T) \cdot P(H) \cdot P(H) = (0.4)(0.6)(0.6) = (0.4)^1(0.6)^2 = 0.144$
$P(H \text{ and } T \text{ and } H) = P(H) \cdot P(T) \cdot P(H) = (0.4)(0.6)(0.6) = (0.4)^1(0.6)^2 = 0.144$
$P(H \text{ and } H \text{ and } T) = P(H) \cdot P(H) \cdot P(T) = (0.4)(0.6)(0.6) = (0.4)^1(0.6)^2 = 0.144$
$P(2) = P\left(THH \text{ or HTH or HHT } \right. = 3 \cdot (0.4)^1(0.6)^2 = 3(0.144) = 0.432$
$P(3) = P(H \text{ and } H \text{ and } H) = P(H) \cdot P(H) \cdot P(H) = (0.6)(0.6)(0.6) = 1 \cdot (0.6)^3 = 0.216$

5. Let's notice some patterns: what do you notice about the probabilities when the number of trials is 3, the probability of success is 0.6, and the probability of failure is 0.4?

$P(0) = 1 \cdot (0.6)^0(0.4)^3$
$P(1) = 3 \cdot (0.6)^1(0.4)^2$
$P(2) = 3 \cdot (0.6)^2(0.4)^1$
$P(3) = 1 \cdot (0.6)^3(0.4)^0$

6. Below are the probabilities when flipping 4 biased coins where the probability of success (landing on heads) is 0.6 and the probability of failure (landing on tails) is 0.4. Fill in the blanks.

$$\begin{aligned} P(0) &= 1 \cdot (0.6)^0(\_\_\_\_)^4 \\ P(1) &= 4 \cdot (\_\_\_\_)^{\_\_\_\_}(0.4)^3 \\ P(2) &= 6 \cdot (0.6)^{\_\_\_\_}(0.4)^{\_\_\_\_} \\ P(3) &= 4 \cdot (\_\_\_\_)^3(\_\_\_\_)^1 \\ P(\_\_\_) &= \_\_\_\_ \cdot (0.6)^4(0.4)^0 \end{aligned} \tag{4.3.1}$$

The Combination Function

Notice that there are 6 combinations of 4 coins that result in 2 heads: {HHTT, HTHT, HTTH, THTH, TTHH, THHT}. This process of listing out all combinations of x successes in n trials is complicated and time consuming. Luckily, mathematicians have derived a formula for such a function. The combination (or choose) function is denoted ${}_nC_x$ ("n choose x"). The number of ways to get x successes in n trials is

$$ {}_nC_x = \frac{n!}{x!(n-x)!}$$

This function involves using the factorial n! (pronounced n factorial) means multiply all the whole numbers n down to 1. For example, $5! = 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 120$. We say that 0!=1.

Let's calculate ${}_4C_2$.

$$ {}_4C_2 = \frac{4!}{2! \cdot 2!} = \frac{4 \cdot 3 \cdot 2 \cdot 1}{(2 \cdot 1)(2 \cdot 1)} = \frac{4 \cdot 3}{2 \cdot 1} = 6$$

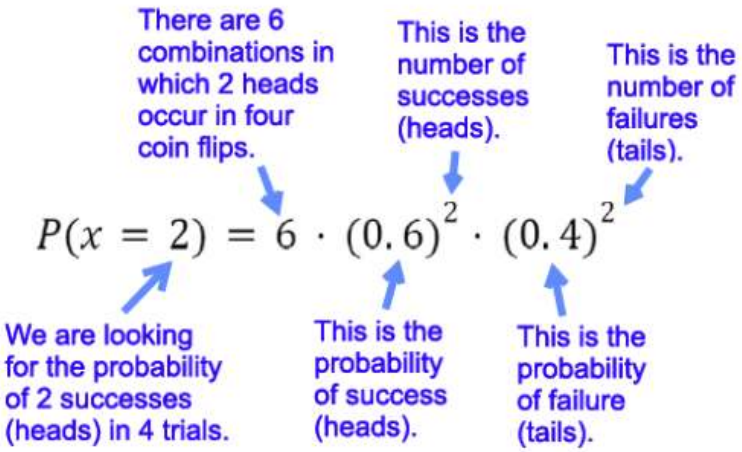
7. Rather than doing this calculation by hand, let's use https://www.desmos.com/calculator to do it for us.
   a. Type in nCr(4,2). Alternatively, you can open the keypad, click on the functions button, scroll to number theory, and select nCr. Enter the number of trials and the number of successes separated by a comma in the empty parentheses. You will see that desmos gives us the answer 6 which matches our calculation above.

   b. Compute ${}_{12}C_9$ by typing nCr(12,9) on line 2.

   c. Compute ${}_{52}C_{49}$ by typing in nCr(52,49) on line 3.

Computing Binomial Probability

Given below is the computation of the probability of flipping exactly two heads in four coin tosses.



We now generalize this formula. In a **binomial** experiment with n independent trials where p is the probability of success, and $q = 1 - p$ is the probability of failure, the probability of exactly x successes in n trials is

$$P(x) = {}_nC_x \cdot p^x \cdot q^{(n-x)}$$

These probabilities form a binomial distribution.

8. The probability of testing positive for COVID 19 in California is approximately 3.2%.
   a. In a group of 20 randomly selected individuals from California, what is the probability that 3 of those people are positive for COVID 19?

   b. In a group of 20 randomly selected individuals from California, what is the probability that less than 3 people are positive for COVID 19?

   c. In a group of 20 randomly selected individuals from California, what is the probability that at least 1 person is positive for COVID 19?

This page titled 4.3: The Binomial Distribution is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Hannah Seidler-Wright.

- **Current page** by Hannah Seidler-Wright is licensed CC BY-NC-SA 4.0.
- **1.2: The Statistical Analysis Process** by Hannah Seidler-Wright is licensed CC BY-NC-SA 4.0.

4.3.1: Exercises

1. In a CDC survey, 37% of US high school students reported regular mental health struggles during the pandemic. An educator is concerned about the mental health of her students. She randomly surveys 8 students. Don tells Isabel that the probability that exactly 6 of the 8 students reported struggling with their mental health during the pandemic can be found using the following formula: $P(6) = (0.37)^6(0.63)^2$. What should Isabel say to Don to help him understand binomial probability better?

2. In a CDC survey, 37% of US high school students reported regular mental health struggles during the pandemic. An educator is concerned about the mental health of her students. She randomly surveys 8 students. Fill in the following table with the appropriate values. Round calculated values to four decimal places.

X = number of students with mental health struggles	${}_nC_x$	$P(X)$
0	${}_8C_0 =$ _____	$P(0) = (0.63)^{\_\_\_\_} \approx 0.0248$
1	${}_8C_{\_\_} = 8$	$P(1) = 8 \cdot (0.37)^{\_\_\_\_} \cdot (0.63)^7 \approx 0.1794$
2	${}_8C_{\_\_} =$ _____	$P(2) =$ ___ $\cdot$ (___)$^2 \cdot (0.63)^6 \approx 0.2397$
3	$\_\_C\_\_ =$ _____	$P(3) =$ ___ $\cdot$ (___)$^3 \cdot (0.63)^{\_\_}$ ≈ 0.2815
4	$\_\_C\_\_ =$ _____	$P(4) =$ ___ $\cdot$ ( ) — $\cdot$ (___) — ≈ 0.2067
5	$\_\_C\_\_ =$ _____	$P($___$) =$ ___ $\cdot$ (___) — $\cdot$ (___) — ≈ 0.0971
6	$\_\_C\_\_ =$ _____	$P($___$) =$ ___ $\cdot$ (___) — $\cdot$ (___) — $\approx$ ___
7	$\_\_C\_\_ =$ _____	$P($___$) =$ ___ $\cdot$ (___) — $\cdot$ (___) — $\approx$ ___
8	$\_\_C\_\_ =$ _____	$P($___$) =$ ___ $\cdot$ (___) — $\cdot$ (___) — $\approx$ ___

3. Use the discrete probability distribution from 1. to find the probability that at most five of the randomly selected students have reported regular mental health struggles during the pandemic.

4. Use the discrete probability distribution from 1. to find the probability that at least six of the randomly selected students have reported regular mental health struggles during the pandemic.

This page titled 4.3.1: Exercises is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Hannah Seidler-Wright.

CHAPTER OVERVIEW

5: Continuous Probability Distributions and The Normal Distribution

5.1: Probability Distributions of Continuous Random Variables

5.1.1: Exercises

5.2: Characteristics of the Normal Distribution and The Empirical Rule

5.2.1: Exercises

5.3: The Standard Normal Distribution

5.3.1: Exercises

5.4: Finding Critical Values from the Normal Distribution

5.4.1: Exercises

This page titled 5: Continuous Probability Distributions and The Normal Distribution is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Hannah Seidler-Wright.

5.1: Probability Distributions of Continuous Random Variables

In unit 4, we organized outcomes and probabilities for a *discrete random variable*. The values of a discrete random variable can be listed in order (can be counted).

Continuous Random Variables

Sometimes the values of a variable cannot be listed in order. This is because, for any given value, it is impossible to list a *next* value. These variables are considered *continuous variables*.

For example, let's try to list the heights of all randomly selected men in order. We are not talking about specific heights, but rather, all possible heights. Consider a man who is 6 feet tall. If we could put these values in order, we should be able to tell what height comes after 6 feet or 6.0000… feet. So what is the next height on the list? Maybe we think 6.1 feet is the next height. But that isn't correct because 6.01 feet is closer to 6 feet. So is the next height 6.01 feet? No, because 6.001 is again closer to 6 feet, and 6.0001 feet is closer to that. We could continue this forever and we would never be able to find the "next height" on the list. This is why the heights of randomly selected men are values of a **continuous random variable**.

A random variable is continuous if it cannot be listed in order, or counted. Usually, continuous random variables are measured (like height, time, or weight).

1. Classify each of the random variables described as either *discrete* or *continuous*.
   a. The number of high school students in a dual enrollment program.

   b. The time it takes to sprint 500 meters.

   c. The number of assignments in an online class.

   d. The time it takes to complete all assignments in an online class.

   e. The distance a planet is from the sun.

   f. The sum of the roll of three dice.

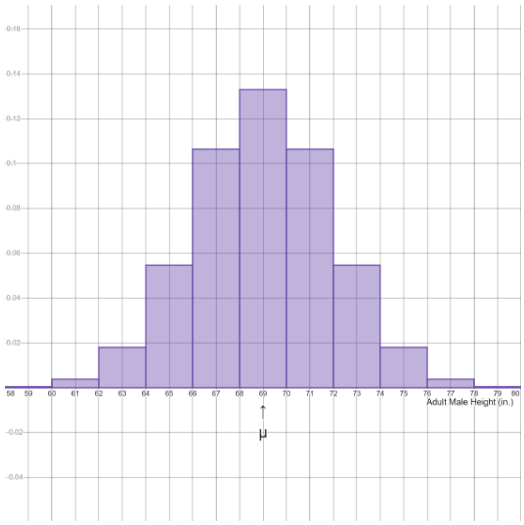
   g. The life expectancy of a person living in a blue zone.

   h. The amount of money in someone's pocket.

Statistical Inference

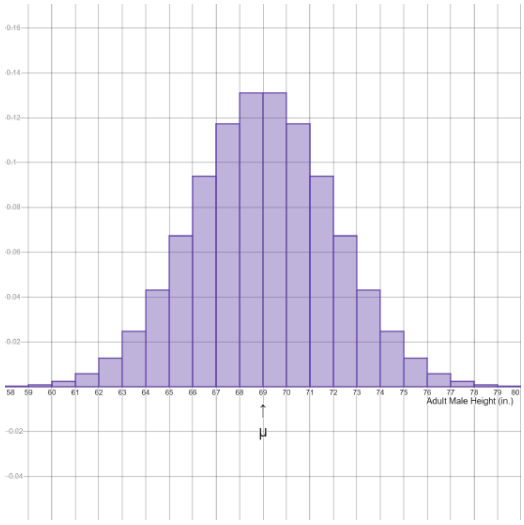
In statistical inference (where we use sample data to make judgements about a population), we often work with continuous random variables, and statistical inference depends on probability. In previous lessons, we visualized probability distributions for discrete random variables using histograms. Each value of a discrete random variable has a probability associated with it, and the probability is the area of the corresponding bar on the probability histogram.

Consider our original example, the distribution of heights of men. Below is a histogram that represents the probabilities that are estimated by relative frequencies from a large random sample of men's heights. The area of each bar represents the probability that a random man's height is within the corresponding range. The mean for all men's heights is 69 inches. Each bar in the histogram has a width of 2 inches.



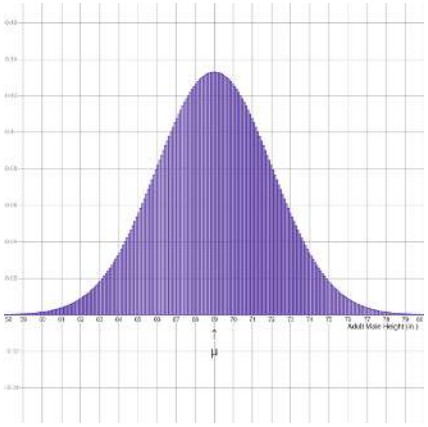
Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

The problem with using this histogram is that it only allows us to estimate probabilities for a predetermined range of values. It does not allow us to find probabilities for other ranges (like the probability that a randomly selected man is between 70 and 71 inches tall). A bar with width 1 would allow us to compute this probability. See the histogram below.



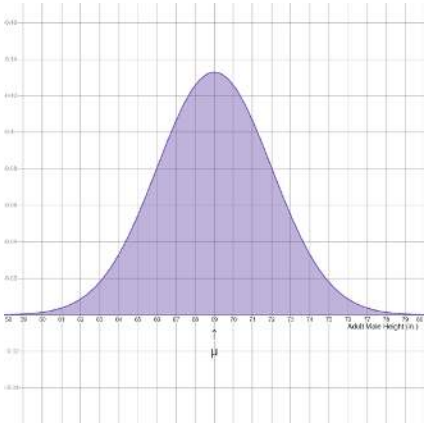
Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

With thinner bars, the graph begins to look closer to a smooth curve. In particular, this histogram is bell-shaped. Unfortunately, we still can't find the probability that a randomly selected man is between 70 and 70.1 inches tall. We can again decrease the width of each bar to be 0.1 inches. See the histogram below.



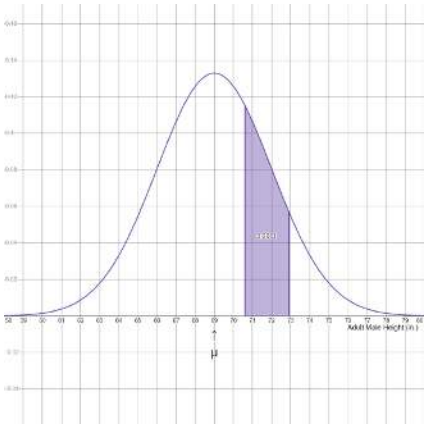
Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

We continue in this way forever until we have achieved a smooth curve. This curve is a **continuous probability distribution.** Mathematicians use Calculus to find areas under this curve which corresponds to probability. The total area under the curve is 1.



Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

Below, the area under the curve from 70.6 to 72.9 inches is shaded.



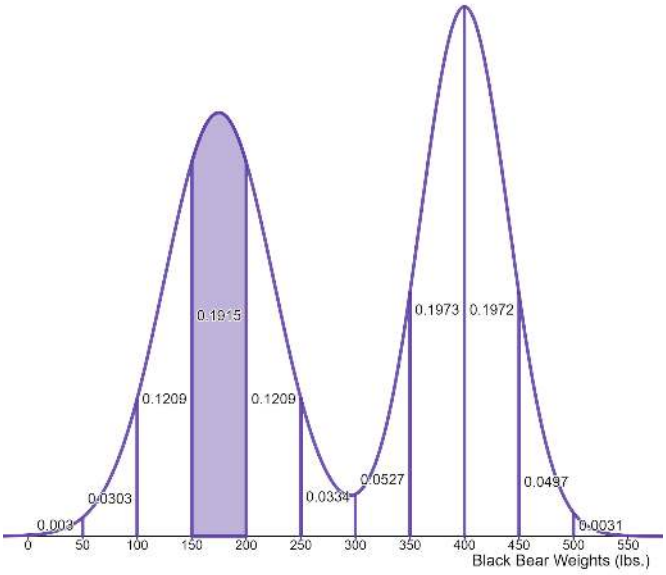
Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

The area of this shaded region is 0.2001. It represents the probability that a randomly selected man is between 70.6 and 72.9 inches tall. If x represents adult male height in inches, we can say this using probability notation: $P(70.6 < x < 72.9) = 0.2001$. The proportion of all adult males that are between 70.6 and 72.9 inches tall is around 20%.

Probabilities from a Bimodal Continuous Probability Distribution

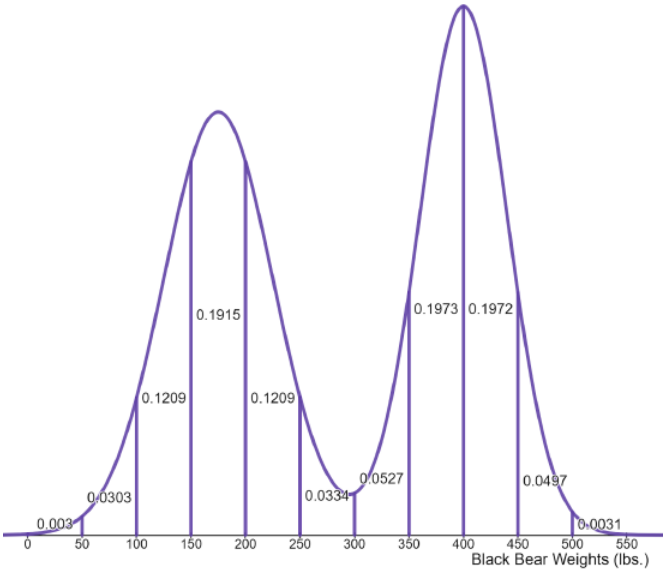
There are many continuous probability distributions. We will use bell-shaped probability curves often in this course. We may see distributions that have other shapes. Sometimes when two populations are merged into one, they form a probability distribution with two peaks. This type of distribution is called **bimodal**.

Black bears weights tend to differ by sex. Female black bears weigh 175 pounds on average, whereas, male black bears weigh 400 pounds on average. Below is the continuous probability distribution of adult black bear weights. Let the continuous random variable take on values of adult black bear weight. The curve has been divided into intervals of equal length. The area of each range is written in each region above the corresponding interval. For example, the area 0.1915 corresponds to the range of values between 150 and 200. In context, this says the probability that a randomly selected black bear weighs between 150 and 200 pounds is 0.1915 or 19.15%.



Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

Using probability notation, we say $P(150 \leq x \leq 200) = 0.1915 = 19.15\%$.



Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

2. Use the continuous probability distribution of randomly selected black bear weights above to answer the following questions.

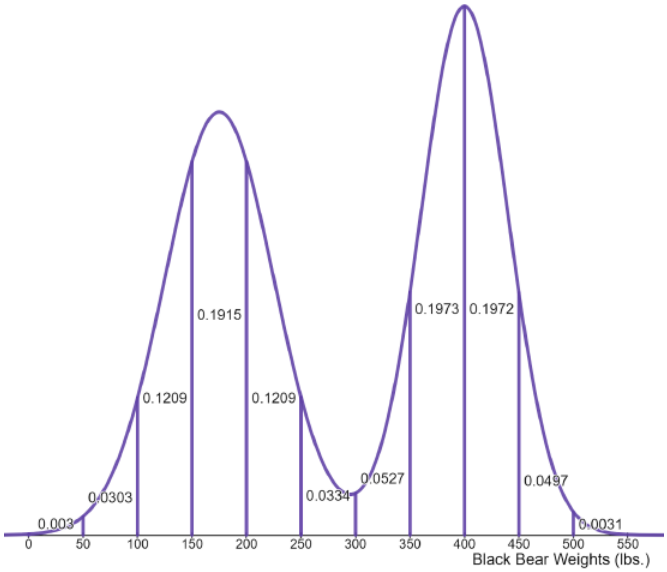
a. Compute the total area under the curve above.

b. Shade the region that represents the probability that a randomly selected black bear weighs between 350 and 500 lbs.

c. The area you shaded can be written using probability notation as $P(350 < x < 500)$. What is this probability?

$P(350 < x < 500) =$

d. Compute the proportion of black bears that weigh between 50 and 250 lbs. Use probability notation in your answer. Shade the area that represents the probability to show your thinking.



Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

3. Use the continuous probability distribution of randomly selected black bear weights above to answer the following questions.
   a. Shade the region that represents the probability that a randomly selected black bear weighs at most 100 lbs.

   b. The area you shaded can be written using probability notation as $P(x \leq 100)$ which is equal to P(x<100) in a continuous probability distribution. What is this probability?

      $P(x \leq 100) =$

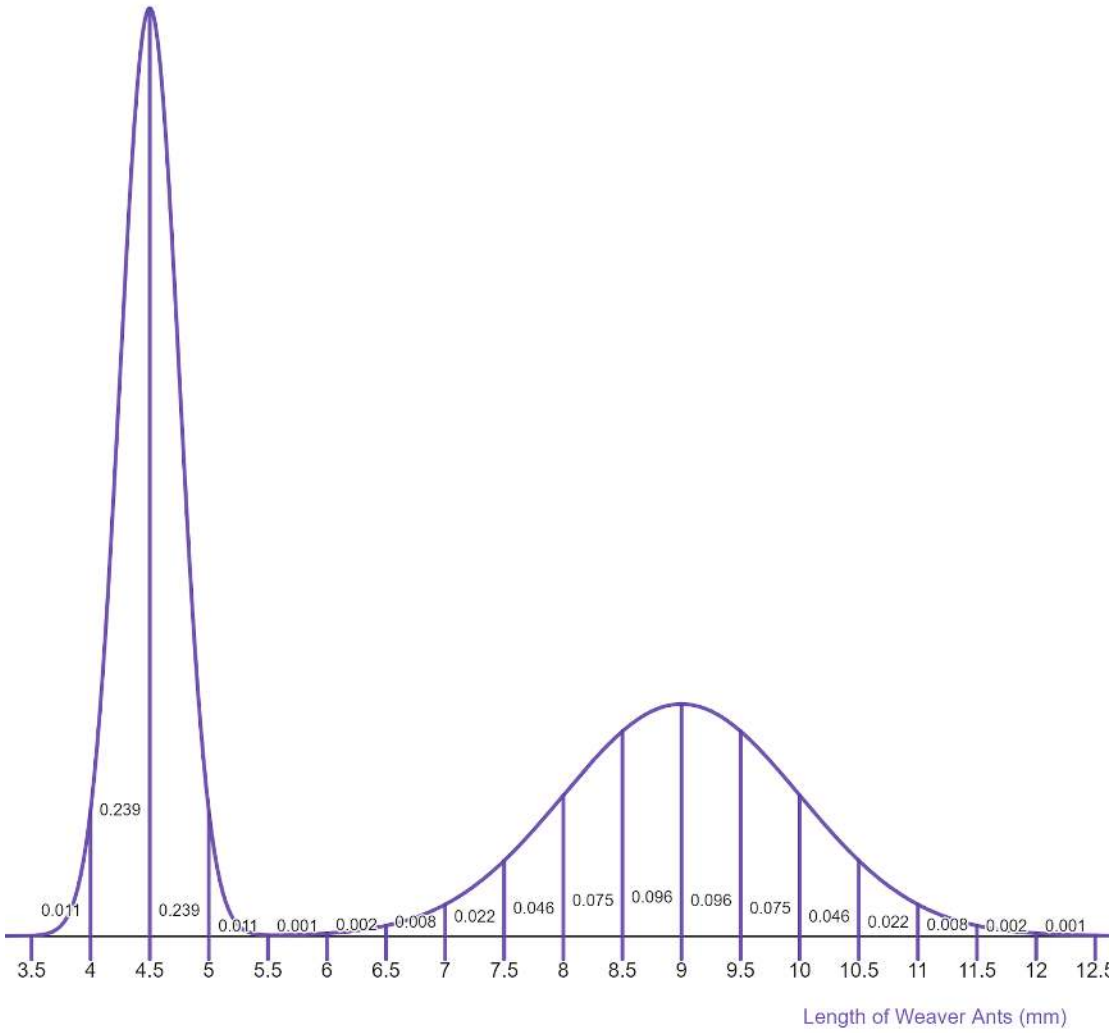
   c. Compute the proportion of black bears that weigh at least 100 lbs. Use probability notation in your answer. *Hint: apply the complement rule*.

This page titled 5.1: Probability Distributions of Continuous Random Variables is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Hannah Seidler-Wright.

- **Current page** by Hannah Seidler-Wright is licensed CC BY-NC-SA 4.0.
- **1.2: The Statistical Analysis Process** by Hannah Seidler-Wright is licensed CC BY-NC-SA 4.0.

5.1.1: Exercises

1. Weaver worker ants have a bimodal length distribution. There are two types of workers, minors and majors. The majors have an average length of 9 mm and minors average half the length. Complete the following probability questions. Include probability notation in your answers.



Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

a. What proportion of weaver ants are between 4.5 and 8 mm long?

b. What proportion of weaver ants are between 9.5 and 11 mm long?

c. Find the probability that a randomly selected weaver ant is at most 4.5 mm long.

d. Find the probability that a randomly selected weaver ant is at most 5 mm long.

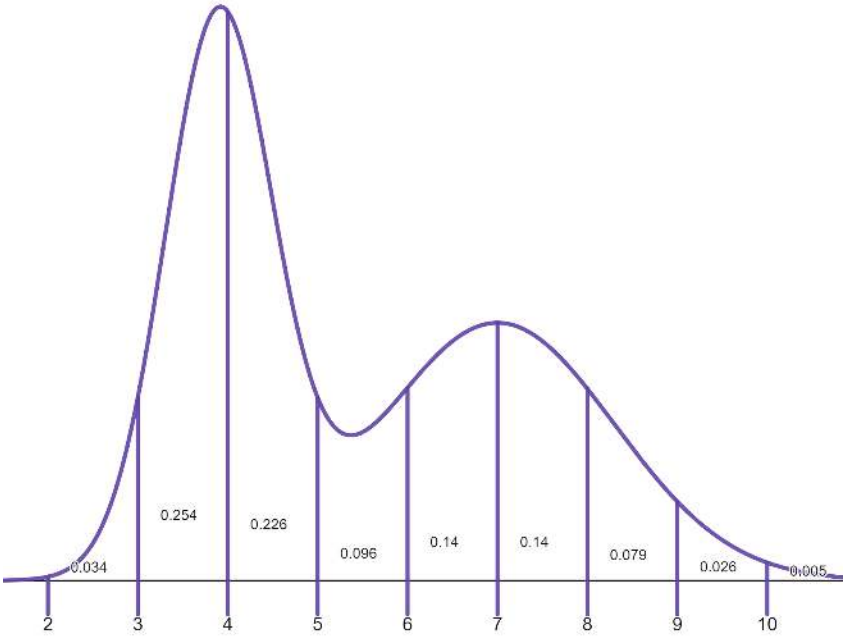
e. What proportion of weaver ants are at most 11 mm long. Use the complement rule in the computation.

f. What proportion of weaver ants are at least 4.5 mm long. Use the complement rule in the computation.

g. What value separates the smallest 25% of weaver ants from the largest 75% of weaver ants?

h. What value separates the largest 25% of weaver ants from the smallest 75% of weaver ants?

2. Lupe uses the following continuous probability distribution to answer questions about probability. Notice any errors she makes and tell her how to fix them.



Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

a. Find the proportion of data that is between 5 and 8.

$P(5 > x > 8) = 0.096 + 0.14 + 0.14 + 0.079 = 45.5$

b. Find the proportion of data that is at most 10.

$P(x > 10) = P(0.005)$

c. Find the proportion of data that is at least 4. Use the complement rule in the calculation.

$P(x \geq 4) = 1 - 0.034 + 0.254 = 0.712$

This page titled 5.1.1: Exercises is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Hannah Seidler-Wright.

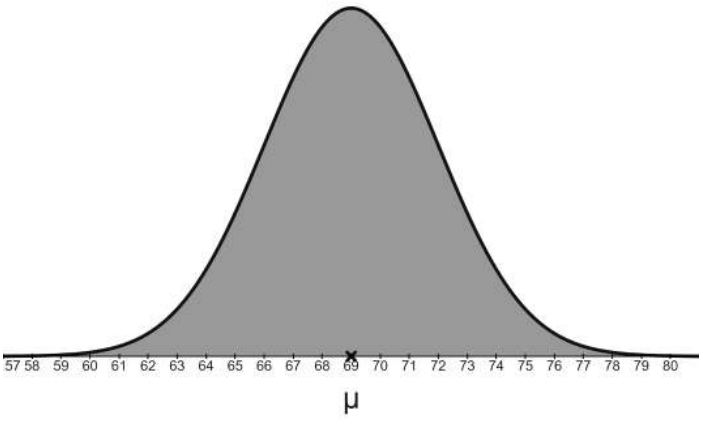
- **Current page** by Hannah Seidler-Wright is licensed CC BY-NC-SA 4.0.
- **1.2: The Statistical Analysis Process** by Hannah Seidler-Wright is licensed CC BY-NC-SA 4.0.

5.2: Characteristics of the Normal Distribution and The Empirical Rule

In the last lesson, we learned about continuous random variables and their probability distributions. A continuous random variable is special because it can take on any of an infinite number of possible values which cannot be listed in any order (or counted). Continuous probability distributions can take on a variety of shapes including uniform, bi- or multi-modal, or bell-shaped, etc., as long as the total area under the curve is 1.

A bell-shaped curve is one that is symmetric, and has a single mode in the center, and has two skinny tails. It is used to represent situations where the random variable is more likely to take on values closer to its average and less likely to take on extreme values. One distribution that we will be using often is called the **normal distribution** for which there is a formula allowing one to find the precise areas (probabilities) for any range of values of the continuous random variable. There are many examples of situations that warrant the use of the normal distribution.

For example, the heights of randomly selected men have a distribution that is approximately normal. The mean of this population is $\mu = 69$ inches, and the population standard deviation is $\sigma = 3$ inches.



Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

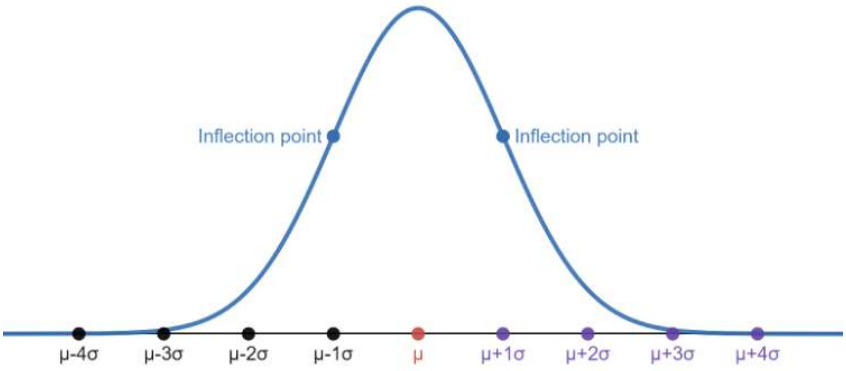
Some symbols to note:

- μ (“mu”) represents a population mean.
- $\bar{x}$ (“x-bar”) represents a sample mean.
- σ (“sigma”) represents a population standard deviation.
- s (“s”) represents a sample standard deviation.

Characteristics of the Normal Distribution

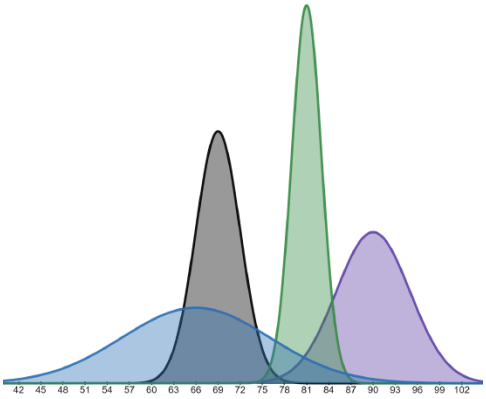
A **normal distribution** with mean μ and standard deviation has the following characteristics:

- The mean, median, and mode are equal. 50% of all values are below the mean and 50% are above it.
- The normal curve is bell-shaped and symmetric about its mean μ.
- The total area under the normal curve is 1.
- Normal curves extend endlessly in both directions, but the curve becomes so close to zero for values of the random variable that are more than 4 standard deviations above or below the mean that the area is negligible.
- The normal curve changes its curvature from a “cup” shape to a “cap” shape or vice-versa at one standard deviation below and above the mean. These points are called inflection points.



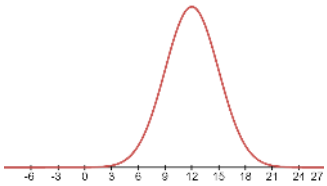
Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

A normal distribution is determined by its **mean** and its **standard deviation** . All normal distributions have the same shape but one distribution might have a mean of 69 whereas another has a mean of 90. This manifests on the graph as a horizontal translation. One normal distribution might have a standard deviation of 10 whereas another might have a standard deviation of 2. Standard deviation is a measure of spread so the higher the standard deviation, the more spread out the data is, and we see a flatter curve. The lower the standard deviation, the less spread there is, resulting in a curve that is narrow with a tall peak in the middle. Standard deviation manifests in the graph as horizontal stretching or shrinking. Explore these characteristics using this desmos graph or using the QR code below. Click the play button left of m to see how changing the mean changes the graph of the normal distribution. Click pause to stop the animation. Click the play button next to s to see how changing the standard deviation changes the graph of the normal distribution.

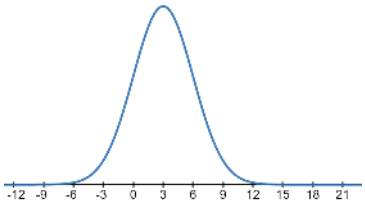


Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

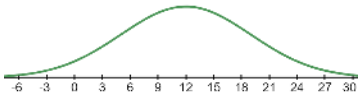
1. Match the following normal distributions with the appropriate means and standard deviations. Explain your reasoning for your choices.



a.



b.



c.



d.

Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

i. $\mu = 12, \sigma = 7$

ii. $\mu = 3, \sigma = 3$

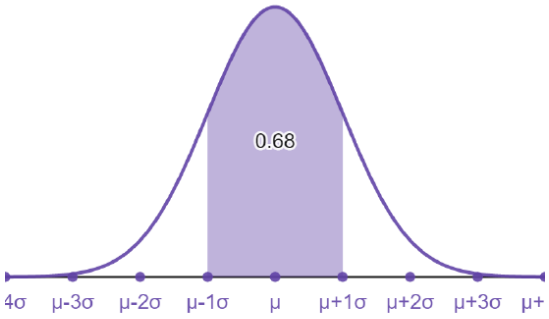
iii. $\mu = 3, \sigma = 2$

iv. $\mu = 12, \sigma = 3$

The Empirical Rule

All normal distributions are the same with respect to their mean and standard deviation. Using the normal distribution as the model, about 68% of values lie within one standard deviation of the mean. About 95% of values lie within two standard deviations from the mean, and about 99.7% of values lie within three standard deviations of the mean. These approximations are collectively known as the **Empirical Rule.**

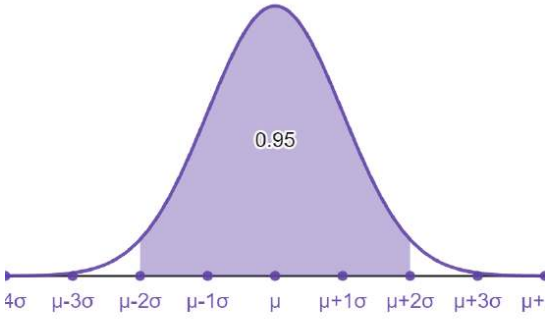
The graph below is of a normal distribution with the mean μ, located in the center of the graph. We see that about 68% of the area under the curve is between $\mu - 1\sigma$ (one standard deviation below the mean) and $\mu + 1\sigma$ (one standard deviation above the mean).



Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

In other words, this tells us that about 68% of all values in a normal population lie within one standard deviation of the mean.

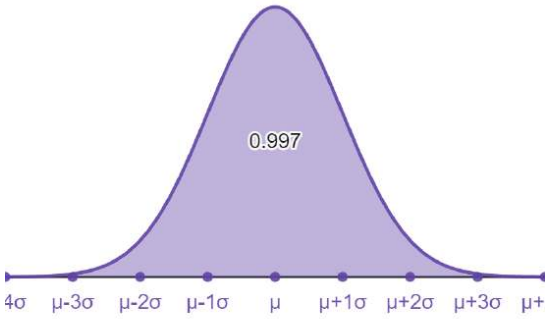
For the region between $\mu - 2\sigma$ (two standard deviations below the mean) and $\mu + 2\sigma$ (two standard deviations above the mean), the area under the curve is about 95%.



Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

In other words, this tells us that about 95% of all values in a normal population lie within two standard deviations of the mean.

For the region between $\mu - 3\sigma$ (three standard deviations below the mean) and $\mu + 3\sigma$ (three standard deviations above the mean), the area under the curve is about 99.7%.



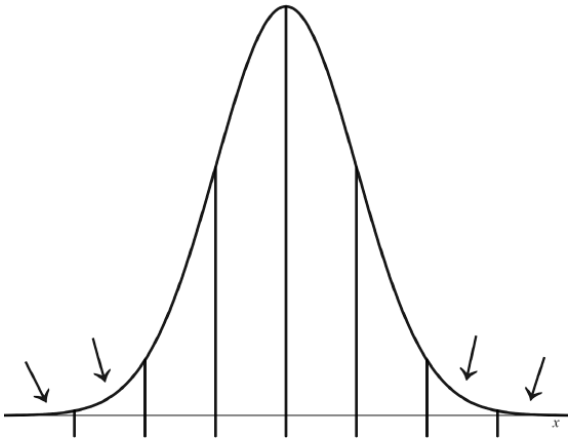
Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

In other words, this tells us that about 99.7% of all values in a normal population lie within three standard deviations of the mean. The remaining 0.3% are split among the more extreme values that lie beyond three standard deviations above or below the mean.

The Empirical Rule tells us that only about 5% of values in a normal distribution are more than two standard deviations from the mean. We will say that any value that is more than two standard deviations from the mean are considered **unusual**.

Computing Probabilities Using the Empirical Rule

2. Recall that adult male heights are approximately normal. The average adult male height is 69 inches and the standard deviation is 3 inches.
   a. Use this information and what you know about the normal distribution to label the tick marks on the horizontal axis, noting that the distance between each tick mark is 1 standard deviation.



Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

b. Remembering that the total area under the probability curve is always 1, find the area of each of the eight regions above.

c. Suppose that x represents a random adult man's height. We want to know the probability that the man is between 72 and 78 inches tall. Using probability notation, this is written as $P(72 < x < 78)$. Use the completed graph above to compute this probability.

d. Find $P(x \geq 75)$

e. Find $P(x < 66)$

f. Find $P(60 \leq x \leq 69)$

g. Find $P(66 < x < 75)$

h. Do we have enough information to find the proportion of adult men who are shorter than 64 inches tall? Explain.

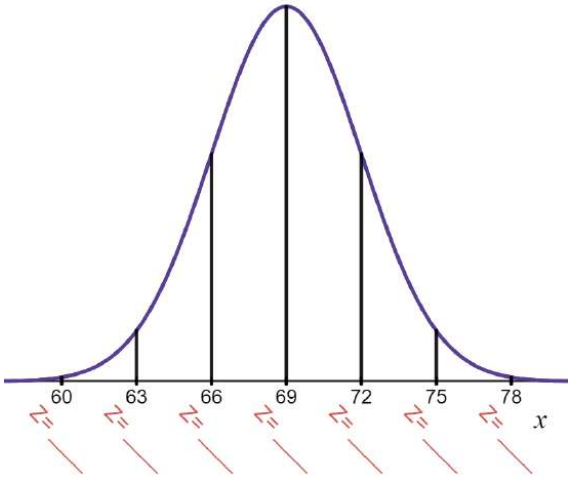
Z-scores

Determining if a value is unusual or not depends on the number of standard deviations it is from the mean. We measure the distance between values using the standard deviation to define the ruler. We define a **Z-score** to be the number of standard deviations a value is from the mean. We compute a Z-score for a given value of x using the following formula:

$$Z = \frac{x - \mu}{\sigma}$$

- When Z is negative, x is below the mean.
- When Z is positive, x is above the mean.
- When Z is zero, x is equal to the mean.

3. Label the horizontal axis with Z-scores below each x on the distribution of adult male heights.



Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

4. We have said that a value in a normal distribution is unusual if it is more than two standard deviations from the mean.
   a. What range of heights belongs to unusually tall men?

   b. What Z-scores correspond to these heights?

   c. What range of heights belongs to unusually short men?

   d. What Z-scores correspond to these heights?

Z-scores provide a standardized way to measure values. For example, when a value has a Z-score of 0.5, we know it is half of a standard deviation above the mean, even if we do not know the values of the mean and standard deviation. Z-scores can be used to compare values from two different populations by measuring their distance relative to each population's mean.

5. A company is hiring a candidate for a high-level research position in a Chemistry lab. Only one position is available. The hiring committee narrows the choice down to two outstanding candidates. Both individuals are highly qualified for the job and have earned a doctorate in Chemistry from two different institutions. Each candidate was required to take a qualifying exam to earn their degrees. Their universities graded the qualifying exams on different scales. Candidate A earned 95 points on their qualifying exam where the average score was 73 and the standard deviation was 9. Candidate B earned 23 points on their qualifying exam where the average score was 18 and the standard deviation was 2.5 points. Using this information and Z-scores, who should be recommended for the position? Justify your answer.

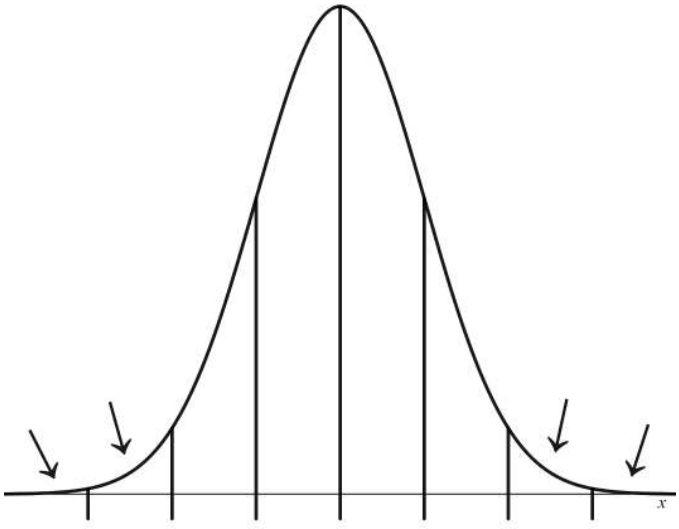
This page titled 5.2: Characteristics of the Normal Distribution and The Empirical Rule is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Hannah Seidler-Wright.

- **Current page** by Hannah Seidler-Wright is licensed CC BY-NC-SA 4.0.

- **1.2: The Statistical Analysis Process** by Hannah Seidler-Wright is licensed CC BY-NC-SA 4.0.

5.2.1: Exercises

1. The patient recovery time from a particular surgical procedure is normally distributed with a mean of 7 days and a standard deviation of 2.1 days.
   a. Use this information and what you know about the normal distribution to label the tick marks on the horizontal axis, noting that the distance between each tick mark is 1 standard deviation.



Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

   b. Remembering that the total area under the probability curve is always 1, label the area of each of the eight regions above according to the empirical rule.

   c. Suppose that x represents a random patient's recovery time from this surgery. Find $P(x < 4.9)$.

   d. Suppose that x represents a random patient's recovery time from this surgery. Find $P(0.7 < x < 4.9)$.

   e. Suppose that x represents a random patient's recovery time from this surgery. We want to know the probability that the patient recovers in between 0.7 and 9.1 days. Compute this probability and include probability notation in your answer.

f. Suppose that x represents a random patient's recovery time from this surgery. What proportion of patients recover in at most 13.3 days? Include probability notation in your answer.

g. Write a sentence interpreting the following probability in context: $P(x \geq 11.2)$.

h. Martha's recovery time has Z-score -2.1. Interpret this Z-score in context. Is Martha's recovery time unusual? Explain.

i. Compute the Z-score for a patient that recovers in 9.5 days. Is this recovery time unusual? Explain.

j. Give the range of unusually long recovery times.

k. Give the range of unusually short recovery times.

2. Weights of adult men and women in the US are normally distributed. The distribution of adult male weights has a mean of 200 lbs and a standard deviation of 9 lbs. The distribution of adult female weights has a mean of 170 lbs and a standard deviation of 7 lbs. Andy is a man who weighs 213 lbs and Anne is a woman who weighs 178 lbs. Who is heavier (relatively)? Justify your answer using Z-scores.

This page titled 5.2.1: Exercises is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Hannah Seidler-Wright.

- **Current page** by Hannah Seidler-Wright is licensed CC BY-NC-SA 4.0.
- **1.2: The Statistical Analysis Process** by Hannah Seidler-Wright is licensed CC BY-NC-SA 4.0.

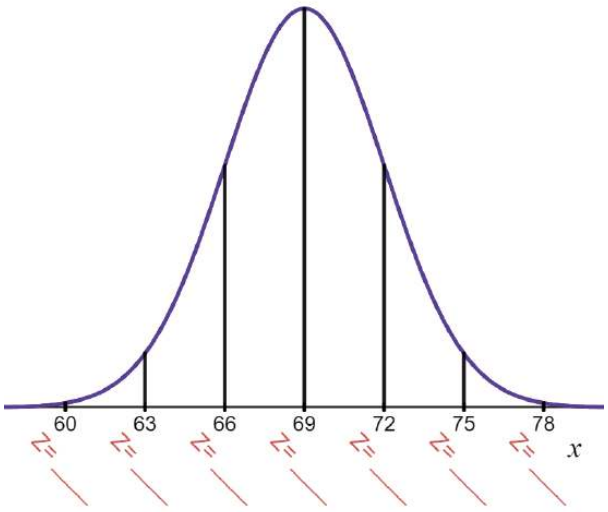
5.3: The Standard Normal Distribution

In the last lesson, we learned that all values from *any* normally distributed data can be *standardized* using Z-scores. In this lesson, we will be looking at the **standard normal distribution** which is the normal distribution that has mean $\mu = 0$ and standard deviation $\sigma = 1$. When using the standard normal distribution, we label the random variable with the letter z.

Accuracy

The empirical rule helps us to *approximate* probabilities of ranges of values that correspond to the eight regions given in the graph below.

1. Label the horizontal axis with the appropriate Z-scores and the appropriate areas for the eight regions given on the graph of the normal distribution of adult male heights below.



Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

The benefit is that we can quickly approximate proportions from a normal distribution for a region between any integer value of z ($z = [-3, -2, -1, 0, 1, 2, 3]$). For example, using the empirical rule, I know that the proportion of adult male heights that are between 66 inches (one standard deviation below the mean) and 75 inches (two standard deviations above the mean) is about 68%+13.5% which sums to 81.5%. Using probability notation,

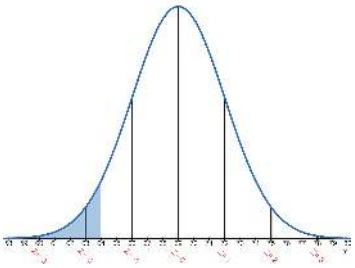
$$P(66 < x < 75) = P(-1 < z < 2) \approx 0.68 + 0.135 = 0.815 = 81.5\%.$$

We do not yet have enough information to find the proportion of adult men who are shorter than 64 inches tall because 64 inches does not correspond to an integer value of z.

2. Compute the Z-score (rounded to two decimal places) for an adult man who is 64 inches tall.

3. Below is a graph of the normal distribution of adult male heights. The proportion of adult males who are shorter than 64 inches has been shaded. In probability notation, this area is represented as $P(x < 64)$. Fill in the blank:

$$P(x < 64) = P(Z < \_\_\_\_\_)$$



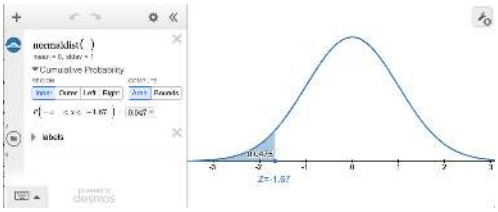
Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

When a continuous random variable is approximately normal, we can use Z-scores to find probabilities for a desired range of values. This is done by translating between x-values and Z-scores and using Calculus to find the corresponding area under the curve. Luckily, mathematicians have found areas for us so that people can learn about statistics without knowing Calculus.

Using desmos to Find Probabilities from The Standard Normal Distribution

We can use desmos to find probabilities from the normal distribution. Here's how to find the probability of an adult male being less than 64 inches tall from the standard normal distribution.

1. Go to https://www.desmos.com/calculator.
2. In the first line type normaldist(). Alternatively, you can click the keypad icon, then the Functions button, scroll to Distributions, and select normaldist. This will graph the standard normal distribution, where the horizontal axis represents z-scores.
3. Click the Zoom Fit button (magnifying glass icon) to adjust the view.
4. Click the small arrow next to Cumulative Probability to expand the options.
5. By default, it shows the probability as $P(-1 \le x \le 1)$. To find the probability that z is less than -1.67, change the upper bound to -1.67 and leave the lower bound blank (desmos will treat it as negative infinity). You will see the probability notation $P(-\infty \le x \le -1.67)$.
6. You should see the shaded area to the left of -1.67, which represents the probability we're looking for.



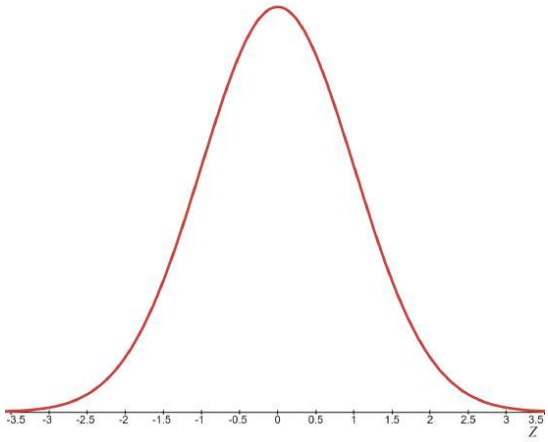
Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

$P(x < 64) = P(z < -1.67) = 0.047$

Therefore, 4.7 of adult men are shorter than 64 inches tall.

You try!

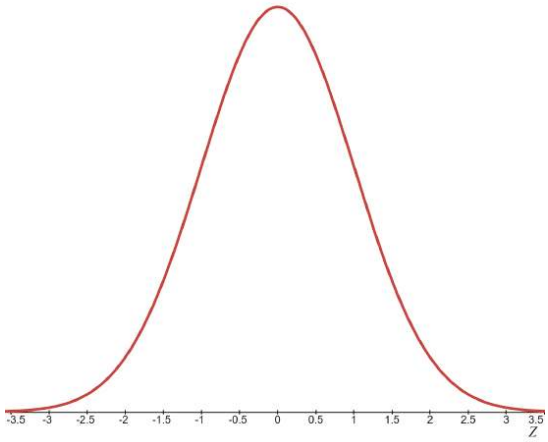
4. Fill in the probability notation with the missing Z-score (rounded to two decimal places) for the x-value 73. Shade the region of the standard normal distribution that represents the proportion of adult males who are at least 73 inches tall. Then use desmos to find the proportion (rounded to four decimal places) of adult males who are at least 73 inches tall.



Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

$$P(x \geq 73) = P(Z \geq \_\_\_\_\_) = \_\_\_\_\_$$

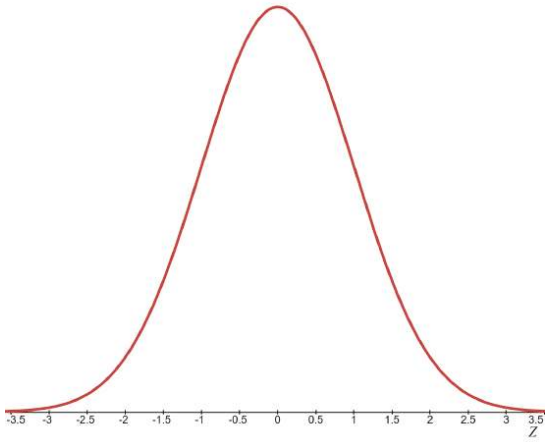
5. Fill in the probability notation with the missing Z-scores (rounded to two decimal places) for the x-values 67.5 and 71.8. Shade the region of the standard normal distribution that represents the proportion of adult males who are between 67.5 and 71.8 inches tall. Then use desmos to find the proportion (rounded to four decimal places) of adult males who are between 67.5 and 71.8 inches tall.



Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

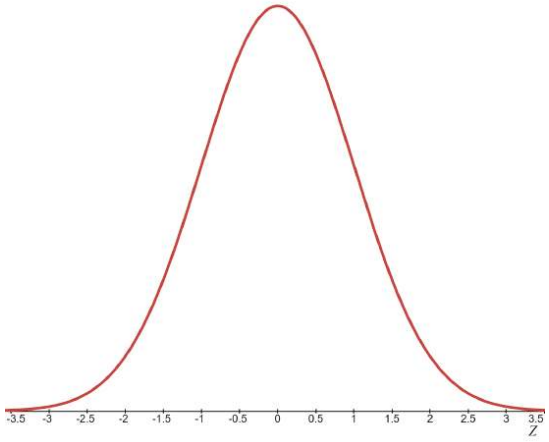
$$P(67.5 \leq x \leq 71.8) = P(\_\_\_\_\_ \leq Z \leq \_\_\_\_\_) = \_\_\_\_\_$$

6. Find the probability that a randomly selected adult male is at most 75.7 inches tall. Round Z-scores to two decimal places, and probability to four decimal places. Include probability notation in your answer and shade the appropriate region on the graph of the standard normal distribution.



Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

7. What proportion of adult men are between 70.5 inches tall and 71.1 inches tall? Round Z-scores to two decimal places, and probability to four decimal places. Include probability notation in your answer and shade the appropriate region on the graph of the standard normal distribution.



Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

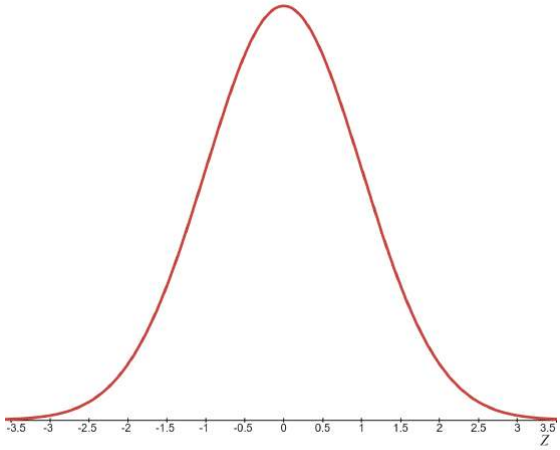
This page titled 5.3: The Standard Normal Distribution is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Hannah Seidler-Wright.

- **Current page** by Hannah Seidler-Wright is licensed CC BY-NC-SA 4.0.
- **1.2: The Statistical Analysis Process** by Hannah Seidler-Wright is licensed CC BY-NC-SA 4.0.

5.3.1: Exercises

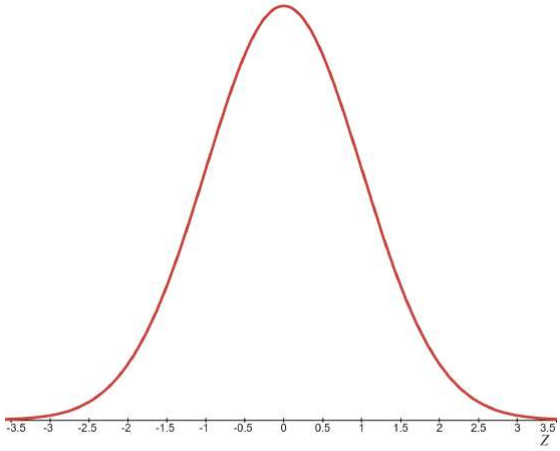
1. The length of time students needed in order to complete a criminal justice test followed a distribution that was approximately normal. The mean was 68 minutes and the standard deviation was 5 minutes.

   a. What proportion of students took more than an hour to complete the test? Use probability notation in your answer. Round your answer to four decimal places.



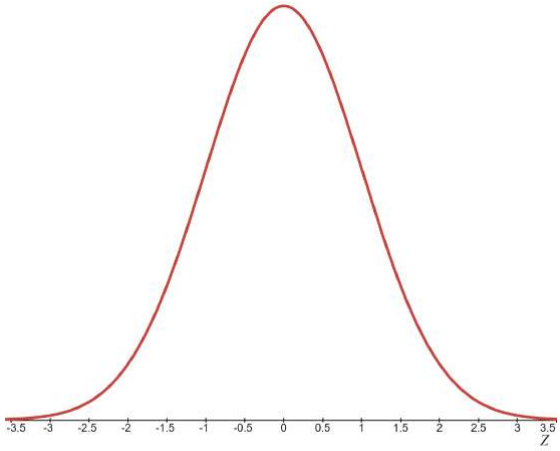
Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

   b. What is the probability that a student needed between 60 and 70 minutes to complete the test? Use probability notation in your answer. Round your answer to four decimal places.



Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

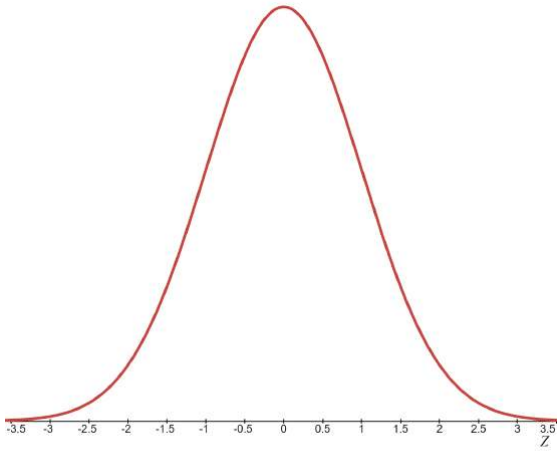
c. What is the probability that a student takes at most 55 minutes to complete the test? Use probability notation in your answer. Round your answer to four decimal places.



Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

d. Daniel took 75 minutes to complete the test. Did Daniel complete the test unusually slow? Explain.

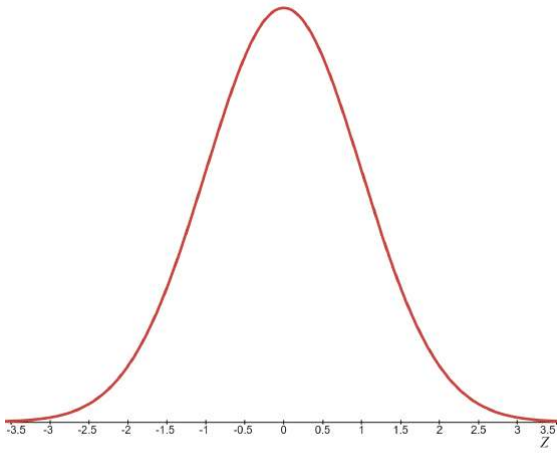
e. What is the proportion of students that complete the test faster than Daniel? Use probability notation in your answer. Round your answer to four decimal places.



Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

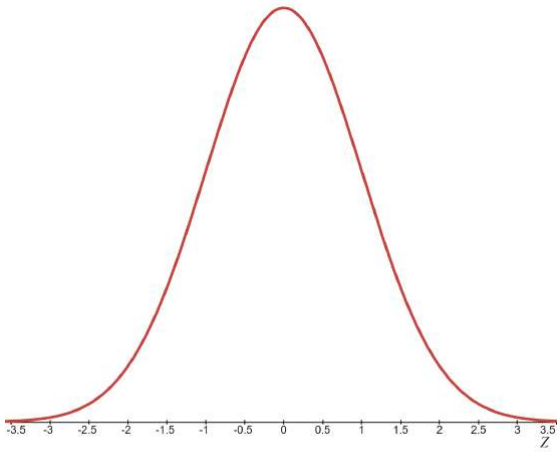
2. The patient recovery time from a particular surgical procedure is normally distributed with a mean of 7 days and a standard deviation of 2.1 days.

a. What is the probability of spending more than 2 days in recovery? Use probability notation in your answer. Round your answer to four decimal places.



Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

b. What is the probability of spending less than 2 days in recovery? Use probability notation in your answer. Round your answer to four decimal places.



Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

c. It takes Reggie 2 weeks to recover. Is this an unusual amount of time for recovery? Explain.

3. Explain the difference between a growth mindset and being positive/optimistic.

This page titled 5.3.1: Exercises is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Hannah Seidler-Wright.

- **Current page** by Hannah Seidler-Wright is licensed CC BY-NC-SA 4.0.
- **1.2: The Statistical Analysis Process** by Hannah Seidler-Wright is licensed CC BY-NC-SA 4.0.

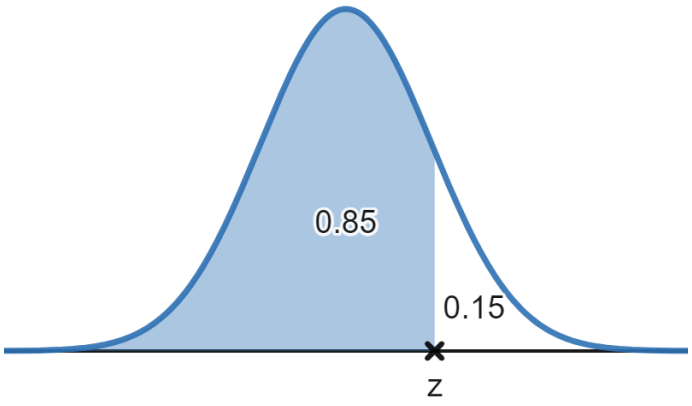
5.4: Finding Critical Values from the Normal Distribution

Sometimes we know the proportion or area that is defined by a range of normally distributed values. But we may not know the cutoff values (also called **critical values**) for the range.

Suppose we want to find the critical value that separates the top 15% of men's heights from the lower 85%. We can use technology to calculate this value from the standard normal distribution. This value is a Z-score.

Using desmos to Find a Critical Value from The Standard Normal Distribution

Suppose we want to find the critical value that separates the top 15% of men's heights from the lower 85%. We can use technology to calculate this value from the standard normal distribution. This value is a Z-score.



Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

To calculate a value of z from a given probability, use the following steps:

1. Go to https://www.desmos.com/calculator.
2. The function we will use takes the area or probability that is less than the critical value. Let's call this area A. It returns a value of z. In the first line, type normaldist().inversecdf(A). Alternatively, you can find the functions normaldist and inversecdf in the Functions menu.

In our example above, we learn that $z = \text{normaldist}().\text{inversecdf}(0.85) \approx 1.036$ is the Z-score that separates the lower 85% from the upper 15%. This is useful information, but we haven't yet found the adult male height that separates the lower 85% from the upper 15%. We need to find the adult male height that corresponds to the Z-score 1.036.

Recall, that a Z-score represents the distance (in standard deviations) a value is above or below the mean. Therefore, the male height that corresponds to this Z-score is 1.036 standard deviations above the mean (since it is positive). x represents random adult male heights, the mean of the distribution is 69 inches, and the standard deviation of the distribution is 3 inches. Let's translate this into a mathematical sentence:

$$\underbrace{\text{This adult male height}}_{x} \text{ is } \underbrace{1.036}_{=1.036} \underbrace{\text{standard deviations}}_{\sigma} \text{ above } \underbrace{\text{the mean}}_{\mu} \tag{5.4.1}$$

$$x = 1.036 \cdot 3 + 69 = 72.108 \text{ inches}$$

Therefore, the adult male height that separates the lower 85% from the top 15% is 72.108 inches. Furthermore, we can say that this adult male height is in the 85th **percentile** because said adult man is taller than 85% of all adult men.

In general, to convert a Z-score to an x-value that is from a normal population that has mean and standard deviation , we use the formula

$$x = z \cdot \sigma + \mu$$

You try!

1. The Welcher Adult Intelligence Test Scale is composed of a number of subtests. On one subtest, the raw scores have a mean of 35 and a standard deviation of 6. Assuming these raw scores form a normal distribution:
   a. What number represents the 65th percentile (what number separates the lower 65% of the distribution)?

   b. What number represents the 90th percentile?

   c. What numbers separate the middle 95% of scores?

2. Kelly's score on the SAT was in the 92nd percentile. Explain what this means about her score relative to all students who took the SAT.

This page titled 5.4: Finding Critical Values from the Normal Distribution is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Hannah Seidler-Wright.

- **Current page** by Hannah Seidler-Wright is licensed CC BY-NC-SA 4.0.
- **1.2: The Statistical Analysis Process** by Hannah Seidler-Wright is licensed CC BY-NC-SA 4.0.

5.4.1: Exercises

1. The patient recovery time from a particular surgical procedure is normally distributed with a mean of 7 days and a standard deviation of 2.1 days.
   a. Martha's recovery time has Z-score -2.1. In how many days did Martha recover?

   b. Hubert's recovery time has a Z-score 1.5. In how many days did Hubert recover?

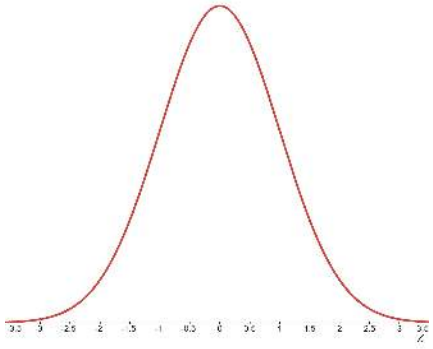
2. The distribution of female baby length at birth is approximately normal with mean 19.3 inches and a standard deviation of 0.6 inches.
   a. What is the range of lengths of unusually short female babies?

   b. What is the range of lengths of unusually long female babies?

   c. A female baby is in the 97th percentile for length. Find this baby's length and interpret your answer in context.

   d. What are the lengths of female babies that correspond to the middle 95%?

3. Suppose the duration of a particular type of criminal trial is known to be normally distributed with a mean of 14 days and a standard deviation of 3 days.
   a. If one of the trials is randomly chosen, find the probability that it lasted at least 19 days. Show your thinking on the graph below, and use probability notation in your answer. Round the probability to four decimal places.



Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

   b. What time frame corresponds to unusually long trials of this type?

   c. 75% of all trials of this type are completed within how many days?

   d. How long did it take the fastest 5% of trials of this type to end?

   e. Jordyn says that the middle 80% of all trials of this type are completed in between 11.48 and 16.52 days. Explain to Jordyn where she made a mistake and how to correct it.

This page titled 5.4.1: Exercises is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Hannah Seidler-Wright.

- **Current page** by Hannah Seidler-Wright is licensed CC BY-NC-SA 4.0.
- **1.2: The Statistical Analysis Process** by Hannah Seidler-Wright is licensed CC BY-NC-SA 4.0.

CHAPTER OVERVIEW

6: Inference Involving a Single Population Proportion

6.1: The Sampling Distribution of Sample Proportions

6.1.1: Exercises

6.2: Estimating a Population Proportion

6.2.1: Exercises

6.3: Introduction to Hypothesis Testing

6.3.1: Exercises

6.4: Hypothesis Tests for a Single Population Proportion

6.4.1: Exercises

6.5: Conclusions (1)

6.5.1: Exercises

This page titled 6: Inference Involving a Single Population Proportion is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Hannah Seidler-Wright.

6.1: The Sampling Distribution of Sample Proportions

Recall the statistical analysis process involves four steps:

1. Ask a question that can be answered by collecting data.
2. Decide what to measure and collect the data.
3. Summarize the data and analyze the data.
4. Draw a conclusion and communicate the results to your audience.

In this unit, we turn our focus to the fourth step: using sample data to draw inferences about a population. In particular, we will focus on categorical/qualitative variables. With categorical variables, individuals in the population fall into some category. We summarize categorical data in a sample by calculating a sample proportion. We will then use sample proportions to draw conclusions about population proportions, which is a proportion (portion, percentage, rate, etc.) for the entire population. This process of drawing conclusions about population parameters from what we observe in a sample (sample statistics) is called statistical inference. In upcoming sections, we will use sample statistics to estimate population parameters, and we will use sample statistics to test claims about population parameters.

What Proportion of the Earth is Covered by Water?

Suppose we want to know the proportion of the Earth's surface which is covered in water. If we collected all points on the surface of the Earth, what proportion of them would be on water? Every point on the surface of the Earth can be described by a pair of coordinates called latitudes and longitudes. Latitudes are horizontal lines that are parallel to the equator. They take on values between -90 and 90 degrees. Latitudes below the equator (0 degrees) are considered negative. Longitudes are vertical lines that are perpendicular to latitudes and take on values between -180 and 180 degrees. Longitudes left of the prime meridian at 0 degrees are considered negative. Each point on the surface of the Earth has an associated latitude and longitude pair.

1. Go to https://www.random.org/geographic-coordinates/ and select a random pair of coordinates by clicking the "Pick Random Coordinates" button. Write the latitude and longitude rounded to the nearest whole degree you found below:

   Latitude:

   Longitude:

   Is this random point on water?

For each point on Earth, the variable is whether or not the point is over water. If a single point is over water, we call it success, and if not, we call it a failure. (Recall we saw this language when computing binomial probability). The proportion of all points that are over water is the total number of points that are over water divided by the total number of points. This population proportion is an example of a parameter. A population proportion is denoted by the symbol p.

We calculate the proportion of points that are over water in a sample by dividing the number of points that are over water in the sample by the number of points in the sample (the sample size). The proportion of points on Earth over water in a sample is an example of a statistic, and is denoted by the symbol $\hat{p}$, pronounced p-hat.

$$\hat{p} = \frac{\text{number of observed successes in the sample}}{\text{sample size}}$$

2. Let's try it! Create a sample of 10 random points and record latitude and longitude rounded to a whole degree, and whether said point is over water in the table below:

	1	2	3	4	5	6	7	8	9	10
latitude/longitude pair										
Over water? (Y/N)										

a. What is the number of observed successes (count of the number of Yes's from the table)?

b. What is the number of failures?

c. Compute the sample proportion.

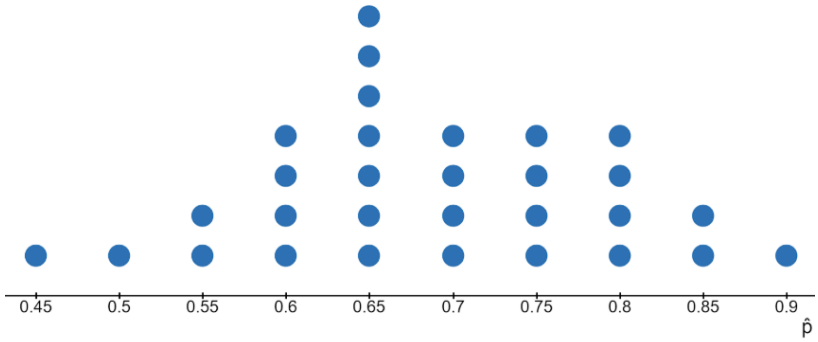
Population	Sample
Collection of all points on Earth	A random sample of 10 points on Earth

Parameter	Statistic
The proportion of points that are over water (p)	Proportion of points that are over water in the sample ($\hat{p}$)

It is important to recognize that there are many samples of points on Earth, each with their own proportion of points over water, but there is only one population proportion. Sample proportions vary from sample to sample, but the population proportion is a single number.

3. Next, we will examine multiple samples of 20 points on Earth. We will calculate a sample proportion $\hat{p}$ for each sample. These proportions are a small part of the collection of all sample proportions. The collection of all sample proportions forms a distribution of values called the sampling distribution of sample proportions.

   The dotplot below shows the results of 30 samples. Each dot represents a sample proportion from a random sample of 20 points on Earth. Use the dotplot to answer the following questions.



Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

a. What does a dot on the graph represent?

b. Describe the shape of the dotplot.

c. Estimate the center of the distribution.

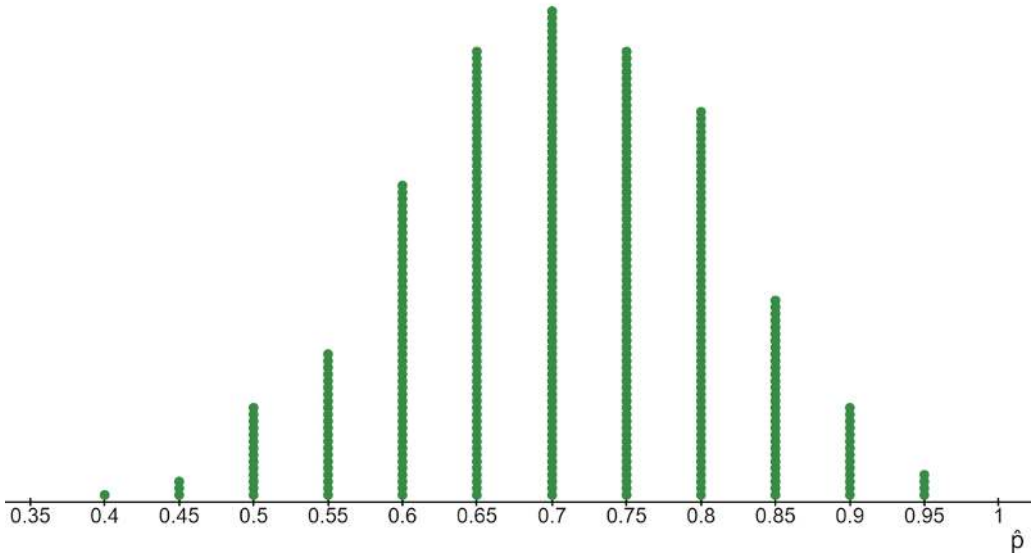
d. What is the range of possible sample proportions?

e. Did each random sample of 20 points on the Earth yield the same sample proportion?

f. Using the dotplot, what is the best estimate of the population proportion of all points on Earth that are over water?

The dotplot above is part of the sampling distribution of sample proportions. A sampling distribution of sample proportions is the distribution of *all* possible sample proportions from samples of a given size.

We use technology to further simulate part of the sampling distribution of sample proportions of points on Earth over water. Counting out points on a map is time consuming and inefficient, but technology can be used to better simulate a distribution of sample proportions. The dotplot below displays 400 sample proportions. Each sample proportion is based on a random sample of 20 points on Earth.



Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

4. Use the above distribution to answer the following questions:
   a. What is the mean of the distribution (visually estimated)?

   b. What is the center, shape, and spread of the distribution?

It turns out that sampling distributions of sample proportions become more normal as the sample size increases. A sampling distribution of sample proportions is the distribution of all possible sample proportions from samples of a given size. If the sample size is large enough, this distribution is approximately normal.

Mean and Standard Error of Sampling Distributions

We denote the mean of sample proportions as $\mu_{\hat{p}}$. We have seen that this mean is equal to the population proportion (p).

$$\text{Mean of sample proportions: } \mu_{\hat{p}} = p$$

A sample proportion is an estimate of the population proportion. When a sample proportion deviates from the population proportion, the deviation is an error in the estimate. Because of this, the standard deviation of sample proportions is called the standard error of sample proportions. We denote the standard error of sample proportions as $\sigma_{\hat{p}}$. The formula for the standard error is:

$$\text{Standard error of sample proportions: } \sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

5. 71% of points on the surface of Earth are on water. So, the population proportion of points on Earth covered over water is 0.71. Use the formulas above to compute the mean and standard error of all sample proportions of points on Earth over water when the sample size is 25. Round the standard error to three decimal places.

   Mean of sample proportions =

   Standard error of sample proportions =

Criteria for Approximate Normality

Statisticians have learned that sampling distributions of sample proportions are approximately normal whenever $np \geq 10$ and $n(1-p) \geq 10$. Since p is the proportion of successes, and n is the sample size, np is the expected, or mean, number of successes in a sample of size n. That is, on average, samples of size n will have np successes. Similarly, $1-p$ is the proportion of failures, and $n(1–p)$ is the expected number of failures. Thus, the sampling distribution of sample proportions is approximately normal if it meets the criteria for approximate normality, which requires there are at least 10 expected successes and 10 expected failures in the sample.

As an example, suppose we sample 50 points on Earth and assume that 71% of all points on Earth are over water. The expected number of successes in a sample is $np = 35.5$ and the expected number of failures is $n(1–p) = 50(0.29) = 14.5$ which could also be calculated as the sample size minus the number of failures.

This means that, on average, samples of 50 points on Earth contain 35.5 points over water and 14.5 points over land. Of course, the number of points over water will vary in individual samples. But since both of the numbers are greater than 10, the normal distribution is a good approximation for the sampling distribution of sample proportions of points on Earth over water in samples of size 50.

You try!

6. When a sampling distribution of sample proportions satisfies the normality criteria we can use the normal distribution properties to find probabilities corresponding to sample proportions.

   The Gallup organization conducts surveys in countries throughout the world to obtain categorical and quantitative data on people and their views about important issues. Gallup surveyed people in the United States in March 2019 to obtain information on U.S. adults' views regarding global warming. They found that 51% of U.S. adults are "concerned believers" who take global warming seriously and believe it poses a serious threat within their lifetime.

   a. For this problem, let's assume that Gallup's result (0.51) is the proportion of all U.S. adults who take global warming seriously. Suppose we sample 120 U.S. adults and determine the proportion who take global warming seriously. We can apply the Central Limit Theorem to describe the sample proportions that are likely to occur given the sample size and assumed population proportion.

      i. What is the sample size (n) and population proportion (p)?

      ii. Sample size, $n =$ _____

      iii. Population proportion, $p =$ _____

   b. Let's explore if the normality criteria are met. First, find the following values. Round answers to one decimal place.

      i. $np =$ _____

      ii. $n(1–p) =$ _____

   c. Are the normality criteria met? Explain.

   d. Find the mean and compute the standard error of the sampling distribution of sample proportions (use three decimal places for the standard error).

      i. Mean $= \mu_{\hat{p}} = p =$ _____

      ii. Standard error $= \sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} =$ _____

   e. Which statement below best describes the standard error of the sampling distribution of sample proportions? Select an answer that is incorrect and explain why.

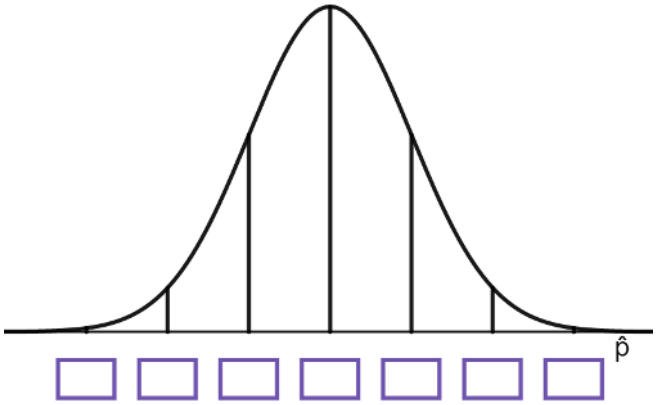
      i. Sample proportions vary, and the difference between the lowest possible sample proportion and highest possible sample proportion is 0.046.

      ii. 0.046 is the typical distance which sample proportions are from the population proportion.

      iii. All sample proportions are within 0.046 from the population proportion.

      iv. No sample proportions equal the population proportion.

f. The boxes under the normal distribution below are one standard error apart, with the center box at the mean. Use the mean and standard error above to enter the correct values into the boxes.



Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

g. Suppose that in a random sample of 120 U.S. adults 48 respond that they take global warming seriously. Compute the sample proportion, $\hat{p}$. Write your answer as a decimal rounded to two decimal places.

h. What is the Z-score of the sample proportion from part g? Use the mean and standard error from part d.

i. Does this Z-score indicate that this sample proportion is unusual? Explain how you know.

j. Use desmos to find the probability of observing a sample proportion that is less than or equal to the value from part g. Round your answer to four places after the decimal. Write the function you used in desmos to find the result.

Summary

- The Central Limit Theorem for sample proportions states that a sampling distribution of sample proportions is approximately normal if the sample size is large enough. Our criteria for determining this are: $np \geq 10$ and $n(1-p) \geq 10$.
- When the criteria for approximate normality are satisfied, the normal distribution may be used to determine probabilities about sample proportions.
- The mean and standard error (or standard deviation) for the sampling distribution of sample proportions are given by:

$$\text{Mean } = \mu_{\hat{p}} = p$$

$$\text{Standard error } = \sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

This page titled 6.1: The Sampling Distribution of Sample Proportions is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Hannah Seidler-Wright.

- **Current page** by Hannah Seidler-Wright is licensed CC BY-NC-SA 4.0.
- **1.2: The Statistical Analysis Process** by Hannah Seidler-Wright is licensed CC BY-NC-SA 4.0.

6.1.1: Exercises

1. In 2022, a study reported that 55% of Americans support cancellation of up to $10,000 per borrower in federal student loans. The study was based on data from a national random survey of 1250 Americans. The sample was representative of all Americans, so researchers used this study to describe characteristics of all Americans.
   a. In this study, what is the sample?

   b. In this study, what is the sample statistic?

   c. In this study, what is the population?

   d. In this study, what is the population parameter that the sample statistic is estimating?

2. In 2022, 74% of student loan borrowers who borrowed for their own education reported having a student loan debt balance of more than $10,000. Go to this sampling distribution applet to complete the following problems. Enter 0.74 in the Probability of success field and change the number of samples to 100. You can use the QR code below to access the applet.

   a. Reduce the sample size (n) to 20, click Draw Samples. Describe the center, shape, and spread of the distribution.

   b. Increase the sample size (n) to 50, click Draw Samples. Describe the center, shape, and spread of the distribution.

   c. Increase the sample size (n) to 1000, click Draw Samples, and check the Show Normal Curve box. Describe the center, shape, and spread of the distribution.

   d. Which feature(s) of the simulated sampling distributions change when you increase the sample size? How do they change?

   e. The sampling distribution contains frequencies of all possible sample proportions from a sample of a fixed size. It has mean, $\mu_{\hat{p}} = p$, and standard error, $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$. Compute the mean and standard error (rounded to three decimal places) for a sample of size 50.

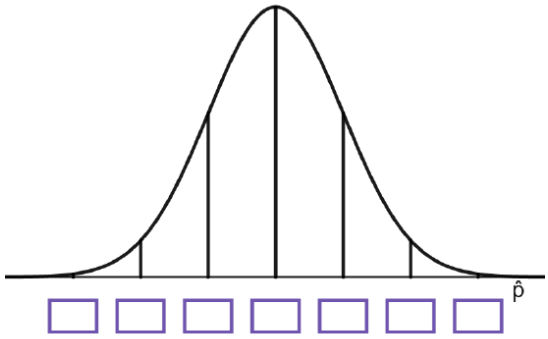
3. As Brazilians head to the polls Sunday to vote to elect their next legislature and president, 67% of Brazilians are not confident in the honesty of their country's elections, according to Gallup. Let's assume this value represents the population proportion. Imagine that a researcher surveys a random sample of 100 adults in Brazil, asking if they are confident in the honesty of their country's elections. The researcher wants to know whether people's opinions have changed. We will use the Central Limit Theorem to think about the possible results.

   a. If samples of size 100 are taken, find the mean and standard error (rounded to three decimal places) of the resulting sampling distribution of sample proportions. Assume the population proportion is $p = 0.67$, the proportion of Brazilians who are not confident in the honesty of their country's elections.

   b. A sampling distribution is a description of all possible values of a statistic. What does this sampling distribution represent?

   c. Is this sampling distribution approximately normal? Explain why or why not.

d. The boxes under the normal distribution below are one standard error apart, with the center box located under the mean. Use the mean and standard error you calculated in a. to label the horizontal axis.



Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

e. Use the empirical rule to find the interval centered at p that contains approximately 95% of all sample proportions.

f. What sample proportions would you consider unusual?

g. In a random sample of 100 adults in Brazil, 74 say they are not confident in the honesty of their country's elections. What is the sample proportion, $\hat{p}$?

h. Find the Z-score for this sample proportion. Round to two decimal places.

i. Use desmos to find the probability that in a random sample of 100 adults in Brazil, 74 or more will say that they do not have confidence in the honesty of their country's elections. Round the probability to four decimal places.

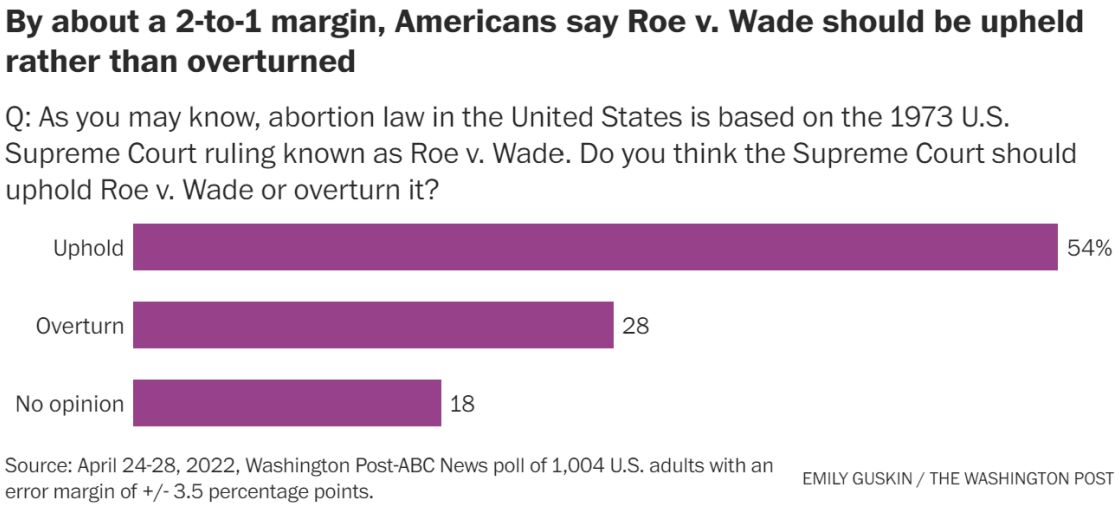
This page titled 6.1.1: Exercises is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Hannah Seidler-Wright.

- **Current page** by Hannah Seidler-Wright is licensed CC BY-NC-SA 4.0.
- **1.2: The Statistical Analysis Process** by Hannah Seidler-Wright is licensed CC BY-NC-SA 4.0.

6.2: Estimating a Population Proportion

On May 3rd, 2022, The Washington Post published an article[4] that included results of a poll that asked US adults whether they think the supreme court should uphold Roe v. Wade or overturn it, after reports of the Supreme Court's plans to overturn the right to abortion were leaked to the public.

Findings from a survey of 1,004 US adults suggest that 54% believe the ruling should be upheld, 28% believe the ruling should be overturned, while 18% have no opinion, with an error margin of +/- 3.5 percentage points.



Identifying Important Information

1. Washington Post-ABC News used sample data to estimate the true population proportion of those adults in the US who believe Roe v. Wade should be upheld.

   a. What is the sample size?

   b. Is 54% a parameter or a statistic? Justify your answer.

c. Do you think the population proportion of adults in the US who believe Roe v. Wade should be upheld is *exactly* 54%? Why or why not?

d. What do you think the "error margin of +/- 3.5 percentage points" tells you about the population proportion?

The sample proportion $\hat{p}$ is a single estimate of the population proportion, and is often referred to as a **point estimate**. An **interval estimate** is a range of values used to estimate a population proportion. Interval estimates are centered at a point estimate. The interval is formed as

$$\text{Point Estimate} \pm \text{Margin of Error}$$

e. Identify the point estimate and the margin of error.

$\hat{p} =$

$E =$

f. Construct the endpoints of the interval given in the results of the poll.

Interval estimates like you found above, give a range of values that *usually* contain the true population proportion, which is the proportion of *all* adults in the US who believe Roe v. Wade should be upheld in the example above.

g. Do you think that it is likely that the true proportion of adults in the US who believe Roe v. Wade should be upheld is as high as 60%? Explain.

h. Do you think it's likely that the population proportion of adults in the US who believe Roe v. Wade should be upheld is as low as 45%? Explain.

We call the interval estimate for a population parameter a **confidence interval**. To construct a confidence interval, we need

- A point estimate (a statistic is a point estimate)
- A margin of error

The interval includes values between the point estimate minus the margin of error and the point estimate plus the margin of error. We write the interval in interval notation as

$$(\text{point estimate - margin of error, point estimate + margin of error})$$

When we estimate a single population proportion, we can translate this interval as

$$(\hat{p} - E, \hat{p} + E)$$

The Sampling Distribution of Sample Proportions and the Margin of Error

To learn about margins of error, we will need to think about the sampling distribution of sample proportions. Recall,

- A sampling distribution of sample proportions is approximately normal if there are at least 10 expected successes and failures in the random sample.
- The mean of the sampling distribution of sample proportions is $\mu_{\hat{p}} = p$.
- The standard error (standard deviation of the sampling distribution of sample proportions) is $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$.

When p is Unknown

When we construct a confidence interval, we *estimate* the true population proportion and we therefore, do not know what the true population proportion is. When the population proportion p is unknown, we use a sample proportion $\hat{p}$ in its place.

- A sampling distribution of sample proportions is approximately normal if there are at least 10 **observed** successes and failures in the random sample. This is because we don't know the population proportion and cannot compute np and $n(1-p)$.
- The mean of the sampling distribution of sample proportions is $\mu_{\hat{p}} \approx \hat{p}$ since we do not know the value of p.
- The standard error (standard deviation of the sampling distribution of sample proportions) is $\sigma_{\hat{p}} \approx \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$.

---

2. Suppose Garrett wants to estimate the proportion of adults in Texas who support upholding Roe v. Wade. He randomly surveys 200 adults in Texas. He finds that 114 surveyed adults say they support upholding Roe v. Wade.

Step 1: Verify normality criteria

a. Are the criteria for the approximate normality of the sampling distribution met?

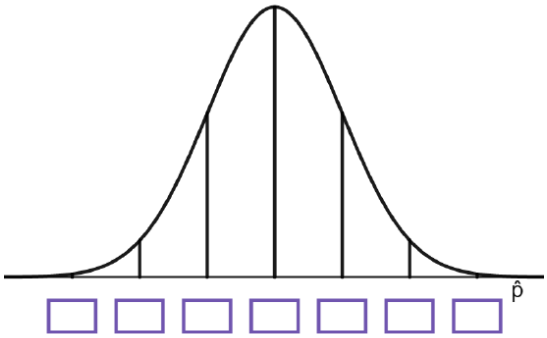
b. Compute the sample proportion $\hat{p}$.

Step 2: Compute the critical value

a. Compute the mean of the sampling distribution of sample proportions, $\mu_{\hat{p}}$.

b. Compute the standard error, $\sigma_{\hat{p}}$, rounded to three decimal places.

c. Label the horizontal axis of the normal sampling distribution of sample proportions below.



Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

d. Garrett would like to be 97% sure or *confident* that the true proportion of all Texans who support upholding Roe v. Wade is in his interval estimate. Shade the middle 97% of the normal sampling distribution above.

e. We want to work backwards from the standard normal distribution to find the lower and upper values of $\hat{p}$ that are limits of the middle 97% of data (all sample proportions). Recall, we find critical values to achieve this. Compute the critical value Z_c that corresponds to the upper limit of the 97% confidence level. Draw a picture to show you thinking.

f. Usually in confidence interval problems, we use three main confidence levels: 90%, 95%, and 99%. Access this desmos graph or the QR code below to compute the critical values Z_c that correspond to these three confidence levels rounded to three decimal places.

Confidence Level	Critical Value
90%	$Z_c = normaldist().inversecdf(0.95) \approx$
95%	$Z_c = normaldist().inversecdf(0.975) \approx$
99%	$Z_c = normaldist().inversecdf(0.995) \approx$

Step 3: Compute the margin of error E.

a. The margin of error is $E = Z_c \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$. Use the information above to compute the margin of error rounded to three decimal places.

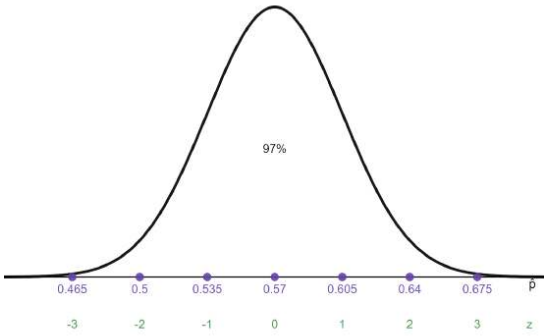
Step 4: Compute the lower and upper limits of the interval

a. Compute the proportion that is the lower limit of the interval $\hat{p} - E$.

b. Compute the proportion that is the upper limit of the interval $\hat{p} + E$.

c. Write the interval in interval notation as $(\hat{p} - E, \hat{p} + E)$.

d. Locate these proportions on the graph below and shade the area in the middle.



Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

Step 5: Interpret the interval in context.

a. Fill in the blanks: We are _____ % confident that the true proportion of all Texans that _____ is between _____% and _____%.

b. Is it likely that the majority of Texans believe Roe v. Wade should be upheld? Use the interval to support your answer.

The Five Step Process for Building a Confidence Interval

Use the following steps when constructing a confidence interval:

1. Verify that the sampling distribution is approximately normal.
   - When estimating with a single population proportion, we check that there are at least 10 observed successes $(n\hat{p})$ and failures $(n(1-\hat{p}))$ in the sample.
2. Compute the critical value.
   - When estimating with a single population proportion, we find the critical value, Z_c, from the standard normal distribution. In desmos, we enter normaldist().inversecdf(A) where A is the area left of the critical value.
3. Compute the margin of error.
   - When estimating with a single population proportion, $E = Z_c \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$.
4. Compute the limits of the interval and write the interval in interval notation.
   - When estimating with a single population proportion, the interval is $(\hat{p} - E, \hat{p} + E)$.
5. Write a conclusion and interpret the interval in context.
   - Include the confidence level, the parameter that you are estimating, and the bounds (with units) in your conclusion.

Reference

[4] "Majority of Americans say Supreme Court should uphold Roe, Post-ABC poll finds," Emily Guskin and Scott Clement, May 3, 2022, accessed May 13, 2022, https://www.washingtonpost.com/politics/2022/05/03/most-americans-say-supreme-court-should-uphold-roe-post-abc-poll-finds/

This page titled 6.2: Estimating a Population Proportion is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Hannah Seidler-Wright.

- **Current page** by Hannah Seidler-Wright is licensed CC BY-NC-SA 4.0.
- **1.2: The Statistical Analysis Process** by Hannah Seidler-Wright is licensed CC BY-NC-SA 4.0.

6.2.1: Exercises

1. In a Washington Post-University of Maryland poll[5] of 1503 randomly selected US adults, 55% said they strongly support gender equity in sports. In the following exercises, you will construct a 99% confidence interval for the proportion of US adults who strongly support gender equity in sports.

   a. Identify the relevant information:

      i. What is the sample size?

      ii. What is the sample proportion?

      iii. Write a sentence to describe what the population parameter is in context.

   b. Step 1: Is the sampling distribution of sample proportions approximately normal? Why or why not?

   c. Step 2: Compute the critical value from the standard normal distribution that corresponds to a confidence level of 99%. Write the function you use in desmos to find a critical value from the standard normal distribution.

   d. Step 3: Compute the margin of error $E \approx Z_c \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$. Round to three decimal places.

   e. Step 4: Give the lower and upper limits of the 99% confidence interval for the population proportion (p). Then write the interval in interval notation.

   f. Step 5: Interpret the interval in context. Is it likely that a majority of US adults strongly support gender equity in sports? Use the interval to support your answer.

2. In a CNN poll[6] of 1002 randomly selected US adults, 371 approve of the supreme court's decision to overturn Roe v. Wade. Construct a 95% confidence interval for the true proportion of US adults who approve of the supreme court's decision to overturn Roe v. Wade.

   a. Write a sentence describing p in context.

   b. Step 1: Verify that the sampling distribution of sample proportions is approximately normal. Justify your answer.

   c. Step 2: Compute the critical value.

   d. Step 3: Compute the sample proportion and the margin of error rounded to three decimal places.

   e. Step 4: Compute the interval in interval notation.

   f. Step 5: Interpret the interval in context.

   g. Is it possible that a majority approves of the supreme court's decision to overturn Roe v Wade? Use the interval to support your answer.

3. According to a poll[7] conducted by Gallup of 800 randomly surveyed US adults, 312 respondents were satisfied with the quality of the environment. Construct a 90% confidence interval using the five step process. Round the margin of error to three decimal places.

Reference

[5] Liz Clarke, Scott Clement, and Emily Guskin, "Most Americans support gender equity in sports scholarships, poll finds," *Washingtonpost.com*, June 22, 2022, accessed September 27, 2022, https://www.washingtonpost.com/sports/2022/06/22/title-ix-poll-americans-support-gender-equity/

[6] Jennifer Agiesta, "About two-thirds of Americans disapprove of overturning Roe v. Wade, see negative effect for the nation ahead," *CNN.com*, July 28, 2022, accessed September 27, 2022, https://www.cnn.com/2022/07/28/politics/cnn-poll-abortion-roe-v-wade/index.html

[7] Jeffrey M. Jones, "Americans Offer Gloomy State of the Nation Report," *Gallup.com*, February 2, 2022, accessed September 27, 2022, https://news.gallup.com/poll/389309/americans-offer-gloomy-state-nation-report.aspx

This page titled 6.2.1: Exercises is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Hannah Seidler-Wright.

6.3: Introduction to Hypothesis Testing

In this section, we begin a new type of statistical inference known as hypothesis testing. Hypothesis testing can seem awkward at first, but when you really understand it, you see that it's actually how your mind makes decisions after being convinced by sufficient evidence. We will consider the outcomes of flipping a coin versus spinning a coin.

1. What is the theoretical probability that a penny is flipped and lands on tails?

2. What if we flipped the coin 100 times? You said the probability of it landing on tails is _____ . If that is true, if you flipped a penny 100 times, would you get EXACTLY 50 tails? Explain.

3. Daquan says, "I flipped a penny 100 times for a school project and the penny came up tails _____ number of times." Fill in the blank with a number that would make you think Daquan is lying or there was something wrong with his experiment. Then explain your thinking.

4. You and Daquan then have this conversation:

   Daquan: "You know spinning a penny is different than flipping a penny?"

   You: "You mean flipping where you throw it in the air, and spinning where you spin it on a table?"

   Daquan: "Yes! If you spin a penny it usually lands tails side up."

   How would you respond?

5. You try it yourself. You spin a penny 100 times and record the results. The penny lands:

   Heads side up 33 times

   Tails side up 67 times.

   Are you surprised by this result? Do you believe Daquan's claim?

You might be surprised by this result or you might think that the observation may have just happened by chance. Either way, you made a decision about a population (all penny spins) which you cannot observe in its entirety. Your decision was based on a small sample (100 penny spins). In statistics, sample data help us make decisions about populations through a process known as hypothesis testing. A hypothesis is an assumption or claim. We need a formal process to test Daquan's claim.

Step 1: Determine the null and alternative hypotheses

Let's see if what we observed was unusual enough to convince us that Daquan is correct. We will test Daquan's claim, that a spinning penny lands tails side up the majority of the time.

We will begin by writing some hypotheses:

The null hypothesis is the statement of no change (the dull hypothesis). In this context, the proportion of coin spins that land tails up is 50% (the same as flipping a penny). In mathematical symbols,

$H_0 : p = 0.5$

Daquan's claim is what we call **the alternative hypothesis**. The proportion of coin spins that land tails up is actually more than 50% (a majority). In mathematical symbols,

$H_a :$ ___ ___ ___

Step 2: Collect Sample Data

We want to know if our observation is unusual, therefore, we want to know if our sample proportion is unusual. We must take a look at the sampling distribution of sample proportions and we hope that it is approximately normal. In our sample, we spin a penny 100 times and it lands tails side up 67 times. This is the number of observed successes in the sample.

a. What is the number of expected successes in our sample (assuming our null hypothesis is true)?

b. What is the number of expected failures in our sample (assuming our null hypothesis is true)?

We want to know if our observation is unusual, therefore, we want to know if our sample proportion is unusual. We must take a look at the sampling distribution of sample proportions. We found that the sampling distribution is approximately normal because there were at least 10 expected successes and failures in our sample.

In our sample, we spin a penny 100 times and it lands tails side up 67 times. What is the sample proportion?

$$\hat{p} = \frac{\text{number of observed successes}}{\text{sample size}} =$$

Step 3: Assess the Evidence

We want to know if our observation is unusual, therefore, we want to know if our sample proportion is unusual. We should find a Z-score. What is the Z-score for the observed sample proportion?

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} = \frac{\_\_ - \_\_}{\sqrt{\frac{\_\_(1-\_\_)}{\_\_}}} =$$

The correct Z-score test value is _____. That means the difference between what I got (67/100 tails) and what most people think I should get (50/100 tails) is 3.4 standard deviations away from what is expected. That is far away from what is expected. Next, we explore probability.

Use the Z-score to find the probability of observing a sample proportion as high or higher than the one we observed. Write the desmos function you used to do the computation.

$$P(\hat{p} \geq \_\_\_\_\_) = P(Z \geq \_\_\_\_\_) =$$

IF the null hypothesis is true (that the probability of tails is 0.5 when spinning a penny) were true, the probability of getting our result of 67 tails out of 100 spins just by chance is _____.

Step 4: State a Conclusion

Statisticians use a rule about how small a probability should be in order for us to consider an event unusual, or **statistically significant**. We often consider an event unusual if the probability of its occurrence is less than or equal to 5%. This is called the **level of significance**. Other levels of significance can be used.

When the P-value is **less than or equal** to the level of significance, we reject the null hypothesis and support the alternative hypothesis.

Since the P-value is less than or equal to the level of significance, we reject the null hypothesis and support the alternative hypothesis. The sample data support the claim that the proportion of spins of a penny that result in tails is more than 50%. We can support the claim that spins land tails up a majority of the time.

What would be some reasons why spinning a penny results in so many more tails than flipping a penny?

The Four Step Hypothesis Testing Process

Step 1. Determine the null and alternative hypotheses

The null hypothesis is a mathematical sentence that makes an assumption of fairness. The alternative hypothesis is a mathematical sentence that represents an opposing or alternative belief.

Step 2. Collect Sample Data

Compute or record the sample statistic and check that the sampling distribution is normally distributed.

Step 3. Assess the Evidence

We determine the strength of our evidence through probability. This probability is called a **P-value**, not to be confused with p which represents a population proportion. The P-value is computed using the assumption made in the null hypothesis. A P-value is the probability of observing a sample statistic that is at least as extreme as the one we observed, assuming the null hypothesis is true.

If our sample proportion differs significantly from the assumed population proportion, then it likely did not occur just by chance.

Step 4. State a Conclusion

Statisticians use a rule about how small a probability should be in order for us to consider an event unusual, or **statistically significant**. We often consider an event unusual if the probability of its occurrence is less than or equal to 5%. This is called the **level of significance**. Other levels of significance can be used.

When the P-value is **less than or equal** to the level of significance, we reject the null hypothesis and support the alternative hypothesis.

When the P-value is **greater than** the level of significance, we do not reject the null hypothesis and we cannot support the alternative hypothesis.

Lastly, write a conclusion in context in plain language.

Tips for Writing Conclusions

1. Notice that *we do not support or accept the null hypothesis*. We assume fairness to begin with.
2. *We do not reject the alternative hypothesis.* We either have strong evidence to support it or not.
3. Always say something that makes it clear that your *evidence is based on sample data*. Always include a word that indicates the conclusion is about a *population parameter*. The parameter (proportion, mean, mean difference, standard deviation, etc.) should be included in the statement of the conclusion.

This page titled 6.3: Introduction to Hypothesis Testing is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Hannah Seidler-Wright.

6.3.1: Exercises

1. According to a poll conducted by Gallup[8] 46% of US adults were satisfied with the quality of the environment in 2020. You wonder if the proportion of US adults who are satisfied with the quality of the environment has decreased. You randomly survey 300 US adults and find that 117 respondents are satisfied with the quality of the environment. Is the sample proportion strong enough evidence to support your claim? Test it at a level of 5% significance.

   Step 1. Determine hypotheses

   a. Let p represent the population proportion of interest (the proportion we are testing a claim about). Write what p represents in words.

   b. The null hypothesis is a statement of no change that we will test our claim against. If the proportion has not changed since 2020, what is it equal to?

   Null hypothesis: $p =$ _____

   c. The alternative hypothesis is the mathematical representation of the claim: "You wonder if the proportion of US adults who are satisfied with the quality of the environment has decreased." Write the alternative hypothesis with one of the following options: $p <$ assumed value, $p >$ assumed value, or $p \neq$ assumed value where the assumed value is given in the null hypothesis. (Fill in the blank with the appropriate inequality).

   Alternative hypothesis: p _____ 0.46

   Step 2. Collect sample data

   a. Compute the *expected* number of successes and failures in the sample assuming the null hypothesis is true (Misplaced '#', Misplaced '#'). Verify that the sampling distribution of sample proportions is approximately normal.

   b. Compute the sample proportion $\hat{p} = \frac{\text{number of observed successes}}{\text{sample size } (n)}$.

   Step 3. Assess the evidence (using probability)

a. Compute the mean and standard error for the sampling distribution of sample proportions.

b. Compute the Z-score for the sample proportion, $\hat{p}$. Round to two decimal places.

c. Sketch the standard normal distribution. Locate the Z-score on the horizontal axis and shade the area to the *left* of the Z-score.

d. The P-value is the probability of observing a test statistic that is at least as extreme as the one we observed. In this hypothesis test, the P-value is the probability of observing a sample proportion as low or lower than the one we observed. This is the area you shaded in c. Use desmos to compute the P-value.

e. The null hypothesis assumes that the proportion of US adults who are satisfied with the quality of the environment is the same as in 2020 (46%). We observed a sample proportion that was lower. Does the P-value indicate that the sample proportion we observed was unusual compared to the assumed population proportion?

Step 4. State the conclusion in context

a. When the P-value is less than the level of significance, the sample proportion is considered statistically significant. Is our observed sample proportion statistically significant?

b. Do we reject or fail to reject the null hypothesis?

c. Can we support the alternative hypothesis or not?

d. Fill in the blanks in the following conclusion:

The sample data _____ the claim that the true _____ of US adults who are satisfied with the quality of the environment has _____ since 2020.

2. Juanda is thinking about buying a new car and is deciding between different colors. She likes the color red but is concerned that red cars get pulled over more often than other cars. She does some research and finds that all cars get pulled over at a rate of around 9%. She wants to know if red cars get pulled over at a higher rate.

a. p represents:

b. Null hypothesis:

c. Alternative hypothesis:

d. To find the P-value, should Juanda use a left-, right-, or two-tailed test based on her alternative hypothesis?

Reference

[8] Jeffrey M. Jones, "Americans Offer Gloomy State of the Nation Report," *Gallup.com*, February 2, 2022, accessed September 27, 2022, https://news.gallup.com/poll/389309/americans-offer-gloomy-state-nation-report.aspx

This page titled 6.3.1: Exercises is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Hannah Seidler-Wright.

6.4: Hypothesis Tests for a Single Population Proportion

In a previous lesson, we were introduced to the four step hypothesis testing process:

Step 1. Determine hypotheses

Step 2. Collect sample data

Step 3. Assess the evidence

Step 4. State a conclusion in context

We will now take a closer look at each of these steps.

Determine the Hypotheses

In order to test a claim about a population parameter, we create two opposing hypotheses. We call these the null hypothesis, H_0, and the alternative hypothesis, H_a. Let p represent a given population proportion.

The Null Hypothesis

In every hypothesis test, we assume that the null hypothesis is true. The null hypothesis is always a statement of equality and therefore, should always contain an equal symbol ($=$). When a test involves a single population proportion, the null hypothesis will be

$$H_0 : p = \text{ value}$$

Since the value is a proportion, it will be a number between 0 and 1 (inclusive).

The Alternative Hypothesis

The alternative hypothesis is a claim implied by the research question and is an inequality. The alternative hypothesis states that population proportion is greater than ($>$), less than ($<$), or not equal ($\neq$) to the assumed value in the null hypothesis.

When a test involves a single population proportion, alternative hypothesis will be one of the following:

$$\begin{aligned} H_a &: p > \text{ value} \\ H_a &: p < \text{ value} \\ H_a &: p \neq \text{ value} \end{aligned} \tag{6.4.1}$$

Example 1

Research on college completion has shown that about 60% of students who begin college eventually graduate. A publication of higher education claims that the proportion for STEM (science, technology, engineering, math) majors is lower.

Solution: We will let p represent the proportion of all STEM majors who begin college and ultimately graduate. The null hypothesis is $H_0 : p = 0.60$. The alternative hypothesis is $H_a : p < 0.60$. The publication authors have the burden of proof and must produce evidence to support their claim that the proportion of college graduates among STEM majors is lower against the assumption that it is not.

You try!

1. The population proportion is represented by the symbol p. In the following questions, write a sentence in words for what p represents. Then determine the null and alternative hypotheses.

   a. About 67% of registered voters voted in the 2020 presidential election. A student claims that less than 67% of students at our college voted in the 2020 presidential election.

   p represents:

   H_0: _____

   H_a: _____

   b. In 2013, the US Department of Defense changed a policy that affected women in the armed forces. Under the new rules, women who met physical requirements could be assigned to combat positions. Many people in the US were opposed to the change. About 18% of women were against it. Researchers want to know if the proportion of US men who were against the decision was different[9].

   p represents:

   H_0: _____

   H_a: _____

c. A pro-life advocate believes that a majority (more than 50%) of unplanned pregnancies resulted from no use of contraceptive methods.

p represents:

H_0: _____

H_a: _____

Collect Sample Data

During a hypothesis test, we work to know if a sample statistic is unusual or not. Therefore, we must think about probabilities from a sampling distribution.

In a previous lesson, we learned about the sampling distribution of sample proportions. The Central Limit Theorem says that a sampling distribution of sample proportions is approximately normal if there are at least 10 expected successes (np) and failures ($n(1-p)$) in the sample. In the second step of a hypothesis test, we verify that the sampling distribution is approximately normal and we identify or compute any sample statistics.

Example 2

Research on college completion has shown that about 60% of students who begin college eventually graduate. A publication of higher education claims that the proportion for STEM (science, technology, engineering, math) majors is lower. Researchers randomly select 100 STEM majors and determine that 51 eventually graduate.

Solution: Recall, the null hypotheses is

$$H_0 : p = 0.60$$

The number of expected successes in the sample is $np = 100(0.60) = 60$. The number of expected failures in the sample is $n(1-p)$ $= n - np = 100 - 60 = 40$. Therefore, the sampling distribution of sample proportions is approximately normal.

The sample proportion is

$$\hat{p} = \frac{x}{n} = \frac{51}{100} = 0.51$$

Assess the Evidence

This step is all about probability. Since the sampling distribution is approximately normal (as determined in step 2), we can compute a Z-score and use the standard normal distribution to find probabilities. The sampling distribution of sample proportions has mean

$$\mu_{\hat{p}} = p$$

and standard error

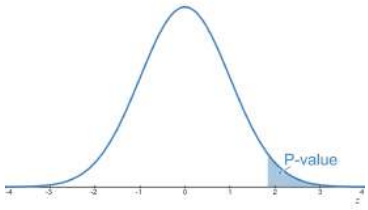
$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

where p is the assumed population proportion, and n is the sample size. The test statistic is

$$z = \frac{x - \mu}{\sigma} \text{ which translates to } z = \frac{\hat{p} - \mu_{\hat{p}}}{\sigma_{\hat{p}}} = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}}$$

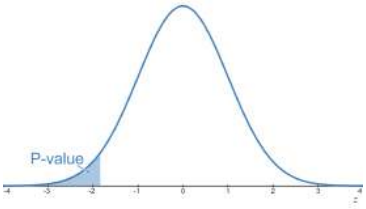
when looking at the sampling distribution of sample proportions.

- When the alternative hypothesis is $H_a : p >$ value , we are conducting a right-tailed test. The P-value is the probability of observing a sample proportion at least as extreme as the one we observed. In this case at least as extreme means “as high or higher”. The P-value is the area to the right of the test statistic (T.S.).



Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

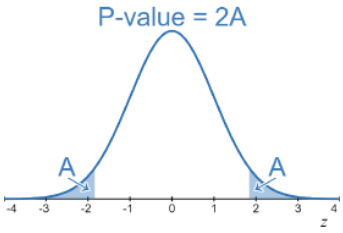
  - To compute the P-value using desmos,
    1. Go to https://www.desmos.com/calculator.
    2. In the first line type normaldist(). Alternatively, you can click the keypad icon, then the Functions button, scroll to Distributions, and select normaldist. This will graph the standard normal distribution, where the horizontal axis represents z-scores.
    3. Click the Zoom Fit button (magnifying glass icon) to adjust the view.
    4. Click the small arrow next to Cumulative Probability to expand the options.
    5. By default, it shows the probability as $P(-1 \leq x \leq 1)$. To find the area in the right-tail, press the Right button. It now shows the probability as $P(x \geq 0)$. Enter the test statistic in the field containing 0. The output is the P-value.
- When the alternative hypothesis is $H_a : p <$ value , we are conducting a left-tailed test. The P-value is the probability of observing a sample proportion at least as extreme as the one we observed. In this case at least as extreme means “as low or lower”. The P-value is the area to the left of the test statistic (T.S.).



Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

- To compute the P-value using desmos,
  1. Go to https://www.desmos.com/calculator.
  2. In the first line type normaldist(). Alternatively, you can click the keypad icon, then the Functions button, scroll to Distributions, and select normaldist. This will graph the standard normal distribution, where the horizontal axis represents z-scores.
  3. Click the Zoom Fit button (magnifying glass icon) to adjust the view.
  4. Click the small arrow next to Cumulative Probability to expand the options.
  5. By default, it shows the probability as $P(-1 \leq x \leq 1)$. To find the area in the left-tail, press the Left button. It now shows the probability as $P(x \leq 0)$. Enter the test statistic in the field containing 0. The output is the P-value.

- When the alternative hypothesis is $H_a : p \neq$ value , we are conducting a two-tailed test, and the P-value is twice the area of either the tail to the right of a positive test statistic (T.S.), or the tail to the left of a negative test statistic (T.S.).



Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

- To compute the P-value using desmos,
  1. Go to https://www.desmos.com/calculator.
  2. In the first line type normaldist(). Alternatively, you can click the keypad icon, then the Functions button, scroll to Distributions, and select normaldist. This will graph the standard normal distribution, where the horizontal axis represents z-scores.
  3. Click the Zoom Fit button (magnifying glass icon) to adjust the view.
  4. Click the small arrow next to Cumulative Probability to expand the options.
  5. By default, it shows the probability as $P(-1 \leq x \leq 1)$. To find the area in the two-tails, press the Outer button. It now shows the probability as $P(x \leq -1 \text{ or } x \geq 1)$. Enter the negative of the test statistic in the field containing -1 and the positive of the test statistic in the field containing 1. The output is the P-value.

Example 3

Research on college completion has shown that about 60% of students who begin college eventually graduate. A publication of higher education claims that the proportion for STEM (science, technology, engineering, math) majors is lower. Researchers randomly select 100 STEM majors and determine that 51 eventually graduate.

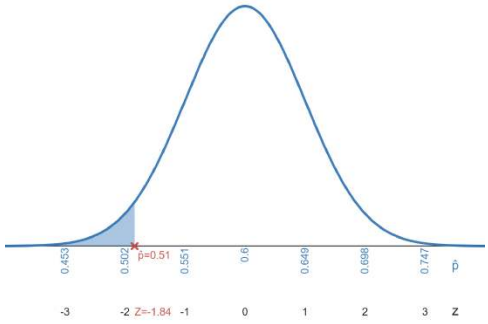
Solution: Recall, the null hypotheses is

$$H_0 : p = 0.60, \hat{p} = \frac{x}{n} = \frac{51}{100} = 0.51$$

The sampling distribution of sample proportions is approximately normal. The mean of the sampling distribution of sample proportions is $\mu_{\hat{p}} = p = 0.60$ and the standard error is

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{(0.60)(1-0.60)}{100}} \approx 0.049$$

The sampling distribution is shown below. The major tick marks have been labeled with values of $\hat{p}$ and the corresponding Z-scores.



Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

We compute the Z-score for the sample statistic,

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} = \frac{0.51 - 0.60}{\sqrt{\frac{(0.60)(1-0.60)}{100}}} \approx -1.84$$

The sample statistic is -1.84 standard errors below the assumed population proportion. We perform a left-tailed test because the alternative hypothesis, $H_a : p < 0.60$, contains a less than inequality. We can now find the P-value, which is the probability of seeing a sample proportion as low or lower than 0.51, by finding the probability from the standard normal distribution. Go to https://www.desmos.com/calculator. In the first line type normaldist(). Alternatively, you can click the keypad icon, then the Functions button, scroll to Distributions, and select normaldist. This will graph the standard normal distribution, where the horizontal axis represents z-scores. Click the Zoom Fit button (magnifying glass icon) to adjust the view. Click the small arrow next to Cumulative Probability to expand the options. By default, it shows the probability as $P(-1 \leq x \leq 1)$. To find the area in the left-tail, press the Left button. It now shows the probability as $P(x \leq 0)$. Enter the test statistic in the field containing 0 so that it

says $P(x \leq -1.84)$. The output of 0.033 is the P-value.

State a Conclusion

Hypothesis tests are all about making decisions. We use the P-value to make a decision about the null and alternative hypotheses.

We compare our P-value to a level of significance. The level of significance, denoted (the Greek letter "alpha"), is how unlikely a sample statistic needs to be to convince us about a claim. It is also the level of risk we accept in being wrong.

We have only two possible conclusions:

- If the $P\text{-value} \leq \alpha$, we reject the null hypothesis and support the alternative hypothesis.
- If the $P\text{-value} > \alpha$, we fail to reject the null hypothesis and cannot support the alternative hypothesis.
  - This does not make the null hypothesis true—we cannot prove the null hypothesis because sample data cannot reveal the true value of the population proportion.

Example 4

Research on college completion has shown that about 60% of students who begin college eventually graduate. A publication of higher education claims that the proportion for STEM (science, technology, engineering, math) majors is lower. Researchers randomly select 100 STEM majors and determine that 51 eventually graduate. Test the claim at a 5% level of significance.

Solution: Recall, the P-value is about 0.033. The level of significance is 5% which is 0.05 as a decimal. $0.033 \leq 0.05$ so we reject the null hypothesis and support the alternative hypothesis.

The sample data support the claim that the proportion of all STEM majors who eventually graduate is less than 60%.

You try!

A pro-life advocate believes that a majority of unplanned pregnancies resulted from no use of contraceptive methods. She randomly surveyed 125 people who had unplanned pregnancies and found that 64 did not use a contraceptive method in the month they became pregnant. Test the claim at a 5% level of significance.

Step 1. Determine the hypotheses

a. p represents:

b. H_0: _____

c. H_a: _____

d. Right-, left-, or two-tailed test? Explain how you know.

Step 2. Collect sample data

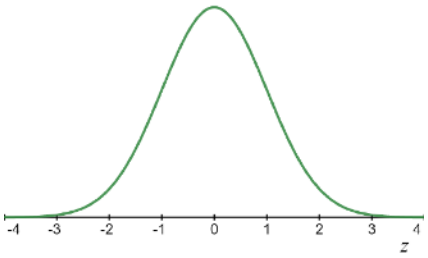
a. Explain why the sampling distribution of sample proportions is approximately normal.

b. Calculate the sample proportion, $\hat{p}$

Step 3 Assess the evidence

a. Compute the Z-score that corresponds to the sample proportion.

b. Locate the Z-score on the horizontal axis of the graph below. Shade the region that represents the P-value.



Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

c. Use the desmos calculator to compute the P-value.

$P(\hat{p} \geq \_\_\_\_\_) = P(z \geq \_\_\_\_\_) = \_\_\_\_\_$

Step 4. State a conclusion

a. What is the level of significance, α?

b. Compare the P-value and the level of significance.

c. Make a decision about the null and alternative hypotheses.

d. State a conclusion in context.

Reference

[9]Alyssa Brown, “Americans Favor Allowing Women in Combat,” Gallup.com, January 25, 2013, accessed May 18, 2022, https://news.gallup.com/poll/160124/americans-favor-allowing-women-combat.aspx

This page titled 6.4: Hypothesis Tests for a Single Population Proportion is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Hannah Seidler-Wright.

- **Current page** by Hannah Seidler-Wright is licensed CC BY-NC-SA 4.0.
- **1.2: The Statistical Analysis Process** by Hannah Seidler-Wright is licensed CC BY-NC-SA 4.0.

6.4.1: Exercises

1. Pearl wonders if the majority of US adults are dissatisfied with the quality of the environment. According to a poll conducted by Gallup[10] of 200 randomly surveyed US adults, 122 respondents were dissatisfied with the quality of the environment. Test Pearl's claim at a 5% level of significance.

   a. p represents the proportion of US adults who are:

   b. H_0:

   c. H_a:

   d. What test should you use to find the P-value? Justify your answer.

   e. Explain why the sampling distribution of sample proportions is approximately normal.

   f. What is the sample proportion, $\hat{p}$? Write your answer as a fraction and decimal.

   g. Compute the Z-score for the sample proportion.

   h. Use desmos to find the P-value from the standard normal distribution. Sketch a graph and shade the area that represents the P-value.

   i. Make a decision about the null and alternative hypotheses. Justify your answer.

   j. State the conclusion in context.

2. In a random sample of 300 Alzheimer's patients taking a new drug, 21 experienced nausea as a side effect. The drug manufacturer claims that fewer than 10% of patients who take its new drug for treating Alzheimer's disease will experience nausea. Test the claim at a 1% level of significance.

   a. Step 1

   b. Step 2

   c. Step 3

   d. Step 4

3. The proportion of smokers among persons who graduated from a four-year university has been widely reported as 22%. A sociologist student wonders if this is still true. They randomly sample 785 four-year university graduates and finds that 157 are smokers. They test the claim at a 5% level of significance. Spot the errors in the students solution below (there is at least one error in each step):

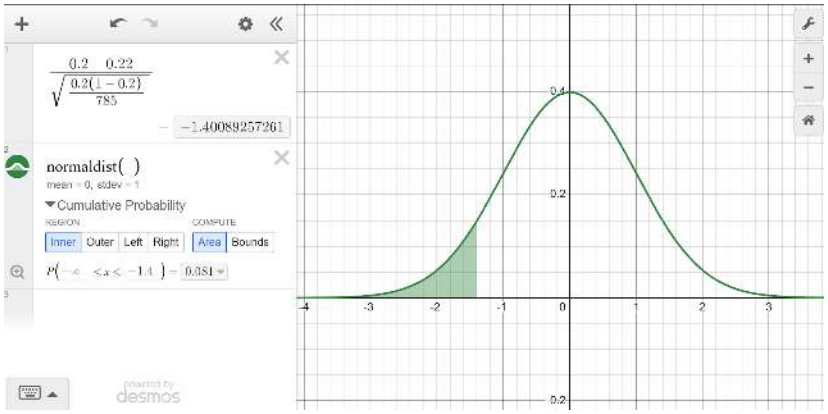
Step 1: p represents four-year university graduates who are smokers.

$H_0 : \hat{p} = 0.22$

$H_a : \hat{p} \neq 0.22$

Step 2: There are 157 successes in the sample and 785-157=628 failures in the sample. These are greater than 10 so it's normal. $\hat{p} = \frac{157}{785} = 0.2$.

Step 3: $Z = \frac{0.2-0.22}{\sqrt{\frac{0.2(1-0.2)}{785}}} \approx -1.40$



Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

The P-value is 0.081.

Step 4: The P-value 0.081 is greater than the level of significance 0.05. We accept the null hypothesis and reject the alternative hypothesis.

The sample data show that the four-year university graduates who are smokers is equal to 22%.

4. The proportion of smokers among persons who graduated from a four-year university has been widely reported as 22%. A sociologist student wonders if this is still true. They randomly sample 785 four-year university graduates and finds that 157 are smokers. Test their claim at a 5% level of significance. Clearly show each step of a hypothesis test.

Reference

[10]Jeffrey M. Jones, "Americans Offer Gloomy State of the Nation Report," *Gallup.com*, February 2, 2022, accessed September 27, 2022, https://news.gallup.com/poll/389309/americans-offer-gloomy-state-nation-report.aspx

This page titled 6.4.1: Exercises is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Hannah Seidler-Wright.

- **Current page** by Hannah Seidler-Wright is licensed CC BY-NC-SA 4.0.
- **1.2: The Statistical Analysis Process** by Hannah Seidler-Wright is licensed CC BY-NC-SA 4.0.

6.5: Conclusions (1)

When we estimate a population parameter or conduct a hypothesis test, our last step is to state a conclusion in context. In this section, we will focus on conclusions of a hypothesis test for a population proportion.

Example 1

A Gallup poll[11] indicates that LGBT (lesbian, gay, bisexual, or transgender) identification has increased significantly for Hispanic adults between 2020 and 2022. In 2020, 8.4% of Hispanic adults identified as LGBT. In a random sample of 500 Hispanic adults, 55 identified as LGBT. Test the claim at a 5% level of significance.

Solution:

Step 1. Let p represent the proportion of Hispanic adults that identify as LGBT in 2022.

$H_0 : p = 0.084$

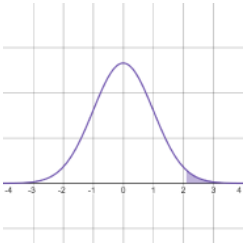
$H_a : p > 0.084$

Right-tailed test because the inequality in the alternative hypothesis is greater than.

Step 2. The number of expected successes in the sample is 500(0.084)=42. The number of expected failures in the sample is 500-42=458. Both are greater than or equal to 10 so the sampling distribution of sample proportions is approximately normal.

$\hat{p} = \frac{55}{500} = 0.11$

Step 3. $Z = \frac{0.11 - 0.084}{\sqrt{\frac{0.084 + (1 - 0.084)}{500}}} \approx 2.10$



P-value $= 0.018$

Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

Step 4. P-value $= 0.018 \leq 0.05 = 5\% = \alpha$

We reject the null hypothesis

We support the alternative hypothesis

There is evidence to support the claim that the proportion of all Hispanic adults who identify as LGBT has increased since 2020.

Errors in Hypothesis Tests

In the above example, we rejected the null hypothesis, meaning we were convinced by the data that the null hypothesis is false. This decision is based on sample data and sometimes, there is a small possibility that we made the wrong decision.

Even if we make no errors in the hypothesis testing process, our conclusion might still be wrong. At the conclusion of a test, we cannot know if we have made an error because we don't know what is actually true. This is why we conduct hypothesis tests, because it is difficult to know the true values of population parameters due to the large size of populations.

There are two possible conclusions to a hypothesis test, which means there are two possible errors:

Type I. We reject the null hypothesis in support of the alternative hypothesis.

- If we made the wrong decision here, what is actually true?

  We decide that the null hypothesis is _____, when actually, it is _____.

Type II. We fail to reject the null hypothesis in support of the alternative hypothesis.

- If we made the wrong decision here, what is actually true?

  We decide that the null hypothesis is _____, when actually, it is _____.

Notice that Type I and II errors are exclusively about the null hypothesis.

The Level of Significance

The level of significance, , is related to the likelihood of these errors. Consider cases where the null hypothesis is true. When α=5%, we reject a null hypothesis for sample data that occur less than 5% of the time by random chance. Thus, when the null hypothesis is true, we reject it 5% of the time, committing a Type I error. Note, this does not mean that a Type I error occurs in 5% of all hypothesis tests because a Type 1 error can only occur if the null hypothesis is true. This probability is conditional in nature.

The level of significance is the probability of rejecting the null hypothesis given that the null hypothesis is true.

$$\alpha = P(\text{ reject } H_0 \mid H_0 \text{ is true })$$

Lower significance levels indicate that you require stronger evidence to reject the null hypothesis.

You try!

1. In example 1, p represented the proportion of Hispanic adults that identify as LGBT in 2022. The null hypothesis was that the proportion of Hispanic adults who identify as LGBT in 2022 is 8.4% ($H_0 : p = 0.084$).

   a. Based on the conclusion of example 1, what type of error might have occurred? Explain.

   b. Describe Type I error in context.

   c. Describe Type II error in context.

Example 2

According to a Gallup poll[12], in 2021, 77% of U.S. adults were generally dissatisfied with the total cost of healthcare in the country. A researcher wants to know if the dissatisfaction rate is different this year. They randomly sampled 1200 U.S. adults and find that the sample proportion was $\hat{p} = 0.78$, so 78% of the adults in the sample were dissatisfied with the total cost of healthcare in the US. This proportion is slightly different from 77%, but with a P-value of around 0.21 (from the Z-score of Z=0.82), the difference is not *statistically significant*.

2. The null hypothesis in this example is $H_0 : p = 0.77$, and the alternative hypothesis is $H_a : p \neq 0.77$.

   a. Based on the conclusion of example 2, what type of error might have occurred? Explain. Describe the error in context.

b. What if we had observed the same difference ($\hat{p} = 0.78$ and $p = 0.77$) from a sample of 10,000 U.S. adults? Compute the Z-score for the revised sample size.

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} = \frac{\_\_\_\_ - \_\_\_\_}{\sqrt{\frac{\_\_\_\_(1-\_\_\_\_)}{\_\_\_\_}}}$$

c. Use https://www.desmos.com/calculator to find the P-value for this two-tailed test.

d. Is the sample proportion statistically significant with this new sample size (using a 5% level of significance)?

e. In one scenario above, the difference between the sample statistic and the population parameter is statistically significant, but in another, it is not. Statistical significance is different from the sort of significance we assign to events through our own experiences and values. Do you feel that the difference between the proportion of U.S. adults in 2021 and the proportion of U.S. adults in the 2022 sample is significant in a real-world or practical sense? Explain.

Statistical and Practical Significance

With very large samples, even extremely small differences can be statistically significant. With this in mind, remember that *statistical significance* is different from *practical significance*. Statistical significance is measured with probability. Practical significance is measured through our own system of personal values and is therefore difficult to measure.

Reference

[11]Values based on summary statistics given in Jeffrey M. Jones, "Growing LGBT ID Seen Across Major U.S. Racial, Ethnic Groups," *Gallup.com*, June 8, 2022, accessed June 21, 2022, https://news.gallup.com/poll/393464/growing-lgbt-seen-across-major-racial-ethnic-groups.aspx

[12]Values based on summary statistics given in *Gallup.com*, accessed June 21, 2022, https://news.gallup.com/poll/4708/Healthcare-System.aspx

This page titled 6.5: Conclusions (1) is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Hannah Seidler-Wright.

- **Current page** by Hannah Seidler-Wright is licensed CC BY-NC-SA 4.0.
- **1.2: The Statistical Analysis Process** by Hannah Seidler-Wright is licensed CC BY-NC-SA 4.0.

6.5.1: Exercises

1. The Department of Motor Vehicles for a large city claimed that 80% of candidates pass driving tests. A local newspaper randomly surveyed 90 teens who had taken the test. They found that only 61 of the teens passed (68%). The researcher at the newspaper conducted a hypothesis test to determine if the passing rate for teens is lower than the proportion the DMV reported. He found a Z-score of -2.85 which corresponded to a P-value of 0.0022.
   a. Write a conclusion to the researcher's hypothesis test. Be sure to make the conclusion easy for the readers of the newspaper to understand.

   b. Based on the conclusion, what type of error may have occurred? Write the error in context.

2. In 2021-2022, 12% of employed people in the US reported belonging to a labor union. Officials in a large city randomly surveyed a sample of 120 and found that 18 reported belonging to a labor union.
   a. State the null hypothesis.

   b. Officials conduct a hypothesis test (at a 5% level of significance) to determine if the union membership rate in the city is different from the national rate. They determine that the P-value is 0.3125. State the conclusion in context.

   c. Based on the conclusion in b, what type of error may have occurred? Write the error in context.

d. A private tech company decides to research labor membership in the large metropolitan area where the company is based. They use their vast resources to conduct a survey of 1200 employed people and get a sample proportion of 0.15. What will happen to the standard error when the sample size increases?

e. The private tech company finds a P-value of 0.0014. They reject the null hypothesis in support of the alternative hypothesis. They conclude that the sample proportion is statistically significant. Do you believe that there is practical significance? Show the difference between the sample proportion (15%) and the assumed population proportion (12%) in your justification.

This page titled 6.5.1: Exercises is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Hannah Seidler-Wright.

CHAPTER OVERVIEW

7: Inference Involving a Single Population Mean

7.1: The Sampling Distribution of Sample Means

7.1.1: Exercises

7.2: The Student’s T-Distribution

7.2.1: Exercises

7.3: Estimating a Population Mean

7.3.1: Exercises

7.4: Hypothesis Tests for a Single Population Mean

7.4.1: Exercises

7.5: Conclusions (2)

7.5.1: Exercises

This page titled 7: Inference Involving a Single Population Mean is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Hannah Seidler-Wright.

7.1: The Sampling Distribution of Sample Means

In the last unit, we used sample proportions to make estimates and test claims about population proportions. In this unit, we will focus on sample means from random samples. We begin by examining the variability of sample means that are randomly obtained from a population. We will use these ideas to learn about sampling distributions of sample means.

Exploring Samples of Acorn Weights

In Texas, oak trees are important to the overall health of the ecosystem. Acorns are the seeds of oak trees. Many birds and small mammals eat the acorns that come from oak trees. A decline in the growth of new oak trees could have a serious impact on local animal and plant life.

Botanists (scientists who study plant life) are especially interested in knowing the weights of acorns in a region. Botanists use information about the weights of acorns to help them predict the future growth of oak trees.

A group of students in Austin, Texas gathered acorns to study their weights. The weights of 400 acorns were measured in grams, and ranged from 0.7g to 6.7g.

Collect the Sample

Here is a list of 9 randomly selected acorn weights (in grams): $W = [4.2, 3.2, 3.9, 4.6, 3.5, 3.9, 4.6, 3.3, 2.6]$. In this sample, we can use desmos to compute the sample mean, $\bar{x}$, by using the $\text{mean}(W)$ function after entering the set labeled W. The sample mean weight for this sample is around 3.67 grams.

Given are 5 samples of 9 acorn weights. Pick one of the samples to compute the mean acorn weight for. Round the sample mean to two decimal places.

Sample 1. [2.5, 3.8, 4, 3.4, 4.5, 5.7, 1, 4.4, 2.6]

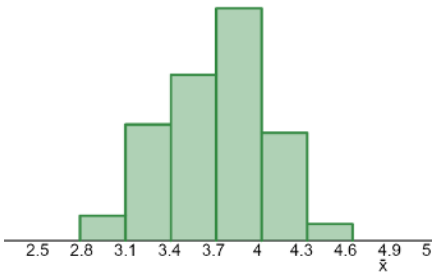
Sample 2. [3.4, 2.4, 3.2, 2, 1.4, 3.9, 0.7, 4.6, 3.8]

Sample 3. [3.8, 2.1, 1.4, 3.9, 4.6, 4.4, 3.4, 4.3, 3.9]

Sample 4. [2.8, 3.2, 3.3, 3.5, 4.8, 3.7, 1, 4.3, 6.6]

Sample 5. [5.7, 3.3, 0.7, 2.5, 4.6, 2.9, 2.6, 4.9, 2.9]

Below is a histogram of mean acorn weight from 80 random samples. Use this distribution of sample means to complete the following questions:



Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

1. What is the largest average acorn weight?

2. What is the range of sample means?

3. What range of sample means occurred most frequently?

4. Describe the center, shape, and spread of the distribution of sample means.

5. Sample means are estimates of the population mean, so a distribution of sample means is a distribution of estimates. Any deviation that exists between a sample mean and the population mean is an error. This is why we call the standard deviation of all sample means the standard error. Estimate the standard error of the distribution of sample means.

Central Limit Theorem for Sample Means

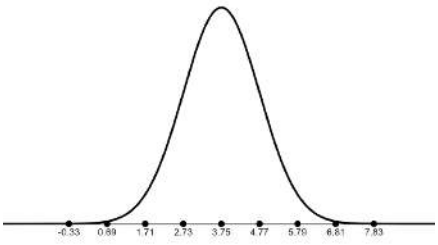
We will now shift our attention from distributions of sample means to the sampling distribution of sample means. The sampling distribution of sample means is the distribution of all possible sample means from random samples of the same size. The Central Limit Theorem for Sample Means states that: Given any population with mean μ and standard deviation σ, the sampling distribution of sample means (sampled with replacement) from random samples of size n will have a distribution that approaches normality with increasing sample size. The mean and standard error of the sampling distribution are:

$$\mu_{\bar{x}} = \mu \text{ and } \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

The criteria for the approximate normality of a sampling distribution are that either the population from which we are sampling is normal, or the sample size is greater than 30. Very non-normal populations may require samples substantially larger than 30. When a sampling distribution of sample means is approximately normal, we can use its mean and standard error to find the Z-score of any particular sample mean. The Z-score of a sample mean $\bar{x}$ from a sample of size n is found by the formula:

$$Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

We explored a population distribution of acorn weights from oak trees. Since we define these 400 acorns as the population, we can calculate the population mean and standard deviation. The population of acorn weights are normally distributed with a mean weight of 3.75 grams (μ) and a standard deviation of 1.02 grams (σ). The population of acorn weights is given as a desmos graph below.



Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

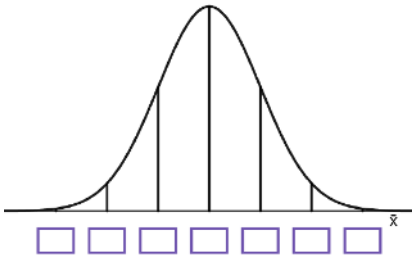
Suppose that we randomly select samples of 40 acorn weights. According to the Central Limit Theorem, since the population is normally distributed, sample means from random samples of size 40 will also be normally distributed.

6. The sample size is $n = 40$. Use the Central Limit Theorem to calculate the mean and standard error of the sampling distribution. Round the standard error to three decimal places.

   a. $\mu_{\bar{x}} = \mu =$ _____

   b. $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} =$ _____

7. We can use a normal curve to represent the sampling distribution of sample means. The boxes under the normal distribution below are one standard error apart, with the center box at the mean. Use the mean and standard error above to enter the correct values into the boxes. Use the standard error, rounded to 3 decimal places.



Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

a. A particular random sample of size 40 has a mean weight of 4 grams. Plot this sample mean on the axis above. Find the Z-score for the sample mean, $\bar{x} = 4$. Round the Z-score to two decimal places.

b. Would you consider this sample mean to be unusual? Explain your answer.

c. If a sample of 40 acorns from this population is randomly selected, what is the probability that its mean weight $\bar{x}$ will be greater than 4 grams? Write your answer as a decimal rounded to three places. Write the desmos function you used to calculate the probability.

Check for Understanding

8. Everyday people watch 1 billion hours of videos on YouTube. That breaks down to every single person on earth watching YouTube videos for about 8.4 minutes per day. For U.S. teens, on any given day, the amount of time spent watching YouTube videos is approximately normal with mean 18.5 minutes and standard deviation 5.3 minutes.

   a. Find the probability that a randomly chosen U.S. teen watches YouTube for more than 25 minutes in a given day. Round to two decimal places. Show your thinking using a sketch and use desmos to do the calculation. Write any functions used in desmos to find your answer.

   b. Suppose we choose a simple random sample of 10 U.S. teens. Let $\bar{x}$ = the mean amount of time spent watching YouTube videos for the sample.

   c. Why is the sampling distribution of $\bar{x}$ normally distributed?

   d. What is the mean of the *sampling distribution* of $\bar{x}$?

   $\mu_{\bar{x}} =$

   e. Calculate the standard deviation of the sampling distribution of x. Round to three decimal places.

   $\sigma_{\bar{x}} =$

   f. Find the probability that the mean amount of time spent watching YouTube for the teens in the sample exceeds 25 minutes. Round to five decimal places. Write the desmos function used to do the calculation.

   $P(\bar{x} \quad 25) = P(Z > \_\_\_) =$

This page titled 7.1: The Sampling Distribution of Sample Means is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Hannah Seidler-Wright.

- **Current page** by Hannah Seidler-Wright is licensed CC BY-NC-SA 4.0.
- **1.2: The Statistical Analysis Process** by Hannah Seidler-Wright is licensed CC BY-NC-SA 4.0.

7.1.1: Exercises

1. In 2022, the average student loan debt for federal loans was $37,358. A student at UCLA wants to know if the average student loan debt for federal loans at his institution is more than $37,358. He randomly surveyed 120 UCLA students with federal loan debt and found the average was $39,532 with a standard deviation of $8,067.

    a. In this study, what is the sample?

    b. In this study, what is the sample statistic?

    c. In this study, what is the population?

    d. In this study, what is the population parameter that the sample statistic is estimating?

2. Go to the website: http://www.rossmanchance.com/applets/OneSample.html?population=model

    a. By selecting “Show Sampling Options,” you can simulate random sampling of the population. The sample means of these simulated samples will be displayed in the “Statistics” plot. Follow the directions below to use the simulation:

        - Select “Show Sampling Options”
        - Set the Number of Samples to 1000
        - Set the Sample size (n) to 10
        - Click “Draw Samples” to plot the distribution of sample means in the plot on the right. The middle plot displays the most recent sample created by the simulation.
            - Write the mean, standard deviation, and shape of the distribution of sample means (the plot on the right).

- Set the Sample size (n) to 20
- Click “Draw Samples” to plot the distribution of sample means in the plot on the right. The middle plot displays the most recent sample created by the simulation.
    - Write the mean, standard deviation, and shape of the distribution of sample means (the plot on the right).

- Set the Sample size (n) to 30
- Click “Draw Samples” to plot the distribution of sample means in the plot on the right. The middle plot displays the most recent sample created by the simulation.
    - Write the mean, standard deviation, and shape of the distribution of sample means (the plot on the right).

    - As the sample size increases, how does the mean, standard deviation (called the standard error when working with a sampling distribution), and shape of the distribution of sample means change?

b. Now we will analyze the sampling distribution of sample means when samples are taken from a **skewed population**:

- Change the drop-down next to “Population Shape” from “Normal” to “Skewed Right” and click the “Set Population” button below
- Set the Sample size (n) to 10
- Click “Draw Samples” to plot the distribution of sample means in the plot on the right. The middle plot displays the most recent sample created by the simulation.
    - Write the mean, standard deviation, and shape of the distribution of sample means (the plot on the right).

- Set the Sample size (n) to 20
- Click “Draw Samples” to plot the distribution of sample means in the plot on the right. The middle plot displays the most recent sample created by the simulation.
    - Write the mean, standard deviation, and shape of the distribution of sample means (the plot on the right).

- Set the Sample size (n) to 40
- Click “Draw Samples” to plot the distribution of sample means in the plot on the right. The middle plot displays the most recent sample created by the simulation.
    - Write the mean, standard deviation, and shape of the distribution of sample means (the plot on the right).

- As the sample size increases, how does the mean, standard deviation (called the standard error when working with a sampling distribution), and shape of the distribution of sample means change?

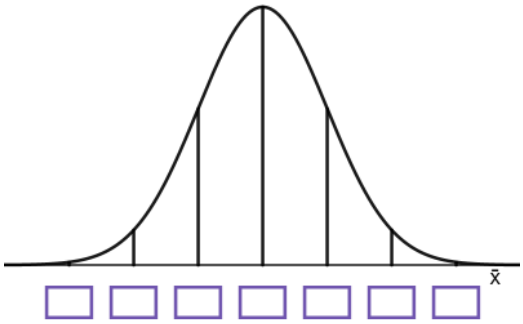
3. The average time it takes corn to germinate is 8 days with a standard deviation of 2.5 days. Germination depends on soil temperature and moisture conditions. A farmer has a variety of corn plots and wants to know about the quality of the soil by measuring the average germination for various plots. She plants 40 corn seeds in each plot.

   a. Find the mean and standard error (rounded to three decimal places) of the sampling distribution of sample means.

   b. A sampling distribution is a description of all possible values of a statistic. What does this sampling distribution represent?

   c. Is this sampling distribution approximately normal? Explain why or why not.

d. The boxes under the normal distribution below are one standard error apart, with the center box located under the mean. Use the mean and standard error you calculated in a. to label the horizontal axis.



Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

e. Use the empirical rule to find the interval centered at that contains approximately 95% of all sample means.

f. What sample means would you consider unusual?

g. In one plot of corn, the average germination time was 9 days. Calculate the Z-score for this sample mean. Round to two decimal places.

h. The farmer can conclude that the soil needs improvement if germination takes too long. What is the probability of observing a plot in which seeds take an average of 9 or more days to germinate?

This page titled 7.1.1: Exercises is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Hannah Seidler-Wright.

- **Current page** by Hannah Seidler-Wright is licensed CC BY-NC-SA 4.0.
- **1.2: The Statistical Analysis Process** by Hannah Seidler-Wright is licensed CC BY-NC-SA 4.0.

7.2: The Student's T-Distribution

In a previous lesson, we saw that under certain conditions, the sampling distribution of sample means is normally distributed. We can use this information to standardize the sampling distribution of sample means and find probabilities and Z-scores from the standard normal distribution. We can do this if we know the population standard deviation. However, it is not common to know a population standard deviation. What happens when we don't know the value of the population standard deviation?

When the Population Standard Deviation, σ, is Unknown

In most cases, we do not know the population standard deviation, σ. The only option we have is to approximate with a known sample standard deviation, s.

1. Review: What conditions should be met to verify that the sampling distribution of sample means is approximately normal?

Recall, the mean of the sampling distribution of sample means is $\mu_{\bar{x}} = \mu$. The standard error of the sampling distribution of sample means is $\mu_{\bar{x}} = \mu$. We substitute s for in the standard error formula since is unknown. So, the standard error of the sampling distribution of sample means is estimated by

$$\frac{s}{\sqrt{n}}$$

The test statistic for a sample mean is $\frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$. The sample standard deviation, s, varies from sample to sample and it only approximates σ. Therefore, this substitution introduces additional variability into the test statistic. This added variability means that the distribution of the test statistic is no longer normal. We instead use the Student's T-distribution.

The T-Distribution

The **T-distribution** describes the variability of the test statistic, $T = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$, when the sampling distribution of sample means is normal, and the sample standard deviation, s, is used to estimate an unknown population standard deviation, σ.

This test statistic is an estimate of how many standard errors the sample mean is from the hypothesized population mean.

A T-distribution is a member of a family of continuous probability distributions. The width of a T-distribution depends on how much a sample standard deviation can vary. The amount of variability in a sample standard deviation depends on how many

deviations *vary freely* when it is computed. Recall, the sample standard deviation is calculated using the formula

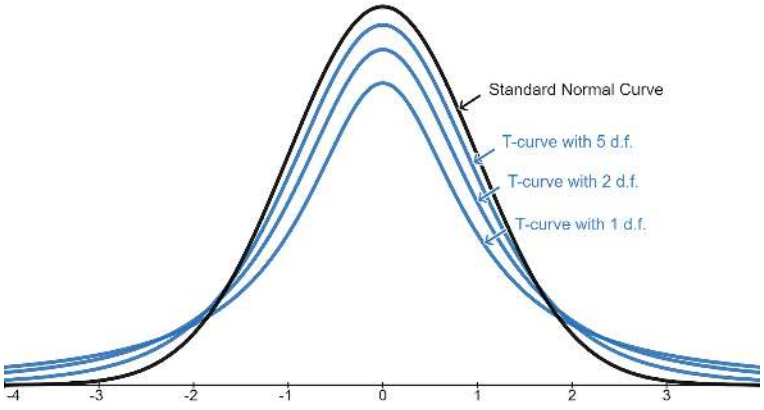
$$s = \sqrt{\frac{\sum_i (x_i - \bar{x})^2}{n-1}}$$

The deviations from the mean are averaged in a sample standard deviation. When the deviations are added together, they always add to zero. Because of this, the last deviation summed in this average is not free – it is always the value that makes the resulting sum zero. There are n deviations from the mean, but only $n-1$ of these are *free* deviations.

The variability of standard deviations depends on the number of free deviations in the sample standard deviations, $n-1$. This quantity is known as the **degrees of freedom (d.f.)**. Each degree of freedom defines a uniquely associated T-distribution.

T-distributions have the following characteristics:

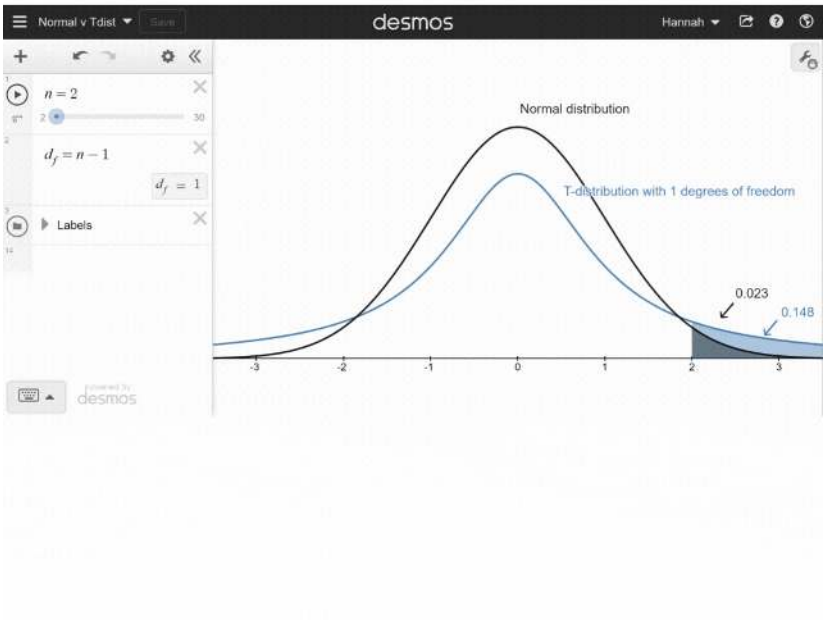
- T-distributions are bell-shaped and symmetric with a mean of 0.
- Each T-distribution depends on the degrees of freedom, d.f. ($d.f. = n-1$).
- T-distributions have heavier tails and narrower peaks than the standard normal distribution.
- The total area under each T-distribution curve is 1.
- As the degrees of freedom increase, the tails become thinner and the curve approaches the standard normal distribution.



Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

With fewer degrees of freedom, the more the sample standard deviation varies. In other words, a smaller sample size corresponds to

greater variability in a T-distribution. In a small sample, it is more likely to observe extreme values. This is reflected in the shape of the T-distribution. As the sample size increases, the T-distribution trends toward the normal distribution.



Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

Check for Understanding

2. The heights of students at Claremont High School is approximately normal. If the average height of a student here is 166 centimeters.

   a. Is the sampling distribution of sample means better represented by the normal distribution or the T-distribution? Justify your answer.

   b. You find the average height and sample standard deviation for 10 randomly selected Claremont High School students. The sample standard deviation is 4 cm. Compute the test statistic for an average height of 168 cm. Round your answer to two decimal places.

   c. We will now use desmos to find the likelihood of seeing an average height or 168 cm or more.

- Go to https://www.desmos.com/calculator
- The degrees of freedom d.f. in this example is $n - 1 = 10 - 1 = 9$. Type tdist(9). In general, type tdist(degrees of freedom).
- Click the "Zoom Fit" button.
- Check the "Find Cumulative Probability (CDF)" box.
- The min and max will default to $-\infty$ and ∞ respectively. Enter the test statistics from b. into the min field. The probability $P(\bar{x} \geq 168) = P(T \geq \_\_\_) \approx 0.0743$.

d. You find the average height and sample standard deviation for 10 randomly selected Claremont High School students. The sample standard deviation is 4 cm. Use the process in c. to compute the probability of observing an average height of 170 cm or more.

e. If the average height of a student at Claremont High School is 166 centimeters, which of the following is more likely:

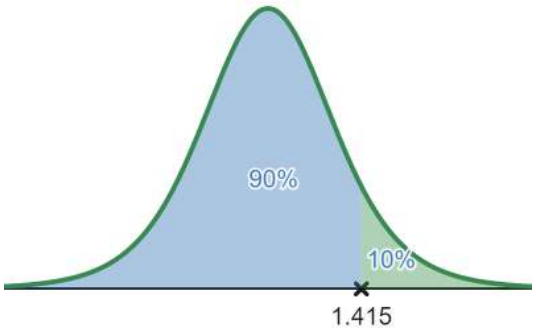
- a random sample of 15 students, with the average height being greater than 170 centimeters? OR
- a random sample of 25 students, with the average height being greater than 170 centimeters?

Justify your answer using what you know about the T-distribution.

Finding a Critical Value from a T-Distribution

3. We will use desmos to find the T critical value that separates the top 10% from the lower 90% from a random sample of size 8.

   a. Go to https://www.desmos.com/calculator
   b. The degrees of freedom $d.f. = n - 1 = 8 - 1 =$_____. The area to the left of the critical value is _____ (90%). Type tdist(7).inversecdf(0.9).
      - In general, you can type tdist(d.f.).inversecdf(L) to find a T critical value where d.f. is the degrees of freedom, and L is the area left of the critical value.



The T critical value here is around 1.415

Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

4. Use desmos to find the T critical value that separates the middle 95% from a random sample of size 8. Show your thinking by sketching a graph.

This page titled 7.2: The Student's T-Distribution is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Hannah Seidler-Wright.

- **Current page** by Hannah Seidler-Wright is licensed CC BY-NC-SA 4.0.
- **1.2: The Statistical Analysis Process** by Hannah Seidler-Wright is licensed CC BY-NC-SA 4.0.

7.2.1: Exercises

1. The length of time students needed in order to complete a criminal justice test follows a distribution that is approximately normal. The mean length of time students take to complete the criminal justice test is 68 minutes. Barbara finds a random sample of 10 criminal justice students and records their test taking times: A= [73.1, 70.4, 68.7, 72.9, 76.7, 78.2, 73.2, 70.2, 64.8, 66.7]

   a. Compute the sample mean and sample standard deviation (rounded to three decimal places) using desmos:

   $\bar{x} = mean(A) =$ _____ $minutes$

   $s = stdev(A) =$ _____ $minutes$

   b. Is the sampling distribution of sample means approximately normal? Explain why or why not.

   c. Which distribution is more appropriate to use to compute probabilities about the sample mean? Justify your selection. If the T-distribution is more appropriate, compute the degrees of freedom.

      i. The standard normal distribution

      ii. The T-distribution

   d. Compute the test statistic for the sample mean of A. Round to two decimal places.

   e. Is the sample mean unusual? Explain why or why not.

   f. Find the probability of observing a sample mean this high or higher. Round to four decimal places.

   g. A criminal justice teacher wants their class to be in the fastest 5% of students' mean length of time to take the criminal justice test. Find the maximum length of time the classes average can be.

2. Franklin wants to know the average female baby birth length in his city. He randomly selects 65 female babies and records their birth length. He finds that the sample mean is 18.8 inches and the sample standard deviation is 0.063 inches.

   a. Is the sampling distribution of sample means approximately normal? Explain why or why not.

   b. Which distribution is more appropriate to use to compute probabilities about the sample mean? Justify your selection. If the T-distribution is more appropriate, compute the degrees of freedom.

      i. The standard normal distribution

      ii. The T-distribution

   c. What is the range of average lengths of unusually short female babies?

   d. What are the average lengths of female babies that correspond to the middle 95%?

This page titled 7.2.1: Exercises is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Hannah Seidler-Wright.

7.3: Estimating a Population Mean

In this section, we will be constructing confidence intervals for population means. You will find that the process is similar to estimating a population proportion, however, we have different criteria, formulae, and frequently use different distributions.

Confidence Interval for a Population Mean

We will estimate the mean number of landfills in states with high poverty rates. We will use the T-distribution to compute margins of error and confidence intervals for a population mean.

Recall the five step process of building a confidence interval:

Step 1. Verify that the sampling distribution is approximately normal.

Step 2. Compute the critical value.

Step 3. Compute the margin of error $E = \text{ critical value } \cdot \text{ standard error }$.

Step 4. Compute the bounds of the interval and write the interval in interval notation.

Step 5. Interpret the interval in context/state a conclusion in context.

Step 1: Verify that the Sampling Distribution of Sample Means is Approximately Normal

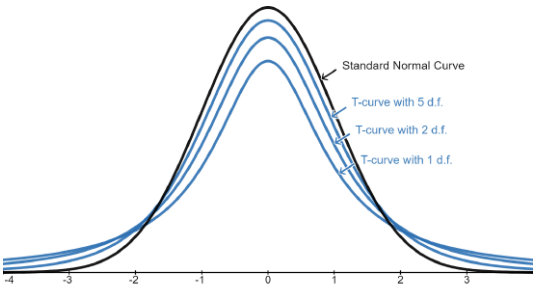
Recall, the normality conditions for the sampling distribution of sample means are **either:**

- The sample size is greater than 30 **OR**
- The sample is from a normally distributed population.

Step 2: Find the Critical Value

If the population standard deviation (σ) is *known*, the critical value comes from the normal distribution. If the population standard deviation (σ)

is *unknown*, the critical value comes from the student's T distribution which varies based on sample size. This is usually the case.



Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

Step 3: Compute the Margin of Error

The margin of error is computed using the equation

$$E = \text{ critical value } \cdot \text{ standard error}$$

Since we often use the sample standard deviation, s, to estimate the unknown population standard deviation, , the estimated margin of error is found using the equation

$$E \approx T_c \cdot \frac{s}{\sqrt{n}}$$

Step 4: Compute the Confidence Interval

Now, we are estimating a population *mean* rather than a population *proportion*. Therefore, the point estimate we use is the sample mean, $\bar{x}$. The confidence interval is

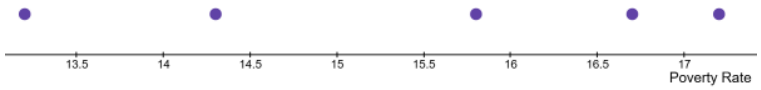
$$(\bar{x} - E, \bar{x} + E)$$

Step 5: Interpret the Interval and State a Conclusion in Context

Arguably the most important step, the five step process always ends with writing the conclusion in plain language. The conclusion should include the confidence level, the parameter that is being estimated, and the bounds of the interval with appropriate units.

Example

A researcher wants to know the average poverty rate in California. They randomly select 5 counties and find that the average poverty rate is 15.44% with a standard deviation of 1.668%. Researchers have determined that the population of poverty rates by county in California is approximately normal. The dotplot of the sample of 5 counties' poverty rates is given below.



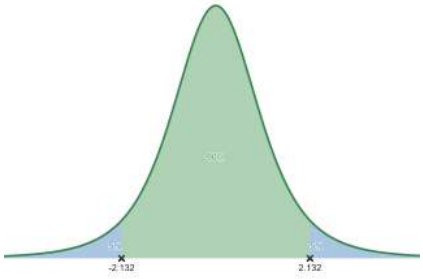
Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

We will construct a 90% confidence interval for the true mean number of landfills in poor counties of California.

Step 1. Although the sample size is less than ____, the problem states that the population of poverty rates by county in California is approximately normal. Therefore, the sample is from a normal population. It follows that the sampling distribution of sample _____ is approximately normal.

Step 2. Since the population standard deviation is unknown, we will find the critical value from the T-distribution. The T-distribution depends on the degrees of freedom. Here, the sample size is 5, so there are _____ degrees of freedom. We will need to know the area left of the critical value. Since the confidence level is 90%=_____, there is 5% or 0.05 in each tail. Therefore, the area left of the critical value is _____.

- In https://www.desmos.com/calculator, we enter tdist(4).inversecdf(0.95) to find the critical value. So the critical value $T_c = 2.132$.

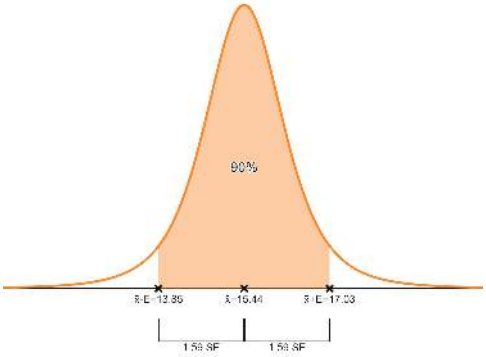


Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

Step 3. To compute the margin of error, we use $E \approx T_c \cdot \frac{s}{\sqrt{n}} = 2.132 \cdot \frac{}{\sqrt{5}} \approx 1.59\%$.

Step 4. The interval is

$$(\bar{x} - E, \bar{x} + E) = (15.44\% - 1.59\%, 15.44\% + 1.59\%) = (13.85\%, 17.03\%)$$



Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

Step 5. We are _____ confident that the _____ poverty rate for all counties in California is between _____% and _____%.

How the Margin of Error Changes

The following questions and animations will help you think about the relationship between sample size, the confidence level, the standard deviation and the margin of error.

1. Fill in the blanks with one of the following: *increases*, *decreases*, or *stays the same* where $E = (\text{ critical value }) \cdot \frac{s}{\sqrt{n}}$. (If you are accessing on a hard copy, go to https://www.desmos.com/calculator/rxiaadmqrm and adjust the indicated variable).

   a. As the sample size (n) increases, the margin of error (E) _____

Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

   b. As the confidence level increases, the margin of error (E) _____

Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

c. As the sample standard deviation (s) increases, the margin of error (E) _____

Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

You try!

2. You work for a consumer advocate agency and want to find the mean repair cost of a washing machine. As part of your study, you randomly select 40 repair costs and find the mean to be $100.00 and the standard deviation $17.50. Construct a 95% confidence interval for the mean repair cost of a washing machine.

Step 1.

Step 2.

Step 3.

Step 4.

Step 5.

This page titled 7.3: Estimating a Population Mean is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Hannah Seidler-Wright.

- **Current page** by Hannah Seidler-Wright is licensed CC BY-NC-SA 4.0.
- **1.2: The Statistical Analysis Process** by Hannah Seidler-Wright is licensed CC BY-NC-SA 4.0.

7.3.1: Exercises

1. Iris wants to know the average number of days for symptoms to develop in people with COVID-19. She randomly samples 35 people with COVID-19, and records the number of days it takes for each subject to develop symptoms. Use the data to answer the following questions to build a 90% confidence interval.

5	5	5	8	3	3	6
2	6	3	7	7	5	8
6	5	6	6	3	5	2
4	2	7	7	4	5	4
8	3	8	2	4	7	7

a. Is the sampling distribution of sample means approximately normal? Explain why or why not.

b. Compute all sample statistics $(\bar{x}, s, n, df)$ and the critical value that corresponds to the confidence level (rounded to three decimal places).

$\bar{x} = \text{mean}(a) \approx$ _____ days

$s = \text{stdev}(a) \approx$ _____ days

$n =$ _____

$df = n - 1 =$ _____ $=$ _____

$T_c = \text{tdist}($_____$).\text{inversecdf}($_____$) \approx$ _____

c. Compute the margin of error $E = T_c \cdot \frac{s}{\sqrt{n}}$ (rounded to three decimal places).

d. Write the interval in interval notation $(\bar{x} - E, \bar{x} + E)$.

e. Interpret the interval in context.

2. The length of time students needed in order to complete a criminal justice test follows a distribution that is approximately normal. Cory finds a random sample of 10 criminal justice students and records their test taking times in minutes: A= [73.1, 70.4, 68.7, 72.9, 76.7, 78.2, 73.2, 70.2, 64.8, 66.7]. Construct a 95% confidence interval for the mean test taking time on the criminal justice test.

a. Step 1

b. Step 2

c. Step 3

d. Step 4

e. Step 5

3. Olu wants to know the average female baby birth length in his city. He randomly selects 65 female babies and records their birth length. He finds that the sample mean is 18.8 inches and the sample standard deviation is 0.063 inches. Construct a 99% confidence interval for the mean female baby birth length in Olu's city.

4. Mars randomly sampled 42 college men and found a mean pulse rate of 70.42 beats per minute with a standard deviation of 9.948 beats per minute. Mars constructs a 95% confidence interval for the mean pulse rate for college men. Their solution is shown below. For each step, identify the error that they made and explain how to correct it and improve their solution.

   Step 1: $(0.95^{*}42) = 39.9 \geq 10$ so the sampling distribution is normal.

   Step 2: $T_c = \text{tdist}(42).\text{inversecdf}(0.95) \approx\ 1.682,\ \bar{x} = 70.42\ \text{bpm}, s = 9.948\ \text{bpm}, n = 42$

   Step 3: $E = 1.682 \cdot \frac{9.948}{\sqrt{42}} = 1.682 \cdot 1.54 = 2.59\ \text{bpm}$

   Step 4: $70.42 - 2.59 < \mu < 70.42 + 2.59,\ \text{so}\ 67.83 < \mu < 73.01.$

   Step 5: The true pulse rate of college men is between 67.83 and 73.01.

This page titled 7.3.1: Exercises is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Hannah Seidler-Wright.

7.4: Hypothesis Tests for a Single Population Mean

In previous lessons, we have learned that there are two fundamental forms of statistical inference: confidence intervals and hypothesis tests. In the last section, we used confidence intervals to estimate a single population mean. We will now apply the four step hypothesis testing process to test hypotheses about the value of a single population mean. To explore this concept, we will examine examples of water contamination across the US.

The Flint Water Crisis

The water crisis in Flint, Michigan, is an example of environmental injustice that took place beginning in 2014. The city decided to switch its drinking water supply from Detroit's system to the water from the Flint River to save money. This switch was made despite inadequate treatment and testing of the water from the river. As a result, the community was deeply impacted. Their water supply turned yellow and began to smell and taste of sewage. Members of the community suffered from skin rashes and hair loss. Later, independent research would reveal that the contaminated water was contributing to dangerous increases of blood lead levels in the city's youth, and the lead level in the water greatly exceeded the EPAs standard.

Eventually, in 2016, with the efforts of community activists and scientists, a federal judge sided with Flint residents and ordered clean bottled water to be delivered to citizens. The following year, the city was ordered to replace the city's lead pipes with funding from the state and the community was provided with resources to support their health and well-being. However, the fight to bring safe access to the water supply is ongoing.



Read more about the water crisis in Flint, Michigan: https://www.nrdc.org/stories/flint-water-crisis-everything-you-need-know

Use Statistics to Defend Environmental Justice

To test whether the water was safe, the scientists likely conducted hypothesis tests. Now you will conduct a hypothesis test, like those scientists who fought for the citizens in Flint.

1. A mean lead level of 15 parts per billion (ppb) is considered safe, while anything above 15 ppb is considered dangerous. You believe the Flint water is unsafe, so you take a random sample of 200 households. You find that the average lead level of the sample is 16.5 ppb, with a standard deviation of 8 ppb. Do we have reason to believe that the true mean lead level for the entire water supply is dangerous, at a significance level of 1%?

Step 1 Determine the hypotheses

Let μ represent the _____ lead level for the entire water supply.

$H_0 : \mu =$ _____ ppb

Select the appropriate alternative hypothesis:

a. $H_a : \mu < 15$ ppb

b. $H_a : \mu > 15$ ppb

c. $H_a : \mu \neq 15$ ppb

Select the appropriate test, and explain why:

a. Left-tailed

b. Right-tailed

c. Two-tailed

Step 2 Collect sample data

The sample mean, $\bar{x}$, is _____ ppb.

The sample standard deviation, s, is _____ ppb.

The sample size, n, is _____ households.

There are _____ degrees of freedom.

Explain why the conditions of the Central Limit Theorem are met:

Step 3 Assess the evidence

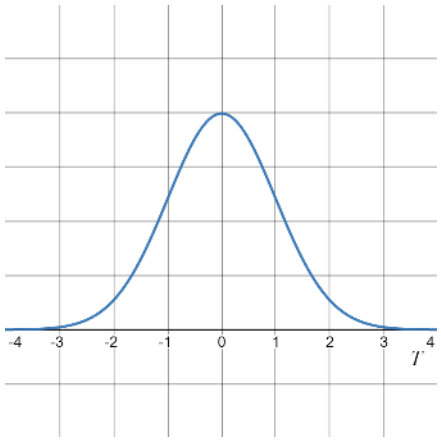
Which distribution will we use to find the P-value?

a. The normal distribution because the population standard deviation is given.
b. The student's T-distribution because the population standard deviation is unknown.

The test statistic, rounded to three decimal places, is

$$T = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} = \frac{\_\_\_\_\_ - \_\_\_\_\_}{\frac{\_\_\_\_\_}{\sqrt{\_\_\_\_\_}}} = \_\_\_\_\_$$

Below is the graph of the T-distribution with _____ degrees of freedom. Label the T-statistic on the horizontal axis. Then shade the region that represents the P-value.



Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

Use https://www.desmos.com/calculator to find the P-value:

1. Enter tdist(199) in the first line.
2. Check the box for Find Cumulative Probability (CDF)
3. The minimum and maximum will default to $-\infty$ and ∞ respectively. Enter the T-score in the min or max so that the graph matches the graph above.

Step 4 State a conclusion in context

The level of significance is $\alpha =$ _____. The P-value (rounded to three decimal places) is 0.004. Fill in the blank with $\leq$ or $>$:

$$0.004 \_\_\_\_\_ 0.01$$

Defend the citizens of Flint:

The evidence supports the claim that the true _____ lead level for the entire water supply in Flint is _____ _____ 15 ppb. Therefore, by the current standards, the lead level in Flint's water supply is dangerously high.

Summary of Hypothesis Testing Process for a Single Population Mean

Step 1: Determine the Hypotheses

In order to test a claim about a population parameter, we create two opposing hypotheses. We call these the null hypothesis, H_0, and the alternative hypothesis, H_a. Let μ represent a given population mean.

The Null Hypothesis

In every hypothesis test, we assume that the null hypothesis is true. The null hypothesis is always a statement of equality and therefore, should always contain an equal symbol (=). When a test involves a single population mean, the null hypothesis will be

$$H_0 : \mu = value$$

The Alternative Hypothesis

The alternative hypothesis is a claim implied by the research question and is an inequality. The alternative hypothesis states that population mean is greater than (>), less than (<), or not equal (≠) to the assumed value in the null hypothesis.

When a test involves a single population mean, alternative hypothesis will be one of the following:

$$\begin{aligned} H_a &: \mu > \text{ value} \\ H_a &: \mu < \text{ value} \\ H_a &: \mu \neq \text{ value} \end{aligned} \tag{7.4.1}$$

Step 2: Collect Sample Data

During a hypothesis test, we work to know if a sample statistic is unusual or not. Therefore, we must think about probabilities from a sampling distribution.

In a previous lesson, we learned about the sampling distribution of sample means. The Central Limit Theorem says that a sampling distribution of sample means is approximately normal if either the sample size, n, is greater than 30 or sampling was performed from a normally distributed population. In the second step of a hypothesis test, we verify that the sampling distribution is approximately normal and we identify or compute any sample statistics.

Step 3: Assess the Evidence

This step is all about probability. Since the sampling distribution is approximately normal (as determined in step 2), and the population standard deviation is likely unknown, we can compute a T-score and use the student's T-distribution to find probabilities. The sampling distribution of sample means has mean

$$\mu_{\bar{x}} = \mu$$

and standard error

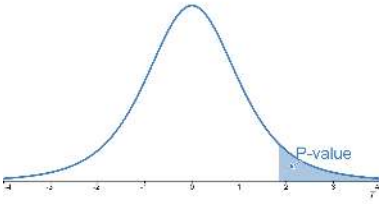
$$\sigma_{\bar{x}} = \frac{s}{\sqrt{n}}$$

where μ is the assumed population mean, s is the sample standard deviation, and n is the sample size. The test statistic is

$$T = \frac{x - \mu}{\sigma} \text{ which translates to } T = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

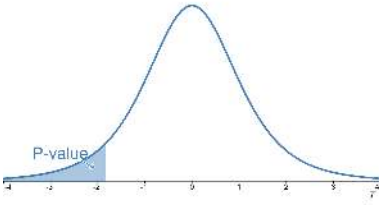
when looking at the sampling distribution of sample means.

- When the alternative hypothesis is $H_a : \mu >$ value, we are conducting a right-tailed test. The P-value is the probability of observing a sample mean at least as extreme as the one we observed. In this case at least as extreme means "as high or higher". The P-value is the area to the right of the test statistic (T.S.).



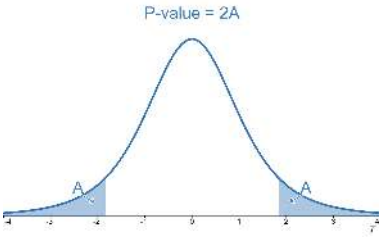
Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

- When the alternative hypothesis is $H_a : \mu <$ value , we are conducting a left-tailed test. The P-value is the probability of observing a sample mean at least as extreme as the one we observed. In this case at least as extreme means "as low or lower". The P-value is the area to the left of the test statistic (T.S.).



Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

- When the alternative hypothesis is $H_a: \mu \neq$ value , we are conducting a two-tailed test, and the P-value is twice the area of either the tail to the right of a positive test statistic (T.S.), or the tail to the left of a negative test statistic (T.S.).



Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

Step 4: State a Conclusion

Hypothesis tests are all about making decisions. We use the P-value to make a decision about the null and alternative hypotheses.

We compare our P-value to a level of significance. The level of significance, denoted α (the greek letter "alpha"), is how unlikely a sample statistic needs to be to convince us about a claim. It is also the level of risk we accept in being wrong.

We have only two possible conclusions:

- If the P-value $\leq \alpha$, we reject the null hypothesis and support the alternative hypothesis.
- If the P-value $> \alpha$, we fail to reject the null hypothesis and cannot support the alternative hypothesis.
    - This does not make the null hypothesis true—we cannot prove the null hypothesis because sample data cannot reveal the true value of the population mean.

You try!

2. A machine is designed to fill jars with 16 ounces of coffee. A consumer suspects that the machine is not filling the jars completely. They randomly sampled 12 jars shown below. Is there enough evidence to support the consumer's claim at a 10% significance level? Assume that the population of volumes in each jar fill is approximately normal.

15	15.4	16.2	16.1	15.8	16.2
15.7	15.6	16	16.3	15.3	15.9

Step 1.

Let μ represent:

H_0:

H_a:

Which type of test and why?

Step 2.

The sampling distribution is approximately normal because:

Go to https://www.desmos.com/calculator. Type $C =$ into the first line and copy and paste the data.

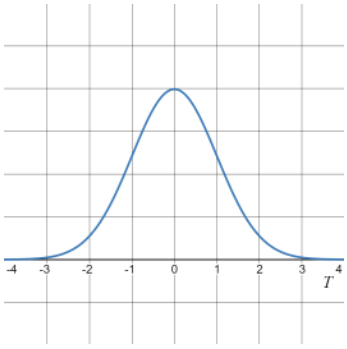
$\bar{x} = \text{mean}(C) =$ _____ ounces

$s = \text{stdev}(C) =$ _____ ounces

$n =$ _____ and $df =$ _____

Step 3.

$T =$



Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

P-value is _____

Step 4.

Compare the P-value and the significance level. Make a decision, and state a conclusion in context:

This page titled 7.4: Hypothesis Tests for a Single Population Mean is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Hannah Seidler-Wright.

- **Current page** by Hannah Seidler-Wright is licensed CC BY-NC-SA 4.0.
- **1.2: The Statistical Analysis Process** by Hannah Seidler-Wright is licensed CC BY-NC-SA 4.0.

7.4.1: Exercises

1. Maryam is curious about the average number of days for COVID-19 symptoms to develop in people ages 70 and older. The average time it takes for someone to develop COVID-19 symptoms is 5 days. She randomly samples 32 people with COVID-19 who are 70, and records the number of days it takes for each subject to develop symptoms. Use the data to test the claim that COVID-19 symptoms develop later than average in people 70 and older. Use the significance level of 1% to test the claim.

2	8	6	15	14	15	2	6
14	7	9	1	3	13	14	7
16	5	12	10	13	1	2	7
2	15	6	5	2	1	6	4

a. μ represents the average number of:

b. H_0:

c. H_a:

d. What test should you use to find the P-value? Justify your answer.

e. Explain why the sampling distribution of sample means is approximately normal.

f. Compute all sample statistics:

$\bar{x} = \text{mean}(a) \approx$ _____ days

$s = \text{stdev}(a) \approx$ _____ days

$n =$ _____

$df = n - 1 =$ _____ $=$ _____

g. Compute the T-score for the sample mean.

h. Use desmos to find the P-value from the T-distribution. Sketch a graph and shade the area that represents the P-value.

i. Make a decision about the null and alternative hypotheses. Justify your answer.

j. State the conclusion in context.

2. A physician claims that jogger's average maximal volume oxygen uptake is greater than that of all adults. Assume that the maximal volume oxygen uptake is approximately normal. A random sample of 40 joggers has a mean of 38.6 ml/kg and a standard deviation of 8 ml/kg. If the average of all adults is 36.7 ml/kg, is there enough evidence to support the physician's claim at a 5% level of significance?

   a. Step 1

   b. Step 2

   c. Step 3

   d. Step 4

3. The Medical Rehabilitation Foundation reports that the average cost of rehabilitation for stroke victims is $24,672. To see if the average cost of rehabilitation is different at a large hospital, a researcher selected a random sample of 35 stroke victims and found that the average cost of their rehabilitation was $25,266 with a standard deviation of $3,251. Can we conclude that the average cost at a large hospital is different at a 1% level of significance?

   a. Step 1

   b. Step 2

   c. Step 3

   d. Step 4

https://stats.libretexts.org/@go/page/49038

4. A researcher wishes to test the claim that the average age of lifeguards in Ocean City is greater than 24 years. She selects a random sample of 26 guards and finds the mean of the sample to be 24.7 years, with a standard deviation of 2 years. Is there evidence to support the claim at a 5% level of significance? Assume that the age of lifeguards is approximately normal.

5. Randall solves the following problem. There is at least one error in each step. Explain to Randall where he made mistakes, and how to fix them.

   A teacher claims that the boys in her classes are taller than the average 69 inches for boys in this age group. The sample of 31 boys had an average height of 69.7 in. If the sample standard deviation is 2.75 inches, is there enough evidence to support the teacher's claim at a 5% level of significance?

   Step 1: μ represents the heights of boys in the teacher's classes.

   $H_0 = 69$

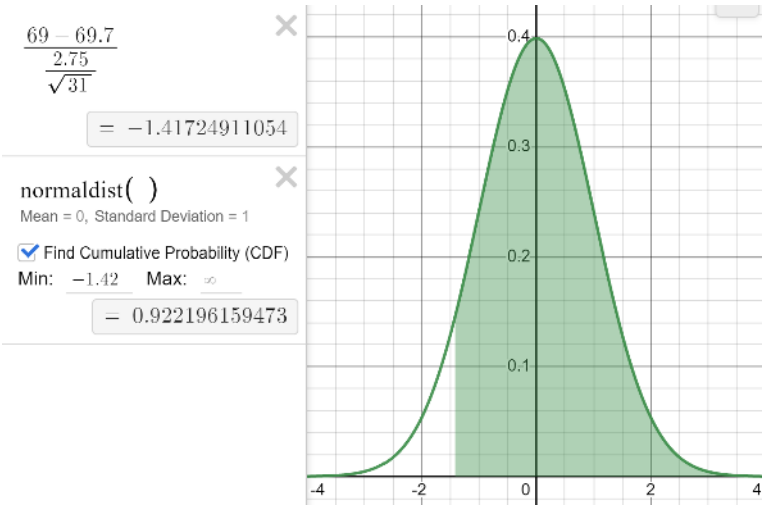
   $H_a > 69$

   We do a right-tailed test because there is a greater than inequality in the alternative hypothesis.

   Step 2: The sample size is 31 which is more than 30 so the population is normal.

   $n = 31, \bar{x} = 69.7, s = 2.75$

   Step 3: $Z = \frac{69-69.7}{\frac{2.75}{\sqrt{31}}} = -1.42$ so the P-value is 0.922



Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

   Step 4: The P-value is greater than the level of significance 0.05 so we support the null hypothesis and reject the alternative hypothesis. There is enough evidence to support the claim that the height of boys in the teachers class is 69 inches.

This page titled 7.4.1: Exercises is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Hannah Seidler-Wright.

7.5: Conclusions (2)

When we estimate a population parameter or conduct a hypothesis test, our last step is to state a conclusion in context. In this section, we will focus on conclusions of a hypothesis test for a population mean. These concepts can be extended to other estimations and hypothesis tests.

Example 1

Economists use average household credit card debt to gauge the financial well-being of families in the U.S. Excessive credit card debt can lead to financial challenges and prevent individuals from saving money or investing money for future expenses. In 2019, U.S. credit card debt hit a record high of $930 billion. On average, Americans carried $6194 in credit card debt in 2019[13].

Suppose we are interested in comparing the mean credit card debt of households from the Baby Boomer generation against the national average. In a random sample of 32 baby boomer households, the mean credit card debt per household was $6043 with a sample standard deviation of $440. At the 1% level of significance, can we conclude that the mean credit card debt of baby boomer households is less than the national average?

Solution:

Step 1. Let μ represent the _____ credit card debt of all baby boomer households.

$H_0 : \mu = \$6194$

H_a : _____

We will perform a _____-tailed test because:

Step 2. Are the criteria for the approximate normality of the sampling distribution of sample means satisfied? Explain.

$n =$ _____

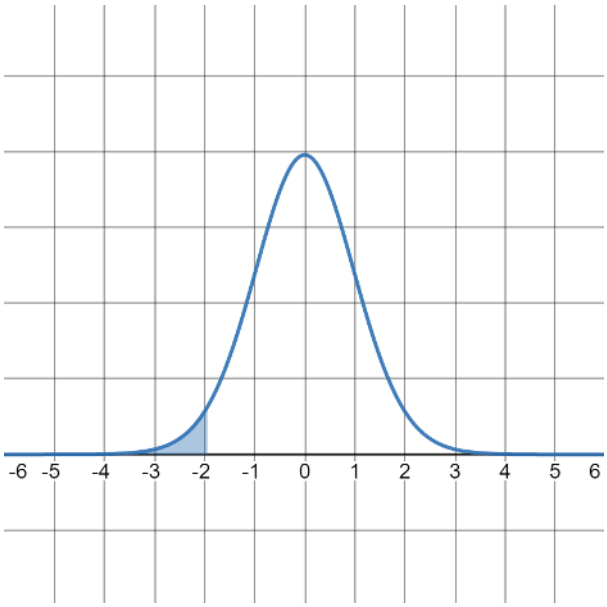
$df =$ _____

$\bar{x} =$ _____

$s =$ _____

Step 3. Calculate the test statistic (rounded to two decimal places) and label it on the graph of the T-distribution below.

$$T = \frac{\qquad\qquad}{\sqrt{\qquad\qquad}} =$$



Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

The P-value is 0.031.

Step 4. The P-value $= 0.031 > 0.01 = 1\% = \alpha$ the level of significance, so we fail to reject the _____ hypothesis and cannot support the _____ hypothesis. Is the sample mean statistically significant? State a conclusion in the context of the problem.

Errors in Hypothesis Tests

In the above example, we failed to reject the null hypothesis, meaning we were not convinced that the null hypothesis was false. This decision is based on sample data and sometimes, there is a small possibility that we made the wrong decision.

Recall, there are two possible conclusions to a hypothesis test, which means there are two possible errors:

Type I. We reject the null hypothesis in support of the alternative hypothesis.

- If we made the wrong decision here, what is actually true?

  We decide that the null hypothesis is _____, when actually, it is _____.

Type II. We fail to reject the null hypothesis in support of the alternative hypothesis.

- If we made the wrong decision here, what is actually true?

  We decide that the null hypothesis is _____, when actually, it is _____.

Notice that Type I and II errors are exclusively about the null hypothesis.

You try!

1. In example 1, μ represented the mean credit card debt for all baby boomer households in 2019.

   The null hypothesis was that the mean credit card debt for all baby boomer households in 2019 was \$6194 $(H_0 : \mu = \$6194)$.

   a. Based on the conclusion of example 1, what type of error might have occurred? Explain.

   b. Describe Type I error in context.

   c. Describe Type II error in context.

2. In example 1, we used data from 32 randomly selected baby boomer households. The mean credit card debt was $6043 with standard deviation $440. We concluded that the sample results based on this small sample were not statistically significant. Can we generalize the results of the hypothesis test to other generations in the U.S.? Explain.

3. Under what conditions would it be appropriate to generalize the results of the hypothesis test?

Statistical and Practical Significance

In example 1, we concluded that the sample mean of $6043 was not statistically significant. The sample did not provide sufficient evidence to conclude that the mean debt of all baby boomer households was less than the national average.

4. The sample mean of $6043 differed from the national average of $6194 by _____ dollars. Do you think the difference between the observed sample mean and the national average is significant in a real-world or practical sense? Explain.

5. How would our assessment of statistical significance change if the sample mean of $6043 was observed for a sample of 100 baby boomer households? Assuming the sample standard deviation is still $440, determine whether the sample mean of $6043 is now statistically significant. Would the conclusion of the hypothesis test change? Explain your answer.

Reference

[13] "Alaskans carry the highest credit card balance–here's the average credit card balance in every state", Alexandria White, May 10 2022, accessed June 21 2022, https://www.cnbc.com/select/average-credit-card-balance-by-state/#:~:text=If%20you%20have%20credit%20card,card%20balance%2C%20on%20average%20%248%2C026.

This page titled 7.5: Conclusions (2) is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Hannah Seidler-Wright.

- **Current page** by Hannah Seidler-Wright is licensed CC BY-NC-SA 4.0.
- **1.2: The Statistical Analysis Process** by Hannah Seidler-Wright is licensed CC BY-NC-SA 4.0.

7.5.1: Exercises

1. Jude was curious if the automated machine at his restaurant was filling drinks with the proper amount. He filled a sample of 20 drinks to test $H_0 : \mu = 530$ mL versus $H_a : \mu \neq 530$ where μ is the mean filling amount. The drinks in the sample contained a mean amount of 528 mL with a standard deviation of 4 mL. These results produced a test statistic of $T = -2.236$ and a P-value of approximately 0.038.

   a. If the significance level is 5%, what can we conclude about the null hypothesis?

   b. If the significance level is 5%, what can we conclude about the alternative hypothesis?

   c. Explain why you made the choices above.

   d. State the conclusion in context.

   e. What type of error might have occurred based on the above conclusion?

   f. Describe the error in context for this example.

2. A quality control engineer is testing the battery life of a new smartphone. The company is advertising that the battery lasts 24 hours on a full charge, but the engineer suspects that the battery life is actually less than that. They take a random sample of 30 of these phones to test $H_0 : \mu = 24$ versus $H_a : \mu < 24$ where μ is the mean battery life of these phones. The sample data had a mean of 21 hours and a standard deviation of 16 hours. These results produced a test statistic of $T \approx -1.03$ and a P-value of approximately 0.156.

   a. If the significance level is 5%, state the conclusion in context.

   b. What type of error might have occurred based on the conclusion above?

   c. Describe the error in context for this example.

   d. Suppose we increase the sample size to 300 and find the same sample mean and sample standard deviation.
      i. Compute the new T-score.

      ii. Is the sample mean statistically significant? Explain why or why not.

      iii. Is the sample mean significant in a practical or real-world sense? Explain why or why not.

3. According to a report from the United States Environmental Protection Agency, burning one gallon of gasoline typically emits about 8.9 kg of CO_2. A fuel company wants to test a new type of gasoline designed to have lower CO_2 emissions. Here are their hypotheses: $H_0 : \mu = 8.9$ kg, $H_a : \mu < 8.9$ kg (where μ is the mean amount of CO_2 emitted by burning one gallon of this new gasoline).

   a. Describe the type I error in context for this example.

   b. Describe the type II error in context for this example.

   c. Suppose Norton writes the following conclusion to the above hypothesis test:

      The P-value is 0.04 which is greater than the level of significance of 1%. Therefore, we accept the null hypothesis and reject the alternative hypothesis. We have proved that the mean amount of CO_2 emitted by burning one gallon of this new gasoline is not less than 8.9 kg.

      Improve Norton's conclusion.

This page titled 7.5.1: Exercises is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Hannah Seidler-Wright.

CHAPTER OVERVIEW

8: Inference Involving Two Population Parameters

8.1: Paired Samples

8.1.1: Exercises

8.2: Distributions of Differences

8.2.1: Exercises

8.3: Inference for a Difference in Two Population Means

8.3.1: Exercises

8.4: Inference for a Difference in Two Population Proportions

8.4.1: Exercises

This page titled 8: Inference Involving Two Population Parameters is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Hannah Seidler-Wright.

8.1: Paired Samples

In this lesson, we will make inferences about the *population mean difference between two quantitative measurements*. We will learn how to construct interval estimates and test hypotheses using paired data.

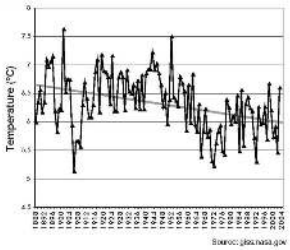
Paired/Dependent and Independent Data

Consider the two scenarios below.

Example 1

The planet's average surface temperature has risen about 2 degrees Fahrenheit (1 degree Celsius) since the late 19th century. Earth's global average surface temperature in 2021 tied with 2018 as the sixth warmest year on record, according to an analysis by NASA[14]. In congress, there are politicians who believe that global warming is a hoax and this belief drives important policy decisions. In 2005, The New York Times reported[15] that Michael Crichton, a novelist, was called to testify before the Senate Committee on Environment and Public Works. The chairman of the committee, Senator James M. Inhofe, who said that global warming is "the greatest hoax ever perpetrated on the American people," had the committee read Crichton's fictional novel "State of Fear" (an environmental thriller that casts doubt on the idea that human activities contribute to global warming). Crichton asserted that cooling observed in the interior of Antarctica shows the lack of reliability of models used for global warming predictions and of climate science in general. The book remains one of the most cited works in climate change skeptic circles.

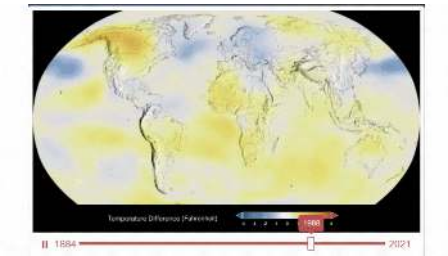
Does the data in "State of Fear" disprove climate change? In the book, a graph of temperatures in Punta Arenas is shown for the last 116 years. The graph has a downward trend and suggests that the temperature is actually dropping over time. To investigate this claim, we randomly sample 32 pairs of latitudes and longitudes, find the nearest station for each set of coordinates, and measure the average temperature over two consecutive blocks of time.



"It's the record from the weather station at Punta Arenas, near here. It's the closest city to Antarctica in the world." He tapped the chart and laughed. **"There's your global warming."**
- from *State of Fear*

Example 2

Los Angeles Daily News reported[16] that “Low-income neighborhoods with higher Black, Hispanic, and Asian populations experience significantly more urban heat than wealthier and predominantly white neighborhoods in Southern California and within a vast majority of populous U.S. counties.” According to a study[17], “roughly 25% of all natural hazard mortality in the U.S. is due to heat exposure (Borden & Cutter, 2008) and heat waves are becoming more frequent, more intense, and are longer in season (Shiva et al.,2019; Wobus et al., 2018); understanding who is affected by urban heating and what drives exposure disparities is therefore critical for crafting just and effective policy responses, particularly under warming climate conditions.”



A statistics student wants to know the mean difference in land surface temperature between poor and affluent communities. They randomly sample 556 counties with high rates of poverty and 500 affluent counties. They find the average temperature and sample standard deviation for each of the groups and use the sample data to estimate the population mean difference in temperature between poor and affluent communities.

Summary

1. Notice any differences and similarities between the two examples.

In both examples, two sets of data were collected. In example 1, the average temperature is measured twice (from 1901-1950 and from 1951-2000) for each randomly chosen station location. In example 2, the average temperature is measured once for each group. In example 1, the two data sets are directly related in pairs. We call such data **paired** or **dependent**. A sample is paired if each subject in the sample is measured twice. In example 2, the two data sets are not directly related. The values of one set have no effect on the values of the other set. Such data is referred to as **independent**.

Identify Paired Samples

In the following questions, identify whether the question is solved by constructing a confidence interval or conducting a hypothesis test, and if this requires data from a paired sample or from two independent samples. Explain how you made your decision.

2. Jason claims that a higher proportion of males pass their drivers test in the first attempt than the proportion of females pass the test in the first attempt.

3. It is believed that the average grade on an English essay in a particular school system for females is higher than for males. A random sample of 31 females had a mean score of 82 with a standard deviation of three, and a random sample of 25 males had a mean score of 76 with a standard deviation of four. Estimate the average difference in grades for females and males.

4. Eight subjects are picked at random and given a new sleep medication. The mean hours slept for each person were recorded before starting the medication and after. Estimate the mean difference in hours slept before and after use of the sleep medication.

5. A new WiFi range booster is being offered to consumers. A researcher tests the native range of 12 different routers under the same conditions. The ranges are recorded. Then the researcher uses the new WiFi range booster and records the new ranges. Does the new WiFi range booster do a better job?

Construct a Confidence Interval for the Mean Difference

In example 1, we want to determine if the earth is warming. We randomly sample 32 pairs of latitudes and longitudes, find the nearest station for each set of coordinates, and measure the average temperature over two consecutive blocks of time.

6. Given below is the average temperature measured from 1901-1950 and the average temperature measured from 1951-2000 for the Punta Arenas station. Does this piece of data support the claim that the earth is not warming? Use a difference to support your answer.

Nearest station	**avg 1901-1950**	**avg 1951-2000**	**Difference**
Punta Arenas	6.5164	6.207959184	

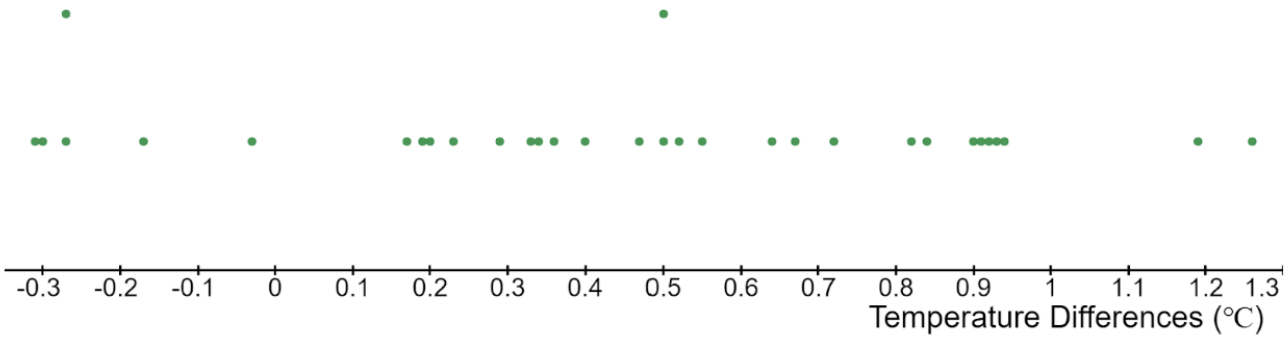
7. The difference in temperature at the Punta Arenas station is _____ which means that the average temperature has decreased. What value for the difference would indicate that there was *no change* in temperature? What does this tell you about the average temperature from 1901-1950 and the average temperature from 1951-2000 at a station that has no change in temperature?

8. If the difference in temperature at a station is _____, then the average temperature has increased. Which measurement would be greater at such a station: the average temperature from 1901-1950, or the average temperature from 1951-2000? Explain why you think this.

9. Given below is the data from the 32 stations in the sample. Highlight any station where the temperature decreased. For how many stations did the temperature decrease?

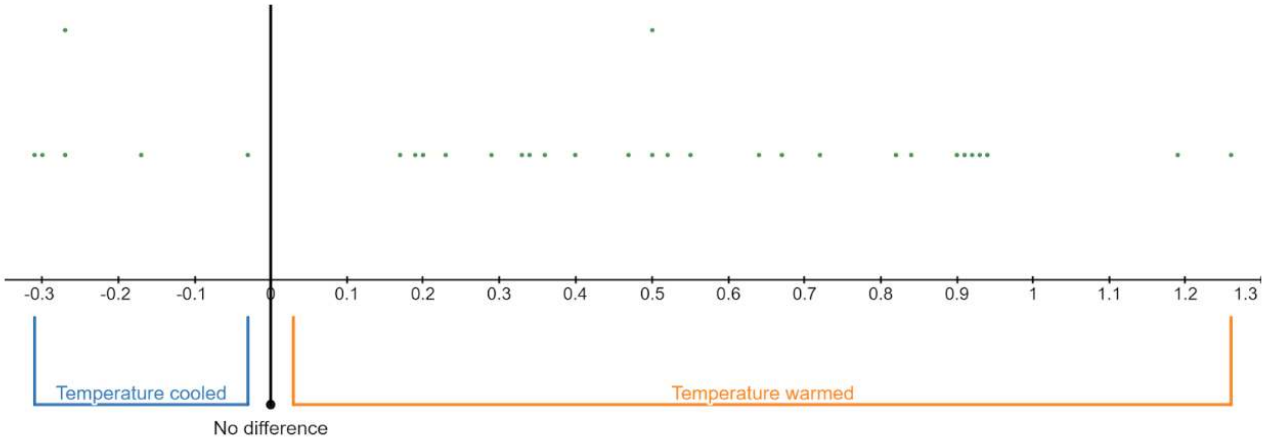
Nearest station	avg 1901-1950	avg 1951-2000	Difference
Tenerife Los Rodeos	14.667	15.927	1.26
Punta Arenas	6.516	6.208	-0.308
Invercargill Airport	9.239	9.784	0.545
Honolulu Intl Ap	24.292	24.486	0.194
Montevideo Prado	15.642	16.037	0.395
Eagle	-5.179	-4.359	0.82
Rarotonga Intl	23.331	23.836	0.505
Accra	26.601	26.767	0.166
Port Elizabeth Intl	17.19	17.164	-0.026
Chatham Islands	10.683	11.324	0.641
Mahe Seychellesbri	26.561	26.896	0.335
Nairobi Dagoretti	17.148	17.649	0.501
Hobart Ellerslie Road	12.299	12.818	0.519
Jask	26.545	26.372	-0.173
Manati	23.694	24.594	0.9
Cape Leeuwin	16.743	17.076	0.333
Cairns Post Office	23.726	24.647	0.921
Durban Intl	19.981	20.704	0.723
Upernavik	-7.012	-7.286	-0.274
Buenos Aires Observ	16.92	18.107	1.187
Auckland Aero Aws	14.46	15.391	0.931
Bahia Blanca Aero	14.441	14.642	0.201
Merced Muni Ap	16.236	15.94	-0.296
Albany	15.608	15.84	0.232
Concepcion	12.226	12.518	0.292
Dakar Yoff	22.727	23.569	0.842
Kazalinsk	8.242	8.912	0.67
Cherdyn	0.3	0.661	0.361
Karlstad	5.373	5.099	-0.274
Vladivostok	3.978	4.919	0.941
Rio De Janeiro	24.36	25.275	0.915
Kirensk	-4.331	-3.861	0.47

10. Mark on the numberline where ‘no difference’ in temperature would be. For how many stations did the temperature warm? For how many stations did the temperature cool?



Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

11. At the Punta Arenas station, the temperature cooled. Do you think that this difference is representative of climate patterns across the world?



Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

12. Let's estimate the mean temperature change for the earth with 95% confidence. The mean difference from our sample, $\bar{x} = 0.452$ degrees celsius, and the sample standard deviation for the differences is s=0.4361 degrees celsius.

Step 1 Is the sampling distribution approximately normal? Explain.

Step 2 Find the critical value (rounded to three decimal places) that corresponds to a ____ % confidence level and _____ degrees of freedom.

$T_c = \text{tdist}(\_\_\_\_\_) \cdot \text{inversecdf}(\_\_\_\_\_) = \_\_\_\_\_$

$n = \_\_\_\_\_$

$df = \_\_\_\_\_$

$\bar{x} = \_\_\_\_\_$

$s = \_\_\_\_\_$

Step 3 The margin of error rounded to three decimal places is

$$E = T_c \cdot \frac{s}{\sqrt{n}} = \_\_\_\_\_ \cdot \frac{\_\_\_\_\_}{\sqrt{\_\_\_\_\_}} = \_\_\_\_\_$$

Step 4 The interval is

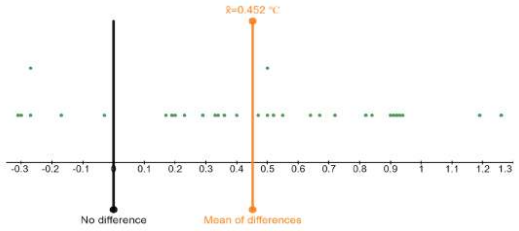
$$(\bar{x} - E, \bar{x} + E) = (\_\_\_\_\_ - \_\_\_\_\_,\ \_\_\_\_\_ + \_\_\_\_\_) = (\_\_\_\_\_, \_\_\_\_\_)$$

Step 5 State the conclusion in context:

Based on the interval, do you think the earth is warming or cooling? Explain.

Conduct a Hypothesis Test about the Mean Difference

Stations showed warming, on average. Is the average temperature change significantly high? Let's test this claim at a 5% level of significance.



Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

Step 1 Let μ represent the mean global temperature change.

$H_0 : \mu = 0$

$H_a :$ _____ _____ _____

We will conduct a right-tailed test.

Step 2 Is the sampling distribution approximately normal? Explain.

$n =$ _____

$df =$ _____

$\bar{x} =$ _____

$s =$ _____

Step 3 Compute the test statistic rounded to three decimal places. Use it to compute the P-value.

$$T = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} = \frac{\_\_\_\_\_}{\frac{\_\_\_\_\_}{\sqrt{\_\_\_\_\_}}} = \_\_\_\_\_$$

P-value is _____

Step 4 State the conclusion in context:

Reference

[14]Climate Change: Vital Signs of the Planet. 2022. *Climate Change Evidence: How Do We Know?* accessed June 28 2022, https://climate.nasa.gov/evidence/

[15] "Michael Crichton, Novelist, Becomes Senate Witness," Michael K Janofsky, Sept 29, 2005, accessed June 28, 2022, https://www.nytimes.com/2005/09/29/books/michael-crichton-novelist-becomes-senate-witness.html

[16]"Poor neighborhoods get up to 7° hotter than rich ones in Southern California, study finds," July 13, 2021, accessed June 28, 2022, https://www.dailynews.com/2021/07/13/poor-southern-california-communities-suffer-more-from-extreme-heat-ucsd-study-finds/

[17]"Widespread Race and Class Disparities in Surface Urban Heat Extremes Across the United States," Susanne Amelie Benz and Jennifer Anne Burney, July 13, 2021, accessed June 28, 2022, https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2021EF002016

This page titled 8.1: Paired Samples is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Hannah Seidler-Wright.

- **Current page** by Hannah Seidler-Wright is licensed CC BY-NC-SA 4.0.
- **1.2: The Statistical Analysis Process** by Hannah Seidler-Wright is licensed CC BY-NC-SA 4.0.

8.1.1: Exercises

Confidence Intervals for Paired Data

Recall the five step process for constructing a confidence interval for a paired mean difference:

1. Verify that the sampling distribution of sample mean differences is approximately normal by showing the number of differences in the set is more than 30 OR the population of differences is normally distributed.
2. Find the critical value from the Student's T distribution that corresponds to the given confidence level. Let $\bar{x}$ represent the mean of sample differences, let s represent the standard deviation of sample differences, let n represent the number of differences, and let $df = n - 1$ represent the degrees of freedom.
3. Compute the margin of error: $E = T_c \cdot \frac{s}{\sqrt{n}}$
4. Compute the upper and lower limit of the interval and write the answer in interval notation: $(\bar{x} - E, \bar{x} + E)$
5. Interpret the interval and state a conclusion in context.

Hypothesis Tests for Paired Data

Recall the four step process testing a hypothesis about a paired mean difference:

1. Define a variable to represent the population mean difference and determine the hypotheses. The null hypothesis will always be $\mu = 0$.
2. Verify that the sampling distribution of sample mean differences is approximately normal by showing the number of differences in the set is more than 30 OR the population of differences is normally distributed.
3. Let $\bar{x}$ represent the mean of sample differences, let s represent the standard deviation of sample differences, let n represent the number of differences, and let $df = n - 1$ represent the degrees of freedom. Compute the T-score and the corresponding P-value. $T = \frac{\bar{x}}{\frac{s}{\sqrt{n}}}$
4. Make a decision and state a conclusion in context.

1. An educational website offers a practice program for the Law School Admissions Test (LSAT). Users of the program take a pretest and posttest. Here are the scores and gains for a random sample of 6 users:

User ID	1	2	3	4	5	6
Pretest	140	152	153	159	150	146
Posttest	150	159	170	164	148	166
Difference (Posttest - Pretest)						

Assume the population of differences in test scores is approximately normal. Construct a 90% confidence interval for the mean difference in test scores. Use the interval to create an advertisement for the educational website.

2. Mariah wants to know if growth-minded language influences performance on mathematics exams. She randomly selects 35 students and gives each student two assessments. One assessment has growth-minded language throughout, and the other does not. She records the scores, finds the difference between the assessment scores (with growth-minded language minus without growth-minded language) for each student. She constructs a 99% confidence interval for the mean difference is scores (out of 10) to be (1.16, 2.725). Improve the following conclusions:
   a. Since 0 is not in the interval, we can conclude that there is a difference, on average.

   b. Since the two numbers in the interval are positive, we can conclude that there is a positive difference on average.

3. Theo wants to know if using melatonin helps with insomnia so he will construct a 95% confidence interval. He randomly selects 10 individuals with insomnia. He records how long each individual sleeps without any intervention. He then gives each subject a dose of melatonin every night for a week. He records the time each subject spent sleeping on the 7th day of the experiment. Research has shown that the population of differences is approximately normal. Below is the data:

Subject ID number	1	2	3	4	5	6	7	8	9	10
Before Melatonin	4.2	3.8	5.1	4.3	4	3.1	2.9	3.4	4.6	5.8
After Melatonin	4.3	4	5	5	6	3.2	3.4	2	4.1	6.3

Below is Theo's solution. Spot any errors. Explain how you would improve/correct each error.

Step 1: $n = 10 \geq 10$ so the sampling distribution is normal.

Step 2: $\bar{x}_{\text{before}} = 4.12$ hours, $\bar{x}_{\text{after}} = 4.33$ hours , so the difference is $\bar{x} = 0.21$ hours . The standard deviation is $s = s_{\text{after}} - s_{\text{before}} = 0.408$ hours . To calculate the critical value, I entered tdist(10).inversecdf(0.95) into desmos and found that $T_c = 1.81$.

Step 3: $E = 1.81 \cdot \frac{0.408}{\sqrt{9}} = 0.246$

Step 4: $(0.21 - 0.246), (0.21 + 0.246) = (-0.036, 0.456)$

Step 5: The true population mean is between -0.036 and 0.456.

Interpretation: Since 0 is contained in the interval, melatonin does not have an effect on sleep for individuals with insomnia.

4. Complete problem 3 correctly.

5. An educational website offers a practice program for the Law School Admissions Test (LSAT). Users of the program take a pretest and posttest. Here are the scores and gains for a random sample of 6 users:

User ID	1	2	3	4	5	6
Pretest	140	152	153	159	150	146
Posttest	150	155	160	159	148	151
Difference (Posttest - Pretest)						

Assume the population of differences in test scores is approximately normal. Test the company's claim that their program improves a customer's average score at a 5% level of significance.

6. Mariah wants to know if growth-minded language influences performance on mathematics exams. She randomly selects 35 students and gives each student two assessments. One assessment has growth-minded language throughout, and the other does not. She records the scores, finds the difference between the assessment scores (with growth-minded language minus without growth-minded language) for each student. She tests the claim that using growth-minded language to frame questions on an assessment improves the average performance on mathematics exams at a 5% level of significance. She calculates the test statistic to be T=2.87 and the corresponding P-value to be 0.0044. Improve the following conclusions.
   a. Since the P-value is less than the level of significance, we reject the null hypothesis and support the alternative hypothesis. Therefore, we have proved that using growth-minded language to frame questions on an assessment improves performance on mathematics exams.

   b. Since the P-value is less than the level of significance, the null hypothesis is false and we accept the alternative hypothesis. Therefore, the data support the claim that the difference in sample means is positive.

7. Theo wants to know if using melatonin helps with insomnia. He randomly selects 10 individuals with insomnia. He records how long each individual sleeps without any intervention. He then gives each subject a dose of melatonin every night for a week. He records the time each subject spent sleeping on the 7th day of the experiment. Research has shown that the population of differences is approximately normal. Below is the data:

Subject ID number	1	2	3	4	5	6	7	8	9	10
Before Melatonin	4.2	3.8	5.1	4.3	4	3.1	2.9	3.4	4.6	5.8
After Melatonin	4.3	4	5	5	6	3.2	3.4	2	4.1	6.3
Difference (After-Before)	0.1	0.2	-0.1	0.7	2	0.1	0.5	-1.4	-0.5	0.5

   Below is Theo's solution. Spot any errors. Explain how you would improve/correct each error.

   Step 1: Let μ represent the population mean difference in amount of sleep after and before using melatonin.

   $H_0 : \bar{x} = 0.21$

   $H_a : \bar{x} > 0.21$

   Step 2: The sample is large enough so it's normal. $\bar{x} = 0.21, s = 0.871, n = 10$

   Step 3: $Z = \frac{0.21}{\frac{0.871}{\sqrt{10}}} \approx 0.76$. The P-value was calculated using normaldist() in desmos and inputting the max as 0.76. The P-value is 0.7764.

   Step 4: The P-value is 0.7764 which is greater than the level of significance 5%. Therefore, we accept the null hypothesis and reject the alternative hypothesis. The data support the claim that the true mean difference in sleep after and before taking melatonin is equal to 0.21 hours. Therefore, melatonin increases sleep.

8. Complete problem 7 correctly.

This page titled 8.1.1: Exercises is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Hannah Seidler-Wright.

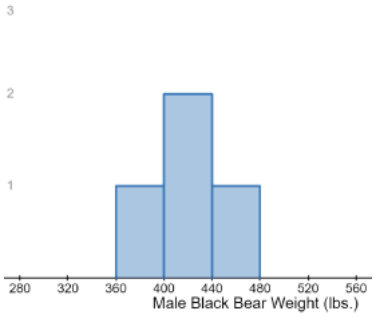
8.2: Distributions of Differences

In this lesson, we will begin to consider statistical methods for comparing independent samples from two populations or experimental treatments. These methods will allow us to compare population means or proportions from two independent groups. To make inferences about differences, we need to use the sampling distribution of differences.

In a previous lesson, we considered the weights of black bears. Black bears weights tend to differ by sex. Female black bears weigh 175 pounds on average, with a standard deviation of 50 pounds, whereas, male black bears weigh 400 pounds on average, with a standard deviation of 40 pounds. Both of the populations are approximately normally distributed.

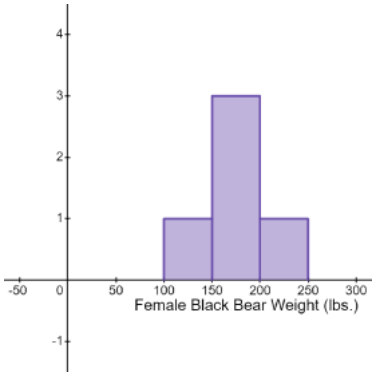
Let's say we are concerned with differences in male and female black bears in a small region in Colorado where the population has dwindled. The population contains only 4 male black bears and 5 female black bears.

Below is a histogram of the 4 male black bear weights. It is very crudely bell-shaped. The mean weight is $\mu_1 = 418.25$ lbs and the standard deviation is $\sigma_1 = 15.943$ lbs. Recall, the variance is the square of the standard deviation. In this case, the variance is $\sigma_1^2 = 254.1875$.



Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

Next is a histogram of the 5 female black bear weights. It is also very crudely bell-shaped. The mean weight is $\mu_2 = 167.6$ lbs and the standard deviation is $\sigma_2 = 36.258$ lbs. Recall, the variance is the square of the standard deviation. In this case, the variance is $\sigma_2^2 = 1314.64$.

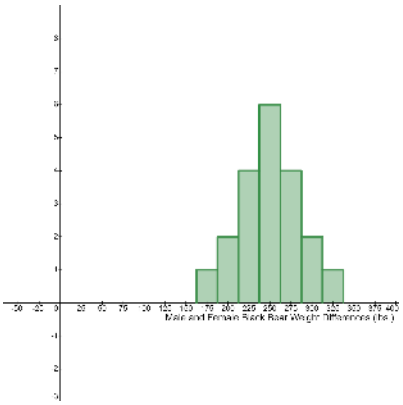


Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

We are interested in the differences in male and female black bear weights for this small population. Let m represent the male black bear weights, and f represent the female black bear weights, then $m - f$ represents the difference in male and female black bear weights. Weight differences are listed for every combination of male and female bear in the table below.

	f values (female black bear weights)					
m values (male black bear weights)		**110**	**164**	**175**	**165**	**224**
	418	308	254	243	253	194
	420	310	256	245	255	196
	395	285	231	220	230	171
	440	330	276	265	275	216

The frequency distribution of weight differences in male and female black bears in the region is given below. It has mean 250.65 lbs, standard deviation 39.608 lbs, and variance 1568.8275.



Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

1. Describe the shape of the distribution of differences.

2. Describe the relationship between the mean difference and the means of male and female black bear weights $(\mu_1 \text{ and } \mu_2)$.

3. Describe the relationship between the *variance* of differences and the *variances* of male and female black bear weights $({\sigma_1}^2 \text{ and } {\sigma_2}^2)$.

Summary

- If two independent quantitative variables are approximately normal, then differences between those variables will be approximately normal as well.
- The mean of all differences between those variables is the difference of their means, respectively.
- The variance of all differences between those variables is the sum of their variances, respectively. The standard deviation of differences is the square root of the variance.

Distribution of Differences Between Population Means

- To understand this sampling distribution for the difference in sample means, we just need to think about the sampling distribution for each population. Recall, the Central Limit Theorem states that the sample size must be greater than 30 or the sample must come from a normal population. If we have two samples, then both sample sizes must be greater than 30 or both samples must come from normal populations to ensure that the distribution of differences in sample means is approximately normal.

Consider the sampling distribution of sample means. It has mean $\mu_{\bar{x}} = \mu$, the population mean, and standard error $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$, and *variance* ${\sigma_{\bar{x}}}^2 = \left(\frac{\sigma}{\sqrt{n}}\right)^2 = \frac{\sigma^2}{n}$. Extending this for two populations, with population means μ_1 and μ_2, standard deviations σ_1 and σ_2, and sample sizes n_1 and n_2 respectively, we find that

- the mean of the sampling distribution of differences is $\mu_{\bar{x}_1 - \bar{x}_2} = \mu_1 - \mu_2$.
- The sampling distribution of differences has *variance* $\sigma^2_{\bar{x}_1 - \bar{x}_2} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$ which implies that the standard error of the sampling distribution of differences is $\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$.

Distribution of Differences Between Population Proportions

- To understand the sampling distribution of the difference in sample proportions, we just need to think about the sampling distribution for each population. Recall, the Central Limit Theorem states that there must be at least 10 expected successes and failures in a sample to ensure its sampling distribution is approximately normal. If we have two samples, then there must be at least 10 expected successes and failures in each sample to ensure that the distribution of differences in sample proportions is approximately normal.

Consider the sampling distribution of sample proportions. It has mean $\mu_{\hat{p}} = p$, the population proportion, standard error $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$, and *variance* ${\sigma_{\hat{p}}}^2 = \left(\sqrt{\frac{p(1-p)}{n}}\right)^2 = \frac{p(1-p)}{n}$.

- Extending this for two populations, with population proportions p_1 and p_2, and sample sizes n_1 and n_2 respectively, we find that the mean of the sampling distribution of differences is $\mu_{\hat{p}_1-\hat{p}_2} = p_1 - p_2$.
- The sampling distribution of differences has *variance* $\sigma^2_{\hat{p}_1-\hat{p}_2} = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}$ which implies that the standard error of the sampling distribution of differences is $\sigma_{\hat{p}_1-\hat{p}_2} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$.

This page titled 8.2: Distributions of Differences is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Hannah Seidler-Wright.

- **Current page** by Hannah Seidler-Wright is licensed CC BY-NC-SA 4.0.
- **1.2: The Statistical Analysis Process** by Hannah Seidler-Wright is licensed CC BY-NC-SA 4.0.

8.2.1: Exercises

1. Weaver worker ants have a bimodal length distribution. There are two types of workers, minors and majors. The majors have an average length of 9 mm and standard deviation of 1 mm. The minors average half the length at 4.5 mm and a standard deviation of 0.25 mm. Suppose an entomologist is interested in examining the difference in sample means from independent random samples of major and minor weaver worker ants. They measure the length of 200 major and 250 minor randomly selected worker ants.

   a. Is the normal distribution an appropriate model for the sampling distribution of the difference in sample means? Explain.

   b. Compute the mean of the sampling distribution of the difference in sample means, $\mu_{\bar{x}_1-\bar{x}_2}$.

   c. Compute the standard error of the sampling distribution of the difference in sample means, $\sigma_{\bar{x}_1-\bar{x}_2}$. Round to three decimal places.

   d. Would it be unusual to see a sample mean difference (in majors minus minors) as low as 4.25 mm? Compute a Z-score (rounded to two decimal places) to justify your answer.

2. Research has shown that 13.8% of all females are left-handed and 16.1% of all males are left-handed. Suppose a sociologist would like to understand the differences in sample proportions from independent random samples of females and males. They randomly surveyed 150 females and 175 males.

   a. Is the normal distribution an appropriate model for the sampling distribution of the difference in sample proportions? Explain.

   b. Compute the mean of the sampling distribution of the difference in sample proportions, $\mu_{\hat{p}_1-\hat{p}_2}$.

   c. Compute the standard error of the sampling distribution of the difference in sample means, $\sigma_{\hat{p}_1-\hat{p}_2}$. Round to three decimal places.

   d. Would it be unusual to see a sample proportion difference of -0.002? Compute a Z-score (rounded to two decimal places) to justify your answer.

This page titled 8.2.1: Exercises is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Hannah Seidler-Wright.

8.3: Inference for a Difference in Two Population Means

Constructing a Confidence Interval for the Difference in Two Population Means

An experiment conducted by a leadership consulting firm, Nextion, was used to investigate the unconscious biases in the workplace at a law firm.[18] Researchers wrote a memo with 22 errors. The errors consisted of minor spelling and grammar errors, major technical writing errors, errors in fact, and errors in analysis of the facts. Fifty-three partners from 22 law firms received copies of the memo. 24 were told the memo was written by an African-American man named Thomas Meyer, and the remaining 29 were told the writer was a Caucasian man named Thomas Meyer. The partners gave the memo from the white Thomas Meyer an average rating of 4.1 out of 5 with a standard deviation of 0.8, while they gave the memo from the black Thomas Meyer an average rating of 3.2 out of 5 and a standard deviation of 1.12. Let's assume that scores for each memo are approximately normal.

1. State $n_1, n_2, \bar{x}_1, \bar{x}_2, s_1, s_2,$ and $\bar{x}_1 - \bar{x}_2$.

2. If the partners do have unconscious bias, which sample mean do you expect to be larger?

- One-sample situations: you compare a statistic in _____ population against a _____ about that population.
  - Example: Is the mean lead level in the Flint, Michigan water supply higher than the acceptable safe level of 15 parts per billion (ppb)?

- Two-sample situations: you measure the _____ _____ in _____ _____ and see if they are significantly different.
  - Example: Are black men assessed more harshly than white men, on average, by partners at law firms?

Let μ_1 represent the mean score given to _____ white Thomas Meyer memos by partners at law firms. Let μ_2 represent the mean score given to _____ black Thomas Meyer memos by partners at law firms.

Five Step Process for Constructing a Confidence Interval for the Difference in Two Population Means

The process we use to build confidence intervals has not changed. The following is a list of familiar steps with the appropriate formulas to use for this situation.

1. Verify that the sampling distribution is approximately normal by checking that *each* sample size is greater than 30 OR *each* sample came from a normal population.
2. Find the critical value from the Student's T-distribution that corresponds to the provided confidence level.
3. Compute the margin of error $E = T_c \cdot \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$
4. Compute the lower and upper limits of the interval, and write the interval in interval notation. $(\bar{x}_1 - \bar{x}_2 - E, \bar{x}_1 - \bar{x}_2 + E)$
5. Write a conclusion in context. Interpret the interval.

Apply this process to the following example: Researchers wrote a memo with 22 errors. The errors consisted of minor spelling and grammar errors, major technical writing errors, errors in fact, and errors in analysis of the facts. Fifty-three partners from 22 law firms received copies of the memo. 24 were told the memo was written by an African-American man named Thomas Meyer, and the remaining 29 were told the writer was a Caucasian man named Thomas Meyer. The partners gave the memo from the white Thomas Meyer an average rating of 4.1 out of 5 with a standard deviation of 0.8, while they gave the memo from the black Thomas Meyer an average rating of 3.2 out of 5 and a standard deviation of 1.12. Let's assume that scores for each memo are approximately normal. We will construct a 95% confidence interval for the true mean difference in memo scores for memos written by white and black men at law firms. Use 51 degrees of freedom.

1. Explain why the sampling distribution of differences in sample means is approximately normal.

2. The confidence level is _____% so the critical value (rounded to three decimal places) is

$$T_c = \text{tdist}(\_\_\_\_\_) \cdot \text{inversecdf}(\_\_\_\_\_) = \_\_\_\_\_ \tag{8.3.1}$$

.

3. $E = T_c \cdot \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \_\_\_\_\_ \cdot \sqrt{\frac{(\_\_\_\_\_)^2}{\_\_\_\_\_} + \frac{(\_\_\_\_\_)^2}{\_\_\_\_\_}} = \_\_\_\_\_$

4. $(\bar{x}_1 - \bar{x}_2 - E, \bar{x}_1 - \bar{x}_2 + E) =$

5. We are _____% confident that the true _____ _____ in memo scores for memos written by white and black men at law firms is between _____ points out of 5 and _____ points out of 5.

   Does the interval suggest there is a difference? Why or why not?

Testing a Claim about the Difference in Two Population Means

Four Step Process for Testing a Claim about the Difference between Two Population Means

The process we use to test a claim about a population parameter has not changed. The following is a list of familiar steps with the appropriate formulas to use for this situation.

1. State the hypotheses. Define μ_1 and μ_2. The null hypothesis will always be $H_0 : \mu_1 = \mu_2$. The alternative will be one of the following $H_a : \mu_1 < \mu_2, H_a : \mu_1 > \mu_2$, or $H_a : \mu_1 \neq \mu_2$ based on the statement of the claim in the problem.

2. Verify that the sampling distribution is approximately normal by checking that *each* sample size is greater than 30 OR *each* sample came from a normal population.

3. Compute the T-score: $T = \dfrac{\bar{x}_1 - \bar{x}_2}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}}$. The degrees of freedom will be provided. Use the Student's T-distribution to find the P-value (the probability of observing a sample difference as extreme or more extreme than the calculated sample difference just by chance).

4. Make a decision about the null and alternative hypotheses and state a conclusion in context.

Apply this process to the following example: Researchers wrote a memo with 22 errors. The errors consisted of minor spelling and grammar errors, major technical writing errors, errors in fact, and errors in analysis of the facts. Fifty-three partners from 22 law firms received copies of the memo. 24 were told the memo was written by an African-American man named Thomas Meyer, and the remaining 29 were told the writer was a Caucasian man named Thomas Meyer. The partners gave the memo from the white Thomas Meyer an average rating of 4.1 out of 5 with a standard deviation of 0.8, while they gave the memo from the black Thomas Meyer an average rating of 3.2 out of 5 and a standard deviation of 1.12. Let's assume that scores for each memo are approximately normal. Do the results give convincing statistical evidence that partners at law firms assess black men more harshly than white men on average? Use a 5% level of significance and 51 degrees of freedom.

1. Let μ_1 represent:

   Let μ_2 represent:

   H_0 :

   H_a :

   Use a _____-tailed test because:

2. Explain why the sampling distribution of differences in sample means is approximately normal.

3. $T = \dfrac{\bar{x}_1 - \bar{x}_2}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}} = \dfrac{\_\_\_\_\_ - \_\_\_\_\_}{\sqrt{\dfrac{(\_\_\_\_\_)^2}{\_\_\_\_\_} + \dfrac{(\_\_\_\_\_)^2}{\_\_\_\_\_}}} =$

   P-value is _____.

4. We _____ the null hypothesis, we _____ the alternative hypothesis.

   Conclusion in context:

Reference

[18]Debra Cassens Weiss. "Partners in study gave legal memo a lower rating when told author wasn't white" *ABA Journal*, April 21, 2014. Accessed July 5, 2022. https://www.abajournal.com/news/article/hypothetical_legal_memo_demonstrates_unconscious_biases

This page titled 8.3: Inference for a Difference in Two Population Means is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Hannah Seidler-Wright.

8.3.1: Exercises

1. What conditions are required to utilize the Student's T-distribution to approximate probabilities when we compare two population means?

2. Suppose you are interested in the difference in average life span between white and black people. You randomly survey death records of 120 white people, the mean life span was 77.6 years with a standard deviation of 12.1 years. Of the 87 black people, the mean life span was 71.8 years with a standard deviation of 16.3 years. Estimate the average difference in lifespan for white and black people using a 90% confidence level and 205 degrees of freedom. Show all five steps of the confidence interval process.

3. You are interested in finding out if mothers and fathers spend different amounts of time on childcare. You randomly survey 32 mothers and find that they spend an average of 14.0 hours per week on childcare with a sample standard deviation of 4.6 hours per week. You randomly survey 31 fathers and find that they spend an average of 10.8 hours per week on childcare with a sample standard deviation of 4.1 hours per week. Test the claim at a 5% level of significance using 61 degrees of freedom. Show all four steps of the hypothesis testing process.

4. Below is a solution that Zed wrote for the following problem: The mean batting average for a sample of eight Rattlers is 0.21 with a sample standard deviation of 0.05, and the mean batting average for a sample of nine Vikings is 0.26 with a sample standard deviation of 0.06. Assume batting averages are normally distributed. Are the batting averages of the Rattlers significantly lower than the Vikings, at a 1% level of significance? Use 15 degrees of freedom.

   Spot any errors in the solutions, and explain to Zed how to correct the error.

   Step 1: Let μ_1 represent the Rattlers and let μ_2 represent the Vikings.

   $H_0 : \mu = 0$

   $H_a : \mu < 0$

   Step 2: There are at least 10 observed successes and failures in the samples so it's normal. $n_1 = 8, \bar{x}_1 = 0.21, s_1 = 0.05, n_2 = 9, \bar{x}_2 = 0.26, s_2 = 0.06$

   Step 3: $T = \frac{0.21-0.26}{\sqrt{\frac{0.05}{8}+\frac{0.06}{9}}} = -0.44$ in desmos: tdist(15) with -0.44 as the min. P-value is 0.667.

   Step 4: The P-value is larger than the level of significance so we accept the null hypothesis and reject the alternative hypothesis. Therefore the batting average for the two teams are the same.

5. The mean batting average for a sample of eight Rattlers is 0.21 with a sample standard deviation of 0.05, and the mean batting average for a sample of nine Vikings is 0.26 with a sample standard deviation of 0.06. Assume batting averages are normally distributed. Are the batting averages of the Rattlers significantly lower than the Vikings, at a 1% level of significance? Use 15 degrees of freedom.

This page titled 8.3.1: Exercises is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Hannah Seidler-Wright.

8.4: Inference for a Difference in Two Population Proportions

Constructing a Confidence Interval for the Difference in Two Population Proportions

In a study,[19] investigators created mock identical resumés, which were sent to job placement ads in Chicago and Boston. Each resumé was randomly assigned either a commonly-white or commonly-black name. In total, 246 out of 2445 commonly-white named resumés received a callback and 164 out of 2445 commonly-black named resumés received a callback.

	Commonly-White Names	Commonly-Black Names	Total
Called back	246	164	410
Not called back	2199	2281	4480
Total	2445	2445	4890

1. Calculate $n_1, n_2, \hat{p}_1, \hat{p}_2, \hat{p}_1 - \hat{p}_2$

2. If there is hiring discrimination, which sample proportion do you expect to be larger?

- One-sample situations: you compare a statistic in _____ population against a _____ about that population.
  - Example: Is the proportion of recent EU migrants who are male actually lower than the claimed 75%?
- Two-sample situations: you measure the _____ _____ in _____ _____ and see if they are significantly different.
  - Example: Is the proportion of callbacks for commonly-white name apps higher than for commonly-black name apps?

Let p_1 represent the proportion of _____ applicants with commonly-white names who'd receive callbacks when applying to jobs like the ones in this study. Let p_2 represent the proportion of _____ applicants with commonly-black names who'd receive callbacks when applying to jobs like the ones in this study.

Five Step Process for Constructing a Confidence Interval for the Difference in Two Population Proportions

The process we use to build confidence intervals has not changed. The following is a list of familiar steps with the appropriate formulas to use for this situation.

1. Verify that the sampling distribution is approximately normal by checking that there are at least 10 observed successes and failures in *each* sample.
2. Find the critical value from the normal distribution that corresponds to the provided confidence level.
3. Compute the margin of error $E = Z_c \cdot \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$
4. Compute the lower and upper limits of the interval, and write the interval in interval notation. $(\hat{p}_1 - \hat{p}_2 - E, \hat{p}_1 - \hat{p}_2 + E)$
5. Write a conclusion in context. Interpret the interval.

Apply this process to the following example: In the Bertrand-Mullainathan race/resumé study, mock identical resumés were sent to job placement ads in Chicago and Boston. Each resumé was randomly assigned either a commonly-white or commonly-black name. In total, 246 out of 2445 commonly-white named resumés received a callback and 164 out of 2445 commonly-black named resumés received a callback. We will construct a 95% confidence interval for the true difference in callback rates for resumés with common white names and common black names.

1. The number of commonly-white named resumés who received a callback was _____.

   The number of commonly-white named resumés who did not receive a callback was _____.

   The number of commonly-black named resumés who received a callback was _____.

   The number of commonly-black named resumés who did not receive a callback was _____.

2. The confidence level is _____% so the critical value (rounded to three decimal places) is

$$Z_c = \text{ normaldist().inversecdf } (\_\_\_\_\_) = \_\_\_\_\_ \tag{8.4.1}$$

   .

3. $E = Z_c \cdot \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} = \_\_\_\_\_ \cdot \sqrt{\frac{\_\_\_\_\_(1-\_\_\_\_\_)}{\_\_\_\_\_} + \frac{\_\_\_\_\_(1-\_\_\_\_\_)}{\_\_\_\_\_}}$

4. $(\hat{p}_1 - \hat{p}_2 - E, \hat{p}_1 - \hat{p}_2 + E) =$

5. We are _____% confident that the true _____ _____ of callbacks for resumés with common white names and common black names is between _____% and _____% (among jobs similar to the ones in this study).

   Does the interval suggest there is a difference? Why or why not?

Testing a Claim about the Difference in Two Population Proportions

Four Step Process for Testing a Claim about the Difference between Two Population Proportions

The process we use to test a claim about a population parameter has not changed. The following is a list of familiar steps with the appropriate formulas to use for this situation.

1. State the hypotheses. Define p_1 and p_2. The null hypothesis will always be $H_0 : p_1 = p_2$. The alternative will be one of the following $H_a : p_1 < p_2, H_a : p_1 > p_2$, or $H_a : p_1 \neq p_2$ based on the statement of the claim in the problem.

2. Verify that the sampling distribution is approximately normal by checking that there are at least 10 observed successes and failures in *each* sample. We do this step using the sample data because we do not assume the population parameters are equal to a value in the null hypothesis so we can't compute the expected number of successes and failures in the samples.

3. Compute the pooled proportion $\bar{p} = \dfrac{x_1 + x_2}{n_1 + n_2}$ and the Z-score $Z = \dfrac{\hat{p}_1 - \hat{p}_2}{\sqrt{\dfrac{\bar{p}(1-\bar{p})}{n_1} + \dfrac{\bar{p}(1-\bar{p})}{n_2}}}$. Use the normal distribution to find the P-value (the probability of observing a sample difference as extreme or more extreme than the calculated sample difference just by chance).

4. Make a decision about the null and alternative hypotheses and state a conclusion in context.

Apply this process to the following example: In the Bertrand-Mullainathan race/resumé study, mock identical resumés were sent to job placement ads in Chicago and Boston. Each resumé was randomly assigned either a commonly-white or commonly-black

name. In total, 246 out of 2445 commonly-white named resumés received a callback and 164 out of 2445 commonly-black named resumés received a callback. Do the results give convincing statistical evidence that employers favored commonly-white name applicants (in terms of callbacks)?

1. Let p_1 represent:

   Let p_2 represent:

   H_0:

   H_a:

   Use a _____-tailed test because:

2. The number of commonly-white named resumés who received a callback was _____.

   The number of commonly-white named resumés who did not receive a callback was _____.

   The number of commonly-black named resumés who received a callback was _____.

   The number of commonly-black named resumés who did not receive a callback was _____.

3. $\bar{p} = \frac{x_1 + x_2}{n_1 + n_2} =$ _____ $=$

   $$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\bar{p}(1-\bar{p})}{n_1} + \frac{\bar{p}(1-\bar{p})}{n_2}}} = \frac{\_\_\_\_\_ - \_\_\_\_\_}{\sqrt{\frac{\_\_\_\_\_(1-\_\_\_\_\_)}{\_\_\_\_\_} + \frac{\_\_\_\_\_(1-\_\_\_\_\_)}{\_\_\_\_\_}}} =$$

   P-value is _____.

4. We _____ the null hypothesis, we _____ the alternative hypothesis.

   Conclusion in context:

Reference

[19]Bertrand, Marianne and Sendhil Mullainathan. "Are Emily And Greg More Employable Than Lakisha And Jamal? A Field Experiment On Labor Market Discrimination," *American Economic Review*, 2004, v94(4,Sep), 991-1013. https://www.nber.org/papers/w9873

This page titled 8.4: Inference for a Difference in Two Population Proportions is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Hannah Seidler-Wright.

8.4.1: Exercises

1. When making inference about population proportions, which distribution is used to find critical values and P-values?

2. Patients and physicians often disagree in their assessment of pain intensity. Research[20] has shown that racial bias plays a large role in pain management. Researchers want to determine whether black patients with extremity fractures are less likely to receive emergency analgesics than similarly injured white patients. Researchers observe 127 black patients and 90 white patients at a particular emergency department with isolated long-bone fractures over a 40-month period. They then recorded analgesic administration. Of the black patients, 72 received analgesics, and of the white patients, 67 received analgesics. Compute a 99% confidence interval for the difference in the proportion of black and white patients who receive analgesics for pain in this emergency department. Show all steps of constructing a confidence interval.

3. Suppose you would like to know if there is a difference between how crime is depicted in the news and what crime looks like in reality. You find that in 148 news reports of crime, 54 involved a black suspect and a white victim. Of the 110 crime reports you randomly survey, only 18 truly involve a black suspect and a white victim. Is there a significant difference between the crime rates of black suspects with white victims depicted in the news and the crime rates of black suspects with white victims, at a 1% level of significance? Show all steps of conducting a hypothesis test.

Reference

[20] Knox H. Todd, Christi Deaton, Anne P. D'Adamo, Leon Goe. Ethnicity and analgesic practice, Annals of Emergency Medicine, Volume 35, Issue 1, 2000, https://doi.org/10.1016/S0196-0644(00)70099-0. (https://www.sciencedirect.com/science/article/pii/S0196064400700990)

This page titled 8.4.1: Exercises is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Hannah Seidler-Wright.

CHAPTER OVERVIEW

9: Linear Regression

9.1: Scatterplots

9.1.1: Exercises

9.2: Quantifying Direction and Strength

9.2.1: Exercises

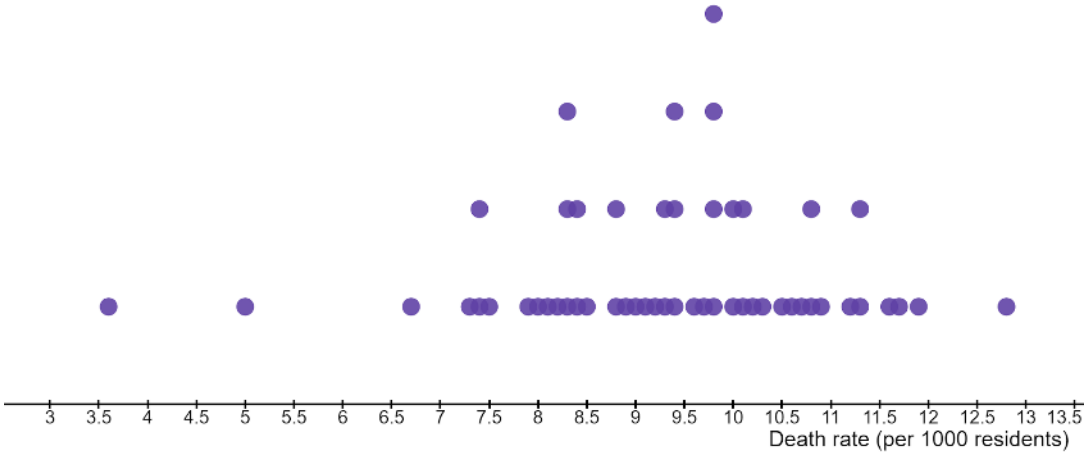
9.3: The Line of Best Fit

9.3.1: Exercises

This page titled 9: Linear Regression is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Hannah Seidler-Wright.

9.1: Scatterplots

Below is the death rate (per 1000 residents) for 53 randomly selected cities. Each dot represents a city. We can describe the shape, center, and spread for the distribution. Using the dotplot, we are able to answer questions about the distribution of death rates for the 53 cities. However, this dotplot does not reveal any possible explanations for the death rates. If we examine these explanations, we may be able to reduce high death rates.



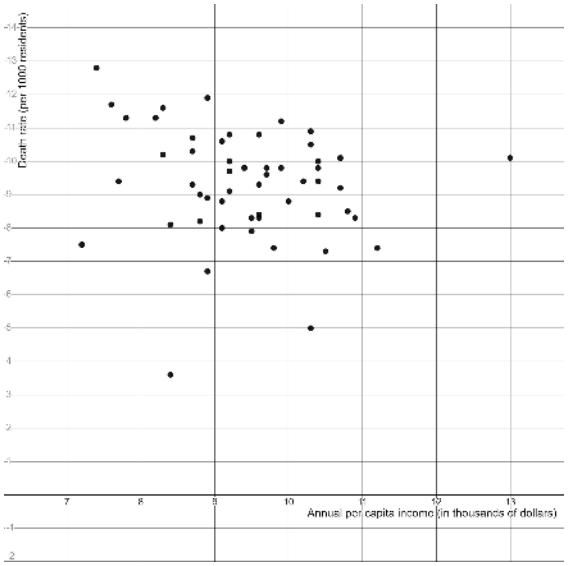
Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

We turn to a new type of graph called a scatterplot. When we have *bivariate data* (data with two quantitative variables), we can express the data in a **scatterplot,** as shown on the graph below. Scatterplots can help us determine relationships between two variables by noticing patterns.

Sometimes, if we observe a strong relationship between two variables, we use the model to make predictions. We try to predict the value of one variable using another variable. The *explanatory variable* is used to make predictions of the *response variable* values, and it is conventional to represent the explanatory variable on the horizontal axis or x-axis of the scatterplot. The response variable is then represented on the vertical axis or y-axis of the scatterplot.

Reading Scatterplots

The scatterplot below describes graphically the relationship between annual per capita income (in thousands of dollars) and the death rate (per 1000 residents) of 53 randomly selected cities.



Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

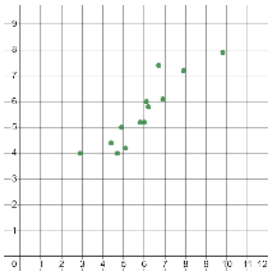
1. For the scatterplot above, what does a point on the scatterplot represent?

2. What are the explanatory and response variables?

3. Estimate the death rate in the city with the highest annual per capita income.

4. Estimate the annual per capita income of the city with the highest death rate.

5. What are two possible categories that could have data/results like are shown in the scatterplot below? Be creative!



Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

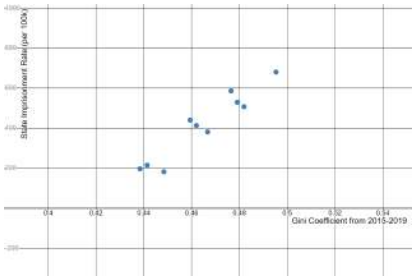
a. A point on the scatterplot could represent:

b. The x-variable could represent:

c. The y-variable could represent:

Recognizing Patterns in Data

The Gini coefficient is a ratio which quantifies the amount of inequality in a population. It is a number between 0 and 1 where 0 represents perfect equality and 1 represents perfect inequality. Below is a scatterplot that shows the Gini coefficient for 2015-2019 and the state imprisonment rate (per 100k) for 10 randomly selected states.



Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

6. Does there appear to be a relationship between the two variables? Explain.

Direction

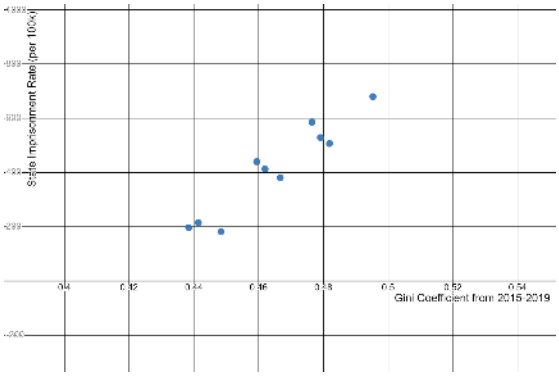
7. Do you expect states with higher Gini coefficients to have lower or higher state imprisonment rates (per 100k)? Explain.

The **direction** of a relationship is a pattern that can be recognized from a scatterplot. If the points trend upward, and as the values of x increase, so do the values of y, we say that the direction is **positive**. If the points trend downward, and as the values of x increase, the values of y decrease, we say that the direction is **negative**.

8. Think of an example of two variables whose scatterplot would have a negative direction.

Strength

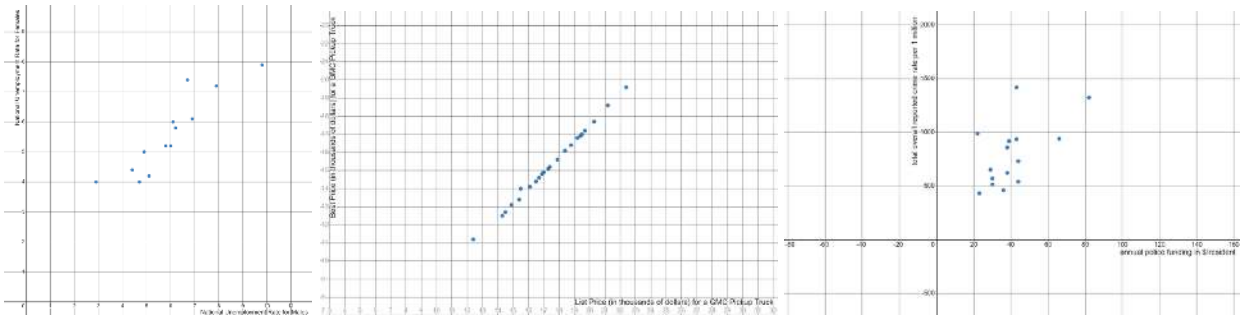
Below is a scatterplot that shows the Gini coefficient for 2015-2019 and the state imprisonment rate (per 100k) for 10 randomly selected states.



Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

9. Use the pattern in the data to predict the state imprisonment rate (per 100k) for a state with a Gini coefficient of 0.52.

10. Of the following three scatterplots, which is the easiest to make predictions from? Explain.



Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

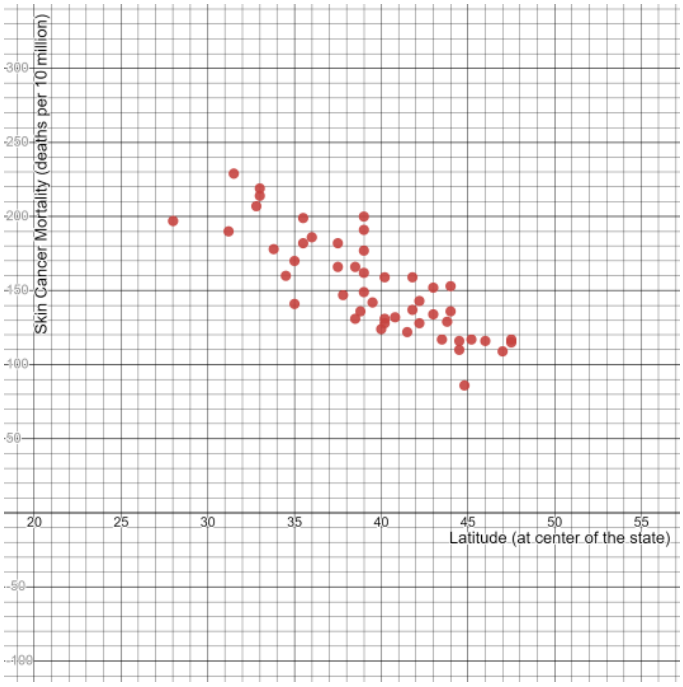
The **strength** of a relationship is another pattern that can be recognized from a scatterplot. The association between two variables is considered **strong** when the points are close to some path or curve. When relationships are strong, it is easier to make predictions using the scatterplot and stronger relationships make for more accurate predictions. The association is **weak** if points are widely scattered from a path or curve.

This page titled 9.1: Scatterplots is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Hannah Seidler-Wright.

- **Current page** by Hannah Seidler-Wright is licensed CC BY-NC-SA 4.0.
- **1.2: The Statistical Analysis Process** by Hannah Seidler-Wright is licensed CC BY-NC-SA 4.0.

9.1.1: Exercises

1. Below is a scatterplot of latitudes (at the center of the state) (degrees north) and skin cancer mortality (deaths per 10 million) for 49 US states from the 1950s.



Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

a. What does a point on the scatterplot represent?

b. What is the explanatory variable? What is the response variable?

c. Estimate the latitude of the state with the highest skin cancer mortality rate.

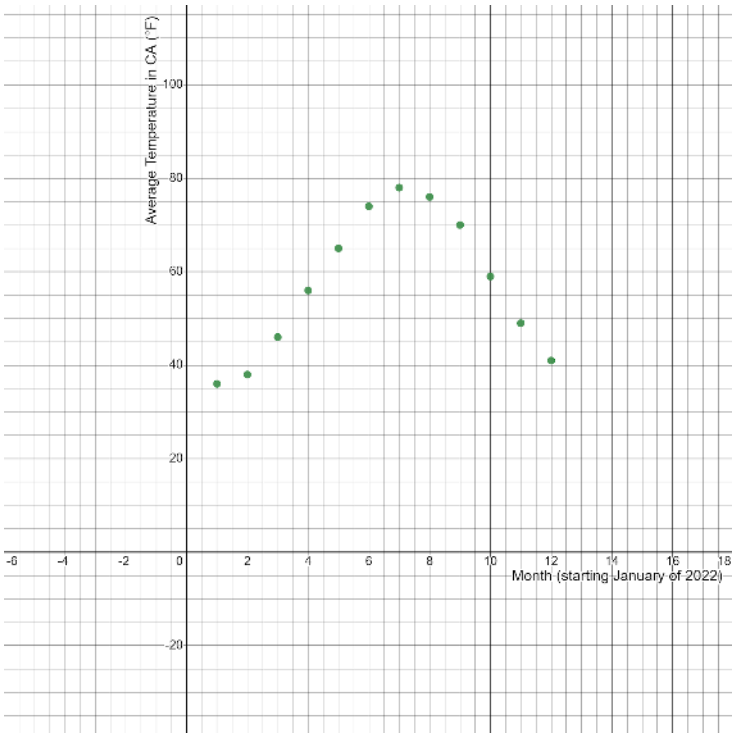
d. Estimate the latitude of the state with the lowest skin cancer mortality rate.

e. Estimate the skin cancer mortality rate of the state with the lowest latitude.

f. What is the direction and strength of the scatterplot? Explain.

g. Predict the skin cancer mortality rate for a state with central latitude of 50 degrees north.

2. Given below is the scatterplot of months (starting January of 2022) and the average temperature in CA in degrees fahrenheit.



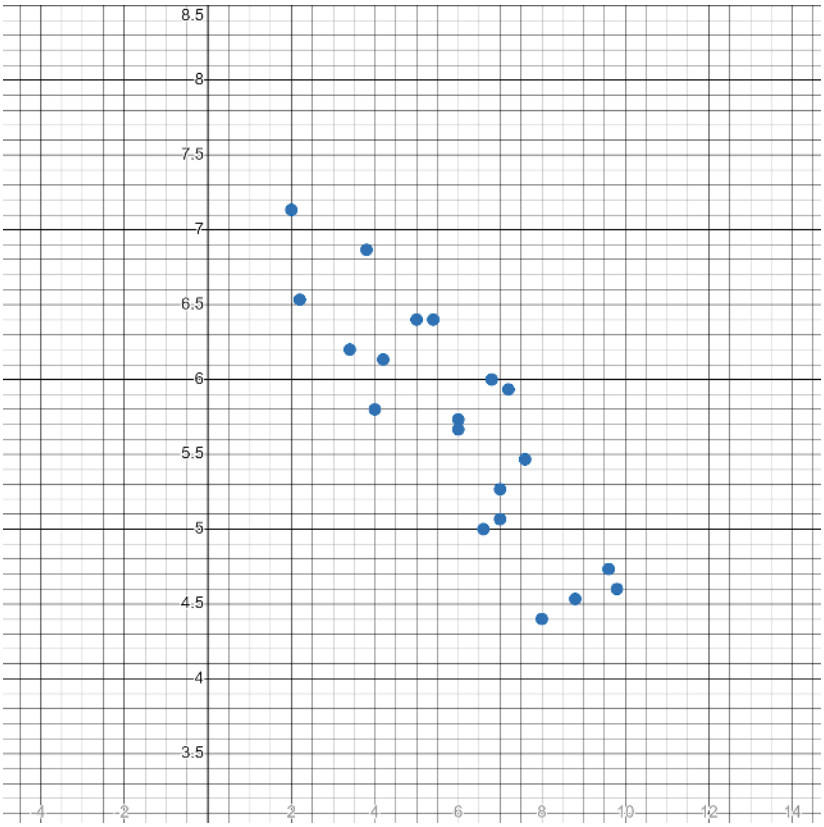
Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

a. Classify the strength of the association. Explain.

b. Use the scatterplot to predict the average temperature in CA in April of 2023 (which corresponds to the x-value 16).

c. Your friend believes that a scatterplot with a linear association (where points follow a path that is a line) is the easiest to make predictions from. Explain why your friend is mistaken.

3. Given below is a scatterplot.



Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

a. What is the strength and direction of the scatterplot? Justify your answer.

b. Think of a possible category that could have data/results like are shown in this scatterplot. What might a point on this scatterplot represent? What does the explanatory variable represent? What does the response variable represent?

This page titled 9.1.1: Exercises is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Hannah Seidler-Wright.

- **Current page** by Hannah Seidler-Wright is licensed CC BY-NC-SA 4.0.
- **1.2: The Statistical Analysis Process** by Hannah Seidler-Wright is licensed CC BY-NC-SA 4.0.

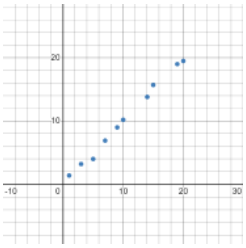
9.2: Quantifying Direction and Strength

In the last lesson, we explored the direction and strength of relationships between two quantitative variables. Now we will begin to explore how to model these relationships. When two variables are related, we say that they **correlate**, and that there is **correlation** between them. Some relationships between the variables in scatterplots can be summarized well by a line. We call such relationships **linear**. Other relationships are better summarized by a curve rather than a line. We call these **nonlinear** or **curvilinear**. Recall that the direction of a relationship can be either positive or negative. Lines and curves can often be fit to the data in a scatterplot. A small amount of deviation away from some line or curve means that the explanatory variable is a good predictor for values of the response variable and we say the relationship is strong. A large amount of deviation from the line or curve indicates that the relationship is weak.

Linear Correlation

One measure of the linear correlation between two variables is called the **linear correlation coefficient**. This correlation coefficient is represented with the letter r. We will use desmos to calculate the value of r for a given scatterplot. Given below is a dataset accompanied by its corresponding scatterplot.

x1	y1
1	1.4
3	3.2
5	4
7	6.9
9	9
10	10.2
14	13.8
15	15.7
19	19
20	19.5



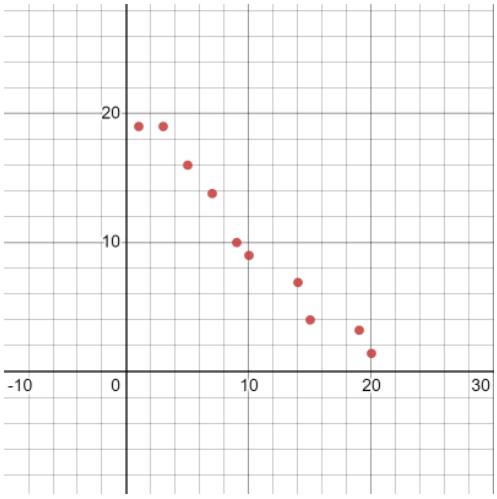
Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

To find the linear correlation coefficient,

1. Go to https://www.desmos.com/calculator.
2. Copy and paste the data set above into line 1, or click the plus icon in the top left corner of the calculator, select table, and enter the values into the table.
3. To find r, type $\text{corr}(x1, y1)$ into line 2.
    a. For the data above, $r = \text{corr}(x1, y1) \approx 0.997$

1. Find the linear correlation coefficient for the following two datasets. The corresponding scatterplots are provided. Round r to three decimal places.
   a. $r = \text{corr}(x2, y2) \approx$ _____

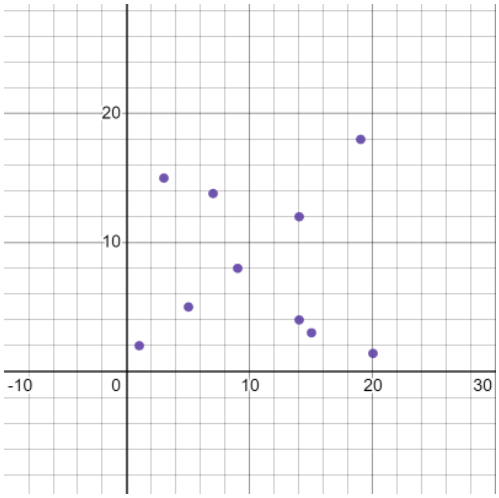
x2	y2
1	19
3	19
5	16
7	13.8
9	10
10	9
14	6.9
15	4
19	3.2
20	1.4



Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

b. $r = \text{corr}(x3, y3) \approx$ _____

x3	y3
1	2
3	15
5	5
7	13.8
9	8
14	4
14	12
15	3
19	18
20	1.4

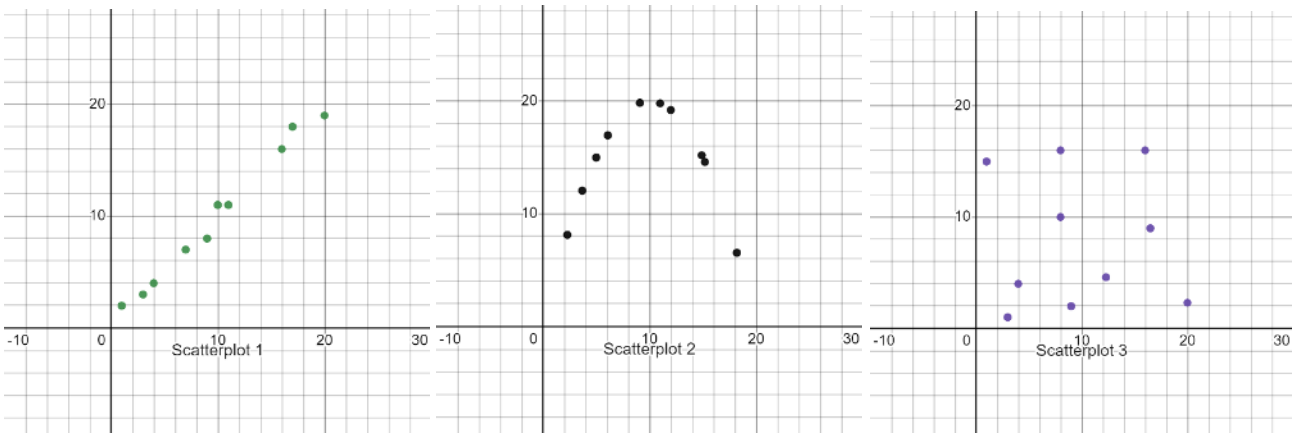


Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

2. What do you think the linear correlation coefficient (r) measures?

3. What is the largest possible value of r? What is the smallest possible value for r?

4. Consider the following scatterplots:

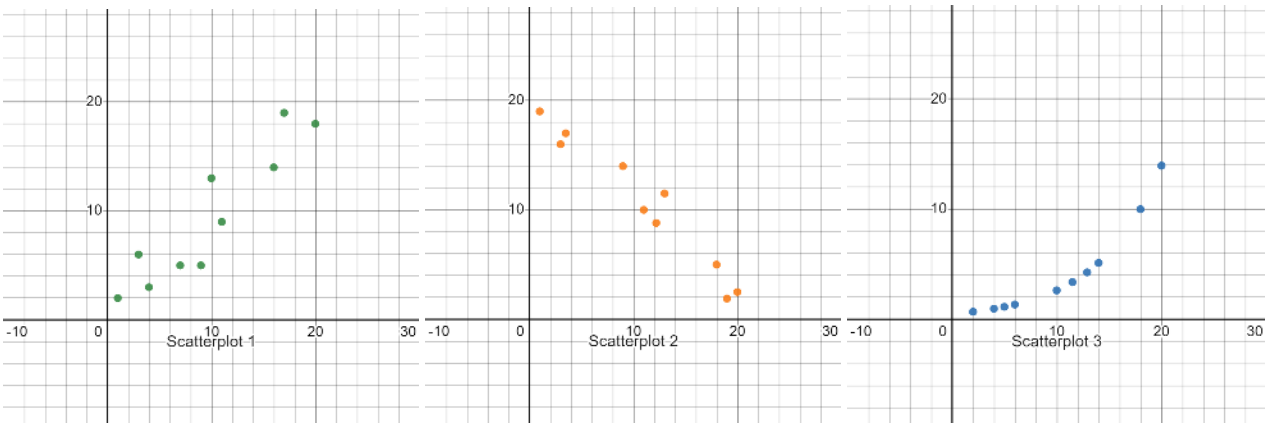


Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

a. Decide which two of these scatterplots have an r-value close to 0 without calculating the r value. Explain why you think this.

b. Do both scatterplots with an r-value close to 0 have a weak relationship? If not, explain why.

5. Consider the following scatterplots:



Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

a. Determine which two scatterplots above have an r-value close to 0.92 without calculating the r-value. Explain why you think this.

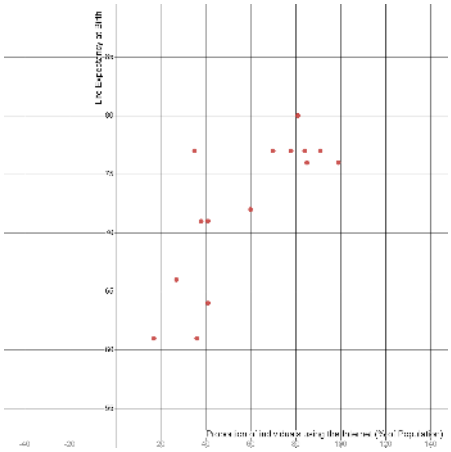
b. Can variables have a nonlinear relationship when r is close to 1? Explain why you think this.

Characteristics of the Linear Correlation Coefficient

- r is between -1 and 1 (inclusive). A scatterplot with an r-value close to -1 or 1 has a strong linear association. Scatterplots with r-values close to 0 have weak linear associations.
- The linear correlation coefficient, r, is a number that describes the direction and strength of a scatterplot with a linear association.
- The sign of r (positive or negative) indicates the direction of a scatterplot with a linear association.

The r-value only gives us reliable information about direction and strength of *linear associations* and should not be used to quantify patterns of a nonlinear association.

6. The scatterplot below shows the rate of individuals using the internet (% of population) and the life expectancy at birth for 15 countries in 2020[21].



Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

a. In this group of 15 countries, does an increase in internet usage tend to be associated with an increase or decrease in life expectancy?

b. The pattern in the scatterplot above indicates a fairly strong, positive, *nonlinear* relationship. Based on this observation, someone might suggest that one way to improve a country's life expectancy would be to get more people online. Is this a reasonable conclusion? Explain.

c. Is the relationship between internet usage and life expectancy one of cause and effect? Consider the type of study that was performed to collect the data.

In an observational study where there is a relationship between two variables, we cannot conclude that the relationship is one of cause and effect. Such conclusions are determined in experimental studies. Sometimes, there is a third variable, possibly unconsidered, that drives changes for both the explanatory and response variables. This additional variable is called a **lurking variable**.

d. Suggest a possible lurking variable that might explain the relationship in internet usage and life expectancy in the data above.

Reference

[21] *World Bank Group*. World Bank. (n.d.). Accessed July 13, 2022, from https://www.worldbank.org/en/home

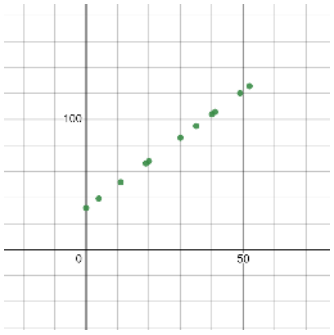
This page titled 9.2: Quantifying Direction and Strength is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Hannah Seidler-Wright.

- **Current page** by Hannah Seidler-Wright is licensed CC BY-NC-SA 4.0.
- **1.2: The Statistical Analysis Process** by Hannah Seidler-Wright is licensed CC BY-NC-SA 4.0.

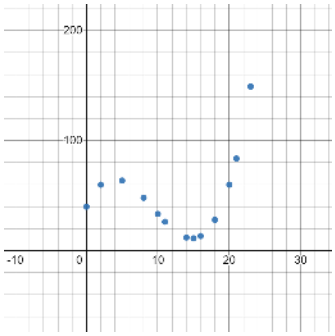
9.2.1: Exercises

1. Match the following scatterplots with the appropriate r value. Classify the association for each r.

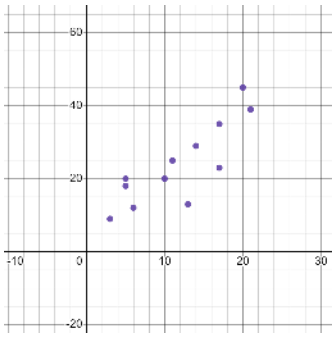
a.



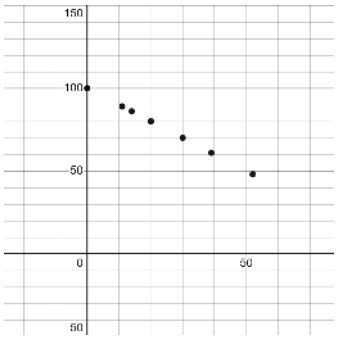
b.



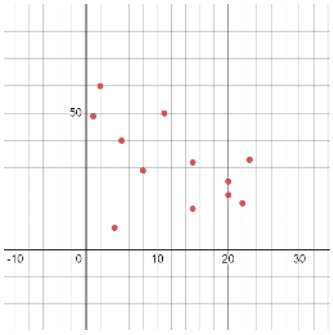
c.



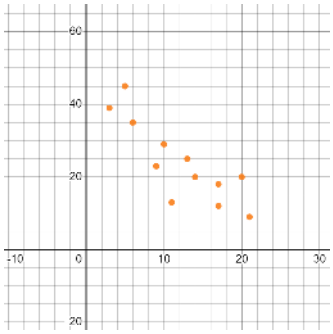
d.



e.



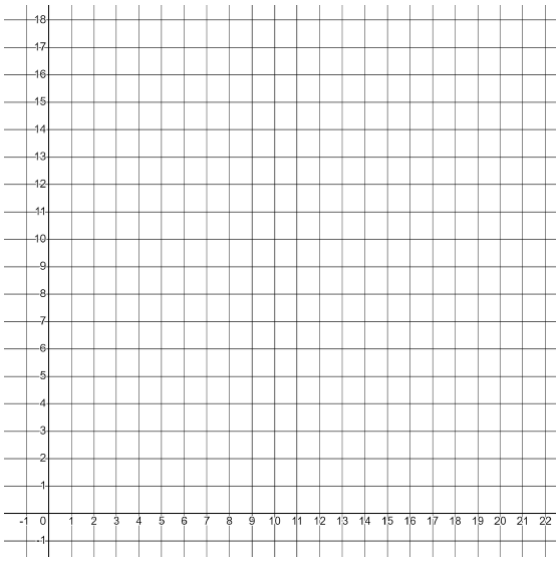
f.



Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

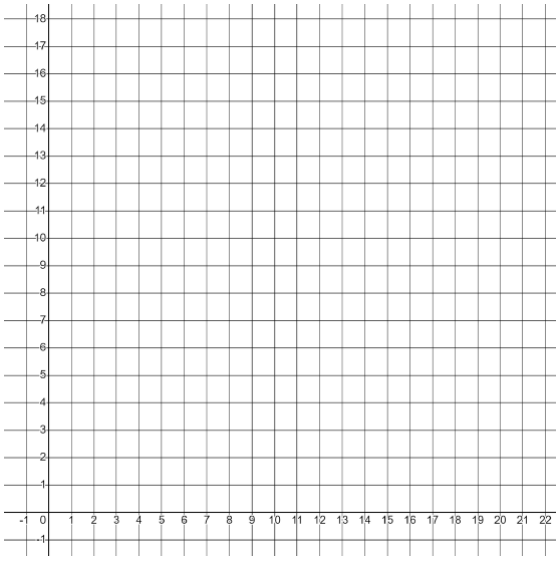
1. $r = -1$ (perfect linear association)
2. $r = 1$ (perfect linear association)
3. $r = 0.304$
4. $r = 0.838$
5. $r = -0.842$
6. $r = -0.483$

2. Sandy says all scatterplots with a linear correlation coefficient close to 0 have a weak relationship. Provide an example of a scatterplot (on the blank graph below) and explain to Sandy why she is mistaken.



Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

3. Travis says all scatterplots with an r-value close to 1 have strong linear associations. Provide an example of a scatterplot (on the blank graph below) and explain to Travis why he is mistaken.



Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

4. Below is data from a local beach that shows the number of ice cream sandwiches sold and the number of drownings on various days during the year.

Number of ice cream sandwiches sold	Number of drownings
57	7
79	6
55	6
23	2
51	4
11	0
30	2
23	3

a. Compute the linear correlation coefficient using desmos. Write the function you used below.

b. Based on the r-value from a, what is the strength and direction of the relationship. Explain.

c. Do you believe that higher ice cream sandwich sales cause more drownings on this beach? If not, what is another possible variable that could explain this relationship?

This page titled 9.2.1: Exercises is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Hannah Seidler-Wright.

- **Current page** by Hannah Seidler-Wright is licensed CC BY-NC-SA 4.0.
- **1.2: The Statistical Analysis Process** by Hannah Seidler-Wright is licensed CC BY-NC-SA 4.0.

9.3: The Line of Best Fit

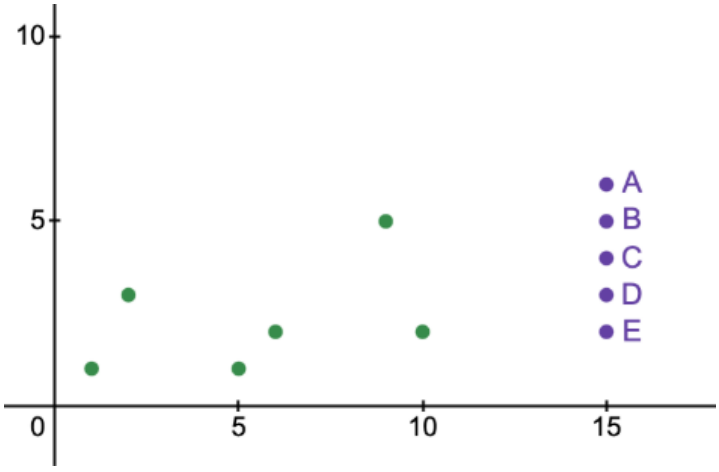
In a previous section, we saw that an r value close to 1 may indicate a strong linear relationship between an explanatory variable and a response variable. In such situations, we may choose to use an equation of a line to summarize the relationship. We can then use a linear model to make predictions about the values of the response variable.

1. Imagine a line that fits the data best. Which of the purple points, A, B, C, D, or E, does the line intersect? Sketch the line below.



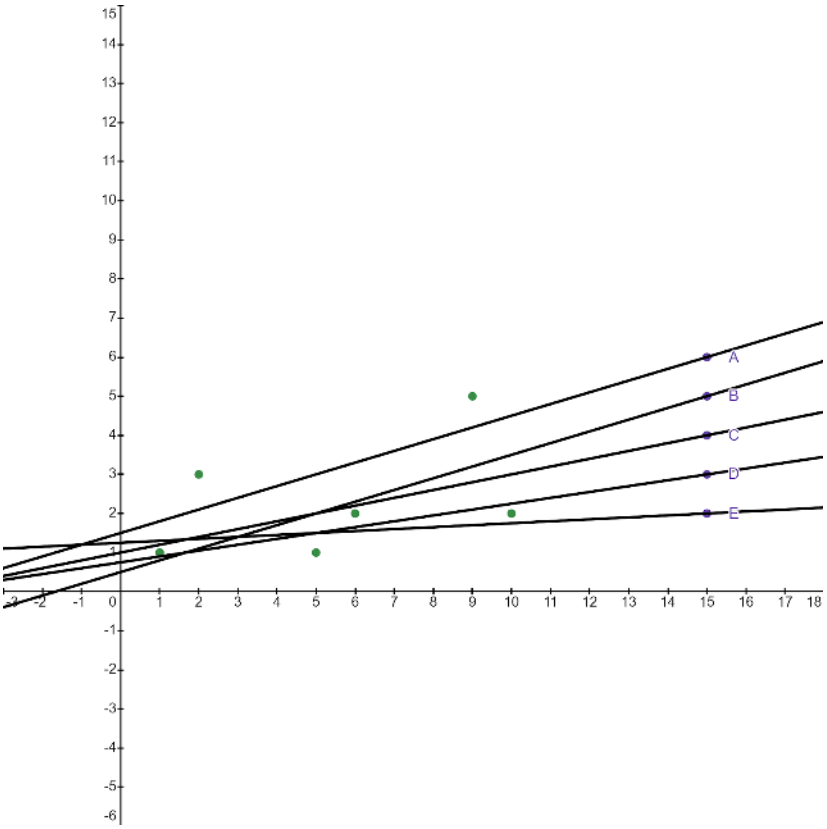
Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

2. Imagine the line in your head continuing out to x=100. Estimate the y-coordinate of the point on the line where x is 100. Explain how you made your guess.



Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

3. Select the line below that corresponds to the choice you made in question 1. Use it's equation to predict the y-coordinate of the point on the line where x is 100.



Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

Line A. $y = 0.3x + 1.5$

Line B. $y = 0.3x + 0.5$

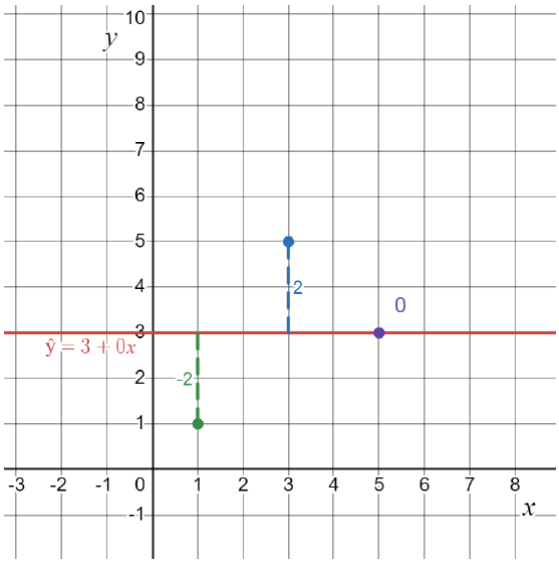
Line C. $y = 0.2x + 1$

Line D. $y = 0.15x + 0.75$

Line E. $y = 0.05x + 1.25$

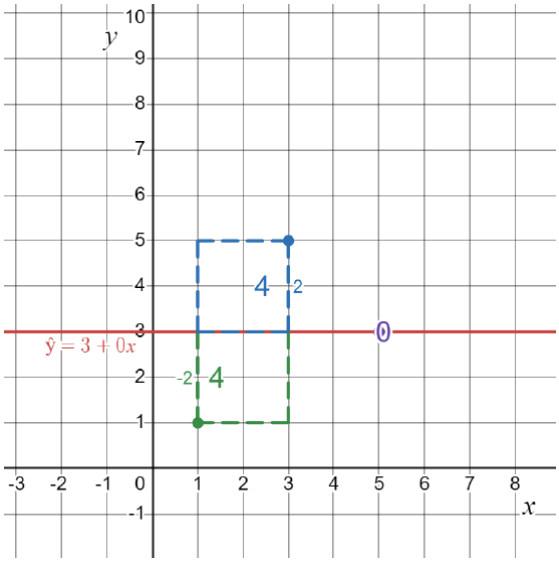
4. Why do you think this line fits the data best?

5. The **least squares regression line (LSR line)** is a linear model that puts roughly half of your scatterplot data above the line and roughly half of your scatterplot data below the line. This line is also called the **line of best fit** as it is the line from which the points in the scatterplot deviate the least from. What do the values -2, 2, and 0 on the graph below represent?



Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

6. What do the values 4, 4, and 0 on the graph represent?



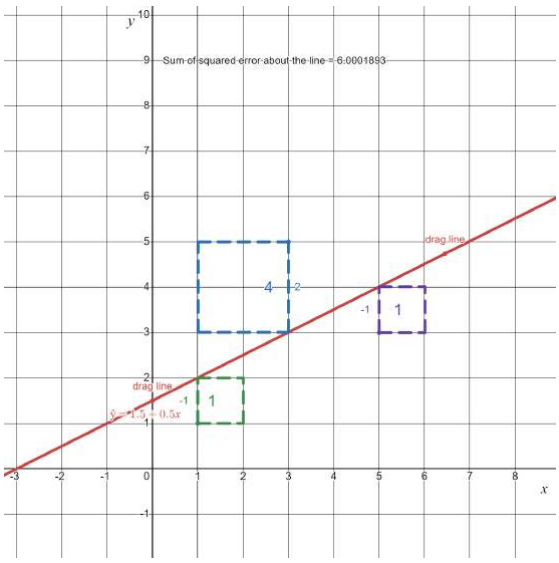
Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

7. Now drag the line by dragging the two red dots at this desmos graph. You can use the QR code below to access the desmos graph.

a. What is the smallest you can make the sum of the squared error about the line?

b. What do you think is true about the line with the smallest sum of squared errors?

For this bivariate set of data, the smallest sum of squared error about the line is 6. The equation of the line of best fit is $\hat{y} = 0.5x + 1.5$.



Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

Residuals

The **residual** is the difference (or vertical distance) between the ACTUAL value of a data set (y) for an x-value and the value that your line PREDICTED $(\hat{y})$ for that x-value.

8. Let's return to the original example. The line of best fit is approximately $\hat{y} = 0.2x + 1$ (the line that passes through point C). Let's look at how much the data deviates from the linear model. The residual for the point (1,1) has been computed and entered in the table below. Complete the table in the same manner.

x	y	$\hat{y}$	$y - \hat{y}$
1	1	$\hat{y} = 0.2(1) + 1 = 1.2$	$1 - 1.2 = -0.2$
2	3		
5	1		
6	2		
9	5		
10	2		

9. Here's one way to measure how well a line fits a data set:

   a. Square all the residuals.

   b. Add up those squares.

The smaller the result, the better the fit. Calculate the square of each residual and enter it in the table. The first squared residual has been computed in the table. Complete the table.

x	Residual: $y - \hat{y}$	Squared Residual
1	-0.2	$(-0.2)^2 = 0.04$
2		
5		
6		
9		
10		

The sum squared residuals is _____.

Calculating and Interpreting Values in the Equation of the LSR Line

10. Use the following steps to calculate the equation of the line of best fit:

a. Open this file that contains randomly collected data comparing the number of cricket chirps and the temperature. Copy the data from the file by highlighting the data and clicking copy.

i. Alternatively, you can access the data set using this QR code:

b. Open https://www.desmos.com/calculator and paste the data into the first line by clicking paste.

c. In the second line, calculate the linear correlation coefficient, r, by typing in $\text{corr}\,(x_1, y_1)$.

$r =$ _____

d. Compute the sample means, $\bar{x} = \text{mean}(x1)$ and $\bar{y} = \text{mean}(y1)$, and sample standard deviations, $s_x = \text{stdev}(x1)$ and $s_y = \text{stdev}(y1)$.

$\bar{x} =$ _____

$\bar{y} =$ _____

$s_x =$ _____

$s_y =$ _____

e. Calculate the slope, $m = \frac{r \cdot s_y}{s_x}$, of the line of best fit.

f. The slope of the line is the predicted change in y for every unit change in x. Interpret the slope of the equation of the line of best fit in context (the units of x are number of chirps, and the units of y are degrees Fahrenheit).

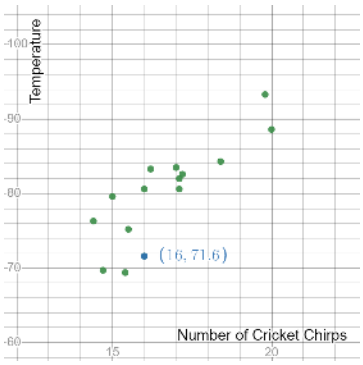
g. Calculate the y-intercept, $b = \bar{y} - m \cdot \bar{x}$, for the line of best fit.

h. The y-intercept of the line of best fit is the predicted y-value when the x-value is 0. Interpret the y-intercept in context.

i. Write the equation you found. Use the line of best fit to predict the temperature near a cricket that chirps 24 times.

$\hat{y} =$ _____.

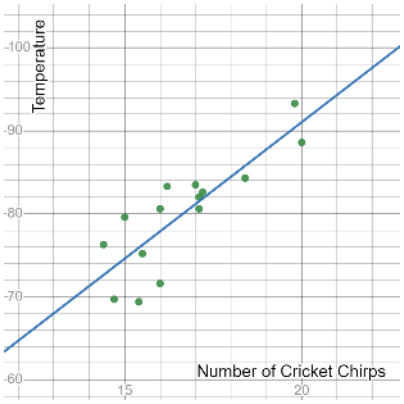
11. One of the crickets chirped 16 times and the temperature near it was 71.6 degrees Fahrenheit. Calculate the residual for this cricket.



Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

12. In desmos, type in y1~mx1+b to check your work.

$\hat{y} =$ _____



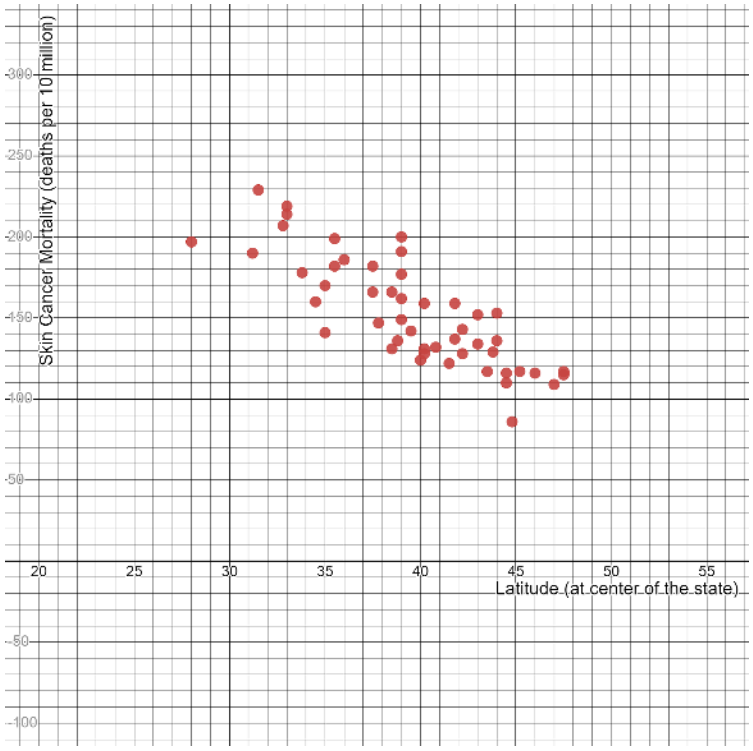
Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

This page titled 9.3: The Line of Best Fit is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Hannah Seidler-Wright.

- **Current page** by Hannah Seidler-Wright is licensed CC BY-NC-SA 4.0.
- **1.2: The Statistical Analysis Process** by Hannah Seidler-Wright is licensed CC BY-NC-SA 4.0.

9.3.1: Exercises

1. Below is a scatterplot of latitudes (at the center of the state) (degrees north) and skin cancer mortality (deaths per 10 million) for 49 US states from the 1950s. Click on the image below or scan the QR code and follow the link to see the set of data in desmos.



Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

a. Compute the linear correlation coefficient, sample mean and sample standard deviation for the explanatory variable, and sample mean and sample standard deviation for the response variable. Round all values to three decimal places.

$r = \text{corr}(x1, y1) =$

$\bar{x} = \text{mean}(x1) =$

$s_x = \text{stdev}(x1) =$

$\bar{y} = \text{mean}(y1) =$

$s_y = \text{stdev}(y1) =$

b. Compute the slope of the line of best fit, $m = r \cdot \frac{s_y}{s_x}$.

c. Compute the y-intercept of the line of best fit, $(0, b = \bar{y} - m \cdot \bar{x})$.

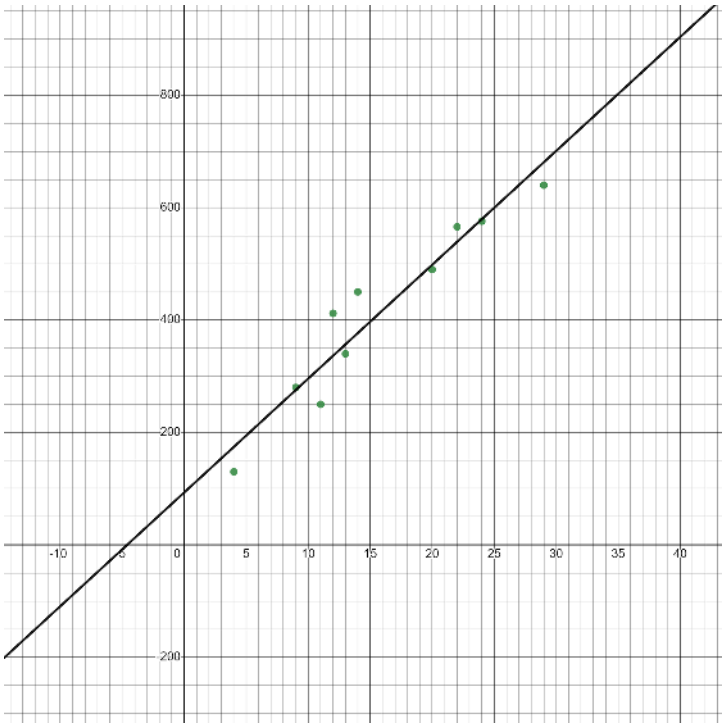
d. Write the equation of the line of best fit, $\hat{y} = mx + b$.

e. Interpret the slope of the line of best fit in context.

f. Interpret the y-intercept of the line of best fit in context.

g. Use the line of best fit to predict the skin cancer mortality rate for a state that has a central latitude of 25 degrees north.

2. Given is a scatterplot of average rainfall and number of plant species for 10 random regions. The line of best fit is shown.



Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

a. The equation of the line of best fit is $\hat{y} = 20.265x + 93.215$. Using the equation, identify the slope and y-intercept of the line of best fit.

b. Interpret the slope and y-intercept in context.

c. Biologists expect the average rainfall to be cut in half in the next year. By how much should they expect the number of plant species to change?

d. Use the line of best fit to predict the number of plant species in a region that gets 60 inches of rainfall on average.

e. In the region that gets 14 inches of rainfall on average, 450 different plant species were observed by biologists. Compute the residual for this region.

f. Based on the residual you found in e, where is the point relative to the line of best fit? Explain.

3. Mathematically, what is the process for creating the line of best fit for bivariate data?

This page titled 9.3.1: Exercises is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Hannah Seidler-Wright.

- **Current page** by Hannah Seidler-Wright is licensed CC BY-NC-SA 4.0.
- **1.2: The Statistical Analysis Process** by Hannah Seidler-Wright is licensed CC BY-NC-SA 4.0.

CHAPTER OVERVIEW

10: Inference Involving More Than Two Parameters

10.1: The Chi-Square Distribution

10.1.1: Exercises

10.2: Goodness-of-Fit

10.2.1: Exercises

10.3: Testing for Independence

10.3.1: Exercises

10.4: ANOVA

10.4.1: Exercises

This page titled 10: Inference Involving More Than Two Parameters is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Hannah Seidler-Wright.

10.1: The Chi-Square Distribution

In previous lessons, we have learned about binomial experiments. A binomial experiment has n independent trials with only two outcomes per trial (success and failure). We will now consider **multinomial experiments** in which we allow two or more outcomes per trial in n independent trials. A binomial experiment is a special case of a multinomial experiment. We will explore multinomial experiments by returning to a hypothesis test for a single population proportion.

College Completion Rates

Research on college completion has shown that about 60% of students who begin college eventually graduate. A publication of higher education claims that the proportion for STEM (science, technology, engineering, math) majors is different. Researchers randomly select 102 STEM majors and determine that 51 eventually graduate.

Step 1.

We will let p represent the proportion of all STEM majors who begin college and ultimately graduate. The null hypothesis is $H_0 : p = 0.60$. The alternative hypothesis is $H_a : p \neq 0.60$. The publication authors have the burden of proof and must produce evidence to support their claim that the proportion of college graduates among STEM majors is different against the assumption that it is not.

Step 2.

Recall, the null hypotheses is

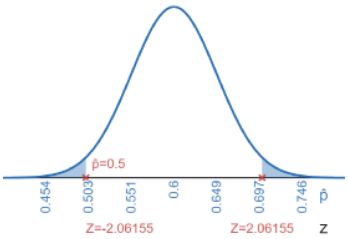
$$H_0 : p = 0.60$$

The number of expected successes in the sample is $np = 102(0.60) = 61.2 \geq 10$. The number of expected failures in the sample is $n(1 - p) = n - np = 102 - 61.2 = 40.8 \geq 10$. Therefore, the sampling distribution of sample proportions is approximately normal.
The sample proportion is

$$\hat{p} = \frac{x}{n} = \frac{51}{102} = 0.5$$

Step 3.

The sampling distribution is shown below. The major tick marks have been labeled with values of $\hat{p}$ and the corresponding Z-scores.



Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

We compute the Z-score for the sample statistic,

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} = \frac{0.5 - 0.60}{\sqrt{\frac{0.60)(1-0.60)}{102}}} \approx -2.06155$$

The sample statistic is 2.06155 standard errors below the assumed population proportion. We perform a two-tailed test because the alternative hypothesis, $H_a : p \neq 0.60$, contains a not equal to inequality. We can now find the P-value, which is the probability of

seeing a sample proportion as extreme or more extreme than 0.5, by finding the probability from the standard normal distribution. The P-value is approximately 0.03925.

Step 4.

The level of significance is 5% which is 0.05 as a decimal. 0.03925<0.05 so we reject the null hypothesis and support the alternative hypothesis. The sample data support the claim that the proportion of all STEM majors who eventually graduate is different than 60%.

Chi-Square Statistic

The test statistic, Z, as in the example above, works well for binomial experiments consisting of only two outcomes per trial. When there are three or more outcomes per trial, things get more complicated and the Z statistics are insufficient. For multinomial experiments, we can use a different test statistic. This statistic is denoted χ^2 (χ is the Greek letter chi and is pronounced like in Cobra Kai, and not like tea). The formula for the test statistic is

$$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}$$

Each outcome corresponds to one term in the sum above. E_i stands for the ith **expected frequency** and O_i stands for the ith **observed frequency**.

Example

To understand the terms in this test statistic, let's apply it to a basic example.

Robin is throwing a party for her daughter's 16th birthday. She has party favors for all the guests to leave with.

Scenario 1. The party is a small gathering and five people attend. Robin only has four party favors.

Scenario 2. The party is huge and 51 people attend. Robin only has 50 party favors.

In which scenario above is Robin's error more noticeable?

We can compute the relative error in each scenario:

Scenario 1. $O = 5, E = 4$, so $\frac{(O-E)^2}{E} = \frac{(5-4)^2}{4} = \frac{1}{4} = 0.25$

Scenario 2. $O = 51, E = 50$, so $\frac{(O-E)^2}{E} = \frac{(51-50)^2}{50} = \frac{1}{50} = 0.02$

When there are fewer people expected to attend, the impact of the error is significantly larger than if the same error is made when many people are expected to attend. The sum of these relative errors make up the chi-square test statistic.

The Normal Distribution vs the Chi-Square Distribution

We return to our original problem: Research on college completion has shown that about 60% of students who begin college eventually graduate. A publication of higher education claims that the proportion for STEM (science, technology, engineering, math) majors is different. Researchers randomly select 102 STEM majors and determine that 51 eventually graduate.

1. What is the observed number of STEM majors who eventually graduate? (These are successes in the experiment).

   $O_1 =$ _____

2. How many STEM majors are expected to graduate based on the null hypothesis?

   $E_1 =$ _____

3. Compute the contribution of this outcome to the χ^2 test statistic.

   $\frac{(O_1-E_1)^2}{E_1} =$ _____

4. What is the observed number of STEM majors who do not eventually graduate? (These are failures in this experiment).

   $O_2 =$ _____

5. How many STEM majors are not expected to graduate based on the null hypothesis?

   $E_2 =$ _____

6. Compute the contribution of this outcome to the χ^2 test statistic.

   $\frac{(O_2-E_2)^2}{E_2} =$ _____

7. Think about the test statistic, where expected frequencies are based on p from the null hypothesis. If the null hypothesis is false, would we expect this statistic to be small or large?

8. The number of STEM majors who did eventually graduate was observed from a random sample, and varies from sample to sample. Once this value is known, is the number of STEM majors who did not eventually graduate random?

9. If **degrees of freedom** represent the number of observed frequencies that vary freely (randomly), how many degrees of freedom are present among the two observed frequencies (of those who did graduate and those that did not)?

10. Add the values from numbers 3 and 6 to find the test statistic (round to three decimal places).

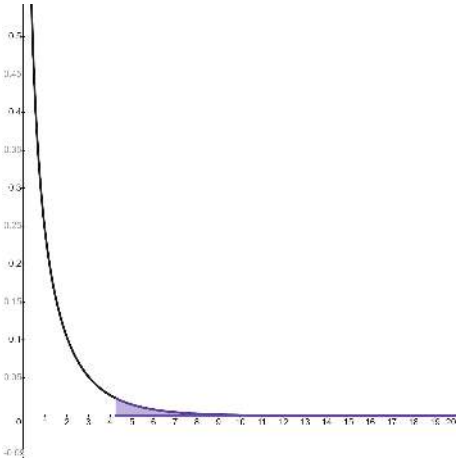
$$\chi^2 = \sum_i \frac{(o_i - E_i)^2}{E_i} = \_\_\_\_\_$$

11. Compute the square of the Z-statistic (round to three decimal places):

$$z^2 = (-2.06155)^2 = \_\_\_\_\_$$

12. How is the χ^2 statistic, with one degree of freedom, related to the Z statistic?

13. Go to this desmos graph, https://www.desmos.com/calculator/bjohldwaym, and enter the degrees of freedom and the test statistic from number 10 to calculate the P-value.



Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

In summary, because the Z test for a population proportion yields the same P-value as the corresponding χ^2 test (with one degree of freedom), the tests are equivalent. We will use this distribution to compare three or more proportions in the next section.

This page titled 10.1: The Chi-Square Distribution is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Hannah Seidler-Wright.

- **Current page** by Hannah Seidler-Wright is licensed CC BY-NC-SA 4.0.
- **1.2: The Statistical Analysis Process** by Hannah Seidler-Wright is licensed CC BY-NC-SA 4.0.

10.1.1: Exercises

1. The proportion of smokers among persons who graduated from a four-year university has been widely reported as 22%. A sociologist student wonders if this is still true. They randomly sample 785 four-year university graduates and finds that 157 are smokers. We test their claim at a 5% level of significance. The beginning of the solution is shown below.

   Step 1: p represents the proportion of four-year university graduates who are smokers.

   $H_0 : p = 0.22$

   $H_a : p \neq 0.22$

   We perform a two-tailed test since there is a not equal to symbol in the alternative hypothesis.

   Step 2: There are 785(0.22)=172.7 expected successes in the sample and 785-172.7=612.3 expected failures in the sample.These are greater than or equal to 10 expected successes and failures in the sample so the sampling distribution of sample proportions is approximately normal.

   $\hat{p} = \frac{157}{785} = 0.2$

   Step 3: $Z = \frac{0.2-0.22}{\sqrt{\frac{0.22(1-0222)}{785}}} \approx -1.35271$

   a. Use the normal distribution to compute the P-value.

   b. Use the Chi-Square distribution to compute the P-value (Use this desmos graph, https://www.desmos.com/calculator/bjohldwaym, to compute the P-value). Remember the test statistic here is the Z-score squared.

   c. What is the benefit of using the Chi-Square distribution instead of the normal distribution?

   d. What are the limitations of using the Chi-Square distribution instead of the normal distribution?

2. In 2022, Governor Ron DeSantis signed a bill into law prohibiting critical race theory from being taught in public schools. A random survey of 300 people found that 183 believe that critical race theory should be a part of public school curricula and the

rest (117 people) did not. We will test the claim that the proportion of people in support of critical race theory being included in curricula is different from the proportion who are not in support. Use a 5% level of significance.

a. The outcome for each of the trials (a randomly selected person) above is the support or non-support of critical race theory being a part of public school curricula. How many possible outcomes (*k*) are there per trial?

b. What are the hypothetical proportions for each outcome (to be used in the null hypothesis)?

H_0 :

$p_1 =$ (The proportion that supports)

$p_2 =$ (The proportion that does not support)

c. Express the alternative hypothesis in a sentence (do not use mathematical symbolism).

H_a :

d. Fill in the expected counts in the table below:

	Observed Counts	Expected Counts
Support	183	
Does not support	117	

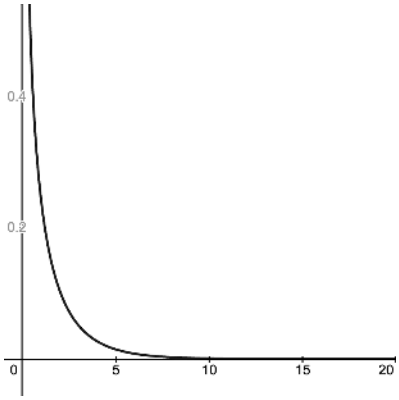
e. Can we use the Chi-Square distribution? Explain.

f. How many degrees of freedom?

g. Compute the test statistic $\chi^2 = \sum \frac{(O-E)^2}{E}$ using the table below:

	Observed Counts O	Expected Counts E	$\frac{(O-E)^2}{E}$
Support	183	150	
Does not support	117	150	
Total:			$\chi^2 =$

h. Below is a desmos graph of the Chi-Square distribution. Label the test statistic and shade the area that represents the P-value.



Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

i. Use this desmos graph, https://www.desmos.com/calculator/bjohldwaym, to compute the P-value.

j. What conclusions can we make about the null and alternative hypothesis?

k. Write a conclusion in context.

This page titled 10.1.1: Exercises is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Hannah Seidler-Wright.

- **Current page** by Hannah Seidler-Wright is licensed CC BY-NC-SA 4.0.
- **1.2: The Statistical Analysis Process** by Hannah Seidler-Wright is licensed CC BY-NC-SA 4.0.

10.2: Goodness-of-Fit

In the last section, we were introduced to the chi-square distribution and its relationship to the normal distribution. We will use the process for χ^2 tests in multinomial experiments (with more than two outcomes per observation). These tests are often referred to as **Goodness-of-Fit tests**.

Step 1: Determine the Hypotheses

The goodness of fit test makes claims about the proportions or probabilities for each outcome of a multinomial experiment. If there are k outcomes per trial, then the null hypothesis would be

$$H_0 : p_1 = \text{ value}_1,\ p_2 = \text{ value}_2, \ldots, p_k = \text{ value}_k$$

The sum of the hypothetical proportions in the null hypothesis must add to 1.

For the null hypothesis to not be true, one or more of the proportions would be incorrect. Thus, the alternative is stated as a sentence, and takes the form

$$H_a : \text{ At least one of these proportions is incorrect.}$$

Step 2: Collect the Data

Using a sample of n independent trials, each having k outcomes, the multinomial experiment is conducted by collecting categorical/qualitative data from a random sample. For each outcome, the observed frequency, O_i, is the number of observed successes. The sum of these observed frequencies is always the sample size,

$$n = \sum_i O_i$$

so the last observed frequency does not vary freely. Therefore, there are $k - 1$ degrees of freedom among the observed frequencies.

To compute the expected frequencies, we take the sample size and multiply by each proportion assumed in the null hypothesis, $E_i = np_i$.

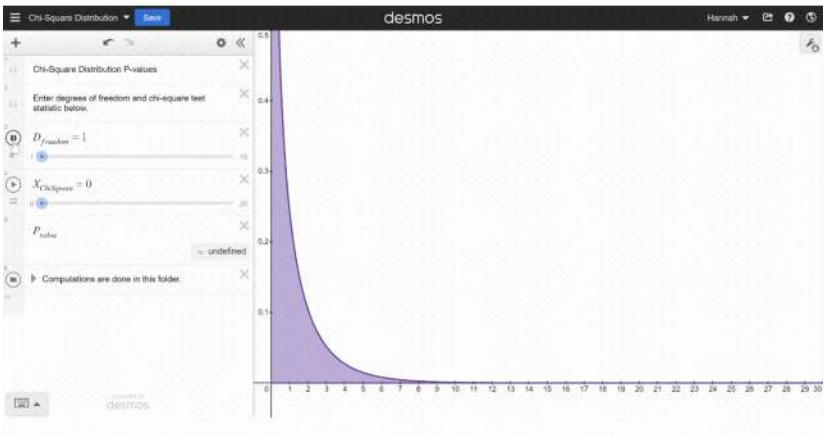
For approximate normality in a binomial experiment, we require at least 10 expected successes and failures in the sample. In a multinomial experiment, the criteria is softened because it can require unreasonably large sample sizes. We say the test statistic is approximately distributed according to the chi-square distribution if each expected frequency, E_i, is at least 5.

Step 3: Assess the Evidence

For goodness of fit tests, the test statistic is

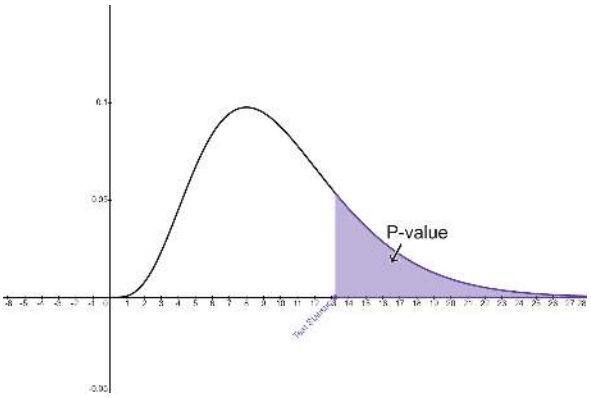
$$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}$$

The shape of the distribution depends on the degrees of freedom. In every case, the distribution starts at 0 and is skewed to the right. The mode (peak) occurs at two less than the number of degrees of freedom.



A goodness of fit test is always right tailed. This is because the test statistic involves squaring an error which will always result in a positive number. If the null hypothesis is false, we expect the test statistic to be large.

We use this desmos graph (https://www.desmos.com/calculator/viuelise2r) to compute the P-value. The P-value is the area of the right tail (starting from the test statistic) under the chi-square distribution with $k-1$ degrees of freedom.



Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

Step 4: Make a Decision and State a Conclusion

Compare the P-value to the level of significance (α). If the P-value is less than or equal to α, we reject the null hypothesis in support of the alternative hypothesis. If the P-value is greater than α, we fail to reject the null hypothesis, and we cannot support the alternative hypothesis.

You try!

Below is the distribution of households total income in 2020 according to the United States Census Bureau:

Income per household	Proportion
Under $35,000	26.2%
Between $35,000 and $100,000	40.3%
Over $100,000	33.5%

You want to know if these proportions are different for black households.

Step 1

1. How many possible outcomes (k) for each black household are there? $k =$ _____

2. State the null hypothesis:

$H_0 : p_1 =$ _____, $p_2 =$ _____, $p_3 =$ _____

Step 2

You randomly survey black households for their household income. The observed frequencies are below.

4. Calculate the sample size and expected frequencies for each category in the table below.

Income per black household	Observed frequency	Expected frequency
Under $35,000	$O_1 = 40$	$E_1 =$ _____
Between $35,000 and $100,000	$O_2 = 41$	$E_2 =$ _____
Over $100,000	$O_3 = 19$	$E_3 =$ _____
Total	$n = \Sigma O_i =$ _____	

5. Do the expected frequencies satisfy the conditions of an approximate chi-square distribution? Explain.

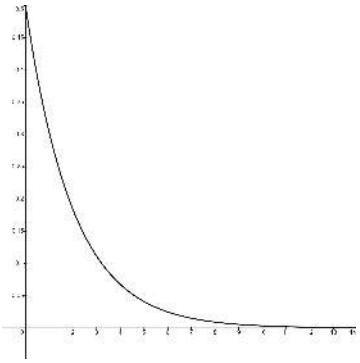
6. What are the degrees of freedom for this test? $df = k - 1 =$ _____

Step 3

7. Fill in the blanks with the appropriate values and compute the χ^2 test statistic rounded to four decimal places:

$$\chi^2 = \sum_{i=1}^{k=3} \frac{(O_i - E_i)^2}{E_i} = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2} + \frac{(O_3 - E_3)^2}{E_3} = \frac{(\_\_\_ - \_\_\_)^2}{\_\_\_} + \frac{(\_\_\_ - \_\_\_)^2}{\_\_\_} + \frac{(\_\_\_ - \_\_\_)^2}{\_\_\_} \approx \_\_\_\_$$

8. Use this desmos graph, https://www.desmos.com/calculator/bjohldwaym, to compute the P-value rounded to four decimal places. Plot the χ^2 statistic and shade the area to the right.



Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

P-value = _____

Step 4

9. Using a 1% level of significance, make a decision about the null and alternative hypotheses.

10. State the conclusion in context.

This page titled 10.2: Goodness-of-Fit is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Hannah Seidler-Wright.

- **Current page** by Hannah Seidler-Wright is licensed CC BY-NC-SA 4.0.
- **1.2: The Statistical Analysis Process** by Hannah Seidler-Wright is licensed CC BY-NC-SA 4.0.

10.2.1: Exercises

1. When performing a Goodness-of-Fit test, what distribution is used to find probabilities?

2. When performing a Goodness-of-Fit test, what type of test is used to find the P-value?

3. performing a Goodness-of-Fit test, what is the alternative hypothesis?

4. Use the Chi-Square distribution desmos graph, https://www.desmos.com/calculator/bjohldwaym, to find the P-value for the test statistic 9.28 using 5 degrees of freedom.

5. Use the Chi-Square distribution desmos graph, https://www.desmos.com/calculator/bjohldwaym, to find the critical value that obtains around 5% in the right tail using 8 degrees of freedom.

6. The marital status distribution of the U.S. male population, ages 15 and older, is as shown in the table below.

Marital Status	Percent
never married	31.3
married	56.1
widowed	2.5
divorced/separated	10.1

Suppose that a random sample of 400 U.S. young adult males, 18 to 24 years old, yielded the following frequency distribution. We are interested in whether this age group of males fits the distribution of the U.S. adult population.

Marital Status	Observed Frequency
never married	140
married	238
widowed	2
divorced/separated	20

a. State the null hypothesis:

$H_0: p_1 =$

$p_2 =$

$p_3 =$

$p_4 =$

b. State the alternative hypothesis:

$H_a:$

c. Calculate the frequency one would expect when surveying 400 people. Fill in the table.

Marital Status	Percent	Expected Frequency
never married	31.3	
married	56.1	
widowed	2.5	
divorced/separated	10.1	

d. Explain why we can use the Chi-Square distribution to compute the P-value.

e. Use the observed frequencies provided in the table below and the expected frequencies you found in part c. to compute each term in the test statistic. Round each term to five decimal places.

Marital Status	Observed Frequency	Expected Frequency	$\frac{(O-E)^2}{E}$
never married	140		
married	238		
widowed	2		
divorced/separated	20		

f. Add all values in the fourth column of the table in e. to find the χ^2 test statistic.

g. How many degrees of freedom are there?

h. Use the Chi-Square distribution desmos graph, https://www.desmos.com/calculator/bjohldwaym, to find the P-value for the test statistic.

i. What can you conclude about the null and alternative hypotheses?

j. State the conclusion in context.

7. Conduct a goodness-of-fit test to determine if the actual college majors of graduating males fit the distribution of their expected majors.Round expected counts to five decimal places. Use this desmos graph, https://www.desmos.com/calculator/bjohldwaym, to

compute the P-value.

Major	Men – Expected Major	Men – Actual Major
Arts & Humanities	11.0%	600
Biological Sciences	6.7%	330
Business	22.7%	1130
Education	5.8%	305
Engineering	15.6%	800
Physical Sciences	3.6%	175
Professional	9.3%	460
Social Sciences	7.6%	370
Technical	1.8%	90
Other	8.2%	400
Undecided	6.6%	340

This page titled 10.2.1: Exercises is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Hannah Seidler-Wright.

10.3: Testing for Independence

In this section, we continue with the χ^2 (chi-square) tests in a special test for independence.

Two-Way Tables

The data used in this problem are based on the article, Attitudes about Marijuana and Political Views[22]. The two-way table below summarizes data on marijuana use and political views from a random sample of 270 adults. Each frequency in an interior cell is a count of the number of adults in the sample that have two specific characteristics. The two-way table is missing several values. We will investigate whether the variables political views and smoking frequency are independent of one-another.

Political Views	Never Smoke	Rarely Smoke	Frequently Smoke	Totals
Liberal	96	35		155
Conservative	43	9		55
Other				60
Totals	173	53	44	270

1. What is the explanatory variable in this study?

2. What is the response variable?

3. Do you think these variables are independent, or dependent? Explain your answer.

4. Enter the missing values into the table above.

5. Compute the following conditional probabilities. Write the probability as a decimal rounded to 2 decimal places.
   a. $P(\text{Conservative} \mid \text{Never Smoke}) =$

   b. $P(\text{Conservative} \mid \text{Frequently Smoke}) =$

   c. Given your previous answer, would you consider marijuana smoking frequency and political views as independent or dependent variables? Explain your answer.

We will now look at the number of degrees of freedom involved in this problem. Below is the two-way table from the start of this section, before values in the “Frequently Smoke” column and “Other” row were entered.

Political Views	Never Smoke	Rarely Smoke	Frequently Smoke	Totals
Liberal	96	35	24	155
Conservative	43	9	3	55
Other	34	9	17	60
Totals	173	53	44	270

6. With the values originally given in the table, are the values you entered free, random values, or dependent on other values? Explain.

7. If the degrees of freedom in the observed frequencies (O) are the number of free, independent observations, how many degrees of freedom are there among the observed frequencies in the table above?

8. Suppose a two-way table has r rows and c columns (don’t count the totals). Make a rule for the degrees of freedom among the observed frequencies in the table. Express this rule as a formula for the degrees of freedom.

   $df =$ _____

Hypothesis Testing Process

Now that we understand the degrees of freedom in a two-way table, it is time to think about a hypothesis test. This test is similar to the goodness-of-fit test already discussed, but the degrees of freedom are different (as discussed) and the expected frequencies have a special formula.

We are conducting a test for independence. In the prior statistical study, the question is, “Are political views independent of marijuana smoking frequency?” The data in the two-way table are from a random sample. By examining conditional probabilities in the sample data, we saw that the variables appear to not be independent. We will now perform a chi-square goodness-of-fit test to examine whether the variables are independent across the entire population.

To conduct a test for independence, we construct expected frequencies based on the assumption that these variables are independent (this will be our null hypothesis).

If events A and B are independent then $P(A \cap B) = P(A) \cdot P(B)$. Suppose $A = \text{liberal}$, and $B = \text{never smoke}$. Refer back to your completed two-way table.

9. Suppose we randomly pick an adult from the sample. If "being liberal” and “never smoking” are independent events, what is the probability that an adult is liberal and never smokes? Round the answer to four decimal places.

$P(\text{liberal } \cap \text{ never smoke}) = P(\text{liberal}) \cdot P(\text{never smoke}) =$ _____

This probability is the proportion of people who should be in the liberal and never smoke category if the events are independent. If this is p, the population proportion of people in this category, then the expected frequency (the number of people in this sample expected to be in this category) is $E = np$. In this formula, n is the sample size (grand total of the two-way table).

10. Find E. Round to two decimal places. $E = np =$ _____

The expected frequency is actually quite close to what was observed, $O = 96$. For this event, the null hypothesis of independence seems to lead to reliable predictions of what actually occurred. Keep in mind, this expected frequency is computed based on our null hypothesis that assumes that our variables are independent. Look once more as we backtrack through the computation which we performed to compute the expected frequency.

$$E = np = 270 \cdot 0.3678 = 270 \cdot \frac{155}{270} \cdot \frac{173}{270} = \frac{155 \cdot 173}{270} = \frac{\text{row total } \cdot \text{ column total}}{\text{grand total}}$$

The key here is that the expected frequency of an event can be found directly from the row, column and grand totals. In general, To compute an expected frequency for an observation in a given row and column of a two-way table, use the formula,

$$E = \frac{\text{row total } \cdot \text{ column total}}{\text{grand total}}$$

Step 1: Determine the Hypotheses

We are now ready to conduct a test for independence using a two-way table. We want to test if political views are independent of marijuana smoking frequency at the 1% significance level.

The hypotheses for this test are:

H_0: The explanatory and response variables are independent.

H_a: The explanatory and response variables are dependent.

11. Write the null and alternative hypotheses in the context of the current problem (naming the explanatory and response variables).

Step 2: Collect the Data

The test for independence between bivariate categorical variables requires that data be summarized in a two way table. Row and column totals should be computed so that expected frequencies can be computed. The expected frequencies are computed with the formula

$$E = \frac{\text{row total} \cdot \text{column total}}{\text{grand total}}$$

As with the previous goodness-of-fit test, we require that each expected frequency is at least 5.

12. Compute the expected frequencies for the cells in the first row. Round the values to two decimal places. The expected frequencies of cells in the 2nd and 3rd rows are provided.

Political Views	Never Smoke	Rarely Smoke	Frequently Smoke	Totals
Liberal	96 $E = \frac{155 \cdot 173}{270} = 99.31$	35	24	155
Conservative	43 $E = \frac{55 \cdot 173}{270} = 35.24$	9 $E = \frac{55 \cdot 53}{270} = 10.8$	3 $E = \frac{55 \cdot 44}{270} = 8.96$	55
Other	34 $E = \frac{60 \cdot 173}{270} = 38.44$	9 $E = \frac{60 \cdot 53}{270} = 11.78$	17 $E = \frac{60 \cdot 44}{270} = 9.78$	60
Totals	173	53	44	270

13. Are the criteria satisfied for the approximate χ^2 distribution for us to perform a test for independence? Explain.

Step 3: Assess the Evidence

The test statistic for a χ^2 test for independence is: $\chi^2 = \sum \frac{(O-E)^2}{E}$ This is approximately distributed according to the χ^2 distribution with degrees of freedom equal to: $df = (r-1) \cdot (c-1)$

Here, r is the number of rows in the table and c is the number of columns. Note – these do not include the total row or column. As before, this test for independence is always a right-tailed test.

14. Enter the expected frequencies next to the corresponding observed frequencies in the table below. For each pair, compute the contribution to the χ^2 statistic, $\frac{(O-E)^2}{E}$ and total these values.

$O =$ Observed Frequency	$E =$ Expected Frequency	$\frac{(O-E)^2}{E}$
96	99.31	$\frac{(96-99.31)^2}{99.31} = 0.1103$
43		
34		
35		
9		
9		
24		
3		
17		
Total (χ^2 test statistic rounded to two decimal places)		

15. What are the degrees of freedom for this test statistic? (Don't count the total row or column)

16. Use this desmos graph https://www.desmos.com/calculator/bjohldwaym to find the P-value for this right-tailed test. Round the value to 4 decimal places.

Step 4: Make a Decision and State a Conclusion

17. What conclusion do you make regarding the null and alternative hypotheses? Why? Recall the level of significance is 1%.

18. Write a brief conclusion in the context of this problem.

Reference

[22] *Psychological Reports*, 1973, pp. 1051 to 1054

This page titled 10.3: Testing for Independence is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Hannah Seidler-Wright.

10.3.1: Exercises

The Pew Research Center studies many different groups in the United States. One of the center's projects is the Pew Internet and American Life Project. In this project, the research center learns how people in the United States use computers and technology.

In one study, researchers asked people, "Do you use a computer at your workplace, at school, at home, or anywhere else on at least an occasional basis?" The possible responses to this question were "Yes" and "No". Researchers also recorded information about each respondent's urbanity, that is whether the respondent lived in an "Urban" area (a city), a "Suburban" area (a neighborhood outside city limits), or a "Rural" area (not in a neighborhood).

Researchers obtained the following results, based on a sample of 8,296 individuals:

		Urbanity		
		Urban	Suburban	Rural
Response	Yes	1946	3533	943
("Do you use a computer?")	No	537	835	502

Do these data support the claim that there is a relationship between a person's response to the question about computer use and the person's urbanity? Execute a complete chi-square test for independence for this case. Use a significance level of $\alpha = 0.01$.

1. Step 1: What are the appropriate hypotheses for this test?

2. Step 2: Collect the Data

   The table below displays the row, column and grand totals, and the expected frequencies for all but one cell. Compute and enter in the missing expected frequency. Round the value to two decimal places.

Computer Usage	Urban	Suburban	Rural	Totals
Yes		3381.30	1118.59	6422
No	560.89	986.70	326.41	1874
Totals	2483	4368	1445	8296

3. Step 3: Assess the Evidence

a. Each pair of observed and expected frequencies are provided in the table below. Compute the missing contribution to the χ^2 test statistic. Round the values to two decimal places.

Pairings of Values	O = Observed Frequency	E = Expected Frequency	$\frac{(O-E)^2}{E}$
Urban / Yes	1946		
Urban / No	537	560.89	1.02
Suburban / Yes	3533	3381.30	6.81
Suburban / No	835	986.70	23.32
Rural / Yes	943	1118.59	27.56
Rural / No	502	326.41	94.46

b. All expected frequencies are greater than 5, so we can proceed with the hypothesis test. What is the value of the χ^2 test statistic? Write the value to two decimal places.

c. Use the desmos graph https://www.desmos.com/calculator/bjohldwaym to determine the P-value.

4. Step 4: Make a Decision

At the 1% significance level, write an appropriate conclusion.

This page titled 10.3.1: Exercises is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Hannah Seidler-Wright.

10.4: ANOVA

In the previous sections, we compared proportions from three or more populations. In this section, we will learn about Analysis of Variance, which allows us to compare three or more means. We abbreviate Analysis of Variance as ANOVA. To use One-Way ANOVA, we must make sure that samples are random, independent, and are drawn from each of k populations that are normal with equal variances. The test is *robust* meaning that moderate departures from these assumptions still yield fairly reliable results.

Dentists use metal alloys to make fillings. Many metals, such as pure gold, are too soft to use for fillings, but they can be hardened by adding other metals at different melting temperatures to make alloys.

Researchers used three different methods to make gold alloys for dental fillings. Vickers hardness numbers are recorded for samples of each method below. A Vickers Hardness Number is a measure of a material's hardness. The higher the number, the harder the material. For example, diamonds are extremely hard and have a Vickers hardness number of 10,000.

The table below displays the Vickers hardness numbers for gold alloys from three different methods.

(A) Method 1	**(B) Method 2**	**(C) Method 3**
805	856	608
675	956	927
702	892	926
793	956	861
689	1214	764
804	724	645
919		743
765		

We will use these data to test the claim that the mean hardness number is the same for each alloy method at a 5% level of significance.

Step 1: Determine the Hypotheses

The null hypothesis is that the population means (three or more) are equal which opposes the alternative hypothesis that at least one population mean is different from one of the others.

1. State the null and alternative hypotheses for this hypothesis test.

H_0 :

H_a :

Step 2: Collect the Data

When conducting the ANOVA test by hand, we must determine the mean and variance of each sample.

2. Calculate the mean for each sample using desmos.

$\bar{x}_1 = \text{mean}(A) =$

$\bar{x}_2 = \text{mean}(B) =$

$\bar{x}_3 = \text{mean}(C) =$

3. Compute the variance (rounded to two decimal places) for each sample using the var function in desmos. Recall that the variance is the square of the standard deviation.

$s_1^2 = \text{var}(A) \approx$

$s_2^2 = \text{var}(B) \approx$

$s_3^2 = \text{var}(C) \approx$

Step 3: Assess the Evidence

To conduct the test, a sample is drawn from each of the k populations. When the sample sizes are all equal, the F-statistic uses two different estimates of the variance (σ^2) that is common to each of the k populations. The test statistic is denoted as F, named for the statistician, Ronald Fisher.

$$F = \frac{n \cdot s_{\bar{x}}^2}{s_p^2} = \frac{n \cdot \text{ variance of the sample means}}{\text{mean of the sample variances}} = \frac{\text{variance between the samples}}{\text{variance within the samples}}$$

When the sample sizes are not all equal these variances become more complicated. Variance is the square of standard deviation. We have seen the formula for standard deviation before.

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

Squaring this gives the formula for variance.

$$s^2 = \frac{\Sigma(x-\bar{x})^2}{n-1}$$

The variance between the samples uses $s_{\bar{x}}^2$, the variance of k sample means.

$$s_{\bar{x}}^2 = \frac{\Sigma(\bar{x}-\overline{\bar{x}})^2}{k-1}$$

We can update the variance between the samples by moving the variable sample size into the sum which is the variance of the sample means.

$$\sigma^2 \approx n s_{\bar{x}}^2 = n\frac{\Sigma(\bar{x}-\overline{\bar{x}})^2}{k-1} = \frac{\Sigma n(\bar{x}-\overline{\bar{x}})^2}{k-1}$$

In this formula, k is the number of samples, and $\overline{\bar{x}}$ is the grand mean. The grand mean can be found in two ways:

- The mean of all the observations. So, you find the sum of all of the values in all of the samples and divide by the total number of observations.

- The weighted mean of the sample means. The weighted mean assigns weights to values based on the sample sizes. The weighted mean assigns weights to values based on the sample sizes. This weighted mean can be expressed by the formula below where n refers to a sample size, $\bar{x}$ refers to a sample mean, and N is the sum of the sample sizes, or total number of observations:

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

4. Compute the grand mean.

   a. Multiply each sample mean by its weight - its sample size. In the table below, for each sample, enter the product $n \cdot \bar{x}$ into the corresponding cell in the third column.

   b. In the Totals row of the table, enter in the total number of observations, N, and the sum of the products $n \cdot \bar{x}$.

Method	n	$\bar{x}$	$n \cdot \bar{x}$
1			
2			
3			
Totals		----------	

   c. The grand mean $\overline{\overline{x}}$ is the sum of the products $n \cdot \bar{x}$ divided by N. Enter in the grand mean below. Round to two decimal places.

$$\overline{\overline{x}} = \frac{\Sigma n\bar{x}}{N} = \_\_\_\_\_$$

The numerator of the F-statistic estimates the common population variance (σ^2) using the variance between the samples. The numerator of the variance between the samples is often referred to as the Sum of Squares between the samples $(\text{SS}_{\text{between}})$,

$$\text{SS}_{\text{between}} = \sum n(\bar{x} - \overline{\overline{x}})^2 = n_1\left(\bar{x}_1 - \overline{\overline{x}}\right)^2 + n_2\left(\bar{x}_2 - \overline{\overline{x}}\right)^2 + \ldots + n_k\left(\bar{x}_k - \overline{\overline{x}}\right)^2$$

To compute this, for each sample mean, we find the square of sample mean minus grand mean, and multiply this value by the sample size. We then add all the products together.

5. Compute $\text{SS}_{\text{between}}$ using the table below.

a. For each sample, find the difference between the sample mean and grand mean, square it, then multiply the squared difference by the sample size. Enter the result into the corresponding cell in the third column in the table below.

Method	n	$\bar{x}$	$\mathrm{n} \cdot (\bar{x} - \overline{\bar{x}})^2$
1			
2			
3			
Totals		----------	

b. Sum of Squares between the samples, $\mathrm{SS}_{\text{between}}$, is the sum of the values in the third column in the previous table. Enter the value below. Round to two decimal places.

$$\mathrm{SS}_{\text{between}} = \_\_\_\_\_$$

Dividing $\mathrm{SS}_{\text{between}}$ by its degrees of freedom $(k–1)$ gives the variance between samples. The variance between the samples is labeled $\mathrm{MS}_{\text{between}}$, because it is the Mean of the Squared deviations between the sample means and the grand mean (denoted $\mathrm{MS}_{\text{between}}$),

$$\mathrm{MS}_{\text{between}} = \frac{\sum n(\bar{x} - \overline{\bar{x}})^2}{k-1} = \frac{\mathrm{SS}_{\text{between}}}{k-1}$$

6. Compute $\mathrm{MS}_{\text{between}}$ by dividing $\mathrm{SS}_{\text{between}}$ by its degrees of freedom. Remember that k is the number of samples.

$$\mathrm{MS}_{\text{between}} = \frac{\mathrm{SS}_{\text{between}}}{k-1} = \_\_\_\_\_$$

The denominator of the F-statistic estimates the common population variance (σ^2) using the variance within the samples. When the sample sizes are different, this is the weighted mean of the sample variances, using the degrees of freedom (n − 1) as weights.

$$s_p^2 = \frac{\Sigma(n-1)s^2}{N-k}$$

The numerator of the variance within the samples is the corresponding sum of squares $(\mathrm{SS}_{\text{within}})$,

$$\mathrm{SS}_{\text{within}} = \sum(n-1)s^2 = (n_1 - 1)\, s_1^2 + (n_2 - 1)\, s_2^2 + \ldots + (n_k - 1)\, s_k^2$$

7. Compute $\mathrm{SS}_{\text{within}}$ using the table below.

a. For each sample, multiply the degrees of freedom by variance. Enter the result into the corresponding cell in the third column in the table below. Round to the nearest whole number.

Method	n	s^2	$(n-1)\cdot s^2$
1			
2			
3			
Totals		----------	

b. Sum of Squares within the samples, $\mathrm{SS}_{\text{within}}$, is the sum of the values in the third column in the previous table. Enter the value below.

$$\mathrm{SS}_{\text{within}} = \_\_\_\_\_$$

The degrees of freedom for $\mathrm{SS}_{\text{within}}$ is the sum of the degrees of freedom for each variance that it uses, $\Sigma(n-1) = N-k$. The variance within the samples is labeled $\mathrm{MS}_{\text{within}}$, because it is the Mean of the Squared deviations within the samples. This is $\mathrm{SS}_{\text{within}}$ divided by its degrees of freedom.

$$\mathrm{MS}_{\text{within}} = \frac{\mathrm{SS}_{\text{within}}}{N-k}$$

8. Compute $\mathrm{MS}_{\text{within}}$ below. Round to two decimal places.

$$\mathrm{MS}_{\text{within}} = \frac{\mathrm{SS}_{\text{within}}}{N-k} = \_\_\_\_\_$$

As before, the F-test statistic is the ratio of the variance between the samples $(\mathrm{MS}_{\text{between}})$ and the variance within the samples $(\mathrm{MS}_{\text{within}})$.

$$F = \frac{\mathrm{MS}_{\text{between}}}{\mathrm{MS}_{\text{within}}}$$

9. Compute the F-test statistic for this hypothesis test. Round to two decimal places.

$$F = \frac{\mathrm{MS}_{\text{between}}}{\mathrm{MS}_{\text{within}}} = \_\_\_\_\_$$

The degrees of freedom for the numerator of the test statistic are $\mathrm{df}_1 = \mathrm{k}-1$. The degrees of freedom for the denominator of the test statistic are $\mathrm{df}_2 = \mathrm{N}-\mathrm{k}$.

10. What are the degrees of freedom for variance in the numerator and in the denominator of the F-test statistic?

$$\mathrm{df}_1 = \_\_\_\_\_$$
$$\mathrm{df}_2 = \_\_\_\_\_ \qquad (10.4.1)$$

Both $\mathrm{MS}_{\text{between}}$ and $\mathrm{MS}_{\text{within}}$ are estimates of the common population variance (σ^2), computed under the assumption that the null hypothesis is true. If the null hypothesis is true, and the samples are from populations with the same mean, the two estimates of the common population variance should be similar, and the F-test statistic should be close to one.

When the null hypothesis is false, the sample means should be very different, and their standard deviation, $s_{\bar{x}}$, will be large, causing the variance between the samples to be larger than the variance within the samples. This will yield a large F-test statistic and a small P-value in the right tail of the distribution of F-statistics. This is why the F-test is always a right-tailed test.

11. We use the desmos graph, https://www.desmos.com/calculator/cmjmq0smlb, to determine P-values for this F-test. Compute the P-value for this hypothesis test.

Step 4: Make a Decision and State a Conclusion

12. What can we conclude about the null and alternative hypotheses?

13. Write a conclusion to the hypothesis test in the context of this problem.

Summary

When conducting a test comparing $k = 3$ or more means using One-Way ANOVA with unequal sample sizes, the null and alternative hypotheses are

$$\begin{aligned} &H_0 : \mu_1 = \mu_2 = ... = \mu_k \\ &H_a : \text{At least one population mean is different from the others.} \end{aligned} \tag{10.4.2}$$

The test statistic is

$$F = \frac{\text{MS}_{\text{between}}}{\text{MS}_{\text{within}}}$$

where

$$\text{MS}_{\text{between}} = \frac{\text{SS}_{\text{between}}}{k-1} = \frac{n_1\left(\bar{x}_1 - \overline{\bar{x}}\right)^2 + n_2\left(\bar{x}_2 - \overline{\bar{x}}\right)^2 + ... + n_k\left(\bar{x}_k - \overline{\bar{x}}\right)^2}{k-1}$$

and

$$\text{MS}_{\text{within}} = \frac{\text{SS}_{\text{within}}}{N-k} = \frac{(n_1-1)\,s_1^2 + (n_1-1)\,s_2^2 + ... + (n_1-1)\,s_k^2}{N-k}$$

In these formulas, $\overline{\bar{x}} = \frac{\Sigma n x}{N}$ and $N = \Sigma n$. The degrees of freedom for the numerator are $\text{df}_1 = \text{k} - 1$ and the degrees of freedom for the denominator are $\text{df}_2 = \text{N} - \text{k}$. This test is always right-tailed.

Use this desmos graph, https://www.desmos.com/calculator/cmjmq0smlb, to compute the P-value.

This page titled 10.4: ANOVA is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Hannah Seidler-Wright.

10.4.1: Exercises

There are many ANOVA technologies available that simplify the testing process. One such way is by using this desmos graph, https://www.desmos.com/calculator/eclxv8cdgq. Use this calculator by entering the data values into the sets labeled l_1, l_2, and l_3.

1. A fourth grade class is studying the environment. One of the assignments is to grow bean plants in different soils. Tommy chose to grow his bean plants in soil found outside his classroom mixed with dryer lint. Tara chose to grow her bean plants in potting soil bought at the local nursery. Nick chose to grow his bean plants in soil from his mother's garden. No chemicals were used on the plants, only water. They were grown inside the classroom next to a large window. Each child grew five plants. At the end of the growing period, each plant was measured, producing the data (in inches) in the table below. Does it appear that the three media in which the bean plants were grown produce the same mean height? Test at a 3% level of significance.

Tommy's Plants	Tara's Plants	Nick's Plants
24	25	23
21	31	27
23	23	22
30	20	30
23	28	20

2. Are the means for the final exams the same for all statistics class delivery types? The table below shows the scores on final exams from several randomly selected classes that used the different delivery types. Assume that all distributions are normal, the four population standard deviations are approximately the same, and the data were collected independently and randomly. Use a level of significance of 5%.

Online	Hybrid	Face-to-Face
72	83	80
84	73	78
77	84	84
80	81	81
81		86
		79
		82

This page titled 10.4.1: Exercises is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Hannah Seidler-Wright.

Index

D

dire

Glossary

Analysis of Variance | also referred to as ANOVA, is a method of testing whether or not the means of three or more populations are equal. The method is applicable if: (1) all populations of interest are normally distributed. (2) the populations have equal standard deviations. (3) samples (not necessarily of the same size) are randomly and independently selected from each population. (4) The test statistic for analysis of variance is the F-ratio. [OpenStax]

Average | a number that describes the central tendency of the data; there are a number of specialized averages, including the arithmetic mean, weighted mean, median, mode, and geometric mean. [OpenStax]

Bernoulli Trials | an experiment with the following characteristics: (1) There are only two possible outcomes called "success" and "failure" for each trial. (2) The probability p of a success is the same for any trial (so the probability $q = 1 - p$ of a failure is the same for any trial). [OpenStax]

Binomial Distribution | a discrete random variable (RV) that arises from Bernoulli trials. There are a fixed number, n, of independent trials. "Independent" means that the result of any trial (for example, trial 1) does not affect the results of the following trials, and all trials are conducted under the same conditions. Under these circumstances the binomial RV X is defined as the number of successes in n trials. The notation is: $X \sim B(n, p) \mu = np$ and the standard deviation is $\sigma = \sqrt{npq}$. The probability of exactly x successes in n trials is $P(X = x) = \binom{n}{x} p^x q^{n-x}$. [OpenStax]

Binomial Experiment | a statistical experiment that satisfies the following three conditions: (1) There are a fixed number of trials, n. (2) There are only two possible outcomes, called "success" and, "failure," for each trial. The letter p denotes the probability of a success on one trial, and q denotes the probability of a failure on one trial. (3) The n trials are independent and are repeated using identical conditions. [OpenStax]

Binomial Probability Distribution | a discrete random variable (RV) that arises from Bernoulli trials; there are a fixed number, n, of independent trials. "Independent" means that the result of any trial (for example, trial one) does not affect the results of the following trials, and all trials are conducted under the same conditions. Under these circumstances the binomial RV X is defined as the number of successes in n trials. The notation is: $X\ B(n, p)$. The mean is $\mu = np$ and the standard deviation is $\sigma = \sqrt{npq}$. The probability of exactly x successes in n trials is $P(X = x) = \binom{n}{x} p^x q^{n-x}$. [OpenStax]

Blinding | not telling participants which treatment a subject is receiving [OpenStax]

Box plot | a graph that gives a quick picture of the middle 50% of the data [OpenStax]

Categorical Variable | variables that take on values that are names or labels [OpenStax]

Central Limit Theorem | Given a random variable (RV) with known mean μ and known standard deviation, σ, we are sampling with size n, and we are interested in two new RVs: the sample mean, $\bar{X}$, and the sample sum, $\sum X$. If the size (n) of the sample is sufficiently large, then $\bar{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$ and $\sum X \sim N(n\mu, (\sqrt{n})(\sigma))$. If the size ($n$) of the sample is sufficiently large, then the distribution of the sample means and the distribution of the sample sums will approximate a normal distributions regardless of the shape of the population. The mean of the sample means will equal the population mean, and the mean of the sample sums will equal n times the population mean. The standard deviation of the distribution of the sample means, $\frac{\sigma}{\sqrt{n}}$, is called the standard error of the mean. [OpenStax]

Central Limit Theorem | Given a random variable (RV) with known mean μ and known standard deviation σ. We are sampling with size n and we are interested in two new RVs - the sample mean, $\bar{X}$, and the sample sum, $\sum X$. If the size n of the sample is sufficiently large, then $\bar{X} - N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$ and $\sum X - N\left(n\mu, \sqrt{n}\sigma\right)$. If the size n of the sample is sufficiently large, then the distribution of the sample means and the distribution of the sample sums will approximate a normal distribution regardless of the shape of the population. The mean of the sample means will equal the population mean and the mean of the sample sums will equal n times the population mean. The standard deviation of the distribution of the sample means, $\frac{\sigma}{\sqrt{n}}$, is called the standard error of the mean. [OpenStax]

Cluster Sampling | a method for selecting a random sample and dividing the population into groups (clusters); use simple random sampling to select a set of clusters. Every individual in the chosen clusters is included in the sample. [OpenStax]

Coefficient of Correlation | a measure developed by Karl Pearson (early 1900s) that gives the strength of association between the independent variable and the dependent variable; the formula is:

$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{\left[n\sum x^2 - (\sum x)^2\right]\left[n\sum y^2 - (\sum y)^2\right]}} \quad (1)$$

where n is the number of data points. The coefficient cannot be more than 1 or less than –1. The closer the coefficient is to ±1, the stronger the evidence of a significant linear relationship between x and y. [OpenStax]

Conditional Probability | the likelihood that an event will occur given that another event has already occurred [OpenStax]

Confidence Interval (CI) | an interval estimate for an unknown population parameter. This depends on: (1) The desired confidence level. (2) Information that is known about the distribution (for example, known standard deviation). (3) The sample and its size. [OpenStax]

Confidence Level (CL) | the percent expression for the probability that the confidence interval contains the true population parameter; for example, if the $CL = 90$, then in 90 out of 100 samples the interval estimate will enclose the true population parameter. [OpenStax]

contingency table | the method of displaying a frequency distribution as a table with rows and columns to show how two variables may be dependent (contingent) upon each other; the table provides an easy way to calculate conditional probabilities. [OpenStax]

Continuous Random Variable | a random variable (RV) whose outcomes are measured; the height of trees in the forest is a continuous RV. [OpenStax]

Control Group | a group in a randomized experiment that receives an inactive treatment but is otherwise managed exactly as the other groups [OpenStax]

Convenience Sampling | a nonrandom method of selecting a sample; this method selects individuals that are easily accessible and may result in biased data. [OpenStax]

Cumulative Relative Frequency | The term applies to an ordered set of observations from smallest to largest. The cumulative relative frequency is the sum of the relative frequencies for all values that are less than or equal to the given value. [OpenStax]

Data | a set of observations (a set of possible outcomes); most data can be put into two groups: qualitative(an attribute whose value is indicated by a label) or quantitative (an attribute whose value is indicated by a number). Quantitative data can be separated into two subgroups: discrete and continuous. Data is discrete if it is the result of counting (such as the number of students of a given ethnic group in a class or the number of books on a shelf). Data is continuous if it is the result of measuring (such as distance traveled or weight of luggage) [OpenStax]

decay parameter | The decay parameter describes the rate at which probabilities decay to zero for increasing values of x. It is the value m in the probability density function $f(x) = me^{(-mx)}$ of an exponential random variable. It is also equal to $m = \frac{1}{\mu}$, where μ is the mean of the random variable. [OpenStax]

Degrees of Freedom (df) | the number of objects in a sample that are free to vary. [OpenStax]

Dependent Events | If two events are NOT independent, then we say that they are dependent. [OpenStax]

Discrete Random Variable | a random variable (RV) whose outcomes are counted [OpenStax]

Double-blinding | the act of blinding both the subjects of an experiment and the researchers who work with the subjects [OpenStax]

Equally Likely | Each outcome of an experiment has the same probability. [OpenStax]

Error Bound for a Population Mean (EBM) | the margin of error; depends on the confidence level, sample size, and known or estimated population standard deviation. [OpenStax]

Error Bound for a Population Proportion (EBP) | the margin of error; depends on the confidence level, the sample size, and the estimated (from the sample) proportion of successes. [OpenStax]

Event | a subset of the set of all outcomes of an experiment; the set of all outcomes of an experiment is called a sample space and is usually denoted by S. An event is an arbitrary subset in S. It can contain one outcome, two outcomes, no outcomes (empty subset), the entire sample space, and the like. Standard notations for events are capital letters such as A, B, C, and so on. [OpenStax]

Expected Value | expected arithmetic average when an experiment is repeated many times; also called the mean. Notations: μ. For a discrete random variable (RV) with probability distribution function $P(x)$,the definition can also be written in the form $\mu = \sum xP(x)$. [OpenStax]

Experiment | a planned activity carried out under controlled conditions [OpenStax]

Experimental Unit | any individual or object to be measured [OpenStax]

Explanatory Variable | the independent variable in an experiment; the value controlled by researchers [OpenStax]

Exponential Distribution | a continuous random variable (RV) that appears when we are interested in the intervals of time between some random events, for example, the length of time between emergency arrivals at a hospital; the notation is $X \sim \text{Exp}(m)$. The mean is $\mu = \frac{1}{m}$ and the standard deviation is $\sigma = \frac{1}{m}$. The probability density function is $f(x) = me^{-mx}$, $x \geq 0$ and the cumulative distribution function is $P(X \leq x) = 1 - e^{mx}$. [OpenStax]

First Quartile | the value that is the median of the of the lower half of the ordered data set [OpenStax]

Frequency | the number of times a value of the data occurs [OpenStax]

Frequency Polygon | looks like a line graph but uses intervals to display ranges of large amounts of data [OpenStax]

Frequency Table | a data representation in which grouped data is displayed along with the corresponding frequencies [OpenStax]

Geometric Distribution | a discrete random variable (RV) that arises from the Bernoulli trials; the trials are repeated until the first success. The geometric variable X is defined as the number of trials until the first success. Notation: $X \sim G(p)$. The mean is $\mu = \frac{1}{p}$ and the standard deviation is $\sigma = \sqrt{\frac{1}{p}\left(\frac{1}{p} - 1\right)}$. The probability of exactly x failures before the first success is given by the formula: $P(X = x) = p(1–p)^{x-1}$. [OpenStax]

Geometric Experiment | a statistical experiment with the following properties: (1) There are one or more Bernoulli trials with all failures except the last one, which is a success. (2) In theory, the number of trials could go on forever. There must be at least one trial. (3) The probability, p, of a success and the probability, q, of a failure do not change from trial to trial [OpenStax]

Hypergeometric Experiment | a statistical experiment with the following properties: (1) You take samples from two groups. (2) You are concerned with a group of interest, called the first group. (3) You sample without replacement from the combined groups. (4) Each pick is not independent, since sampling is without replacement. (5) You are not dealing with Bernoulli Trials. [OpenStax]

Hypergeometric Probability | a discrete random variable (RV) that is characterized by: (1) A fixed number of trials. (2) The probability of success is not the same from trial to trial. We sample from two groups of items when we are interested in only one group. X is defined as the number of successes out of the total number of items chosen. Notation: $X \sim H(r, b, n)$, where $r =$ the number of items in the group of interest, $b =$ the number of items in the group not of interest, and $n =$ the number of items chosen. [OpenStax]

Hypothesis | a statement about the value of a population parameter, in case of two hypotheses, the statement assumed to be true is called the null hypothesis (notation H_0) and the contradictory statement is called the alternative hypothesis (notation H_a). [OpenStax]

Hypothesis Testing | Based on sample evidence, a procedure for determining whether the hypothesis stated is a reasonable statement and should not be rejected, or is unreasonable and should be rejected. [OpenStax]

Independent Events | The occurrence of one event has no effect on the probability of the occurrence of another event. Events A and B are independent if one of the following is true: (1) $P(\text{A}|\text{B}) = P(\text{A})$, (2) $P(\text{B}|\text{A}) = P(\text{B})$, (3) $P(\text{A AND B}) = P(\text{A})P(\text{B})$ [OpenStax]

Inferential Statistics | also called statistical inference or inductive statistics; this facet of statistics deals with estimating a population parameter based on a sample statistic. For example, if four out of the 100 calculators sampled are defective we might infer that four percent of the production is defective. [OpenStax]

Informed Consent | Any human subject in a research study must be cognizant of any risks or costs associated with the study. The subject has the right to know the nature of the treatments included in the study, their potential risks, and their potential benefits. Consent must be given freely by an informed, fit participant. [OpenStax]

Institutional Review Board | a committee tasked with oversight of research programs that involve human subjects [OpenStax]

Interval | also called a class interval; an interval represents a range of data and is used when displaying large data sets [OpenStax]

Level of Significance of the Test | probability of a Type I error (reject the null hypothesis when it is true). Notation: α. In hypothesis testing, the Level of Significance is called the preconceived α or the preset α. [OpenStax]

Lurking Variable | a variable that has an effect on a study even though it is neither an explanatory variable nor a response variable [OpenStax]

Mean | a number that measures the central tendency; a common name for mean is "average." The term "mean" is a shortened form of "arithmetic mean." By definition, the mean for a sample (denoted by $\bar{x}$) is $\bar{x} = \frac{\text{Sum of all values in the sample}}{\text{Number of values in the sample}}$, and the mean for a population (denoted by μ) is $\mu = \frac{\text{Sum of all values in the population}}{\text{Number of values in the population}}$. [OpenStax]

Mean of a Probability Distribution | the long-term average of many trials of a statistical experiment [OpenStax]

Median | a number that separates ordered data into halves; half the values are the same number or smaller than the median and half the values are the same number or larger than the median. The median may or may not be part of the data. [OpenStax]

memoryless property | For an exponential random variable X, the memoryless property is the statement that knowledge of what has occurred in the past has no effect on future probabilities. This means that the probability that X exceeds $x + k$, given that it has exceeded x, is the same as the probability that X would exceed k if we had no knowledge about it. In symbols we say that $P(X > x + k|X > x) = P(X > k)$ [OpenStax]

Midpoint | the mean of an interval in a frequency table [OpenStax]

Mode | the value that appears most frequently in a set of data [OpenStax]

Mutually Exclusive | Two events are mutually exclusive if the probability that they both happen at the same time is zero. If events A and B are mutually exclusive, then $P(\text{A AND B}) = 0$. [OpenStax]

Nonsampling Error | an issue that affects the reliability of sampling data other than natural variation; it includes a variety of human errors including poor study design, biased sampling methods, inaccurate information provided by study participants, data entry errors, and poor analysis. [OpenStax]

Normal Distribution | a continuous random variable (RV) with pdf $f(x) = \frac{1}{\sigma\sqrt{2\pi}}e^{\frac{-(x-\mu)^2}{2\sigma^2}}$, where μ is the mean of the distribution and σ is the standard deviation; notation: $X \sim N(\mu, \sigma)$. If $\mu = 0$ and $\sigma = 1$, the RV is called a standard normal distribution. [OpenStax]

Numerical Variable | variables that take on values that are indicated by numbers [OpenStax]

One-Way ANOVA | a method of testing whether or not the means of three or more populations are equal; the method is applicable if: (1) all populations of interest are normally distributed. (2) the populations have equal standard deviations. (3) samples (not necessarily of the same size) are randomly and independently selected from each population. (4) The test statistic for analysis of variance is the F-ratio. [OpenStax]

Outcome | a particular result of an experiment [OpenStax]

Outlier | an observation that does not fit the rest of the data [OpenStax]

p-value | the probability that an event will happen purely by chance assuming the null hypothesis is true. The smaller the p-value, the stronger the evidence is against the null hypothesis. [OpenStax]

Paired Data Set | two data sets that have a one to one relationship so that: (1)both data sets are the same size, and (2) each data point in one data set is matched with exactly one point from the other set. [OpenStax]

Parameter | a number that is used to represent a population characteristic and that generally cannot be determined easily [OpenStax]

Parameter | a numerical characteristic of a population [OpenStax]

Placebo | an inactive treatment that has no real effect on the explanatory variable [OpenStax]

Point Estimate | a single number computed from a sample and used to estimate a population parameter [OpenStax]

Poisson distribution | If there is a known average of λ events occurring per unit time, and these events are independent of each other, then the number of events X occurring in one unit of time has the Poisson distribution. The probability of k events occurring in one unit time is equal to $P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$. [OpenStax]

Poisson Probability Distribution | a discrete random variable (RV) that counts the number of times a certain event will occur in a specific interval; characteristics of the variable: (1) The probability that the event occurs in a given interval is the same for all intervals. (2) The events occur with a known mean and independently of the time since the last event. The distribution is defined by the mean μ of the event in the interval. Notation: $X \sim P(\mu)$. The mean is $\mu = np$. The standard deviation is $\sigma = \sqrt{\mu}$. The probability of having exactly x successes in r trials is $P(X = x) = (e^{-\mu}) \frac{\mu^x}{x!}$. The Poisson distribution is often used to approximate the binomial distribution, when n is "large" and p is "small" (a general rule is that n should be greater than or equal to 20 and p should be less than or equal to 0.05). [OpenStax]

Pooled Proportion | estimate of the common value of p_1 and p_2. [OpenStax]

Population | all individuals, objects, or measurements whose properties are being studied [OpenStax]

Probability | a number between zero and one, inclusive, that gives the likelihood that a specific event will occur [OpenStax]

Probability | a number between zero and one, inclusive, that gives the likelihood that a specific event will occur; the foundation of statistics is given by the following 3 axioms (by A.N. Kolmogorov, 1930's): Let S denote the sample space and A and B are two events in S. Then: (1) $0 \leq P(\mathrm{A}) \leq 1$, (2) If A and B are any two mutually exclusive events, then $P(\mathrm{A\ OR\ B}) = P(\mathrm{A}) + P(\mathrm{B})$ and (3) $P(\mathrm{S}) = 1$. [OpenStax]

Probability Distribution Function (PDF) | a mathematical description of a discrete random variable (*RV*), given either in the form of an equation (formula) or in the form of a table listing all the possible outcomes of an experiment and the probability associated with each outcome. [OpenStax]

Proportion | the number of successes divided by the total number in the sample [OpenStax]

Qualitative Data | See Data. [OpenStax]

Quantitative Data | See Data. [OpenStax]

Random Assignment | the act of organizing experimental units into treatment groups using random methods [OpenStax]

Random Sampling | a method of selecting a sample that gives every member of the population an equal chance of being selected. [OpenStax]

Random Variable (RV) | a characteristic of interest in a population being studied; common notation for variables are upper case Latin letters X, Y, Z,...; common notation for a specific value from the domain (set of all possible values of a variable) are lower case Latin letters x, y, and z. For example, if X is the number of children in a family, then x represents a specific integer 0, 1, 2, 3,.... Variables in statistics differ from variables in intermediate algebra in the two following ways. (1) The domain of the random variable (RV) is not necessarily a numerical set; the domain may be expressed in words; for example, if X = hair color then the domain is {black, blond, gray, green, orange}. (2) We can tell what specific value x the random variable X takes only after performing the experiment [OpenStax]

Relative Frequency | the ratio of the number of times a value of the data occurs in the set of all outcomes to the number of all outcomes to the total number of outcomes [OpenStax]

Representative Sample | a subset of the population that has the same characteristics as the population [OpenStax]

Response Variable | the dependent variable in an experiment; the value that is measured for change at the end of an experiment [OpenStax]

Sample | a subset of the population studied [OpenStax]

Sample Space | the set of all possible outcomes of an experiment [OpenStax]

Sampling Bias | not all members of the population are equally likely to be selected [OpenStax]

Sampling Distribution | Given simple random samples of size n from a given population with a measured characteristic such as mean, proportion, or standard deviation for each sample, the probability distribution of all the measured characteristics is called a sampling distribution. [OpenStax]

Sampling Error | the natural variation that results from selecting a sample to represent a larger population; this variation decreases as the sample size increases, so selecting larger samples reduces sampling error. [OpenStax]

Sampling with Replacement | Once a member of the population is selected for inclusion in a sample, that member is returned to the population for the selection of the next individual. [OpenStax]

Sampling without Replacement | A member of the population may be chosen for inclusion in a sample only once. If chosen, the member is not returned to the population before the next selection. [OpenStax]

Simple Random Sampling | a straightforward method for selecting a random sample; give each member of the population a number. Use a random number generator to select a set of labels. These randomly selected labels identify the members of your sample. [OpenStax]

Skewed | used to describe data that is not symmetrical; when the right side of a graph looks "chopped off" compared the left side, we say it is "skewed to the left." When the left side of the graph looks "chopped off" compared to the right side, we say the data is "skewed to the right." Alternatively: when the lower values of the data are more spread out, we say the data are skewed to the left. When the greater values are more spread out, the data are skewed to the right. [OpenStax]

Standard Deviation | a number that is equal to the square root of the variance and measures how far data values are from their mean; notation: *s* for sample standard deviation and σ for population standard deviation. [OpenStax]

Standard Deviation of a Probability Distribution | a number that measures how far the outcomes of a statistical experiment are from the mean of the distribution [OpenStax]

Standard Error of the Mean | the standard deviation of the distribution of the sample means, or $\frac{\sigma}{\sqrt{n}}$. [OpenStax]

Standard Normal Distribution | a continuous random variable (RV) $X \sim N(0, 1)$; when X follows the standard normal distribution, it is often noted as \(Z \sim N(0, 1)\. [OpenStax]

Statistic | a numerical characteristic of the sample; a statistic estimates the corresponding population parameter. [OpenStax]

Stratified Sampling | a method for selecting a random sample used to ensure that subgroups of the population are represented adequately; divide the population into groups (strata). Use simple random sampling to identify a proportionate number of individuals from each stratum. [OpenStax]

Student's *t*-Distribution | investigated and reported by William S. Gossett in 1908 and published under the pseudonym Student. The major characteristics of the random variable (RV) are: (1) It is continuous and assumes any real values. (2) The pdf is symmetrical about its mean of zero. However, it is more spread out and flatter at the apex than the normal distribution. (3) It approaches the standard normal distribution as n gets larger. (4) There is a "family" of t-distributions: every representative of the family is completely defined by the number of degrees of freedom which is one less than the number of data items. [OpenStax]

Systematic Sampling | a method for selecting a random sample; list the members of the population. Use simple random sampling to select a starting point in the population. Let k = (number of individuals in the population)/(number of individuals needed in the sample). Choose every kth individual in the list starting with the one that was randomly selected. If necessary, return to the beginning of the population list to complete your sample. [OpenStax]

The AND Event | An outcome is in the event A AND B if the outcome is in both A AND B at the same time. [OpenStax]

The Complement Event | The complement of event A consists of all outcomes that are NOT in A. [OpenStax]

The Conditional Probability of *A* GIVEN *B* | $P(\mathrm{A}|\mathrm{B})$ is the probability that event A will occur given that the event B has already occurred. [OpenStax]

The Conditional Probability of One Event Given Another Event | *P*(*A*|*B*) is the probability that event *A* will occur given that the event *B* has already occurred. [OpenStax]

The Law of Large Numbers | As the number of trials in a probability experiment increases, the difference between the theoretical probability of an event and the relative frequency probability approaches zero. [OpenStax]

The Or Event | An outcome is in the event A OR B if the outcome is in A or is in B or is in both A and B. [OpenStax]

The OR of Two Events | An outcome is in the event *A* OR *B* if the outcome is in *A*, is in *B*, or is in both *A* and *B*. [OpenStax]

Treatments | different values or components of the explanatory variable applied in an experiment [OpenStax]

Tree Diagram | the useful visual representation of a sample space and events in the form of a "tree" with branches marked by possible outcomes together with associated probabilities (frequencies, relative frequencies) [OpenStax]

Type 1 Error | The decision is to reject the null hypothesis when, in fact, the null hypothesis is true. [OpenStax]

Type 2 Error | The decision is not to reject the null hypothesis when, in fact, the null hypothesis is false. [OpenStax]

Uniform Distribution | a continuous random variable (RV) that has equally likely outcomes over the domain, $a < x < b$; it is often referred as the rectangular distribution because the graph of the pdf has the form of a rectangle. Notation: $X \sim U(a, b)$. The mean is $\mu = \frac{a+b}{2}$ and the standard deviation is $\sigma = \sqrt{\frac{(b-a)^2}{12}}$. The probability density function is $f(x) = \frac{1}{b-a}$ for $a < x < b$ or $a \leq x \leq b$. The cumulative distribution is $P(X \leq x) = \frac{x-a}{b-a}$. [OpenStax]

Uniform Distribution | a continuous random variable (RV) that has equally likely outcomes over the domain, \(a < x < b\); often referred as the Rectangular Distribution because the graph of the pdf has the form of a rectangle. Notation: $X \sim U(a, b)$. The mean is $\mu = \frac{a+b}{2}$ and the standard deviation is $\sigma = \sqrt{\frac{(b-a)^2}{12}}$. The probability density function is $f(x) = \frac{a+b}{2}$ for $a < x < b$ or $a \leq x \leq b$. The cumulative distribution is $P(X \leq x) = \frac{x-a}{b-a}$. [OpenStax]

Variable | a characteristic of interest for each person or object in a population [OpenStax]

Variable (Random Variable) | a characteristic of interest in a population being studied. Common notation for variables are upper-case Latin letters X, Y, Z,... Common notation for a specific value from the domain (set of all possible values of a variable) are lower-case Latin letters x, y, z,.... For example, if X is the number of children in a family, then x represents a specific integer 0, 1, 2, 3, ... Variables in statistics differ from variables in intermediate algebra in the two following ways. (1) The domain of the random variable (RV) is not necessarily a numerical set; the domain may be expressed in words; for example, if X = hair color, then the domain is {black, blond, gray, green, orange}. (2) We can tell what specific value *x* of the random variable X takes only after performing the experiment. [OpenStax]

Variance | mean of the squared deviations from the mean; the square of the standard deviation. For a set of data, a deviation can be represented as $x - \bar{x}$ where x is a value of the data and $\bar{x}$ is the sample mean. The sample variance is equal to the sum of the squares of the deviations divided by the difference of the sample size and one. [OpenStax]

Venn Diagram | the visual representation of a sample space and events in the form of circles or ovals showing their intersections [OpenStax]

z-score | the linear transformation of the form $z = \frac{x-\mu}{\sigma}$; if this transformation is applied to any normal distribution $X \sim N(\mu, \sigma$ the result is the standard normal distribution $Z \sim N(0, 1)$. If this transformation is applied to any specific value x of the RV with mean μ and standard deviation σ, the result is called the z-score of x. The z-score allows us to compare data that are normally distributed but scaled differently. [OpenStax]

- **Glossary** has no license indicated.

Detailed Licensing

Overview

Title: Introductory Statistics (Hannah Seidler-Wright)

Webpages: 100

Applicable Restrictions: Noncommercial

All licenses found:

- CC BY-NC-SA 4.0: 99% (99 pages)
- Undeclared: 1% (1 page)

By Page

- Introductory Statistics (Hannah Seidler-Wright) - *CC BY-NC-SA 4.0*
  - Front Matter - *CC BY-NC-SA 4.0*
    - TitlePage - *CC BY-NC-SA 4.0*
    - InfoPage - *CC BY-NC-SA 4.0*
    - Table of Contents - *Undeclared*
    - Licensing - *CC BY-NC-SA 4.0*
  - 1: Designs of Statistical Studies - *CC BY-NC-SA 4.0*
    - 1.1: Welcome to Statistics - *CC BY-NC-SA 4.0*
      - 1.1.1: Exercises - *CC BY-NC-SA 4.0*
    - 1.2: The Statistical Analysis Process - *CC BY-NC-SA 4.0*
      - 1.2.1: Exercises - *CC BY-NC-SA 4.0*
    - 1.3: Research Questions, Types of Statistical Studies, and Stating Reasonable Conclusions - *CC BY-NC-SA 4.0*
      - 1.3.1: Exercises - *CC BY-NC-SA 4.0*
    - 1.4: Random Sampling and Bias - *CC BY-NC-SA 4.0*
      - 1.4.1: Exercises - *CC BY-NC-SA 4.0*
    - 1.5: Experiments and Random Assignment - *CC BY-NC-SA 4.0*
      - 1.5.1: Exercises - *CC BY-NC-SA 4.0*
  - 2: Descriptive Statistics - *CC BY-NC-SA 4.0*
    - 2.1: Descriptive Statistics - Dotplots and Histograms - *CC BY-NC-SA 4.0*
      - 2.1.1: Exercises - *CC BY-NC-SA 4.0*
    - 2.2: Quantifying the Center of a Distribution - *CC BY-NC-SA 4.0*
      - 2.2.1: Exercises - *CC BY-NC-SA 4.0*
    - 2.3: Quantifying Variability Relative to the Median - *CC BY-NC-SA 4.0*
      - 2.3.1: Exercises - *CC BY-NC-SA 4.0*
    - 2.4: Quantifying Variability Relative to the Mean - *CC BY-NC-SA 4.0*
      - 2.4.1: Exercises - *CC BY-NC-SA 4.0*
  - 3: Probability - *CC BY-NC-SA 4.0*
    - 3.1: Introduction to Probability - *CC BY-NC-SA 4.0*
      - 3.1.1: Exercises - *CC BY-NC-SA 4.0*
    - 3.2: Marginal, Joint, and Conditional Probability - *CC BY-NC-SA 4.0*
      - 3.2.1: Exercises - *CC BY-NC-SA 4.0*
    - 3.3: The Addition and Complement Rules - *CC BY-NC-SA 4.0*
      - 3.3.1: Exercises - *CC BY-NC-SA 4.0*
  - 4: Discrete Probability Distributions - *CC BY-NC-SA 4.0*
    - 4.1: Discrete Random Variables - *CC BY-NC-SA 4.0*
      - 4.1.1: Exercises - *CC BY-NC-SA 4.0*
    - 4.2: The Geometric Distribution - *CC BY-NC-SA 4.0*
      - 4.2.1: Exercises - *CC BY-NC-SA 4.0*
    - 4.3: The Binomial Distribution - *CC BY-NC-SA 4.0*
      - 4.3.1: Exercises - *CC BY-NC-SA 4.0*
  - 5: Continuous Probability Distributions and The Normal Distribution - *CC BY-NC-SA 4.0*
    - 5.1: Probability Distributions of Continuous Random Variables - *CC BY-NC-SA 4.0*
      - 5.1.1: Exercises - *CC BY-NC-SA 4.0*
    - 5.2: Characteristics of the Normal Distribution and The Empirical Rule - *CC BY-NC-SA 4.0*
      - 5.2.1: Exercises - *CC BY-NC-SA 4.0*
    - 5.3: The Standard Normal Distribution - *CC BY-NC-SA 4.0*
      - 5.3.1: Exercises - *CC BY-NC-SA 4.0*
    - 5.4: Finding Critical Values from the Normal Distribution - *CC BY-NC-SA 4.0*
      - 5.4.1: Exercises - *CC BY-NC-SA 4.0*
  - 6: Inference Involving a Single Population Proportion - *CC BY-NC-SA 4.0*

- 6.1: The Sampling Distribution of Sample Proportions - *CC BY-NC-SA 4.0*
  - 6.1.1: Exercises - *CC BY-NC-SA 4.0*
- 6.2: Estimating a Population Proportion - *CC BY-NC-SA 4.0*
  - 6.2.1: Exercises - *CC BY-NC-SA 4.0*
- 6.3: Introduction to Hypothesis Testing - *CC BY-NC-SA 4.0*
  - 6.3.1: Exercises - *CC BY-NC-SA 4.0*
- 6.4: Hypothesis Tests for a Single Population Proportion - *CC BY-NC-SA 4.0*
  - 6.4.1: Exercises - *CC BY-NC-SA 4.0*
- 6.5: Conclusions (1) - *CC BY-NC-SA 4.0*
  - 6.5.1: Exercises - *CC BY-NC-SA 4.0*

7: Inference Involving a Single Population Mean - *CC BY-NC-SA 4.0*
- 7.1: The Sampling Distribution of Sample Means - *CC BY-NC-SA 4.0*
  - 7.1.1: Exercises - *CC BY-NC-SA 4.0*
- 7.2: The Student's T-Distribution - *CC BY-NC-SA 4.0*
  - 7.2.1: Exercises - *CC BY-NC-SA 4.0*
- 7.3: Estimating a Population Mean - *CC BY-NC-SA 4.0*
  - 7.3.1: Exercises - *CC BY-NC-SA 4.0*
- 7.4: Hypothesis Tests for a Single Population Mean - *CC BY-NC-SA 4.0*
  - 7.4.1: Exercises - *CC BY-NC-SA 4.0*
- 7.5: Conclusions (2) - *CC BY-NC-SA 4.0*
  - 7.5.1: Exercises - *CC BY-NC-SA 4.0*

8: Inference Involving Two Population Parameters - *CC BY-NC-SA 4.0*
- 8.1: Paired Samples - *CC BY-NC-SA 4.0*
  - 8.1.1: Exercises - *CC BY-NC-SA 4.0*
- 8.2: Distributions of Differences - *CC BY-NC-SA 4.0*
  - 8.2.1: Exercises - *CC BY-NC-SA 4.0*
- 8.3: Inference for a Difference in Two Population Means - *CC BY-NC-SA 4.0*
  - 8.3.1: Exercises - *CC BY-NC-SA 4.0*
- 8.4: Inference for a Difference in Two Population Proportions - *CC BY-NC-SA 4.0*
  - 8.4.1: Exercises - *CC BY-NC-SA 4.0*

9: Linear Regression - *CC BY-NC-SA 4.0*
- 9.1: Scatterplots - *CC BY-NC-SA 4.0*
  - 9.1.1: Exercises - *CC BY-NC-SA 4.0*
- 9.2: Quantifying Direction and Strength - *CC BY-NC-SA 4.0*
  - 9.2.1: Exercises - *CC BY-NC-SA 4.0*
- 9.3: The Line of Best Fit - *CC BY-NC-SA 4.0*
  - 9.3.1: Exercises - *CC BY-NC-SA 4.0*

10: Inference Involving More Than Two Parameters - *CC BY-NC-SA 4.0*
- 10.1: The Chi-Square Distribution - *CC BY-NC-SA 4.0*
  - 10.1.1: Exercises - *CC BY-NC-SA 4.0*
- 10.2: Goodness-of-Fit - *CC BY-NC-SA 4.0*
  - 10.2.1: Exercises - *CC BY-NC-SA 4.0*
- 10.3: Testing for Independence - *CC BY-NC-SA 4.0*
  - 10.3.1: Exercises - *CC BY-NC-SA 4.0*
- 10.4: ANOVA - *CC BY-NC-SA 4.0*
  - 10.4.1: Exercises - *CC BY-NC-SA 4.0*

Back Matter - *CC BY-NC-SA 4.0*
- Index - *CC BY-NC-SA 4.0*
- Glossary - *CC BY-NC-SA 4.0*
- Detailed Licensing - *CC BY-NC-SA 4.0*