INTRODUCTION TO STATISTICS (RRCC)



Mathematics Department
Red Rocks Community College

LibreTexts™

Introduction to Statistics (RRCC)

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 04/19/2026

Required Course Learning Outcomes

1. Communicate the language of statistics using the appropriate terminology.
2. Analyze numerical summaries and graphical displays of sample data.
3. Analyze bivariate data by applying the concepts of correlation and regression.

4. Apply the basic rules of probability.
5. Utilize the appropriate probability distribution.
6. Make inferences about one or more populations using sample data.
7. Utilize technology to further statistical understanding and reasoning.

Recommended Topical Outline

I. The Language and Terminology of Statistics

- A. Introduction to experimental vs observational studies
- B. Introduction to qualitative vs quantitative studies
- C. Introduction to the research process
- D. Introduction to sources of bias
- E. Population vs samples
- F. Parameters vs statistics
- G. Types of data (discrete/continuous, quantitative/qualitative)
- H. Types of variables (lurking, confounding, explanatory, response)
- I. Sampling techniques (random, systematic, stratified, cluster, convenience)
- J. Relevance of statistics to scientific and other real-world problems

II. Numerical Summaries and Graphical Displays of Sample Data

- A. Frequency distribution tables including construction
- B. Graphical displays of qualitative and quantitative data including construction
- C. Distribution shapes (modality and skew)
- D. Measures of central tendency (including calculations with or without technology)
- E. Measures of variation (including calculations with or without technology)
- F. Measures of relative position (including calculations with or without technology)

III. Bivariate Data: Correlation and Regression

- A. Scatter plots including construction
- B. Correlation including calculations with or without technology
- C. Linear regression
- D. Analysis of residuals

IV. Fundamentals of Probability

- A. Law of Large Numbers
- B. Addition and multiplication rules
- C. Introduction to permutations and combinations
- D. Complementary events
- E. Conditional probability

V. Probability Distributions

- A. Introduction to discrete probability distributions
- B. Binomial distributions
- C. Normal distributions

VI. Inference for One or More Populations Using Sample Data

- A. The Central Limit Theorem
- B. Sample size for mean and proportion
- C. Confidence intervals for one and two populations
- D. Hypothesis tests for one and two populations
- E. Chi-square and goodness-of-fit tests
- F. Introduction to Analysis of Variance (ANOVA)

Course Alignment Reference Table

This table shows how each chapter in the textbook supports our course objectives and learning outcomes. Use this to reflect on your progress and see how each skill builds toward the full understanding of statistical reasoning with real-world data.

Textbook Chapter	Primary Course Learning Outcomes (CLOs)	Relevant Topical Outline Elements (TOs)
1. Introduction to Data	CLO 1, CLO 6	I.A – I.J, I.E, I.F, I.G, I.I
2. Exploring and Summarizing Data	CLO 2, CLO 7	II.A – II.F
3. Visualizing Data	CLO 2, CLO 7	II.A – II.C
4. Introduction to Probability	CLO 4, CLO 7	IV.A – IV.E
5. Probability Distributions	CLO 5, CLO 6, CLO 7	V.A – V.C
6. Sampling Distributions	CLO 6, CLO 7	VI.A
7. Confidence Intervals	CLO 6, CLO 7	VI.B – VI.C
8. Hypothesis Testing	CLO 6, CLO 7	VI.D, VI.E, VI.F
9. Correlation and Regression	CLO 3, CLO 7	III.A – III.D
10. Appendix and Tables	CLO 7 (supporting)	All through VII

Legend

- **CLO = Course Learning Outcome** (Official campus-level outcome)
- **TO = Topical Outline Objective** (Recommended content coverage by section)

To see the full list of CLOs and outline objectives, refer to the course syllabus or the section below.

TABLE OF CONTENTS

Licensing

1: Introduction to Data

- 1.1: What is Statistics?
- 1.2: Data Types- Categorical vs. Numerical
- 1.3: Populations, Samples, and Sampling Methods
- 1.4: Sources of Bias and Randomness
- 1.5: Exploring Statistical Questions
- 1.6: Designing a Statistical Study
  - 1.6.1: Project Introduction
- 1.7: Vocabulary (Chapter 1)
- 1.8: Practice (Chapter 1)

2: Exploring and Summarizing Data

- 2.1: Measures of Central Tendency- Mean, Median, and Mode
- 2.2: Measures of Spread- Range, Variance, and Standard Deviation
  - 2.2.1: Project Check-In 1
- 2.3: The Five-Number Summary and Interquartile Range
  - 2.3.1: Project Check-In 2
  - 2.3.2: Percentage Review
- 2.4: Identifying Outliers with Boxplots
- 2.5: Vocabulary (Chapter 2)
- 2.6: Practice (Chapter 2)

3: Visualizing Data

- 3.1: Graphical Displays for Categorical Data- Bar Charts, Pie Charts
- 3.2: Graphical Displays for Quantitative Data- Dot Plots and Stemplots
- 3.3: Frequency Tables and Relative Frequency
- 3.4: Shapes of Distributions
- 3.5: Time Series Plots and Trends Over Time
  - 3.5.1: Project Check-In 3
- 3.6: Vocabulary (Chapter 3)
- 3.7: Practice (Chapter 3)

4: Introduction to Probability

- 4.1: What is Probability?
- 4.2: Introduction to Counting
- 4.3: Foundational Probability Rules
- 4.4: Compound Events, Independence and Conditional Probability
- 4.5: Vocabulary (Chapter 4)
- 4.6: Practice (Chapter 4)

5: Probability Distributions

- 5.1: Random Variables- Discrete and Continuous
- 5.2: Expected Value of Discrete Distributions
- 5.3: Discrete Distributions- Bernoulli and Binomial
- 5.4: The Normal Distribution
- 5.5: The Empirical Rule and Standard Normal (Z) Distribution
- 5.6: Application of the Normal Distribution
- 5.7: Vocabulary (Chapter 5)
- 5.8: Practice (Chapter 5)

6: Sampling Distributions

- 6.1: Sampling Distributions
- 6.2: Mean and Standard Deviation of Sampling Distributions
- 6.3: Central Limit Theorem- Meaning and Implications
- 6.4: CLT for Means and Proportions
- 6.5: Approximating Binomial with the Normal Distribution
- 6.6: Vocabulary (Chapter 6)
- 6.7: Practice (Chapter 6)

7: Confidence Intervals

- 7.1: Point Estimates vs. Interval Estimates
- 7.2: Confidence Intervals for a Population Proportion
- 7.3: Confidence Intervals for a Population Mean
- 7.4: Effect of Sample Size and Confidence Level
- 7.5: Minimum Sample Size
- 7.6: Interpreting Confidence Intervals
  - 7.6.1: Project Check-In 4
- 7.7: Vocabulary (Chapter 7)
- 7.8: Practice (Chapter 7)

8: Hypothesis Testing

- 8.1: Why hypothesis testing and Type I and Type II Errors
- 8.2: Hypothesis Testing Framework
- 8.3: Tests for a Single Mean (z and t-tests)
- 8.4: Comparing Two Means (Independent Samples)
  - 8.4.1: Project Checkpoint
- 8.5: Tests for a Single Proportion
- 8.6: Interpreting p-values and Significance
- 8.7: Vocabulary (Chapter 8)
- 8.8: Practice (Chapter 8)

9: Bivariate Data, Correlation, and Regression

- 9.1: Scatterplots and Direction of Association
  - 9.1.1: Datasaurus Dozen
- 9.2: Correlation Coefficient (r)
- 9.3: Least-Squares Regression Line

- 9.4: Interpreting Slope, Intercept, and Residuals
- 9.5: Cautions with Correlation and Causation
- 9.6: Vocabulary (Chapter 9)
- 9.7: Practice (Chapter 9)

10: Project, Labs, Appendix and Tables

- 10.1: Project Assignments
  - 10.1.1: Sampling and Data Web Page
  - 10.1.2: Quantitative Variable Analysis
  - 10.1.3: Qualitative Variable Analysis
  - 10.1.4: Interval Estimates
  - 10.1.5: Hypothesis Testing
  - 10.1.6: Correlation
- 10.2: Tech Labs
  - 10.2.1: TL 1
  - 10.2.2: TL 2
  - 10.2.3: TL 3
  - 10.2.4: TL 4
  - 10.2.5: TL 5
  - 10.2.6: TL 6
  - 10.2.7: TL 7
- 10.3: Summary Statistic Cards
- 10.4: Project Reflection
- 10.5: z-Table
- 10.6: t-Table

11: Applets and Other Stuff

Index

Glossary

Detailed Licensing

Detailed Licensing

Licensing

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

This textbook by Red Rocks Community College Mathematics Department is licensed under CC BY 4.0

Practice problems include parts of:

Introductory Statistics 2e by OpenStax is licensed under CC BY 4.0

Access for free at https://openstax.org/books/introductory-statistics-2e/pages/1-introduction

and

OpenIntro Statistics by David Diez, Christopher Barr, & Mine Çetinkaya-Rundel is licensed under CC BY 3.0

1: Introduction to Data

A Data Study Example – Tracking Tree Health

Before we define what data is or what statistics means, let's look at a real-world example of how people use data to answer important questions. To introduce the kinds of decisions and steps that are involved in statistics, let's imagine a research project that tracks the health of trees over time in a large urban park.

Asking a Question

Suppose city ecologists want to understand whether the health of old-growth trees in the Central City Park has changed over the past 10 years. They want to find out:

> *How has tree health changed in Central City Park from 2013 to 2023, and what factors might be related to that change?*

This question is important for managing urban green spaces and planning tree preservation strategies. It's specific and measurable and can be answered by collecting and analyzing data.

Collecting Accurate and Relevant Data

To answer this question, the researchers need a dataset that includes information about individual trees over time. They might decide to select a sample of 200 trees within the park and assess each one once a year from 2013 to 2023.

For each tree, they could record:

- **Tree ID:** A unique number or code for each tree
- **Species:** Common/species name (e.g., Oak, Maple)
- **Location:** GPS coordinates or zone within the park
- **Year:** Year of observation
- **Health Score:** A rating from 0 to 10 based on visible signs (e.g., leaf color, bark condition, dead branches)
- **Comments:** Notes about the health score (this could be useful for tracking specific diseases or blights)
- **Canopy Size (ft^2):** Estimated tree cover area
- **Trunk Diameter (ft):** Measured by circumference and then calculated to quantify tree growth.
- **Presence of Disease:** Yes or No
- **Nearby Construction:** Yes or No

This data can be collected using field surveys, aerial imagery, and environmental sensors.

Organizing and Summarizing the Data

Once the data is collected, it must be organized into a consistent format. Below you can see an example of four records from two trees. This is a small picture of a much larger data set.

Each horizontal row in the dataset represents a measurement of one tree in one year. Each column represents one type of recording, called a *variable*. Since we have 10 variables, we will have 10 columns. Since we have sampled 200 trees over 10 years, we expect to have 2000 rows. Why might we end up with fewer rows than that?

Table describes exerpt from table of trees.

Tree ID	Species	Year	Health Score
45	Common Hackberry (Celtis occidentalis)	2013	10
119	Montana Oak (Quercus montana)	2017	6

Researchers could start by investigating and exploring the data by summarizing each variable. For example:

- What is the average health score across all trees each year?
- How many trees show signs of disease each year?
- How many different tree species are represented?

They might create graphs and charts, for example showing average health score by year, or the number of diseased trees per species.

Analyzing the Data to Find Patterns

The researchers now look for trends and relationships. Some possible investigates are:

- Is there a downward trend in tree health over the decade?
- Are trees near construction sites more likely to have lower health scores?
- Are there areas in the park whose trees exhibit slower growth or higher rates of disease?

They will use the statistical methods we will explore in the course to make sure their results are supported mathematically and rigorously.

Interpreting the Results

Based on the analysis, the research team might discover, for instance, that tree health scores have declined slightly over the decade, and that this decline is strongly associated with nearby construction. They interpret this to mean that urban development is likely affecting tree health.

Understanding the *why* behind the numbers is a key part of every data investigation.

Communicating the Findings

The final step is to share results in a clear and useful way. The team might prepare:

- A report with graphs, maps, and summary statistics for city officials
- A public presentation with visuals and plain-language findings
- Publicly releasing data and calculations to allow for other scientists, students or reporters to continue to investigate

Clear communication ensures the data can influence policy—such as protecting key areas, planting new trees, increasing tree care, or limiting construction in sensitive locations.

What's Next?

Now that you've seen how statistics is used in action, we'll step back and define some core ideas. In the next section, we'll look more closely at what **data** is, and how we describe the different kinds of information that can be collected.

This page titled 1: Introduction to Data is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Mathematics Department.

1.1: What is Statistics?

Introduction to Statistics and Data

We live in a world driven by data. From the moment we wake up and check the weather app to the instant we scroll through our streaming feed, we are interacting with data systems. Understanding how to think about, collect, analyze, and draw conclusions from data is at the heart of statistics and helps us make better decisions in everyday life and in professional fields ranging from public health to business to environmental science.

What is Data?

Definition: Data

Data refers to recorded information about individuals, objects, or events. These pieces of information are typically collected through observation, measurement, or experimentation.

A single recorded value is referred to as an *observation*, and a group of related observations forms a *dataset*.

Examples: Datasets That Fit the Definition of Data

1. Student Enrollment Dataset

What it records: Information about students at a college or university.

Example variables: Student ID, Major, Credits completed, GPA, Enrollment status

Each observation: One student's record for a particular semester.

2. Fitness Tracker Dataset

What it records: Daily activity data from a wearable device.

Example variables: Date, Step count, Active minutes, Resting heart rate, Calories burned

Each observation: One person's activity summary for one day.

3. Restaurant Health Inspection Dataset

What it records: Health inspection results from restaurants in a city.

Example variables: Restaurant name, Date of inspection, Number of violations, Score (out of 100), Status (pass/fail)

Each observation: One inspection event at a single restaurant.

4. Used Car Listings Dataset

What it records: Listings for vehicles posted on a used car website.

Example variables: Make and model, Year, Mileage, Listed price, City

Each observation: One vehicle currently listed for sale.

What is Statistics?

Definition: Statistics

Statistics is the science of learning from data in the presence of variation. It provides tools and methods for collecting, organizing, visualizing, analyzing, and interpreting data in order to make decisions, test hypotheses, and better understand the world.

Statistics is more than just "number crunching." It is a process that helps us make sense of information and uncertainty by understanding *how* the data were collected, *what* patterns they reveal, and *whether* those patterns are likely to be meaningful or due to chance.

Working with Data

When working with data, we usually follow a cycle that includes:

- **Asking a question** that can be answered with data
- **Collecting** accurate and relevant data
- **Organizing and summarizing** the data to understand its structure
- **Analyzing** the data to find patterns or test relationships
- **Interpreting** the results in context
- **Communicating** findings clearly and effectively

Data doesn't speak for itself it needs interpretation and responsible analysis.

Descriptive vs. Inferential Statistics

Descriptive statistics help us summarize and understand the data right in front of us (like averages, medians, and graphs).
Inferential statistics help us go further making predictions or generalizations about a larger population based on sample data.

What's Next?

In the next section, we'll take a closer look at two basic ways to describe variables in a dataset. Some variables describe *qualities, labels, or categories*, while others measure *counts, measurements, or amounts*. Developing an understanding of these fundamental categories **categorical** and **quantitative** variables will help us decide which statistical tools are appropriate for different types of data.

This page titled 1.1: What is Statistics? is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Mathematics Department.

1.2: Data Types- Categorical vs. Numerical

Classifying Data

Now that we've seen what a dataset might look like, it's time to develop the language we need to describe data accurately and consistently. At the heart of most datasets are **variables**. In this section, we'll define what a variable is, explain the main types of variables you'll encounter in statistics, and introduce key subcategories that will help you prepare for more advanced topics later on.

What Is a Variable?

> #### Definition: Variable
>
> A **variable** is a characteristic or attribute that can take on different values for different individuals or observations. In a dataset, each column usually represents a variable, and each row provides the values for a single observation.

Variables help describe individuals in our dataset whether they are people, plants, countries, experiments, or ideas. We classify variables based on the kind of information they represent.

Examples of Variables

Here are just a few examples of variables you might encounter in different kinds of data:

- **In a housing dataset:** Price of the home, square footage, number of bedrooms, ZIP code, listing date
- **In a student survey:** Age, preferred learning method (online/in-person), major, number of credits completed, satisfaction with campus resources
- **In a fitness app:** Daily step count, heart rate, type of workout, duration of workout, calories burned
- **In a food delivery service:** Order total, payment method, delivery time, customer tip, restaurant type
- **In a wildlife field study:** Species observed, habitat type, weight of animal, date of observation, number of sightings

In each case, **variables describe the features or measurements** of the observations — whether those observations are people, houses, animals, or transactions.

As we move forward, we'll learn to classify these variables based on the kind of information they represent: names or labels vs. counts and measurements. This classification guides which graphs and summaries we use later on.

Main Types of Variables

Variables generally fall into one of two main types:

- **Categorical variables** describe qualities, descriptive labels, or characteristics. They place individuals into distinct *groups* or *categories*.
- **Numerical variables** represent amounts or quantities. They are used to *measure* or *count* something.

Categorical Variables

Categorical variables (also known as qualitative variables) take on values that are labels or names. These values reflect categories, not quantities. Even if we use numbers to code them (like 1 = Red, 2 = Blue), the numbers are just labels, not measurements.

Example: The tree species in the Tree Health Study ("Maple", "Oak", "Birch") is a categorical variable.

Nominal vs. Ordinal

Categorical variables can be further classified into nominal or ordinal types:

- **Nominal** variables have categories with *no inherent order*.
- **Ordinal** variables have categories with a *meaningful order*, but the distances between categories are not necessarily equal.

Examples of Nominal Categorical Variables

- Tree species: oak, maple, pine, birch
- Favorite color: red, green, red
- City of birth: Atlanta, Boston, Seattle
- Marital status: single, married, divorced
- Blood type: A, B, AB, O

Examples of Ordinal Categorical Variables

- Tree health rating: poor, fair, good, excellent
- Spicy level on a menu: mild, medium, hot
- Education level: High School, Bachelor's, Master's, PhD
- Customer satisfaction: very unsatisfied to very satisfied
- Fitness level: beginner, intermediate, advanced

Numerical Variables

Numerical (also known as quantitative) variables represent a measurable quantity. These values are numbers where basic math *makes sense*: you can average them, compare differences, and perform meaningful computations.

A helpful question to ask: *If we added the values together, would the result make any sense?*

Discrete vs. Continuous

Numerical variables can also be further classified into discrete or continuous types.

- **Discrete** variables can only take on certain values, often only whole numbers.
  - Often these are *counts* but this could also include examples like shoe size where we have "half" sizes but nothing in between.
- **Continuous** variables can take on any value on a number line within a range.
  - Typically these are are based on *measurements*. An example would be the length of an object, which could be 1 inch, 2 inches, or anything in between like 1.5482675 inches. Even though we typically will round continuous data values when they are recorded, these variables *could* have any value, not just whole numbers.

Examples of Discrete Variables

- Number of trees in a plot
- Number of visitors to a park on a given day
- Number of siblings a person has
- Number of pets in a household
- Number of books read in a year

Examples of Continuous Variables

- Canopy area in square feet
- Air temperature in Celsius
- Wind speed in miles per hour
- Time it takes a seed to sprout (days)
- Human height in centimeters

Sometimes, it may be tricky to tell which category a variable belongs to right away. For example, *age* could be described as a continuous variable (like 21.7 years old), but if grouped into age ranges ("Under 18", "18–34", "35–49", etc.), it becomes *categorical* and ordinal.

Why This Matters

Knowing the type of each variable is essential for choosing the right statistical tools later. Some graphs and summaries only make sense for numerical data (like histograms), while others are designed specifically for categorical data (like bar charts). Statistical

tests also depend heavily on whether the variables involved are categorical or numerical.

Activity: Think of Your Own Examples

Now it's your turn! Try to come up with at least one new example for each of the four subtypes of variables below:

Examples of four data types.

Variable Subtype	Your Example
Nominal (Categorical)	_____
Ordinal (Categorical)	_____
Discrete (Numerical)	_____
Continuous (Numerical)	_____

Bonus challenge: Think of a variable you might need to reclassify depending on context (like "age" or "income level"). Describe *both* ways it could be used.

Quick Quiz: What Type of Variable Is It?

For each variable below, select the most appropriate type.

1. Type of music a person prefers (e.g., rock, jazz, classical)

--Choose--

2. Number of books borrowed from a library each month

--Choose--

3. Height of tree seedlings (in centimeters)

--Choose--

4. Customer satisfaction rating (poor, fair, good, excellent)

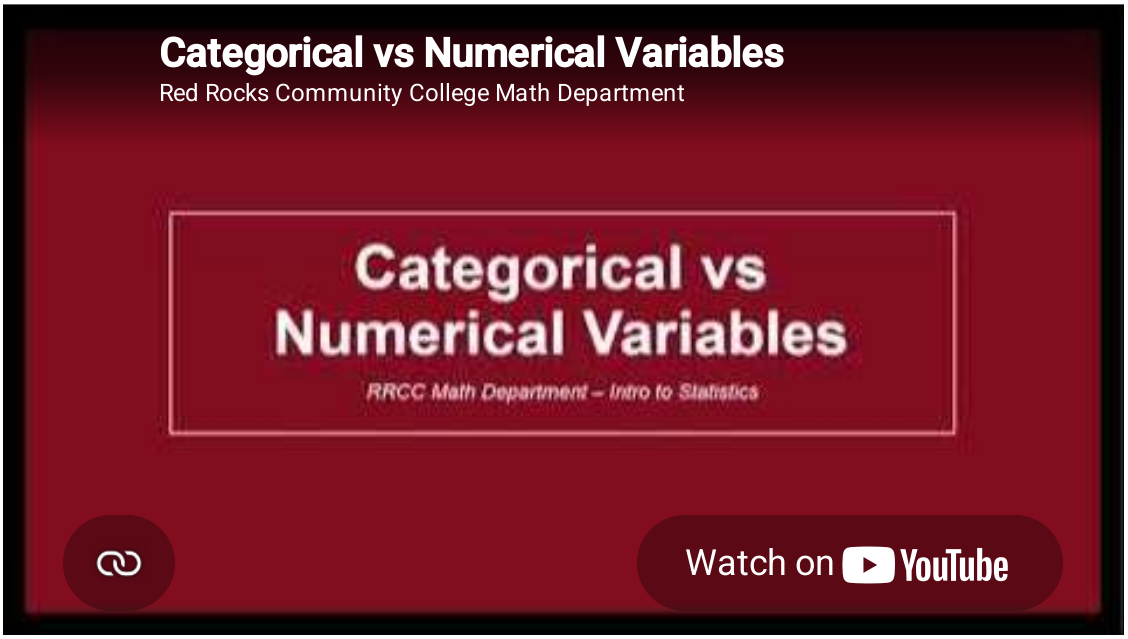
--Choose--

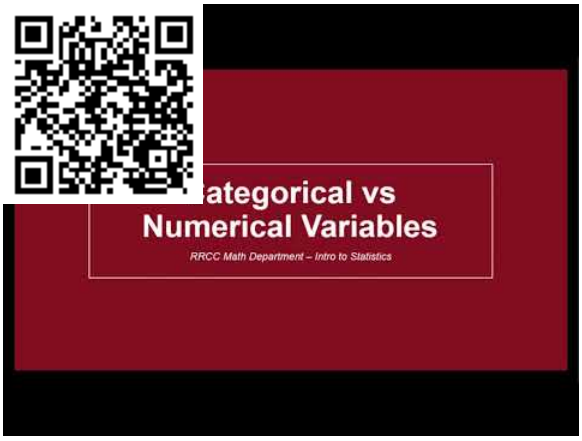
5. Blood type (A, B, AB, O)

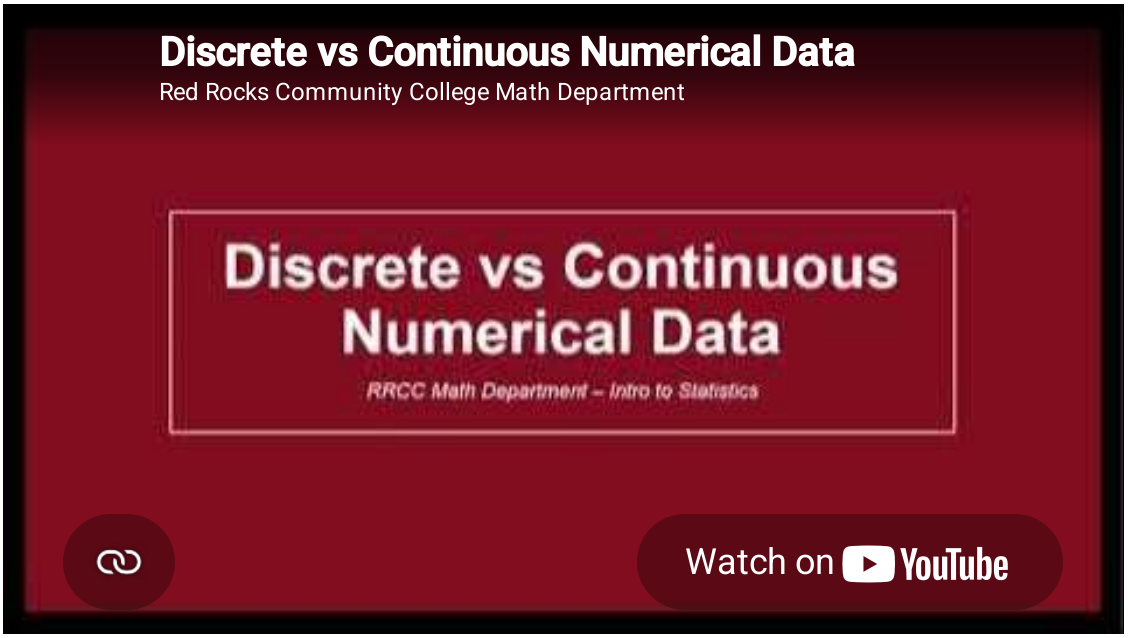
--Choose--

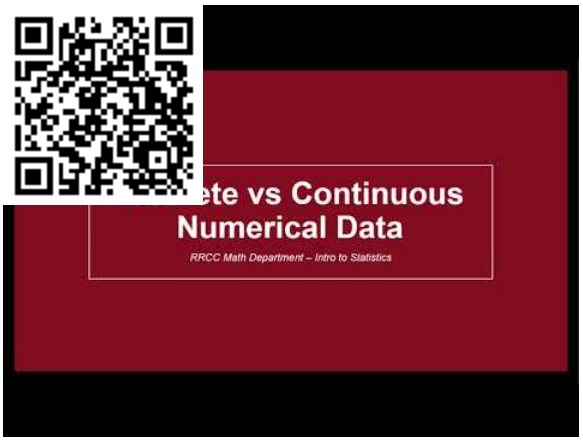
Check Answers

Related Videos


Categorical vs Numerical Variables
Red Rocks Community College Math Department
Categorical vs Numerical Variables
RRCC Math Department – Intro to Statistics
Watch on YouTube


ategorical vs
Numerical Variables
RRCC Math Department – Intro to Statistics


Discrete vs Continuous Numerical Data
Red Rocks Community College Math Department
Discrete vs Continuous Numerical Data
RRCC Math Department – Intro to Statistics
Watch on YouTube



What's Next?

Now that we understand what different types of variables are and how to describe them, we're ready to think about how we collect data. In the next section, we'll explore the ideas of **populations** and **samples**, and how we design studies using **sampling methods** that help us draw meaningful and fair conclusions.

This page titled 1.2: Data Types- Categorical vs. Numerical is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Mathematics Department.

1.3: Populations, Samples, and Sampling Methods

Populations, Samples, and Sampling Methods

In many cases, we're interested in learning something about a large group all the voters in a country, all the trees in a national forest, or all the customers who bought a certain product last year. But collecting data from *every member* of a big group isn't always realistic. Instead, we collect information from a **subset** and use that data to draw conclusions about the larger group.

Populations and Samples

In statistics, we use precise terms to talk about who or what we're studying:

- **Population:** The entire group we want to understand or describe.
- **Sample:** A smaller group selected from the population that we actually observe or measure.

The population might be enormous (e.g., all registered voters in a state), while the sample might include only a few hundred of them. If the sample is chosen well, insights from the sample can be extended to the entire population.

Example: Tree Study Revisited

Let's return to our earlier tree health project in Central City Park:

- **Population:** All mature trees in Central City Park
- **Sample:** The 200 trees the researchers chose to monitor from 2013–2023

It would be difficult and costly to regularly check every single tree in the park. Instead, researchers track a smaller number of trees through time. Their goal is to learn something about the *entire forest* by closely watching just a *portion of it*.

Why the Sample Matters

For conclusions drawn from a sample to be useful for the population, the sample must be **representative**. This means:

- The sample includes a mix of individuals or objects that resemble the full group in key ways
- The method used to choose those individuals is fair and not systematically biased

If the sample is too small or poorly chosen, it may not reflect the diversity or characteristics of the population and conclusions based on that sample can be misleading or just plain wrong.

Example: A Non-Representative Sample

Suppose our tree study only included trees located near park entrances because they were easier to reach. If these areas receive more pollution or foot traffic, then our sample may overestimate tree health problems across the whole park. We'd be studying a biased slice of the forest.

Alternatively, suppose that researchers chose only to measure smaller trees because it was easier to check their leaves and canopy sizes. This would not be representative of the diversity of trees in the park.

The Role of Randomness

To guard against bias and help ensure a sample is representative, statisticians often rely on **random sampling**.

A **random sample** is one where each member of the population has a known and typically equal chance of being included. This helps eliminate systematic overrepresentation or underrepresentation of certain types of individuals.

Although random selection doesn't guarantee a perfect sample, it does help control for unseen influences and makes it more reasonable to generalize from the sample to the population.

Common Sampling Methods

There are several techniques used to select a sample. Each has strengths and weaknesses depending on the goal and constraints of a study.

- **Simple Random Sample (SRS):** Every individual in the population has an equal chance of being selected. Often done via random numbers or software.
- **Stratified Sample:** The population is divided into meaningful subgroups (strata), and a random sample is taken from each group. Helps ensure representation across key characteristics like species, age, or region.
- **Cluster Sample:** The population is divided into naturally occurring groups (such as zones, classrooms, or plots). A random selection of these entire groups is observed.
- **Systematic Sample:** Select every nth individual from an ordered list (e.g., every 10th tree along a trail).
- **Convenience Sample:** Individuals are chosen based on ease of access. This type is not random and often introduces bias. It should be used with caution.

In practice, the method chosen depends on available resources, population structure, and the goals of the study. The important thing is to avoid unintentionally favoring certain individuals or groups in the selection process. It is also possible to combine these methods into more complex sampling designs, such as starting with clusters and then stratifying them.

Comparison of Sampling Methods

Sampling Method	Pros	Cons
Simple Random Sample (SRS)	Every individual has an equal chance of being selected; easy to understand; minimizes bias.	Can be difficult and time-consuming to implement for large or geographically dispersed populations.
Stratified Sample	Ensures representation across key subgroups; increases precision if groups differ.	Requires knowledge of population strata; more complex to design and implement.
Cluster Sample	More practical when population is spread out; reduces cost and time.	May be less representative if clusters differ greatly from one another (higher sampling error).
Systematic Sample	Simple to execute and can work well if there's no underlying pattern.	Risk of bias if there's a hidden pattern in the data that aligns with the interval.
Convenience Sample	Easy, quick, and inexpensive to gather.	Often biased; not generalizable to the population; may exclude important subgroups.

Looking Ahead

A carefully selected sample is an essential part of any good statistical study. But even a well-planned sampling method can still be vulnerable to problems especially if we're not careful about how the data is collected or who is represented.

In the next section, we will explore the idea of **bias** systematic errors in data collection or sampling and learn how to spot problems that could affect the accuracy and fairness of our conclusions.

Quick Reflection

Think about a group you care about your school, your city, a community that you are a member of. How could you gather a sample to understand something about that group?

What might go wrong if your sample misses certain types of individuals?

Quick Quiz: Which Sampling Method Is Described?

Choose the sampling method utilized in the following scenarios:

1. A researcher uses a random number generator to choose 100 students from a list of all 5,000 students at a university.
Select one:

--Choose--

2. A polling company divides a city into neighborhoods, randomly selects 5 neighborhoods, and interviews *all* households in those neighborhoods.

--Choose--

3. A teacher gives a survey to the first 20 students who walk into the cafeteria.

--Choose--

4. A health researcher selects 25 men and 25 women from each age category to ensure balance in gender and age.

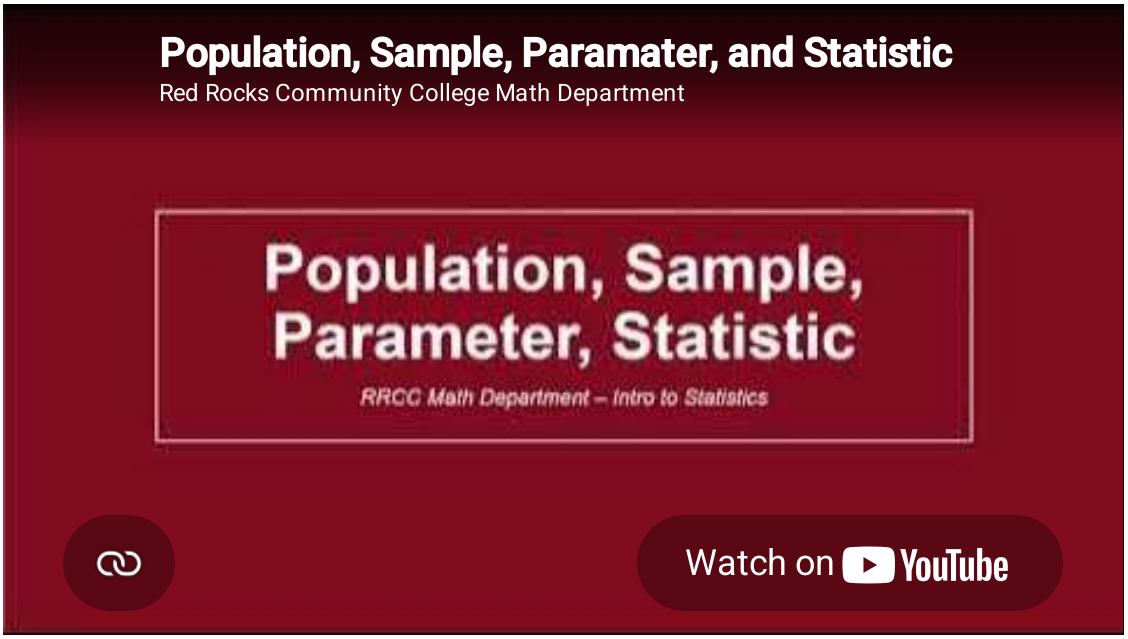
--Choose--

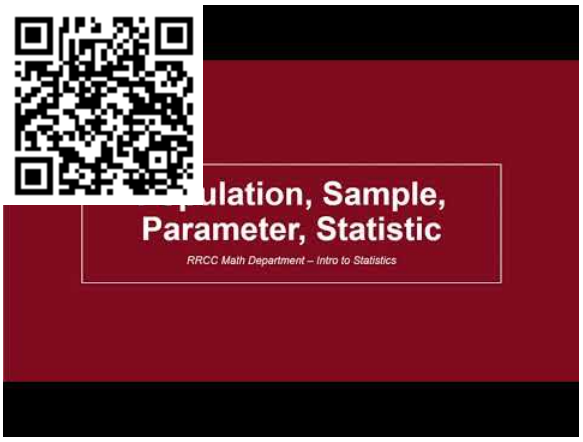
5. A company chooses every 10th product off an assembly line for quality inspection.

--Choose--

Check Answers

Related Videos









This page titled 1.3: Populations, Samples, and Sampling Methods is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Mathematics Department.

1.4: Sources of Bias and Randomness

Bias and the Importance of Randomness

Now that we understand how we often use a **sample** to learn about a larger **population**, let's talk about a key idea that can interfere with our conclusions: **bias**.

What Is Bias?

Bias occurs when the method used to collect data causes certain individuals or outcomes to be favored or excluded, making the sample systematically different from the population.

Bias can lead us to incorrect conclusions about a population because the information in our sample doesn't fairly or fully represent the entire group. While some variability is natural in any sample, *biased* samples introduce **systematic errors** that can't be explained by random chance.

Examples of Bias in Sampling

- **Convenience Bias:** A teacher wants to know how much students are studying on campus. She surveys the students in the library at 9am. But those students might be more prepared or studious than the general student population.
- **Nonresponse Bias:** A health department emails a survey on exercise habits, but only 20% of people respond. If more active people are more likely to answer, the final summary may overestimate how much exercise the population gets.
- **Undercoverage Bias:** A city survey collects responses entirely via smartphones. This may exclude older adults or those with limited internet access—leaving out important parts of the population.
- **Voluntary Response Bias:** An online poll asks if customers are "satisfied with service." People who feel extremely positive or negative are more likely to respond, skewing the results.

In each case, the problem is that the *method* causes misrepresentation. If we base policies or decisions on biased data, we may end up hurting the very communities we aim to serve.

How Randomness Helps

Random sampling is one of the most important defenses against bias.

By using a **random process** to select individuals from the population, giving each one a known and fair chance of being included, we reduce the influence of personal judgment, patterns, convenience, or external factors.

Randomness also lets us make statistical claims later, because it gives us a defensible reason to treat our sample as representative of the full population *within the limits of chance*.

Example: Comparing Two Approaches

A researcher wants to know the average commuting time for people in a city. They consider two approaches:

- **Approach A:** They survey 300 people in the downtown train station at 8:30am on a weekday.
- **Approach B:** They randomly select 300 residents from city voter rolls and call them at home.

Approach A is likely to lead to bias those passing through the station during rush hour might have different commuting patterns than city residents as a whole. Some might not even live in the city. Approach B has a better chance of reaching people across neighborhoods, schedules, and lifestyles — especially if non-response is addressed properly.

Beyond Sampling: Measurement and Design Bias

Bias isn't limited to sampling. It can also creep in when:

- **Survey questions** are worded in leading or confusing ways.
- **Instruments** (like scales or tests) aren't well-calibrated or fair.
- **Researchers** influence responses consciously or unconsciously.

We'll talk more about these issues when we learn how to design good studies and interpret study results fairly.

Can We Avoid Bias Completely?

Bias is often invisible unless we actively look for it. Even with good intentions, human judgments and systems can introduce unfair patterns. While we may not eliminate *all* sources of bias, awareness and thoughtful design give us the best possible chance to reduce it.

Reflection: Can You Spot the Bias?

- A school district sends out a digital survey asking about lunch program satisfaction. Who might be unknowingly excluded?
- A company lets employees “opt in” to a survey on job satisfaction. What kind of responses might they get?
- A sports article collects Twitter replies about a controversial call in a playoff game. Is that representative of all fans?

Quick Quiz: Identify the Type of Bias

For each example below, choose the type of sampling bias that is most likely present.

1. A news website runs an online poll asking readers if they're satisfied with government policy. Only extremely positive and negative responses pour in.
Select one:

--Choose--

2. A public health survey is distributed only through a smartphone app. Many older adults never see it and aren’t included in the data.

--Choose--

3. A school researcher surveys students who are already standing in front of the cafeteria to get quick answers about lunch preferences.

--Choose--

4. A market research firm emails a survey to 1,000 customers, but only 210 respond — most of whom are satisfied repeat buyers.

--Choose--

Check Answers

What's Next?

Now that we understand the importance of fair sampling and the dangers of bias, we’re ready to think about **how to ask good statistical questions** and **design reliable studies**. In the next section, we’ll learn how to plan an investigation that collects data in a way that’s focused, useful, and ethical from the start.

This page titled 1.4: Sources of Bias and Randomness is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Mathematics Department.

1.5: Exploring Statistical Questions

Thinking Like a Statistician – A Guided Experience

Now that you've seen the building blocks of a statistical investigation including data types, sampling, randomness, and bias, it's time to bring it all together. In this section, you'll engage with a real question and work through the full cycle of a statistical study.

This will give you practice using the language and structure you've learned so far, and prepare you for more complex analyses in future chapters.

Step-by-Step: The Statistical Process

Six major steps that can guide statistical thinking:

1. **Ask a question** that can be answered with data
2. **Collect** accurate and relevant data
3. **Organize and summarize** the data to understand its structure
4. **Analyze** the data to find patterns or test relationships
5. **Interpret** the results in context
6. **Communicate** your findings clearly and effectively

Below are five sample statistical questions representing different real-world domains. Choose one that interests you most, and walk through the six steps, using the guiding prompts provided.

Choose One Question from the List Below

1. **Education:** Do students who participate in after-school tutoring programs perform better on standardized math tests?
2. **Environment:** Is there a relationship between air quality and asthma-related hospital visits in major cities?
3. **Health:** How does average daily screen time differ between high school students who sleep 6 or fewer hours and those who sleep 8 or more?
4. **Business:** Does the size of a restaurant's menu affect the average customer rating on review platforms?
5. **Sports:** Are baseball players with higher batting averages more likely to have higher on-base percentages?

Work Through the Process

Take your chosen question and apply the steps below. Use your notebook, a worksheet, or small group discussion to organize your thinking. Some of the ideas, vocab or concepts below may be unfamiliar, but take some time thinking about each of the steps as well as the process as a whole. Throughout the book we will reflect back on these steps with the goal of filling in all the concepts and details.

---

1. Ask a Clear Question

Rephrase the topic into a well-formed statistical question.

- Does your question involve variability?
- Can it be answered by collecting real data (not just opinions)?
- What exactly are you comparing or examining?

2. Collect Accurate and Relevant Data

Imagine or sketch out a plan to collect meaningful data.

- Who or what will you study (your population)?
- Will you take a sample, and if so, how will you select it?
- Where will the data come from? A survey, experiment, database, or sensors?
- Could any forms of bias affect your results?

3. Organize and Summarize the Data

Think about how to make sense of what you collect.

- Would tables, averages, or percentages be useful?
- What types of graphs or displays might help explain the data?
- How will you account for trends, groups, or outliers?

4. Analyze the Data

What relationship or pattern are you looking for?

- Are you comparing two groups? Looking for a correlation?
- Will you be calculating differences, percent changes, or association measures?
- Would a statistical test or confidence interval help support your findings later?

5. Interpret the Results in Context

What do your calculations and patterns actually mean?

- What is your overall takeaway?
- How confident are you in your result? What limitations exist?
- Could there be alternative explanations?

6. Communicate Clearly and Effectively

Imagine explaining this to someone unfamiliar with the data.

- How would you explain your findings in plain English?
- What visuals or summaries would support your answer?
- What key information must be included to avoid misinterpretation?

Looking Ahead: Designing a Statistical Study

Next we will look at how to synthesis ideas of questions, sampling, and data collection into the framework of a statistical study.

This page titled 1.5: Exploring Statistical Questions is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Mathematics Department.

1.6: Designing a Statistical Study

Launching Your Statistical Study

As we move forward in the course, you'll be applying the skills you're learning by designing and carrying out your own **statistical study**. This will be a semester-long project where you take on the role of a statistician asking a real-world question, collecting or analyzing relevant data, uncovering patterns, and sharing what you've found.

In this short section, we'll outline the structure of a statistical study and identify what you'll need to include as you begin planning your investigation.

What Is a Statistical Study?

A **statistical study** uses data to answer a question about a population or process. A good study is built on clear thinking, purposeful design, and awareness of bias and variability.

Statistical studies can take different forms, including:

- Surveys —> collecting responses from individuals using questions
- Observational studies —> recording observations without interference
- Experiments —> intentionally applying treatments and comparing outcomes
- Data analyses —> working with existing datasets to explore a question

You'll choose a question and approach that fits your interests and access to data.

Elements of a Well-Planned Study

As you begin, here are the major components you'll need to outline (many of which connect to ideas from Chapter 1):

- **A good statistical question** —> something that involves variability and can be answered with data
- **A target population** —> the group you want to learn about
- **A sampling strategy or data source** —> how you'll choose participants or find your data
- **A plan for data collection or acquisition** —> what variables you'll record, and how
- **Ways to minimize bias** —> including your role in the study and any design choices
- **A timeline or check-ins** —> since this will unfold over several weeks

What Happens Next?

We will look at the premise of a semester long study investigating housing affordability in your area.

This page titled 1.6: Designing a Statistical Study is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Mathematics Department.

1.6.1: Project Introduction

Semester Project: Investigating Housing Affordability

Throughout this course, you will work on a semester-long project that gives you the opportunity to apply what you're learning in a meaningful, hands-on way. The focus of the project is a topic you've likely heard a lot about: **housing affordability**.

You'll take on the role of a data analyst investigating a key question about housing in your own region. Along the way, you'll learn the statistical skills needed to collect, describe, analyze, and communicate your results step by step, as they're introduced in each chapter. By the end of the course, you'll produce either a written report or a short presentation that brings all your findings together.

Project Theme: Housing Affordability

Each student (or team) will choose one foundational focus area:

- Single-family home prices
- Rental prices
- Condominium (condo) prices

This will define the type of properties you'll be collecting data about in your region. You'll identify sources of housing data online (public listings, real estate websites, government resources, etc.) and gather data to analyze over the course of the semester.

Project Roadmap

This project will evolve in stages throughout the course. The full study includes:

1. **Statistical question:** Define a focused, researchable question related to housing affordability. For example:
    - What is the average cost of renting a 2-bedroom apartment in my city?
    - Have condo listing prices increased more in downtown neighborhoods vs. suburbs?
    - What is the variation in home prices across ZIP codes?
2. **Data collection plan:** Identify where and how you'll collect your data. Consider:
    - What websites or databases can you gather data from?
    - What information (variables) should you record? Examples: location, price, size (sq. ft), number of bedrooms, zip code, listing date, etc.
    - How can you avoid introducing bias into the sample?
3. **Descriptive statistics:** As we cover these in Chapter 2, you will describe your dataset by summarizing key variables numerically (e.g., average prices, ranges, medians) and noting patterns or outliers.
4. **Data visualization:** In Chapter 3, you'll create graphs and plots to help visualize your housing data (such as histograms, box plots, or scatterplots).
5. **Inference and confidence intervals:** Later in the course, you'll estimate parameters like average price using confidence intervals to reflect uncertainty about the population.
6. **Hypothesis testing:** You'll form a testable claim (e.g., "rents are higher in zip code A than zip code B") and test it using appropriate methods (t-tests, etc.). We'll provide step-by-step guidance when we reach those chapters.
7. **Final report or presentation:** At the end of the semester, you'll compile your findings — including cleaned and summarized data, visuals, written interpretation, and conclusions — into a report or shared class presentation.

This Week's Assignment: Start Collecting Data

To get started, you'll complete an initial project assignment:

1. Choose your **focus area**: home sales, rental listings, or condos.
2. Define your **geographic area**: local city, county, region, or ZIP code.
3. Find at least **15–20 listings** from a trusted source (e.g. Zillow, Realtor.com, Rentals.com, Craigslist, or government housing data).
4. Record the following variables for each listing (as available):
    - Price

- Price
- Location (address or ZIP)
- Property type
- Number of bedrooms
- Square footage
- Year built (if available)
- Listing date
- Any other variables of interest

5. Save your data into a spreadsheet or table (Google Sheets, Excel, CSV).

Submission Instructions:

Your instructor will provide submission details (how/where to upload your spreadsheet or report this information).

What to Expect

You do **not** need to clean, summarize, or analyze your data yet. We'll do this together, one skill at a time, starting in Chapter 2.

As the course continues, we'll use your dataset in:

- Examples and case studies in class
- Practice problems and assignments
- Midterm and final assessments

By the end of the course, you'll have completed a full statistical investigation with real-world data and have something meaningful to show for it.

This page titled 1.6.1: Project Introduction is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Mathematics Department.

1.7: Vocabulary (Chapter 1)

Chapter 1 Vocabulary

Below is a list of key vocabulary terms introduced in Chapter 1. These words form the foundation of your statistical thinking and will appear throughout the rest of the course. Keep this glossary handy as a reference!

Data

Information collected about individuals, objects, events, or conditions. Each data point represents an observation.

Observation

A single row or record in a dataset; one instance of data collected from an individual or object.

Dataset

A structured collection of data, typically organized in rows (observations) and columns (variables).

Variable

A characteristic that is recorded about each observation and can vary from one observation to another.

Statistics

The science of collecting, analyzing, interpreting, and communicating data in the presence of variability.

Variability

The tendency for data values to differ from one another; natural differences observed in real-world data.

Categorical Variable

A variable that places observations into non-numeric groups or categories.

Numerical Variable

A variable that represents quantifiable amounts or measurements; values are numeric and meaningful.

Nominal Variable

A categorical variable with no natural order (e.g., color, gender, blood type).

Ordinal Variable

A categorical variable with a meaningful order but inconsistent spacing between values (e.g., satisfaction level, fitness difficulty).

Discrete Variable

A numerical variable based on counts; values are whole numbers (e.g., number of pets, number of bedrooms).

Continuous Variable

A numerical variable based on measurements; values can include decimals and fall along an interval (e.g., height, income, temperature).

Population

The entire group of individuals or observations that you're interested in studying or learning about.

Sample

A subset of the population that is actually studied or measured. Used to make inferences about the population.

Sampling

The process of selecting individuals or units from a population to create a sample.

Representative Sample

A sample that accurately reflects the characteristics of the population.

Simple Random Sample (SRS)

A sample in which every member of the population has an equal chance of being selected.

Stratified Sample

A sample created by dividing the population into subgroups (strata) and randomly sampling from each subgroup.

Cluster Sample

A sample created by dividing the population into groups (clusters), then randomly selecting entire clusters.

Systematic Sample

A sample chosen by selecting every nth individual from a list or process.

Convenience Sample

A non-random sample chosen based on ease of access, often leading to bias.

Bias

Systematic error in the way a sample is collected or data is measured, leading to results that do not accurately reflect the population.

Sampling Bias

A specific type of bias introduced when some members of the population are less likely to be included in the sample than others.

Convenience Bias

Bias introduced when individuals are selected simply because they are easy to reach.

Nonresponse Bias

Bias that occurs when individuals selected for a survey do not respond, and their views differ from those who did respond.

Undercoverage Bias

Bias caused by systematically missing part of the population during sampling.

Voluntary Response Bias

Bias that occurs when participants choose to respond — often those with strong opinions are overrepresented.

Random Sampling

A method of sampling where each member of the population has a known and typically equal chance of being selected.

Statistical Question

A question that can be answered using data and involves variability in the answers.

Observational Study

A study in which researchers observe and measure variables without affecting or influencing subjects.

Survey

A method of collecting information from individuals by asking structured questions.

Experiment

A study in which researchers deliberately apply a treatment and observe the outcomes, often using random assignment.

Treatment Group

The group that receives the experimental condition or change in an experiment.

Control Group

The group in an experiment that does not receive the treatment; used as a baseline comparison.

Confounding Variable

An outside factor that affects both the explanatory and response variable, potentially obscuring results.

Random Assignment

Assigning individuals to groups in an experiment by chance to ensure fairness and reduce bias.

Blinding

A method used in experiments to prevent subjects or researchers from knowing which treatment was given, to reduce bias.

Statistical Process

A step-by-step approach to answering questions with data: ask a question, collect data, organize/summarize, analyze, interpret, communicate.

Data Collection Plan

A plan that outlines what data will be gathered, how it will be obtained, and what variables will be included.

Variable of Interest

A specific characteristic or measurement that is central to a particular study or research question.

Semester Project

A long-term project where students apply the full statistical process to a real-world topic — in this case, housing affordability.

Tip: Don't try to memorize everything! Instead, refer back to these terms as we dive deeper. You'll become familiar with each one as we use it in real examples, calculations, and discussions.

This page titled 1.7: Vocabulary (Chapter 1) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Mathematics Department.

1.8: Practice (Chapter 1)

1.1 What is Statistics?

1. What is the definition of data? Give an example from everyday life.
2. Explain the difference between a dataset and a single observation.
3. What are statistics?
    a. A question with a variety of answers.
    b. A way to measure the entire population
    c. The science of collecting, organizing, analyzing and interpreting data.
    d. A question from a survey.
4. Describe the process of statistics in your own words.
5. What are some examples of sources of data that are commonly used in research?
6. Describe one way you've encountered data in your life outside of school.
7. What is the goal of statistics as a discipline?
8. Why is variability important when working with data?
9. Give one real-world example of a statistical investigation. State what was studied and what the researchers were trying to learn.
10. A cognitive psychologist is interested in comparing two ways of presenting stimuli on subsequent memory. Twelve subjects are presented with each method and a memory test is given. What would be the roles of descriptive and inferential statistics in the analysis of these data?
11. A teacher wishes to know whether the full time students have better study habits than the part time students. A questionnaire is distributed assessing study skills and the full time and part time students are compared. Is this an example of descriptive or inferential statistics?
12. Describe the difference between descriptive statistics and inferential statistics. Illustrate with an example.
13. When inferential statistics are used, the goal is to _____ results from samples to the populations from which they were drawn.
14. _____ statistics include calculations such as the mean, median, and mode to summarize data from samples univariately.

1.2 Data Types: Categorical vs. Quantitative

1. Define the term "variable" in a statistical context.
2. For each of the following, say whether the variable is categorical or numerical:
    a. Age
    b. Favorite color
    c. Number of siblings
    d. Grade level
3. State whether the variable is discrete or continuous.
    a. A cat's weight.
    b. The height of a building.
    c. A person's age.
    d. The number of floors of a skyscraper.
    e. The number of clothing items available for purchase.
4. State whether the variable is discrete or continuous.
    a. Temperature in degrees Celsius.
    b. The number of cars for sale at a car dealership.
    c. The time it takes to run a marathon.
    d. The amount of mercury in a tuna fish.
    e. The weight of a hummingbird.

5. For each of the following, describe as numerical discrete, numerical continuous, or categorical, and give an example value of the variable.
    a. Household income
    b. Movie rating (G, PG, PG-13, R)
    c. Eye color
    d. Number of classes taken
    e. Daily temperature
    f. Age of executives in Fortune 500 companies
    g. Number of tickets sold to a concert
    h. Favorite baseball team
    i. Time in line waiting to buy groceries
    j. Number of students enrolled at Red Rocks Community College
    k. Most-watched television show
    l. Brand of toothpaste
    m. Distance to the closest movie theater
    n. Number of competing computer spreadsheet software packages

Use the following information to answer the next two exercises: A study was done to determine the age, number of times per week, and the duration (amount of time) of resident use of a local park in Lakewood. The first house in the neighborhood around the park was selected randomly and then every 8th house in the neighborhood around the park was interviewed.

6. “Number of times per week” is what type of data?
    a. categorical
    b. numerical discrete
    c. numerical continuous
7. “Duration (amount of time)” is what type of data?
    a. categorical
    b. numerical discrete
    c. numerical continuous
8. The Well-Being Index is a survey that follows trends of U.S. residents on a regular basis. There are six areas of health and wellness covered in the survey: Life Evaluation, Emotional Health, Physical Health, Healthy Behavior, Work Environment, and Basic Access. Some of the questions used to measure the Index are listed below. Identify the type of data obtained from each question used in this survey: categorical, numerical discrete, or numerical continuous.
    a. Do you have any health problems that prevent you from doing any of the things people your age can normally do?
    b. During the past 30 days, for about how many days did poor health keep you from doing your usual activities?
    c. In the last seven days, on how many days did you exercise for 30 minutes or more?
    d. Do you have health insurance coverage?
9. Select the measurement scale Nominal, Ordinal, Interval or Ratio for each scenario.
    a. A person’s age.
    b. A person’s race.
    c. Age groupings (baby, toddler, adolescent, teenager, adult, elderly).
    d. Clothing brand.
    e. A person’s IQ score.
    f. Temperature in degrees Celsius.
    g. The amount of mercury in a tuna fish.
10. Select the measurement scale Nominal, Ordinal, Interval or Ratio for each scenario.
    a. Temperature in degrees Kelvin.
    b. Eye color.
    c. Year in school (freshman, sophomore, junior, senior).

d. The weight of a hummingbird.
e. The height of a building.
f. The amount of iron in a person's blood.
g. A person's blood type.

11. Find a data set on the kaggle page (opens in new window) that includes both categorical and numerical variables. Describe the dataset as a whole and then list all the categorical and all the numerical variables. For the numerical variables, include if they are continuous or discrete.
12. What is the difference between a categorical variable and a numerical variable?
13. Explain the difference between a discrete and continuous variable. Give an example of each.
14. Explain the difference between a nominal and ordinal variable. Give an example of each.
15. "Star ratings" on a product review site are what kind of variable? Explain.
16. Someone collects data on the breed, age, and weight of 50 dogs. Classify each variable.
17. Give your own example for each type of variable:
    a. Nominal
    b. Ordinal
    c. Discrete
    d. Continuous
18. Why does classifying your variables correctly matter when performing statistical analysis?
19. Can a variable be categorical in one context and numerical in another? Give an example.
20. Describe a situation where categorical and numerical variables appear together in a study.

1.3 Populations, Samples, and Sampling Methods

For each of the following exercises, identify: a. the population, b. the sample, c. the parameter, d. the statistic, e. the variable, and f. the data. Give examples where appropriate.

1. A fitness center is interested in the mean amount of time a client exercises in the center each week.
2. Ski resorts are interested in the mean age that children take their first ski and snowboard lessons. They need this information to plan their ski classes optimally.
3. A marriage counselor is interested in the proportion of clients she counsels who stay married.
4. A cardiologist is interested in the mean recovery period of their patients who have had heart attacks.
5. Insurance companies are interested in the mean health costs each year of their clients, so that they can determine the costs of health insurance.
6. A politician is interested in the proportion of voters in their district who think the politician is doing a good job.
7. Political pollsters may be interested in the proportion of people who will vote for a particular cause.
8. A marketing company is interested in the proportion of people who will buy a particular product.

Use the following information to answer the next three exercises: A Red Rocks Community College instructor is interested in the mean number of days Red Rocks Community College math students are absent from class during a quarter.

9. What is the population she is interested in?
    a. all Red Rocks Community College students
    b. all Red Rocks Community College English students
    c. all Red Rocks Community College students in the instructor's classes
    d. all Red Rocks Community College math students
10. Consider the following: $\boldsymbol{X}$ = number of days a Red Rocks Community College math student is absent. In this case, X is an example of:
    a. variable
    b. population
    c. statistic
    d. data

11. The instructor's sample produces a mean number of days absent of 3.5 days. This value is an example of:
    a. parameter
    b. data
    c. statistic
    d. variable
12. Suppose you want to determine the mean number of students per statistics class in your state. Describe a possible sampling method in three to five complete sentences. Make the description detailed.
13. Suppose you want to determine the mean number of cans of soda drunk each month by students in their twenties at your school. Describe a possible sampling method in three to five complete sentences. Make the description detailed.
14. The instructor takes a sample by gathering data on five randomly selected students from each Red Rocks Community College math class. The type of sampling used is
    a. cluster sampling
    b. stratified sampling
    c. simple random sampling
    d. convenience sampling
15. A study was done to determine the age, number of times per week, and the duration (amount of time) of residents using a local park in San Jose. The first house in the neighborhood around the park was selected randomly and then every eighth house in the neighborhood around the park was interviewed. The sampling method was:
    a. simple random
    b. systematic
    c. stratified
    d. cluster
16. Name the sampling method used in each of the following situations:
    a. A person in the airport is handing out questionnaires to travelers asking them to evaluate the airport's service. The person does not ask travelers who are hurrying through the airport with their hands full of luggage, but instead asks all travelers who are sitting near gates and not taking naps while they wait
    b. A teacher wants to know if students are doing homework so they randomly select rows two and five and then call on all students in row two and all students in row five to present the solutions to homework problems to the class
    c. The marketing manager for an electronics chain store wants information about the ages of its customers. Over the next two weeks, at each store location, 100 randomly selected customers are given questionnaires to fill out asking for information about age, as well as about other variables of interest
    d. The librarian at a public library wants to determine what proportion of the library users are children. The librarian has a tally sheet on which they mark whether books are checked out by an adult or a child. The librarian records this data for every fourth patron who checks out books
    e. A political party wants to know the reaction of voters to a debate between the candidates. The day after the debate, the party's polling staff calls 1,200 randomly selected phone numbers. If a registered voter answers the phone or is available to come to the phone, that registered voter is asked whom they intend to vote for and whether the debate changed their opinion of the candidates.
17. Give an example of a population and two different characteristics that may be of interest.
18. Identify which value represents the sample mean and which value represents the claimed population mean.
    a. A recent article in a college newspaper stated that college students get an average of 5.5 hrs of sleep each night. A student who was skeptical about this value decided to conduct a survey by randomly sampling 25 students. On average, the sampled students slept 6.25 hours per night
    b. American households spent an average of about $52 in 2007 on Halloween merchandise such as costumes, decorations and candy. To see if this number had changed, researchers conducted a new survey in 2008 before industry numbers were reported. The survey included 1,500 households and found that average Halloween spending was $58 per household.
    c. The average GPA of students in 2001 at a private university was 3.37. A survey on a sample of 203 students from this university yielded an average GPA of 3.59 in Spring semester of 2025.

19. In a study, the sample is chosen by separating all cars by size, and selecting 10 of each size grouping. What is the sampling method? Is the sample representative?
20. In a study, the sample is chosen by writing everyone's name on a playing card, shuffling the deck, then choosing the top 20 cards. What is the sampling method? Is the sample representative?
21. In a study, the sample is chosen by asking people on the street. What is the sampling method? Is the sample representative?
22. In a study, the sample is chosen by selecting a room of the house, and appraising all items in that room. What is the sampling method? Is the sample representative?
23. In a study, the sample is chosen by surveying every third driver coming through a tollbooth. What is the sampling method? Is the sample representative?
24. Define the term population in a statistical study.
25. What's a sample? How is it different from the population?
26. Why do we usually collect data from a sample, not the entire population?
27. Explain the concept of a representative sample. Why is it important?
28. Which sampling method is being used in each case:
    a. A random number generator selects 200 voters from a list
    b. A company emails a survey to all its customers and records whoever replies
    c. A school surveys every third student entering the cafeteria
    d. A study randomly picks five neighborhoods, then surveys every household there
29. Compare and contrast simple random sampling and stratified sampling.
30. When might cluster sampling be more practical than other methods?
31. Which sampling method might best ensure racial diversity in your city's resident sample?
32. A teacher surveys students in her own class to evaluate school-wide stress. Identify any concerns.
33. You need to survey employees at a nationwide company. Suggest a good sampling method and explain why.
34. What are the characteristics of a "bad" sample?

1.4 Sources of Bias and Randomness

1. Define bias in the context of statistics.
2. Give an example of a biased sample and explain the effect it might have on conclusions.
3. Define each of the following:
    a. Selection bias
    b. Convenience bias
    c. Nonresponse bias
    d. Voluntary response bias
4. A researcher collects data by placing a QR code on a receipt that links to a survey. What types of bias might occur?
5. How does random sampling help reduce bias?
6. What's the difference between random error and bias?
7. Explain the value of randomness when selecting a sample. What are we trying to protect against?
8. Airline companies are interested in the consistency of the number of babies on each flight, so that they have adequate safety equipment. Suppose an airline conducts a survey. Over Thanksgiving weekend, it surveys six flights from Denver to Salt Lake City to determine the number of babies on the flights. It determines the amount of safety equipment needed by the result of that study.
    a. Using complete sentences, list three things wrong with the way the survey was conducted.
    b. Using complete sentences, list three ways that you would improve the survey if it were to be repeated.
9. A "random survey" was conducted of 3,274 people of the "microprocessor generation" (people born since 1971, the year the microprocessor was invented). It was reported that 48% of those individuals surveyed stated that if they had $2,000 to spend, they would use it for computer equipment. Also, 66% of those surveyed considered themselves relatively savvy computer users.
    a. Do you consider the sample size large enough for a study of this type? Why or why not?

b. Based on your “gut feeling,” do you believe the percents accurately reflect the U.S. population for those individuals born since 1971? If not, do you think the percents of the population are actually higher or lower than the sample statistics? Why?

10. Additional information: The survey, reported by Intel Corporation, was filled out by individuals who visited the Los Angeles Convention Center to see the Smithsonian Institute's road show called “America’s Smithsonian.”
    a. With this additional information, do you feel that all demographic and ethnic groups were equally represented at the event? Why or why not?
    b. With the additional information, comment on how accurately you think the sample statistics reflect the population parameters.
11. In advance of the 1936 Presidential Election, a magazine titled Literary Digest released the results of an opinion poll predicting that the republican candidate Alf Landon would win by a large margin. The magazine sent post cards to approximately 10,000,000 prospective voters. These prospective voters were selected from the subscription list of the magazine, from automobile registration lists, from phone lists, and from club membership lists. Approximately 2,300,000 people returned the postcards. Think about the state of the United States in 1936. Explain why a sample chosen from magazine subscription lists, automobile registration lists, phone books, and club membership lists was not representative of the population of the United States at that time.
    a. What effect does the low response rate have on the reliability of the sample?
    b. Are these problems examples of sampling error or non-sampling error?
    c. During the same year, George Gallup conducted his own poll of 30,000 prospective voters. These researchers used a method they called "quota sampling" to obtain survey answers from specific subsets of the population. Quota sampling is an example of which sampling method described in this module?
12. A scholarly article about response rates begins with the following quote: "Crime-related and demographic statistics for 47 US states in 1960 were collected from government agencies, including the FBI's *Uniform Crime Report*. One analysis of this data found a strong connection between education and crime indicating that higher levels of education in a community correspond to higher crime rates". Which of the potential problems with samples could explain this connection?
13. Imagine you work for a polling company and a member of your team has proposed the following survey question: “Do you feel happy paying your taxes while some politicians are allowed to use loopholes and avoid paying their fair share of taxes?”. As part of preliminary data collection, 11 people responded to this question. Each participant answered “NO!”. Which of the potential problems with samples could explain this connection?
14. “Declining contact and cooperation rates in random digit dial (RDD) national telephone surveys raise serious concerns about the validity of estimates drawn from such research.” (Scott Keeter et al., “Gauging the Impact of Growing Nonresponse on Estimates from a National RDD Telephone Survey,” Public Opinion Quarterly 70 no. 5 (2006), link(opens in new window) [poq.oxfordjournals.org] (accessed May 1, 2013)). The Pew Research Center for People and the Press admits: “The percentage of people we interview – out of all we try to interview – has been declining over the past decade or more.” (Frequently Asked Questions, Pew Research Center for the People & the Press, source(opens in new window) [www.people-press.org] (accessed May 1, 2013)).
    a. What are some reasons for the decline in response rate over the past decade.
    b. Explain why researchers are concerned with the impact of the declining response rate on public opinion polls.
15. The Gallup Poll uses a procedure called random digit dialing, which creates phone numbers based on a list of all area codes in America in conjunction with the associated number of residential households in each area code. Give a possible reason the Gallup Poll chooses to use random digit dialing instead of picking phone numbers from the phone book.
16. A statistics student who is curious about the relationship between the amount of time students spend on social networking sites and their performance at school decides to conduct a survey. Three research strategies for collecting data are described below. In each, name the sampling method proposed and any bias you might expect.
    a. They randomly sample 40 students from the study's population, gives them the survey, asks them to fill it out and bring it back the next day.
    b. They give out the survey only to their friends, and makes sure each one of them fills out the survey.
    c. They post a link to an online survey on their social media feed and asks friends to fill out the survey.

17. Suppose we want to estimate family size, where family is defined as one or more parents living with children. If we select students at random at an elementary school and ask them what their family size is, will our average be biased? If so, will it overestimate or underestimate the true value?
18. Identify the flaw in reasoning in the following scenarios. Explain what the individuals in the study should have done differently if they wanted to make such strong conclusions.
    a. Students at an elementary school are given a questionnaire that they are required to return after their parents have completed it. One of the questions asked is, "Do you nd that your work schedule makes it difficult for you to spend time with your kids after school?" Of the parents who replied, 85% said "no". Based on these results, the school officials conclude that a great majority of the parents have no difficulty spending time with their kids after school.
    b. A survey is conducted on a simple random sample of 1,000 women who recently gave birth, asking them about whether or not they smoked during pregnancy. A follow-up survey asking if the children have respiratory problems is conducted 3 years later, however, only 567 of these women are reached at the same address. The researcher reports that these 567 women are representative of all mothers.
    c. A orthopedist administers a questionnaire to 30 of his patients who do not have any joint problems and finds that 20 of them regularly go running. He concludes that running decreases the risk of joint problems.
19. 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.
20. Which of the methods above would produce the most representative sample and the best estimate? Explain.
    a. 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.
    b. 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.
21. Compare and contrast random sampling and random assignment.
22. In practice, wording effects are often an extremely strong influence on the answers people give when surveyed. Suppose you were doing a survey of American voters opinions of the president. Think of a way of asking a question which would tend to maximize the number of people who said they approved of the job he is doing. Then think of another way of asking a question which would tend to minimize that number [who say they approve of his job performance].
23. Think of a survey question you could ask in a survey of the general population of Americans in response to which many [most?] people would lie. State what would be the issue you would be investigating with this survey question, as a clearly defined, formal variable and parameter on the population of all Americans. Also tell exactly what would be the wording of the question you think would get lying responses.

1.5 Exploring Statistical Questions

1. What is a statistical question?
    a. A question where you expect to get a variety of answers and you are interested in the distribution and tendency of those answers
    b. A question using reported statistics.
    c. A question on a survey.
    d. A question on a census.
2. Which of the following are statistical questions? Select all that apply.

a. How old are you?
b. What is the weight of a mouse?
c. How tall are all 3-year-olds?
d. How tall are you?
e. What is the average blood pressure of adults?

3. Which of the following are statistical questions? Explain your reasoning:
   a. How tall is Mount Everest?
   b. How much sleep do teenagers get on weekdays?
   c. What is my shoe size?
   d. Do students who drink coffee study more?
4. Identify the explanatory and response variables for a study of whether playing music affects test scores.
5. What makes a good study design? List three key features.
6. In your own words, why is it important to have a plan before collecting data?
7. Identify the steps in the statistical process in the context of the studies described below.
   a. Researchers collected data to examine the relationship between pollutants and preterm births in Southern California. During the study air pollution levels were measured by air quality monitoring stations. Specifically, levels of carbon monoxide were recorded in parts per million, nitrogen dioxide and ozone in parts per hundred million, and coarse particulate matter (PM_{10}) in $\mu g = m^3$. Length of gestation data were collected on 143,196 births between the years 1989 and 1993, and air pollution exposure during gestation was calculated for each birth. The analysis suggested that increased ambient PM10 and, to a lesser degree, CO concentrations may be associated with the occurrence of preterm births.
   b. The Buteyko method is a shallow breathing technique developed by Konstantin Buteyko, a Russian doctor, in 1952. Anecdotal evidence suggests that the Buteyko method can reduce asthma symptoms and improve quality of life. In a scientific study to determine the effectiveness of this method, researchers recruited 600 asthma patients aged 18-69 who relied on medication for asthma treatment. These patients were split into two research groups: one practiced the Buteyko method and the other did not. Patients were scored on quality of life, activity, asthma symptoms, and medication reduction on a scale from 0 to 10. On average, the participants in the Buteyko group experienced a significant reduction in asthma symptoms and an improvement in quality of life.
   c. In a study of the relationship between socio-economic class and unethical behavior, 129 University of California undergraduates at Berkeley were asked to identify themselves as having low or high social-class by comparing themselves to others with the most (least) money, most (least) education, and most (least) respected jobs. They were also presented with a jar of individually wrapped candies and informed that they were for children in a nearby laboratory, but that they could take some if they wanted. Participants completed unrelated tasks and then reported the number of candies they had taken. It was found that those in the upper-class rank condition took more candy than did those in the lower-rank condition.
8. The Buteyko method is a shallow breathing technique developed by Konstantin Buteyko, a Russian doctor, in 1952. Anecdotal evidence suggests that the Buteyko method can reduce asthma symptoms and improve quality of life. In a scientific study to determine the effectiveness of this method, researchers recruited 600 asthma patients aged 18-69 who relied on medication for asthma treatment. These patients were split into two research groups: one practiced the Buteyko method and the other did not. Patients were scored on quality of life, activity, asthma symptoms, and medication reduction on a scale from 0 to 10. On average, the participants in the Buteyko group experienced a significant reduction in asthma symptoms and an improvement in quality of life. [McGowan. “Health Education: Does the Buteyko Institute Method make a difference?” In: Thorax 58 (2003).] Which of the following is the main research question?
   a. The Buteyko method causes shallow breathing.
   b. The Buteyko method can reduce asthma symptoms and an improvement in quality of life.
   c. Effectiveness of the Buteyko method.
   d. The patients score on quality of life, activity, asthma symptoms and medication reduction.

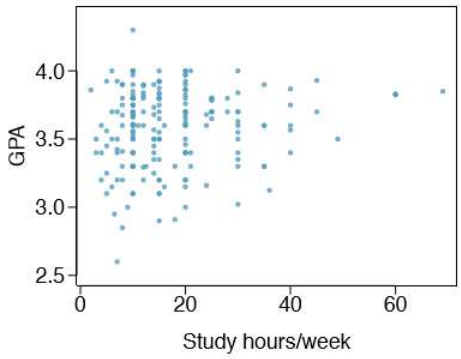
Read the following study description and answer the following questions: The United States Government recommends that to stay physically fit, middle-aged adults (ages 40 to 60) need to burn 150 to 400 calories per day doing exercise. Researchers at Minnesota State University, Mankato, wanted to learn 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.

9. Which question below is a reasonable research question for this investigation? Explain how you came to this answer.
    a. Do people think that playing the Wii Fit video game burns calories?
    b. Does the Wii Fit video game burn enough calories to be considered suitable exercise?
    c. Does the Wii Fit video game burn more calories than traditional exercise?
    d. What is the average amount of time that middle-aged adults spend playing Wii Fit video games?
10. What data did the researchers collect to answer the research question?
    a. The amount of time that the adults exercised.
    b. The name of the adults.
    c. The type of exercise the adults completed.
    d. The total amount of calories the adults burned through exercising.
11. How are the data summarized?
    a. Researchers found the proportion of adults who exercise using the Wii Fit video game.
    b. Researchers found the proportion of adults who prefer exercising with the Wii Fit video game over traditional exercises.
    c. Researchers found the average amount of calories that adults consumed through exercising using the Wii Fit video game.
    d. Researchers found the average amount of time that adults spent exercising using the Wii Fit video game.
12. What did the researchers conclude?
    a. The Wii Fit video game is a preferred exercise for some middle-aged adults.
    b. 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.
    c. 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.
    d. The sample size is too small to make any reasonable conclusions about all middle-aged adults.
13. Give an example of a research question that involves estimating a characteristic about the population of Registered Nurses in California.
14. Improve this poorly stated research question: Do registered nurses work a lot?
15. Create a cause-and-effect research question.
16. For each of the following topics, provide a research question that involves that topic.
    a. Air quality in the Denver metro area.
    b. The bike paths of Lakewood.
    c. Access to fresh produce in Wheat Ridge.
    d. Cost of food in Colorado.
    e. Pokemon cards.
17. Suppose that a research question has been formed. In 6 steps, describe the process of exploring this question. Think about what data we might need, how to collect that data and what we might do with that data. (This question is what this semester is about. At this point we are looking for some initial plans and we will refine this answer throughout the class)
18. Find a public data set from the website www.Kaggle.com. Describe the data set and provide an example of a research question you could explore using this data.

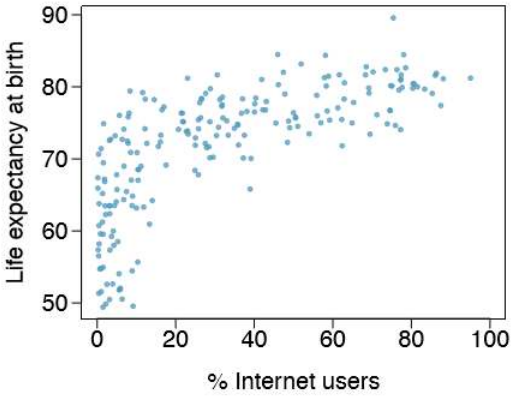
1.6 Designing a Statistical Study

1. List some practical difficulties involved in getting accurate results from a telephone survey.
2. List some practical difficulties involved in getting accurate results from a mailed survey.
3. Describe the difference between an observational study and an experiment.
4. How is a survey different from an experiment? What limitations do surveys have?
5. With your classmates, brainstorm some ways you could overcome these problems if you needed to conduct a phone or mail survey.
6. A survey was conducted on 218 undergraduates from Duke University who took an introductory statistics course in Spring 2012. Among many other questions, this survey asked them about their GPA and the number of hours they spent studying per week. The scatterplot below displays the relationship between these two variables.



   a. What is the explanatory variable and what is the response variable?
   b. Describe the relationship between the two variables. Make sure to discuss unusual observations, if any.
   c. Is this an experiment or an observational study?
   d. Can we conclude that studying longer hours leads to higher GPAs?

7. A large college class has 160 students. All 160 students attend the lectures together, but the students are divided into 4 groups, each of 40 students, for lab sections administered by different teaching assistants. The professor wants to conduct a survey about how satis ed the students are with the course, and he believes that the lab section a student is in might affect the student's overall satisfaction with the course.
   a. What type of study is this?
   b. Suggest a sampling strategy for carrying out this study.
8. The scatterplot below shows the relationship between estimated life expectancy at birth as of 201260 and percentage of internet users in 201061 in 208 countries.
   a. Describe the relationship between life expectancy and percentage of internet users.
   b. What type of study is this?
   c. State a possible confounding variable that might explain this relationship and describe its potential effect.



9. In order to assess the effectiveness of taking large doses of vitamin C in reducing the duration of the common cold, researchers recruited 400 healthy volunteers from staff and students at a university. A quarter of the patients were assigned a placebo, and the rest were evenly divided between 1g Vitamin C, 3g Vitamin C, or 3g Vitamin C plus additives to be taken at onset of a cold for the following two days. All tablets had identical appearance and packaging. The nurses who handed the prescribed pills to the patients knew which patient received which treatment, but the researchers assessing the patients when they were sick did not. No significant differences were observed in any measure of cold duration or severity between the four medication groups, and the placebo group had the shortest duration of symptoms.[65]
    a. Was this an experiment or an observational study? Why?
    b. What are the explanatory and response variables in this study?
    c. Were the patients blinded to their treatment?
    d. Was this study double-blind?
    e. Participants are ultimately able to choose whether or not to use the pills prescribed to them. We might expect that not all of them will adhere and take their pills. Does this introduce a confounding variable to the study? Explain your reasoning.
10. You would like to conduct an experiment in class to see if your classmates prefer the taste of regular Coke or Diet Coke. Briefly outline a design for this study.
11. A researcher is interested in the effects of exercise on mental health and he proposes the following study: Use stratified random sampling to ensure representative proportions of 18-30, 31-40 and 41-55 year olds from the population. Next, randomly assign half the subjects from each age group to exercise twice a week, and instruct the rest not to exercise. Conduct a mental health exam at the beginning and at the end of the study, and compare the results.
    a. What type of study is this?
    b. What are the treatment and control groups in this study?
    c. Does this study make use of blocking? If so, what is the blocking variable?
    d. Does this study make use of blinding?
    e. Comment on whether or not the results of the study can be used to establish a causal relationship between exercise and mental health, and indicate whether or not the conclusions can be generalized to the population at large.
    f. Suppose you are given the task of determining if this proposed study should get funding. Would you have any reservations about the study proposal?
12. Chia Pets - those terra-cotta figurines that sprout fuzzy green hair - made the chia plant a household name. But chia has gained an entirely new reputation as a diet supplement. In one 2009 study, a team of researchers recruited 38 individuals and divided them evenly into two groups, and they randomly placed half of these participants into the treatment group and the other half into the control group. One group was given 25 grams of chia seeds twice a day, and the other was given a placebo. The subjects volunteered to be a part of the study. After 12 weeks, the scientists found no significant difference between the groups in appetite or weight loss.[66]
    a. What type of study is this?
    b. What are the experimental and control treatments in this study?
    c. Has blocking been used in this study? If so, what is the blocking variable?
    d. Has blinding been used in this study?
    e. Comment on whether or not we can make a causal statement, and indicate whether or not we can generalize the conclusion to the population at large.
13. State whether each study is observational or experimental.
    a. You want to determine if cinnamon reduces a person's insulin sensitivity. You give patients who are insulin sensitive a certain amount of cinnamon and then measure their glucose levels.
    b. A researcher wants to evaluate whether countries with lower fertility rates have a higher life expectancy. They collect the fertility rates and the life expectancies of countries around the world.
    c. A researcher wants to determine if diet and exercise together helps people lose weight over just exercising. The researcher solicits volunteers to be part of the study, and then randomly assigns the volunteers to be in the diet and exercise group or the exercise only group.

d. You collect the weights of tagged fish in a tank. You then put an extra protein fish food in water for the fish and then measure their weight a month later.

14. State whether each study is cross-sectional, retrospective or prospective.
    a. To see if there is a link between smoking and bladder cancer, patients with bladder cancer are asked if they currently smoke or if they smoked in the past.
    b. The Nurses Health Survey was a survey where nurses were asked to record their eating habits over a period of time, and their general health was recorded.
    c. A new study is underway to track the eating and exercise patterns of people at different time-periods in the future, and see who is afflicted with cancer later in life.
    d. The prices of generic items are compared to the prices of the equivalent named brand items.
15. “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 design. 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?
16. Complete the design-layout table.

A student would like to know which of two possible routes is faster for the daily trips to school. Route 1 is shorter but has many traffic lights. Route 2 is a little longer but doesn’t have traffic lights. Each morning, a coin flip will be used to determine the route taken to school. The time it takes for the commute will be measured with a stopwatch. After approximately 15 trials on each route, the average time for each will be compared.

Research design table.

Research Design Table	
Research Question:	
Type of Research	Observational Study Observational Experiment Manipulative Experiment
What is the response variable?	
What is the parameter that will be calculated?	Mean Proportion
List potential confounding variables.	
Grouping/explanatory Variable (if present)	Levels:

17. Many parents believe that their small children get a bit hyperactive when they eat or drink sweets (candies, sugary sodas, etc.), and so do not let their kids have such things before nap time, for example. A pediatrician at Euphoria State University Teaching Hospital [ESUTH] thinks instead that it is the parents’ expectations about the effects of sugar which cause their children to become hyperactive, and not the sugar at all.

a. Describe a randomized, placebo-controlled, double-blind experiment which would collect data about this ESUTH pediatrician's hypothesis. Make sure you are clear about both which part of your experimental procedure addresses each of those important components of good experimental design.
b. Is the experiment you described in the previous exercise an ethical one? What must the ESUTH pediatrician do before, during, and after the experiment to make sure it is ethical? Make sure you discuss (at least) the checklist of ethical guidelines from this chapter and how each point applies to this particular experiment.

Practice problems include parts of:

Introductory Statistics 2e by OpenStax is licensed under CC BY 4.0

Access for free at https://openstax.org/books/introductory-statistics-2e/pages/1-introduction

and

OpenIntro Statistics by David Diez, Christopher Barr, & Mine Çetinkaya-Rundel is licensed under CC BY 3.0

This page titled 1.8: Practice (Chapter 1) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Mathematics Department.

- **Current page** by Mathematics Department is licensed CC BY 4.0.
- **1.E: Introduction to Data (Exercises)** by David Diez, Christopher Barr, & Mine Çetinkaya-Rundel is licensed CC BY-SA 3.0. Original source: https://www.openintro.org/book/os.
- **Licensing** by Mathematics Department has no license indicated.

2: Exploring and Summarizing Data

Summarizing

Once we've collected data, the next step is to make sense of it. Datasets often include dozens, hundreds, or even millions of values. Instead of inspecting every individual data point, we use **summary statistics** to describe what the data looks like overall. These summaries help us identify patterns, compare groups, and make informed decisions.

In this chapter, we'll begin learning how to **process raw data** by calculating useful statistics. These tools simplify complex datasets into concise, meaningful numbers that help us interpret what's typical, what's unusual, and how things vary across a group.

What Are Summary Statistics?

Summary statistics are numerical values that give us insight into a dataset. Different statistics describe different features of the data. Some focus on its center, what's typical or average, while others describe how spread out the data is, or how it's shaped and grouped.

- **Measures of Center:** Tell us what a "typical" value looks like. Examples include the *mean* (average), *median*, and *mode*.
- **Measures of Spread:** Show how much the data varies or how spread out the values are. These include the *range*, *interquartile range (IQR)*, *standard deviation*, and *variance*.
- **Measures of Grouping and Shape:** Help us understand how the data is distributed such as whether values are *skewed*, follow a *bell shape*, or are grouped in clusters.

Together, these types of summaries give us a "big picture" view of the data. Rather than memorizing all the values, we can work with a small number of statistics that tell a story.

Why Is Summarizing Data Important?

Summary statistics allow us to:

- Compare two groups at a glance, for example comparing average home prices in different neighborhoods
- Identify outliers or unusual values that stand out
- Spot patterns or trends that might not be obvious in a full dataset
- Create visuals like boxplots and histograms based on calculated values
- Prepare for probability models and statistical inference later on

In real-life data problem, including our semester-long housing project, these summaries are often the first and most important step in exploring a dataset. Whether you're evaluating apartment rental trends, test scores, or survey responses, descriptive statistics help turn raw data into insight.

What You'll Learn in This Chapter

In the pages ahead, we'll cover how to:

- Calculate and interpret measures of center: mean, median, and mode
- Calculate and interpret measures of spread: range, IQR, and standard deviation
- Create and interpret data summaries using the five-number summary
- Identify and explain outliers

Whether you're working by hand or using software, these tools will give you a deep understanding of the shape and substance of your data — and allow you to start answering real questions.

Coming up:

We begin with measures of center, including averages and medians, and how they differ in the picture of typicalness they show us.

This page titled 2: Exploring and Summarizing Data is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Mathematics Department via source content that was edited to the style and standards of the LibreTexts platform.

2.1: Measures of Central Tendency- Mean, Median, and Mode

Imagine a local high school is holding a fundraiser to support a new science lab. Each student donates what they can; some give a few dollars, while others contribute much more. At the end of the campaign, the principal wants to share the results with the school community. But what number should she report? Should she present the average donation? The amount most students gave? The amount exactly in the middle?

These are all ways of summarizing data using a single number; what statisticians call **measures of center**. These values help us describe what is "typical" in a dataset and give us a meaningful way to compare data, report insights, or support decisions. Choosing the right measure often depends on the shape of the data and on what story you want to tell.

In this section, you'll explore the three most common summary statistics for center: the **mean** (average) [opens in new window], the **median** (middle value) [opens in new window], and the **mode** (most frequent value) [opens in new window]. Each of these captures a different perspective on what might be considered "typical" and contributes to understanding the greater story.

You'll learn what each measure means, how to calculate it, when it's useful, and when it can be misleading. After exploring interactive summary cards for each statistic, you'll examine the school fundraiser in more detail and compare how each measure helps tell the story. Finally, you'll see how to calculate these statistics in Excel; a skill you can apply in your own data projects.

Mean (Arithmetic Average)

Description: Add up all the values and divide by how many there are. This gives us the "center of gravity" of the data.

Symbols:

- $\bar{x}$: Sample mean
- $x_1, x_2, \ldots, x_n$: Individual data values
- n: Sample size (number of values)

Formula:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{2.1.1}$$

Notes: The mean is *not robust*. It is sensitive to outliers and skewed values.

Example: What is the mean of the values {5, 10, 12}?

- Sum: $5 + 10 + 12 = 27$
- Count: $n = 3$
- Mean: $\bar{x} = \frac{27}{3} = 9$

Quick Check: What is the mean of {3, 4, 8, 10}?

Mean formula

—Choose—

Check Answer

Median (Middle Value)

Description: The median is the middle number when the data is sorted. If there are an odd number of values, it's the one in the center. If there's an even number, it's the average of the two middle numbers.

Notation: Usually referred to as “median”. Abbreviation: **Med**

How to Find the Median:

- Sort the values from least to greatest
- If n is odd: take the middle value
- If n is even: average the two middle values

Notes: The median is *robust*. It is not affected by outliers or extreme values.

Example: What is the median of {4, 6, 9, 10, 11}?

- Ordered: {4, 6, 9, 10, 11}
- Middle value = **9**

Another Example: What is the median of {6, 8, 10, 14}?

- Ordered: {6, 8, 10, 14}
- Middle two values: 8 and 10
- Median: $\frac{8+10}{2} = 9$

Quick Check: What is the median of: {1, 3, 3, 7, 10, 12, 13}?

Steps to find the Mean

—Choose—

Check Answer

Mode (Most Frequent)

Description: The mode is the value that appears most often in a dataset. A set can have one mode (unimodal), more than one (bimodal or multimodal), or no mode at all. On a smooth distribution (shown below), the mode is the highest point on the graph.

Notation: Mode has no standard symbol; just use the word “mode.”

How to Find:

- Tally the frequency of each value
- Pick the value(s) that occur(s) the most

Notes: Mode is useful for categorical data and discrete data. It may not describe “center” well unless repeated values dominate the dataset.

Example 1: Mode of {1, 2, 2, 3, 4} is **2**

Example 2: Mode of {5, 6, 7, 8} → No repetitions → *No mode*

Example 3: Mode of {2, 3, 3, 4, 4, 5} → **Bimodal**: 3 and 4

Quick Check: What is the mode of {6, 2, 6, 3, 3, 3, 6}?

Steps to find the Mode

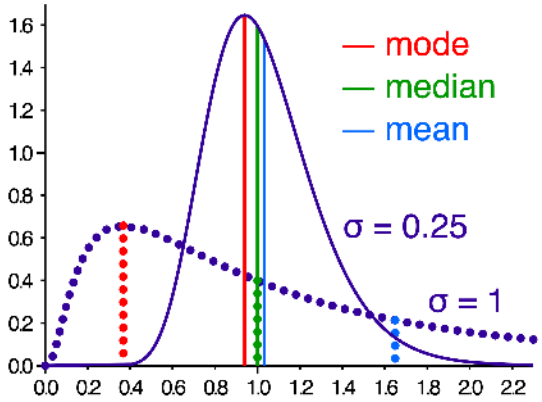
—Choose—

Check Answer

Visual Comparison: Mean, Median, and Mode

This visual shows a powerful comparison of how the **mean** (blue), **median** (green), and **mode** (red) behave within two different probability distributions. Both distributions are unimodal (one peak) but have different shapes and spreads.

The taller, narrower curve represents a dataset with smaller variability (standard deviation $\sigma = 0.25$) — the values are tightly centered. Notice how in this case, the mean, median, and mode are all very close together, reinforcing how symmetric distributions tend to align these three statistics.



The shorter, wider curve shows a more spread-out dataset ($\sigma = 1$). In this less symmetric shape, the mean, median, and mode become notably separated. The mode remains near the peak (most frequent value), the median stays near the center of mass, and the mean shifts toward the longer tail — pulled by extreme values.

This visualization helps illustrate why it's important to **examine the shape of your data before choosing which measure of center to report**. When your data is skewed, has outliers, or is irregularly distributed, these summary values can tell very different stories. We will go into this more in the next few sections.

As you analyze sample data (like in our upcoming example), remember to reflect on the overall distribution — not just compute a summary automatically. These differences also affect how we generalize from samples back to the larger **population**. Understanding the shape and the behavior of each summary statistic helps us make deeper, more credible conclusions from our data.

Example: Telling a Data Story with Measures of Center

Olive Grove High School recently held a district-wide fundraiser: each student was asked to contribute any amount they could toward a new science lab. The funds were collected anonymously, and the principal received a list of how much each of the 20 students donated. As the school newspaper reports on the results, they want to highlight a summary of the donations made.

Here's the dataset showing dollars donated by each student:

```
[0, 5, 5, 5, 10, 10, 10, 10, 15, 15, 20, 20, 20, 25, 30, 35, 40, 50, 90, 200]
```

Calculate the Measures

- **Mean = 32.75.** This is pulled upward by a few very large donations, especially the $200 gift (an outlier).
- **Median = 17.5.** Half the students gave less than $17.50 and half gave more. This better reflects what most students gave.
- **Mode = 10.** More students gave $10 than any other amount. This is the most "popular" donation level.

Storytelling with the Right Statistic

Each of these measures tells a slightly different story:

- The newspaper might highlight the **median** to report what a "typical" student gave and avoid distortion from unusually large donations.
- The principal might focus on the **mean** to emphasize overall generosity and total impact.
- A student campaign reflection might mention the **mode** to show which amount students most commonly felt comfortable donating.

> ***Which measure is best?*** *That depends on your audience and purpose. Choosing between mean, median, and mode is often less about math, and more about what story you're trying to tell. Usually, multiple measures of center should be reported to give a clearer picture of the data.*

Example: Computing Measures of Center in Excel

Let's apply what we've learned using Microsoft Excel, a useful tool for quickly computing summary statistics. Suppose you're tracking the number of hours 10 students reported studying during the past week. You've entered the data in an Excel column like this:

```
A  Hours Studied
1  8
2  5
3  7
4  10
5  4
6  6
7  12
8  8
9  7
10 9
11 20
```

Goal:

Use Excel to compute the mean, median, and mode of the data in column A.

Step-by-Step Instructions

1. **Open Excel** and type your data into cells A2 through A11 (under a header like "Hours Studied").
2. In any empty cell, compute the **mean** using a

   ```
   =AVERAGE(A2:A11)
   ```

3. Compute the **median** using:

   ```
   =MEDIAN(A2:A11)
   ```

4. Compute the **mode** using:

   ```
   =MODE.SNGL(A2:A11)
   ```

   Note: Use `=MODE.MULT` for multiple modes (in array formulas)

Interpretation:

Results you might get (based on the above data):

- **Mean:** 8.8 (skewed slightly upward by the 20-hour outlier)

- **Median:** 7.5 (a better “typical” value)
- **Mode:** 7 (most common number of study hours reported)

> *Which one would you highlight when reporting class study habits and why?*

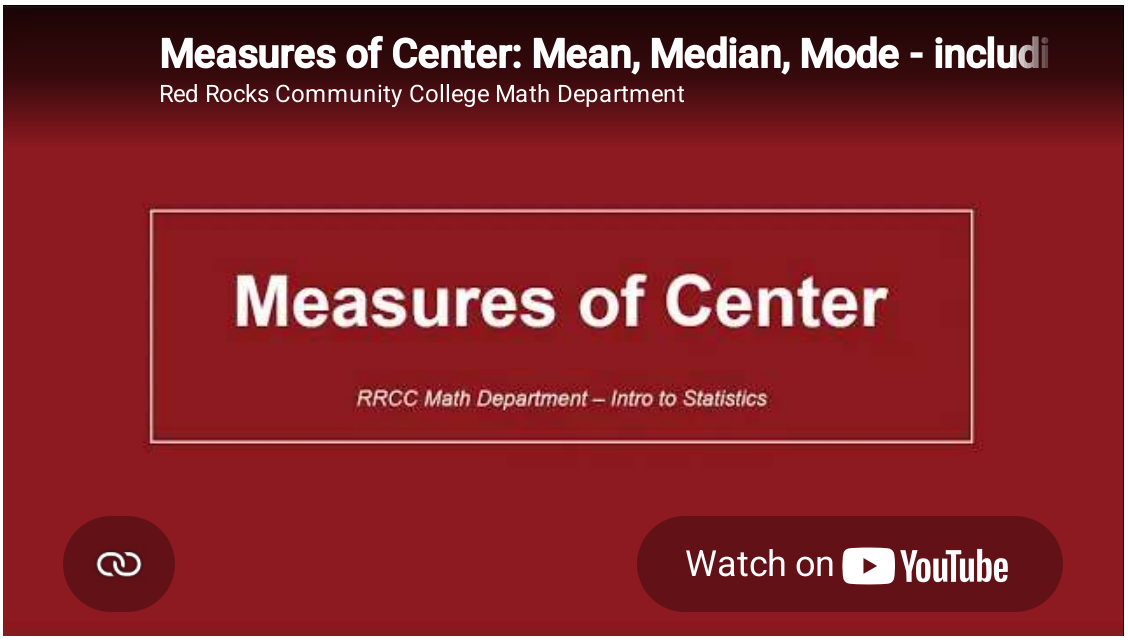
Optional: Add a Chart

1. Select your data in A2:A11.
2. Go to the **Insert** tab → choose a **Histogram** or **Column Chart**.
3. Add a title and observe how the visual distribution supports your values.

You can apply these same steps to your semester-long housing dataset to compute summary statistics for price, square footage, number of bedrooms, and more!

Next Up: We'll explore how data spreads out, including how to calculate the range, interquartile range (IQR), standard deviation, and more.

Related Videos





This page titled 2.1: Measures of Central Tendency- Mean, Median, and Mode is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Mathematics Department.

2.2: Measures of Spread- Range, Variance, and Standard Deviation

Measures of Spread: Range and Standard Deviation

In the previous section, we learned how to describe what's *typical* in a dataset using measures of center. But that doesn't tell us everything. Two datasets can have the same average while behaving very differently overall.

Measures of spread help us describe how much the values in a dataset vary or deviate from the center. They show how consistent your data is, how wide it is spread out, and how dependable a summary like the mean or median will be in representing the whole.

In this section, you'll learn two essential ways to measure spread:

- **Range** – the simple difference between the highest and lowest values
- **Standard Deviation** – the average distance each value is from the mean

Framing Example: Pizza Delivery Times

Imagine you're comparing two local pizza restaurants that advertise "fast delivery." You decide to test them and record the delivery times (in minutes) for 10 orders from each shop:

```
Shop A: 27, 28, 28, 29, 29, 30, 30, 30, 31, 31
Shop B: 20, 22, 25, 28, 30, 32, 35, 38, 40, 45
```

Both shops might have similar averages, but one shows much more consistent delivery times. After reviewing the three cards below, you'll come back to this example and calculate how spread helps you choose the more reliable option.

Range

Description: The range is the difference between the largest and smallest values in a dataset. It gives a quick sense of how wide the data is spread.

Formula:

$$\text{Range} = \text{Maximum} - \text{Minimum} \tag{2.2.1}$$

Notes:

- Very sensitive to outliers
- Use for a quick check, but not reliable for distributions with extreme values

Example: Dataset: $\{2, 4, 5, 6, 10\}$ → Range = $10 - 2 = 8$

Standard Deviation (SD)

Description: The standard deviation measures how spread out values are in a dataset. Specifically, it tells us how far each value typically is from the mean.

Formula (sample data):

$$s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}} \tag{2.2.2}$$

Symbols:

- x_i: Individual data value
- $\bar{x}$: Mean of the sample
- n: Number of values in the sample

Notes:

- Very sensitive to outliers
- Gives a complete view of spread for symmetric, bell-shaped data

How to Calculate Standard Deviation (by hand)

To help you understand the formula, let's break it down into six clear, repeatable steps:

1. Calculate the mean $\bar{x}$
2. Find the difference between each data point and the mean $(x_i - \bar{x})$
3. Square each difference: $(x_i - \bar{x})^2$
4. Add up all the squared differences: $\sum (x_i - \bar{x})^2$
5. Divide by $n-1$ (sample) or n (population) → this gives you the variance
6. Take the square root of the variance → this gives you the standard deviation

Example: Small Dataset

Let's follow the steps for the dataset: **{4, 6, 8, 10}**

1. Mean:

$$\bar{x} = \frac{4+6+8+10}{4} = \frac{28}{4} = 7 \tag{2.2.3}$$

2. Differences from the mean:

$$(4-7) = -3, \quad (6-7) = -1, \quad (8-7) = 1, \quad (10-7) = 3 \tag{2.2.4}$$

3. Squares of differences:

$$9,\ 1,\ 1,\ 9 \tag{2.2.5}$$

4. Sum of squares:

$$9 + 1 + 1 + 9 = 20 \tag{2.2.6}$$

5. Divide by $n - 1 = 3$ (sample):

$$\frac{20}{3} \approx 6.67 \tag{2.2.7}$$

   This is the **variance**

6. Take the square root:

$$s = \sqrt{6.67} \approx 2.58 \tag{2.2.8}$$

Interpretation: The values in this dataset vary, on average, about 2.58 units from the mean.

> Reminder:
>
> When you divide by $n-1$ instead of n, you're using the Bessel's correction(opens in new window). It makes the sample standard deviation a better estimate of the population value — we'll mostly use the sample version in this course.

In this course, we'll primarily use the **sample standard deviation**, because we usually work with data collected from a sample, not an entire population.

Coefficient of Variation

To consider the standard deviation in context of the mean is using the **Coefficient of Variation**, often abbreviated as CV The CV is simply defined as:

$$CV = \frac{s}{\bar{x}}, \tag{2.2.9}$$

or the ratio of the standard deviation to the mean. For example, if standard deviation is $s = 2.58$ and mean $\bar{x} = 7$, then the coefficient of variation is $CV = 2.58/7 = 0.368$ or 36.8%. The CV allows us to compare how wide a dataset is to its mean, which is particularly useful when looking at two similar datasets side by side.

Analyzing the Pizza Delivery Example

Now let's return to our earlier question: **Which pizza shop is more consistent with delivery times?**

From first glance, both shops might appear to deliver in roughly the same amount of time. But measures of spread can reveal important differences in reliability that basic averages might miss.

Here are the two sets of delivery times again:

```
Shop A: 27, 28, 28, 29, 29, 30, 30, 30, 31, 31
Shop B: 20, 22, 25, 28, 30, 32, 35, 38, 40, 45
```

Let's compute the summary statistics for each shop. These include:

- Measures of Center — Mean, Median, Mode
- Measures of Spread — Range, Standard Deviation, Interquartile Range (IQR, discussed in the next section)
- Additional values — Minimum and Maximum

Summary Statistics Table

This table give an overview of the Shop A and Shop B statistics

Statistic	Shop A	Shop B
Mean	29.3	31.5
Median	29.5	31.0
Mode	30	None
Range	31 − 27 = 4	45 − 20 = 25
Standard Deviation	~1.4	~7.9
Minimum	27	20
Maximum	31	45
CV	0.048	0.251

Interpretation

The results confirm what we sensed visually:

- **Shop A**'s delivery times are clustered tightly around 29–30 minutes it's highly consistent, with **low variation** (small range, small standard deviation). Its CV is very small.
- **Shop B**, while having a similar median and mean, shows much wider delivery times overall from **20 to 45 minutes**. It has a larger range and a much higher standard deviation, and particular a much larger CV, indicating less reliability.

This helps us answer our original question: **Shop A is the more consistent choice**. Even though both shops may advertise “fast” delivery, Shop A's low variability makes it more predictable an important factor for customers deciding where to order from.

This example shows how **measures of spread work alongside measures of center** to tell the full story. In real-world data, it's rarely enough to just look at one number like the mean without understanding context, consistency, and variation.

Z-Scores: Measuring Distance from the Mean

Now that you understand standard deviation, we can use it to introduce one of the most useful ideas in statistics: the **z-score**.

A **z-score** tells you how many standard deviations a value is from the mean. It helps compare values from different datasets or distributions and identifies whether a specific value is typical or unusual.

Formula for Z-Score:

$$z = \frac{x - \bar{x}}{s} \tag{2.2.10}$$

x: the individual data value

$\bar{x}$: the sample mean

s: the sample standard deviation

What Z-Scores Tell Us:

- A **z-score of 0** means the value is exactly at the mean.
- A **positive z-score** means the value is *above* the mean.
- A **negative z-score** means the value is *below* the mean.

Because z-scores are based on standard deviation, they're a useful way to compare apples to oranges or pizza deliveries to grocery bills. For example, a delivery time with a z-score of 1.5 is 1.5 standard deviations longer than the typical delivery time.

We'll use z-scores often in future chapters when working with normal distributions, confidence intervals, and hypothesis tests. For now, just remember that **z-scores use the mean and standard deviation to put values into context.**

Calculating Summary Statistics in Excel

Once you understand how to compute each of these values by hand, Excel can make these calculations much faster and easier especially when working with larger datasets, like your semester-long project. Below is a list of spreadsheet functions you can use to compute measures of spread (and center) directly within Excel or Google Sheets.

Excel Formulas for Summary Statistics

This give us a reference for how to find the summary statistics with Excel

Statistic	Excel Formula	How It Works
Mean	`=AVERAGE(A2:A11)`	Averages the values in cells A2 through A11
Median	`=MEDIAN(A2:A11)`	Returns the middle value of the dataset
Mode	`=MODE.SNGL(A2:A11)`	Returns the most frequent number (use `=MODE.MULT()` for multiple modes)
Minimum	`=MIN(A2:A11)`	Returns the smallest value
Maximum	`=MAX(A2:A11)`	Returns the largest value
Range	`=MAX(A2:A11)-MIN(A2:A11)`	Calculates the spread from smallest to largest
Standard Deviation (sample)	`=STDEV.S(A2:A11)`	Estimates SD for a sample using Bessel's correction
Standard Deviation (population)	`=STDEV.P(A2:A11)`	Standard deviation if you have the full population

Example Setup

To get started:

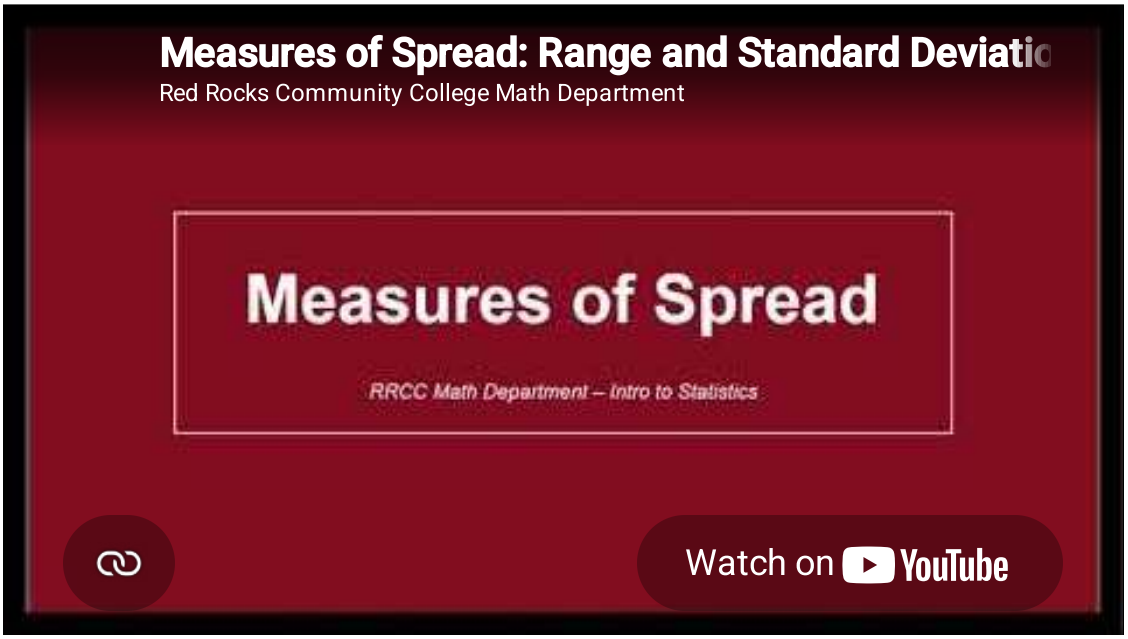
1. Enter your dataset (e.g., delivery times, prices, square footage) into a single column, say from `A2` to `A11` .
2. Select blank cells next to your data and enter the formulas above.
3. Let Excel do the calculations for you! You can even use conditional formatting to highlight outliers and trends.

Notes:

- All formulas can be used directly in **Google Sheets** as well.
- Make sure not to include column headers (like "Price") inside function ranges.
- You can also use `=VAR.S()` and `=VAR.P()` to compute the variance directly.

We'll continue using spreadsheets throughout the semester in our project analysis and examples, so this is a great time to get comfortable with these tools!

Related Videos





This page titled 2.2: Measures of Spread- Range, Variance, and Standard Deviation is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Mathematics Department.

2.2.1: Project Check-In 1

Semester Project Check-In: Describing Your Housing Data

Now that you've collected data as part of your semester-long investigation into housing affordability, you're ready to start **summarizing your own dataset**.

In this check-in, you'll choose **one continuous variable** from your dataset, most likely the listing **price**, and calculate key summary statistics using the tools we've just covered in Chapter 2.

This is the beginning of your full data overview analysis. You'll build on this work in Chapter 3, where we'll add graphs and visualizations to further explore your variable and comparisons within your dataset. For now, focus on understanding what your data looks like without graphing it.

What to Do

1. Choose one continuous (numerical) variable from your semester housing dataset. We suggest starting with **price**, but square footage or lot size could also work.
2. Use Excel or Google Sheets to calculate the following:
   - Mean ($\bar{x}$)
   - Median
   - Minimum
   - Maximum
   - Range
   - Standard Deviation
3. Record these values in a short labeled table.
4. Answer the reflection questions below in a few sentences.

Questions for Reflection

- What do these summary statistics suggest about the typical value of the variable you chose?
- Is there a large spread in the data, or does it appear to be consistently centered around a particular value?
- Does the mean or median better represent the "typical" value in your data? Why?
- What might be important to know about potential outliers or extremes?
- If someone were new to your dataset, what key observations would help them immediately understand it?

Create Your One-Page Data Overview

Use this check-in to begin a short written summary of your dataset. This will become part of your final project deliverable and evolve in Chapter 3 when you add visualizations. For now, your overview should include:

- A short paragraph introducing your dataset (what you collected, where it's from, what region it covers)
- The name and definition of the variable you're analyzing
- Your summary statistics, clearly labeled and formatted for clarity (use a simple table)
- Brief, thoughtful written responses to at least two reflection questions

> ***Tip:*** *Keep your data overview focused but polished. Think of it as the introduction slide to your future presentation or executive summary in your project report.*

You'll revisit and revise this page later as we add **visuals** in Chapter 3 and explore **comparisons and relationships** in Chapters 4–7.

This page titled 2.2.1: Project Check-In 1 is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Mathematics Department.

2.3: The Five-Number Summary and Interquartile Range

Example Scenario: Test Scores

Imagine a teacher administered a statistics quiz and got back 25 student scores, ranging from very low to very high. She wants to understand how students performed overall: How many did really well? How many struggled? Where's the "typical" performance?

This is where **percentiles** and the **five-number summary** can help break down the distribution of the data and make outliers easier to spot. Let's explore the tools you need to do just that.

What Is the Five-Number Summary?

The **five-number summary** gives a quick, informative snapshot of how a set of data is distributed. It consists of five key statistics that help describe the position and spread of a numerical variable. Think of it as an overview, like a short story, that explains what your data looks like.

The five numbers include the lowest and highest values, the middle value, and two quartile markers that show where the middle 50% of the values fall. These numbers are especially useful when comparing groups or spotting skewed distributions and outliers.

This is a reference for the five number summary.

Statistic	What It Tells You
Minimum	Smallest value in the dataset
Q1 (1st Quartile)	Value that marks the bottom 25% of the data
Median (Q2)	The middle value; 50% of data is above, 50% is below
Q3 (3rd Quartile)	Value that marks the top 25% of the data
Maximum	Largest value in the dataset

Each of these points corresponds to meaningful locations in your data's distribution — and together, they form the backbone of the **box-and-whisker plot** you'll learn about in the next section.

Percentiles & Quartiles

Description: A percentile is the value below which a given percentage of data falls. The **quartiles** divide the data into four equal parts:

- **Q1** (25th percentile): 25% of values are below this point
- **Q2** (Median / 50th percentile): middle point
- **Q3** (75th percentile): 75% of values are below this point

Use cases:

- Classifying scores on tests or surveys
- Describing household income, rent, or prices in a market
- Identifying values at known thresholds (e.g. top 10%)

Example: If your exam score is in the 80th percentile, you scored better than 80% of students.

How to Calculate Percentiles (by hand)

To estimate a specific percentile (e.g., the 25th percentile), use the following approach:

1. Sort the dataset from lowest to highest.
2. Use the formula to estimate the rank position:

$$R = \frac{p}{100} \times (n+1) \tag{2.3.1}$$

   where p is the desired percentile (e.g., 25 for the 25th), and n is the number of data points.
3. If R is a whole number, that item is your percentile value.
   If R is a decimal, interpolate between the two closest items.

Example: To find the 25th percentile of a dataset with 10 ordered values:

- $R = \frac{25}{100} \times (10+1) = 2.75$
- The 25th percentile will lie between the 2nd and 3rd values — 75% of the way between them.

How to Find Quartiles

Quartiles follow the same logic:

- **Q1 (25th percentile)**: what value is one-quarter of the way through the dataset?
- **Q2 (Median, 50%)**: middle value
- **Q3 (75th percentile)**: value three-quarters of the way through the dataset

Excel / Google Sheets Shortcut

To easily find quartiles and percentiles in Excel:

- `=PERCENTILE.INC(A2:A26, 0.25)` → 25th percentile
- `=PERCENTILE.INC(A2:A26, 0.75)` → 75th percentile
- `=QUARTILE.INC(A2:A26, 1)` → Q1
- `=QUARTILE.INC(A2:A26, 3)` → Q3

These functions save time and are especially helpful when working with larger datasets.

Interquartile Range (IQR)

Description: The IQR describes the range of the middle 50% of data — between the 1st and 3rd quartiles. It's a common measure of spread and helps identify unusually high or low values.

Formula:

$$\text{IQR} = Q3 - Q1 \tag{2.3.2}$$

Why it matters: The IQR is **resistant** to extreme values and is often used to find skew or irregularities in the tails of a distribution.

Example: If Q1 = 65 and Q3 = 85, then IQR = 85 - 65 = 20 points.

Outliers (IQR Rule)

Description: An **outlier** is a data point that lies far outside the rest of the distribution. One common rule uses the IQR to flag values:

Outlier criteria:

- Lower outlier: $\text{value} < Q1 - 1.5 \times \text{IQR}$
- Upper outlier: $\text{value} > Q3 + 1.5 \times \text{IQR}$

Note: This isn't a universal rule, but it's good for detecting extremes in a consistent way.

Example: If Q1 = 40 and Q3 = 80 → IQR = 40
Any value below $40 - 60 = -20$ or above $80 + 60 = 140$ may be flagged.

Applied Example: Student Quiz Scores

Let's pretend the teacher has the following quiz scores (out of 100) from 25 students:

```
[42, 58, 61, 63, 66, 67, 69, 70, 72, 74, 75, 76, 77, 78, 80, 82, 84, 85, 88, 90, 92, 9
```

From this data, she computes:

- Min: 42
- Q1: 66
- Median (Q2): 76
- Q3: 88
- Max: 98

Five-Number Summary: [42, 66, 76, 88, 98]

IQR = 88 - 66 = 22 → Outliers would be below $66 - 1.5 \times 22 = 33$ or above $88 + 33 = 121$ → no outliers in this case.

The five-number summary helps us quickly understand how the class performed overall. In this case, we can see that the lowest score (42) is quite far from the rest of the class, which may indicate a struggling student or an outlier. Half the class scored between 66 and 88 which suggests a relatively strong, consistent middle group. The median score of 76 tells us that most students performed reasonably well. The maximum score (98) shows that at least one student nearly aced the quiz. Altogether, the summary gives us a simple but powerful story: most students scored in a tight middle band with a few scores stretching toward the extremes.

As you analyze your own data, keep in mind how you might use the five-number summary to describe the distribution of your chosen variable, especially when comparing across groups, highlighting trends, or spotting unusual values or outliers.

Using Excel to Find Five-Number Summary

Use these spreadsheet functions to quickly generate a five-number summary:

Reference functions for the Excel for the 5 number summary.

Statistic	Function in Excel
Minimum	`=MIN(A2:A26)`
Q1 (25th percentile)	`=QUARTILE.INC(A2:A26, 1)`
Median (Q2)	`=MEDIAN(A2:A26)`
Q3 (75th percentile)	`=QUARTILE.INC(A2:A26, 3)`
Maximum	`=MAX(A2:A26)`

Project Sidequest: Adding a Summary Story

Now that you've learned how to calculate the five-number summary and identify outliers, it's time to return to your semester housing dataset and tell a story with your numbers.

TODO: Add the five-number summary for one continuous variable, most likely **price**, to your project's executive summary write-up. This will provide an overview of how housing costs are distributed and give your audience a clearer picture of the housing market in your region.

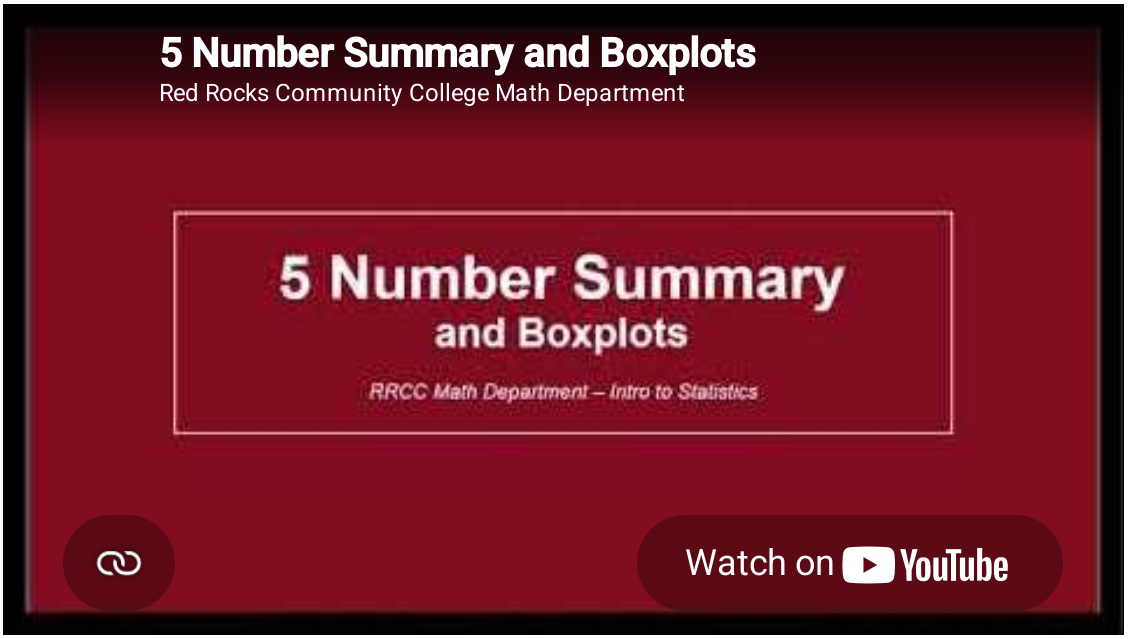
What to include:

- The full five-number summary (Min, Q1, Median, Q3, Max) clearly labeled
- A few sentences interpreting what the summary says about your dataset
- If applicable, mention any potential outliers (based on the IQR rule)

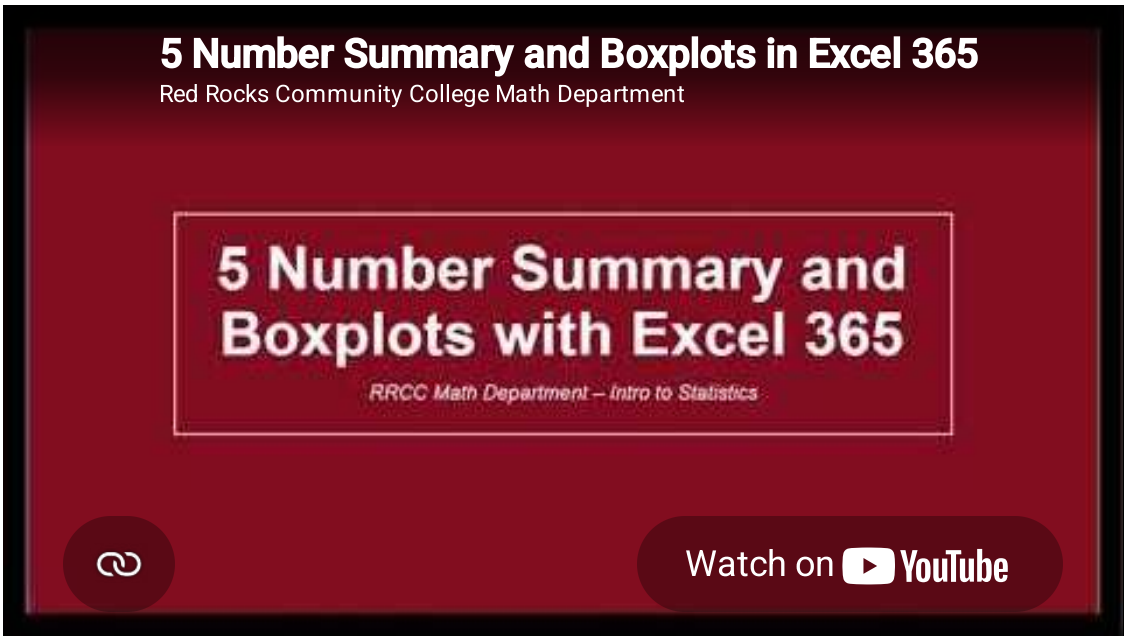
You'll use this in the next section when we create boxplots and start comparing distributions. Your write-up is slowly transforming into a full data story!

> 💬 ***Need inspiration?*** *Revisit the quiz score example in this chapter — notice how just five numbers can tell a powerful story about performance, consistency, and extremes.*

Related Videos









What's Next: Box-and-Whisker Plots

Now that you know how to calculate the five-number summary and define outliers, you're ready to visualize it. In the next section, we'll introduce **box-and-whisker plots**, a powerful visual that displays a full distribution at a glance using the five-number summary.

This page titled 2.3: The Five-Number Summary and Interquartile Range is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Mathematics Department.

2.3.1: Project Check-In 2

Project Check-In 2: Adding a Summary Story

Now that you've learned how to calculate the five-number summary and identify outliers, it's time to return to your semester housing dataset and tell a story with your numbers.

TODO: Add the five-number summary for one continuous variable, most likely **price**, to your project's executive summary write-up. This will provide an overview of how housing costs are distributed and give your audience a clearer picture of the housing market in your region.

What to include:

- The full five-number summary (Min, Q1, Median, Q3, Max) clearly labeled
- A few sentences interpreting what the summary says about your dataset
- If applicable, mention any potential outliers (based on the IQR rule)

You'll use this in the next section when we create boxplots and start comparing distributions. Your write-up is slowly transforming into a full data story!

> ***Example to Follow*** *Revisit the quiz score example in section 2.3 — notice how just five numbers can tell a powerful story about performance, consistency, and extremes.*

This page titled 2.3.1: Project Check-In 2 is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Mathematics Department.

2.3.2: Percentage Review

Before we work more with **percentiles** and the five-number summary, let's take a quick review of how **percentages, decimals, and fractions** relate to one another.

A **percent** means "per 100" — it's a way to express a portion or part of a whole using 100 as the reference point.

> **Definition: Percent**
>
> A **percent** is a ratio that compares a number to 100. It can be converted to a fraction or decimal as needed.
> Example: 60% = $\frac{60}{100} = 0.60$

Common Conversions

Here are some important examples of how to convert between fractions, decimals, and percents:

Reference for common conversions for percent, decimal and fractions.

Fraction	Decimal	Percent
1/2	0.50	50%
1/4	0.25	25%
3/4	0.75	75%
1/5	0.20	20%
1/10	0.10	10%
2/5	0.40	40%
5/8	0.625	62.5%

Conversion Rules

- **Fraction → Decimal:** Divide the top by the bottom.
  Example: $\frac{1}{4} = 0.25$
- **Decimal → Percent:** Multiply by 100 and add a % sign.
  Example: $0.25 = 25\%$
- **Percent → Decimal:** Divide by 100 or move the decimal two places left.
  Example: $75\% = 0.75$
- **Decimal → Fraction:** Use place value (0.40 = 40/100 = 2/5)

Try It Yourself

1. Convert 0.3 to a percent and a fraction.
2. Convert 12.5% to a decimal and a simplified fraction.
3. Convert $\frac{3}{5}$ to a decimal and a percent.

Why This Matters for Percentiles

Percentiles are essentially a series of benchmark **percent positions** in a dataset. For example:

- The 25th percentile (Q1) marks the value where 25% of data lies below it.
- The 50th percentile is the median (half of data below, half above).
- The 75th percentile (Q3) marks that 75% of data lies below it.

If you're comfortable moving between fractions, decimals, and percentages, you'll be better equipped to:

- Estimate values in a dataset
- Understand percentile rank questions
- Describe what part of a distribution each quartile captures

> **Tip:** Whenever you see “%,” ask: “Out of what?” The answer is usually 100 — and the concept of relative position starts from that idea.

This page titled 2.3.2: Percentage Review is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Mathematics Department.

2.4: Identifying Outliers with Boxplots

A **box-and-whisker plot**, or simply a **boxplot**, is a compact visual that displays a full distribution using just five numbers. It helps you quickly see how data is spread out, where it centers, and whether any potential outliers exist.

What makes the boxplot powerful is that it combines elements you already know — the **five-number summary** and **interquartile range (IQR)** — into an easy-to-scan visualization that's useful for comparing groups or identifying odd values.

What a Boxplot Shows

- **Box:** spans from Q1 to Q3 (interquartile range). It captures the middle 50% of the dataset.
- **Line inside the box:** the median (Q2), showing the "typical" center of the data.
- **Whiskers:** extend to the smallest and largest values *within* the expected range (i.e., not outliers).
- **Dots or stars (often):** mark potential outliers beyond $1.5 \times \text{IQR}$ from the quartiles.

Together, these parts create a visual "map" of your dataset's shape. One glance at a boxplot can tell you: is the data skewed? Is it tightly grouped? Are there surprising highs or lows?

Example: Student Quiz Scores

Let's walk through the full process using the dataset from the previous section, only this time the teacher includes a score of zero for the student who missed the quiz. We'll compute all the important summary statistics, identify any outliers, and construct our boxplot step-by-step. This hands-on example will give you a model to apply with your own project data.

Here's the dataset we'll use:

```
[0, 42, 58, 61, 63, 66, 67, 69, 70, 72, 74, 75, 76, 77, 78, 80, 82, 84, 85, 88, 90, 92
```

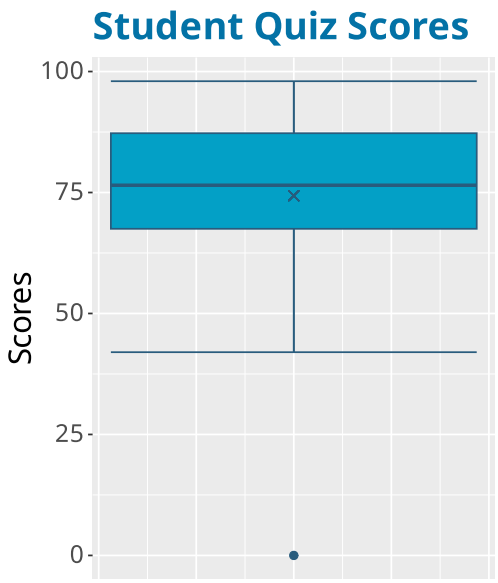
Calculated Summary Statistics

Example for how to present summary statistics.

Statistic	Value
Minimum	0
Q1	67.0
Median	76.5
Q3	88.0
Maximum	98
IQR	21
Mean	74.346
Standard Deviation	20.272

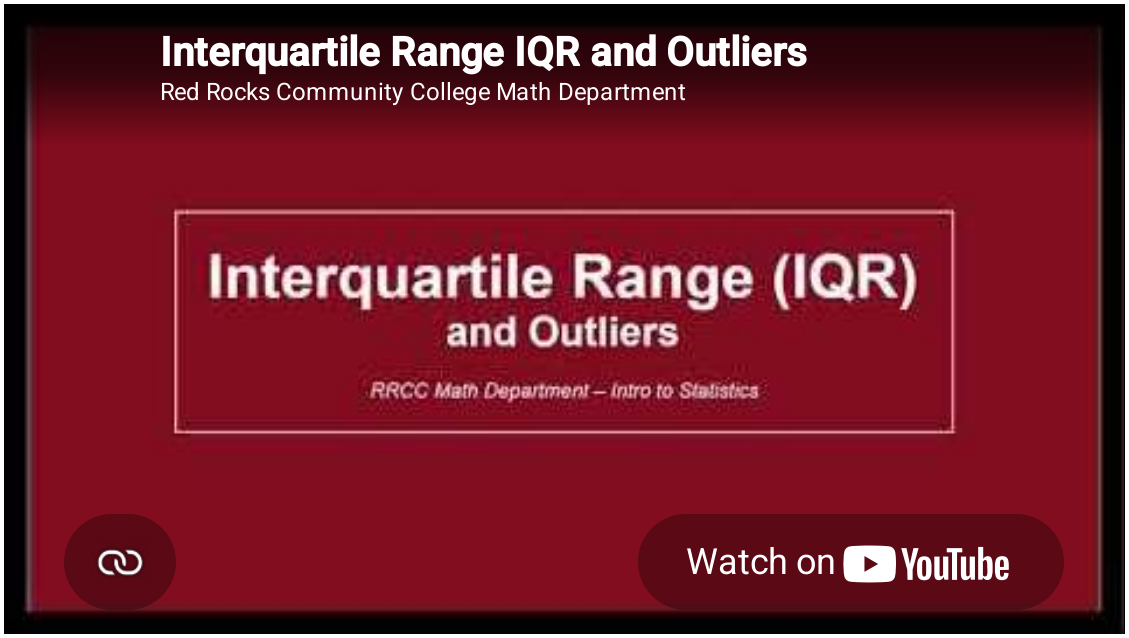
Building the Box-and-Whisker Plot

1. Draw a number line that covers the full range of your data.
2. Mark the five-number summary: Min, Q1, Median, Q3, Max.
3. Draw a box from Q1 to Q3 and place a line at the median.
4. Optionally add an X at the mean to help identify skew.
5. Determine outliers as as points that fall beyond 1.5 × IQR from Q1 or Q3. These can be plotted as dots below or above the graph if it does not significantly affect scaling.
6. Extend whiskers to the lowest and highest values *that are not outliers.*



> ***Boxplot tip:*** *If the median is closer to the bottom or top of the box, the data might be skewed. Skewness also shows if one whisker is much longer than the other.*

Related Videos





What's Coming Next

In Chapter 3, we'll formally introduce the world of data visualizations — histograms, bar graphs, scatterplots, and more. This box-and-whisker plot is your first step into that world. You can use it to describe full distributions, compare groups, or amplify your project's story with a clean, simple summary.

This page titled 2.4: Identifying Outliers with Boxplots is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Mathematics Department.

2.5: Vocabulary (Chapter 2)

Chapter 2 Vocabulary

Here is a list of key terms introduced throughout Chapter 2. These concepts help you describe data numerically and visually — forming the bridge between raw values and interpretation.

Measure of Center

A single number that represents the "typical" or central value in a dataset. Includes the mean, median, and mode.

Mean (Arithmetic Average)

The sum of all values divided by the number of values. Sensitive to outliers. Symbol: $\bar{x}$

Median

The middle value when the data are sorted. If there's an even number of values, it's the average of the two middle ones. Resistant to outliers.

Mode

The value (or values) that occur most often in a dataset. There can be no mode, one mode, or multiple modes.

Measure of Spread

A statistic that describes how much the values vary around the center. Includes range, IQR, and standard deviation.

Range

The difference between the maximum and minimum values in a dataset. Sensitive to outliers.

Interquartile Range (IQR)

The range of the middle 50% of values. Calculated as $Q3 - Q1$; resistant to extreme values.

Standard Deviation

The average distance from each data point to the mean. Uses all values in the dataset and is sensitive to outliers.

Variance

The squared average distance from the mean. Used as an intermediate step when calculating standard deviation.

Percentile

A value that separates a certain percent of the data. For example, the 80th percentile is the value that 80% of the dataset lies below.

Quartiles

Special percentiles that divide the data into quarters:

- **Q1:** 25th percentile
- **Q2:** 50th percentile / median
- **Q3:** 75th percentile

Five-Number Summary

A descriptive summary of a dataset using five values: minimum, Q1, median, Q3, and maximum.

Outlier

A data point that is unusually far from the others. Often flagged if it lies beyond 1.5 × IQR below Q1 or above Q3.

Bessel's Correction

The practice of dividing by $n - 1$ instead of n when calculating the sample variance or standard deviation. Helps make estimates less biased when working with sample data.

Box-and-Whisker Plot (Boxplot)

A standardized graph that displays the five-number summary. Visually shows a dataset's spread, skew, and potential outliers.

Whiskers

Lines on a boxplot that extend from Q1 to the minimum and from Q3 to the maximum — unless outliers are present, in which case they stop at the nearest non-outlier.

Skew (in a boxplot)

Asymmetry in the distribution. A longer whisker on one side suggests skew in that direction.

Tip: As you work on your semester project, you can refer back to these definitions to describe your data clearly and efficiently.

This page titled 2.5: Vocabulary (Chapter 2) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Mathematics Department.

2.6: Practice (Chapter 2)

2.1: Measures of Central Tendency — Mean, Median, and Mode

1. Define the following in your own words:
    a. mean
    b. median
    c. mode
2. Which measure of center is most affected by outliers?
3. What symbol is used for the sample mean?
4. What symbol is used or the population mean?
5. Calculate the mean, median, and mode of the following dataset:
   {2, 5, 5, 7, 10, 14, 14, 14, 17, 20}
6. A group of students reports how many hours they study per week: {10, 12, 12, 14, 16, 20}.
    a. Find the mean and the median.
    b. Which one better represents the group?
7. Give an example of a small dataset (5–6 numbers) where:
    a. The mean is greater than the median
    b. The median is equal to the mode
8. Explain which measure of center is most appropriate for this dataset and why:
   {28, 30, 31, 34, 36, 38, 300}True or False: "The median is always one of the values in the dataset." Explain your reasoning.
9. Describe a real-world situation where the **mode** might be more useful than the mean or median.
10. Identify which value represents the sample mean and which value represents the claimed population mean.
    a. A recent article in a college newspaper stated that college students get an average of 5.5 hrs of sleep each night. A student who was skeptical about this value decided to conduct a survey by randomly sampling 25 students. On average, the sampled students slept 6.25 hours per night.
    b. American households spent an average of about $52 in 2007 on Halloween merchandise such as costumes, decorations and candy. To see if this number had changed, researchers conducted a new survey in 2008 before industry numbers were reported. The survey included 1,500 households and found that average Halloween spending was $58 per household.
    c. The average GPA of students in 2001 at a private university was 3.37. A survey on a sample of 203 students from this university yielded an average GPA of 3.59 in Spring semester of 2012.
11. Workers at a particular mining site receive an average of 35 days paid vacation, which is lower than the national average. The manager of this plant is under pressure from a local union to increase the amount of paid time off. However, he does not want to give more days off to the workers because that would affect profits. Instead he decides to fire 10 employees in such a way as to raise the average number of days off that are reported by the site. In order to achieve this goal, should he fire employees who have the most number of days off, least number of days off, or those who have about the average number of days off?
12. A factory quality control manager decides to investigate the percentage of defective items produced each day. Within a given work week (Monday through Friday) the percentage of defective items produced was 2%, 1.4%, 4%, 3%, 2.2%. Calculate the mean for these data.
13. The following data show the lengths of boats moored in a marina. The data are ordered from smallest to largest: 16; 17; 19; 20; 20; 21; 23; 24; 25; 25; 25; 26; 26; 27; 27; 27; 28; 29; 30; 32; 33; 33; 34; 35; 37; 39; 40
    a. Calculate the mean.
    b. Identify the median.
    c. Identify the mode.
14. Sixty-five randomly selected car salespersons were asked the number of cars they generally sell in one week. Fourteen people answered that they generally sell three cars; nineteen generally sell four cars; twelve generally sell five cars; nine generally sell six cars; eleven generally sell seven cars. Calculate the following:
    a. mean = _____

b. median = _____

c. mode = _____

15. A group of 10 children are on a scavenger hunt to find different color rocks. The results are shown in the table below. The column on the right shows the number of colors of rocks each child has. What is the mean number of rock colors?

Table for the rocks.

Child	Rock colors
1	5
2	5
3	6
4	2
5	4
6	3
7	7
8	2
9	1
10	10

16. A group of children are measured to determine the average height of the group. The results are in the table below. What is the mean height of the group to the nearest hundredth of an inch?

Table of the heights of the children

Child	Height in inches
Adam	45.21
Betina	39.45
Chen	43.78
Donna	48.76
Edhas	37.39
Fran	39.90
George	45.56
Heather	46.24

17. A person compares prices for five automobiles. The results are in the table below. What is the mean price of the cars the person has considered?

Prices in the table.

Price
$20,987
$22,008
$19,998
$23,433
$21,444

18. When is it best to use the mode as a measure of center? Describe what type of data would lead you to choosing the mode over the mean or median.
19. Suppose you are representing the employees at a large corporation during contract negotiations. You have a list of the salaries of all the employees at the corporation (The salaries include the many lower level employees and the few high paid management employees) and you plan to find a measure of center (e.g. mean, median, or mode). Which measure or center would you use to represent the employees in an effort to support your claim that the average salary of lower level employees is much lower than the national average (for lower level employees) and thus should be increased? Why is this the better choice?
20. A standardized test is given to ten people at the beginning of the school year with the results given in the table below. At the end of the year the same people were again tested.
    a. What is the average improvement?
    b. Does it matter if the means are subtracted, or if the individual values are subtracted?

Table of the scores for the students

Student	Beginning score	Ending score
1	1100	1120
2	980	1030
3	1200	1208
4	998	1000
5	893	948
6	1015	1030
7	1217	1224
8	1232	1245
9	967	988
10	988	997

21. A small class of 7 students has a mean grade of 82 on a test. If six of the grades are 80, 82, 86, 90, 90, and 95, what is the other grade?
22. A class of 20 students has a mean grade of 80 on a test. Nineteen of the students has a mean grade between 79 and 82, inclusive.
    a. What is the lowest possible grade of the other student?
    b. What is the highest possible grade of the other student?
23. If the mean of 20 prices is $10.39, and 5 of the items with a mean of $10.99 are sampled, what is the mean of the other 15 prices?
24. In a recent issue of the *IEEE Spectrum*, 84 engineering conferences were announced. Four conferences lasted two days. Thirty-six lasted three days. Eighteen lasted four days. Nineteen lasted five days. Four lasted six days. One lasted seven days. One lasted eight days. One lasted nine days. Let X = the length (in days) of an engineering conference. If you were planning

an engineering conference, which would you choose as the length of the conference: mean; median; or mode? Explain why you made that choice.

25. Create a sample data set of size n =3 for which the mean is greater than the median.
26. Create a sample data set of size n =3 for which the mean is less than the median.
27. Create a sample data set of size n =4 for which the mean, the median, and the mode are all identical.
28. Create a sample data set of size n =4 for which the median and the mode are identical but the mean is different.
29. Which measure of central tendency is most often used for returns on investment?

2.2: Measures of Spread — Range, Variance, and Standard Deviation

1. Define the following in your own words:
    a. range
    b. variance
    c. standard deviation
2. What symbol is used for the sample standard deviation?
3. What symbol is used or the population population standard deviation?
4. Find the range, variance, and standard deviation of the following samples:
    a. {4, 6, 9, 10, 11}
    b. {7.8, 25.7, 5.5, 14.7, 20.7, 17.6, 20.8, 27.3}
    c. {4585, 34587, 62689, 7147}
    d. {15 oz, 14 oz, 14, oz, 17 oz}
5. Use the formula to calculate the sample standard deviation for {2, 4, 6, 8}.
6. A small dataset consists of the values {5, 7, 8, 8, 10}. Without calculating, which will be larger: the standard deviation or the range? Why?
7. How does increasing an outlier in a dataset affect the standard deviation? Give a conceptual explanation.
8. Why do we divide by (n - 1) when calculating the sample variance?
9. What does a standard deviation of zero mean? Give an example of a dataset with standard deviation equal to zero.
10. In a class of 25 students, 24 of them took an exam in class and 1 student took a make-up exam the following day. The professor graded the first batch of 24 exams and found an average score of 74 points with a standard deviation of 8.9 points. The student who took the make-up the following day scored 64 points on the exam.
    a. Does the new student's score increase or decrease the average score?
    b. What is the new average?
    c. Does the new student's score increase or decrease the standard deviation of the scores?
11. A random sample of data from 5 smokers is provided below.

Sample of the collected information.

age	maritalStatus	grossIncome	annual weekend days
51	Married	$2,600 to $5,200	2
24	Single	$10,400 to $15,600	0
33	Married	$10,400 to $15,600	0
17	Single	$5,200 to $10,400	1
76	Married	$5,200 to $10,400	0

age	maritalStatus	grossIncome	amtweekdays/day

a. Find the mean amount of cigarettes smoked on weekdays and weekends by these 5 respondents.
b. Find the standard deviation of the amount of cigarettes smoked on weekdays and on weekends by these 5 respondents. Is the variability higher on weekends or on weekdays?

12. A factory quality control manager decides to investigate the percentage of defective items produced each day. Within a given work week (Monday through Friday) the percentage of defective items produced was 2%, 1.4%, 4%, 3%, 2.2%. Calculate the range and standard deviation for these data.
13. For each part, compare distributions (1) and (2) based on their ranges and standard deviations. Make sure to explain your reasoning.
    a. (1) 3, 5, 5, 5, 8, 11, 11, 11, 13; (2) 3, 5, 5, 5, 8, 11, 11, 11, 20
    b. (1) -20, 0, 0, 0, 15, 25, 30, 30; (2) -40, 0, 0, 0, 15, 25, 30, 30
    c. (1) 0, 2, 4, 6, 8, 10; (2) 20, 22, 24, 26, 28, 30
    d. (1) 100, 200, 300, 400, 500; (2) 0, 50, 300, 550, 600
14. If a set of grades for a class has a large range and a small standard deviation, what can you say about the class? Include an interpretation that is specific to grades in a class.
15. Suppose you wish to invest money safely and are trying to decide between stock options in two companies. The average rates of return are the same for both companies but, one company has a much larger standard deviation of its rates of return. Which company should you invest in? Why is this the better choice?
16. The population parameters below describe the full-time equivalent number of students (FTES) each year at Red Rocks Community College over a 29-year period. μ = 1000 FTES, median = 1,014 FTES, σ = 474 FTES, n = 29 years
    a. A sample of 11 years is taken. About how many are expected to have a FTES of 1014 or above? Explain how you determined your answer.
    b. The population standard deviation = _____
    c. How many standard deviations away from the mean is the median?
17. Three students were applying to the same graduate school. They came from schools with different grading systems. Which student had the best GPA when compared to other students at their school? Explain how you determined your answer.

Table of student progress for the question

Student	GPA	School Average GPA	School Standard Deviation
Thuy	2.7	3.2	0.8
Vichet	87	75	20
Kamala	8.6	8	0.4

18. A music school has budgeted to purchase three musical instruments. They plan to purchase a piano costing $3,000, a guitar costing $550, and a drum set costing $600. The mean cost for a piano is $4,000 with a standard deviation of $2,500. The mean cost for a guitar is $500 with a standard deviation of $200. The mean cost for drums is $700 with a standard deviation of $100. Which cost is the lowest, when compared to other instruments of the same type? Which cost is the highest when compared to other instruments of the same type. Justify your answer.
19. An elementary school class ran one mile with a mean of 11 minutes and a standard deviation of three minutes. Rachel, a student in the class, ran one mile in eight minutes. A junior high school class ran one mile with a mean of nine minutes and a standard deviation of two minutes. Kenji, a student in the class, ran 1 mile in 8.5 minutes. A high school class ran one mile with a mean of seven minutes and a standard deviation of four minutes. Nedda, a student in the class, ran one mile in eight minutes.
    a. Why is Kenji considered a better runner than Nedda, even though Nedda ran faster?
    b. Who is the fastest runner with respect to their class? Explain why.
20. Create a sample data set of size n =3 for which the range is 0 and the sample mean is 2.
21. Create a sample data set of size n =3 for which the sample variance is 0 and the sample mean is 1.

2.3: The Five-Number Summary and Interquartile Range

1. List the five numbers that make up the five-number summary and explain what each one represents.
2. Given the dataset {45, 50, 52, 55, 58, 60, 63, 65, 70, 75}, find the five-number summary and the IQR.
3. Describe a situation where knowing the five-number summary would be more helpful than knowing just the mean and standard deviation.
4. Why is the IQR considered a "robust" measure of spread?
5. A dataset has an IQR of 20. If Q1 = 40, what is Q3?
6. Explain how to find the IQR using only the sorted dataset and no formulas.
7. True or False: "The median is always the average of Q1 and Q3." Explain your reasoning.
8. For each part, compare distributions (1) and (2) based on their medians and IQRs. Make sure to explain your reasoning.
    a. (1) 3, 5, 6, 7, 9; (2) 3, 5, 6, 7, 20
    b. (1) 3, 5, 6, 7, 9; (2) 3, 5, 8, 7, 9
    c. (1) 1, 2, 3, 4, 5; (2) 6, 7, 8, 9, 10
    d. (1) 0, 10, 50, 60, 100; (2) 0, 100, 500, 600, 1000
9. For each of the following, specify whether the mean, median, or mode would best represent a typical observation in the data, and whether the variability of observations would be best represented using the standard deviation, range, or IQR.
    a. Housing prices in a country where 25% of the houses cost below $350,000, 50% of the houses cost below $450,000, 75% of the houses cost below $1,000,000 and there are a meaningful number of houses that cost more than $6,000,000.
    b. Housing prices in a country where 25% of the houses cost below $300,000, 50% of the houses cost below $600,000, 75% of the houses cost below $900,000 and very few houses that cost more than $1,200,000.
    c. Number of alcoholic drinks consumed by college students in a given week.
    d. Annual salaries of the employees at a Fortune 500 company.
10. For each of the quartiles Q1, Q2, & Q3 write a sentence using the definition of a percentile to interpret it. (Hint: A number x is at the kth percentile if k% of the data in the set is less than or equal to x).
11. Jesse was ranked 37th in his graduating class of 180 students. What percentile does this correspond to?
12. For runners in a race, a low time means a faster run. The winners in a race have the shortest running times.
    a. Is it more desirable to have a finish time with a high or a low percentile when running a race?
    b. Is it more desirable to have a speed with a high or a low percentile when running a race?
13. The 20th percentile of run times in a particular race is 5.2 minutes. Write a sentence interpreting the 20th percentile in the context of the situation.
14. A bicyclist in the 90th percentile of a bicycle race completed the race in 1 hour and 12 minutes. Are they among the fastest or slowest cyclists in the race? Write a sentence interpreting the 90th percentile in the context of the situation.
15. The 40th percentile of speeds in a particular race is 7.5 miles per hour. Write a sentence interpreting the 40th percentile in the context of the situation.
16. On an exam, would it be more desirable to earn a grade with a high or low percentile? Explain.

17. Mina is waiting in line at the Department of Motor Vehicles (DMV). Her wait time of 32 minutes is the 85th percentile of wait times. Is that good or bad? Write a sentence interpreting the 85th percentile in the context of this situation.
18. In a survey collecting data about the salaries earned by recent college graduates, Li found that her salary was in the 78th percentile. Should Li be pleased or upset by this result? Explain.
19. In a study collecting data about the repair costs of damage to automobiles in a certain type of crash tests, a certain model of car had $1,700 in damage and was in the 90th percentile. Should the manufacturer and the consumer be pleased or upset by this result? Explain and write a sentence that interprets the 90th percentile in the context of this problem.
20. The University of Colorado has two criteria used to set admission standards for students to be admitted to a college in the UC system:
    a. Students' GPAs and scores on standardized tests (SATs and ACTs) are entered into a formula that calculates an "admissions index" score. The admissions index score is used to set eligibility standards intended to meet the goal of admitting the top 12% of high school students in the state. In this context, what percentile does the top 12% represent?
    b. Students whose GPAs are at or above the 96th percentile of all students at their high school are eligible (called eligible in the local context), even if they are not in the top 12% of all students in the state. What percentage of students from each high school are "eligible in the local context"?
21. Suppose that you are buying a house. You and your realtor have determined that the most expensive house you can afford is the 34th percentile. The 34th percentile of housing prices is $240,000 in the town you want to move to. In this town, can you afford 34% of the houses or 66% of the houses?
22. Use the following information to find the statistics below. Sixty-five randomly selected car salespersons were asked the number of cars they generally sell in one week. Fourteen people answered that they generally sell three cars; nineteen generally sell four cars; twelve generally sell five cars; nine generally sell six cars; eleven generally sell seven cars.
    a. First quartile = _____
    b. Second quartile = median = 50th percentile = _____
    c. Third quartile = _____
    d. Interquartile range (*IQR*) = _____ – _____ = _____
    e. 10th percentile = _____
    f. 70th percentile = _____
    g. Outliers =
23. Your younger brother comes home one day after taking a science test. He says that someone at school told him that "60% of the students in the class scored above the median test grade." What is wrong with this statement? What if he said "60% of the students scored below the mean?"
24. What rule do we use to identify outliers using the IQR? Write out both conditions.
25. Given that the first and third quartiles of a data set are Q_1=97 and Q_3=133, find the following:
    a. Interquartile range (IQR) =
    b. Lower bound for finding outliers =
    c. Upper bound for finding outliers =
    d. If the minimum data value is 51 and the maximum data value is 203, are there outliers in the data set?
26. Which measure of variation is most affected by outliers? Which is most resistant to outliers? Why?
27. Why is it important to identify outliers?
28. You work for a large farm with many fields of corn. You are investigating the mass in grams of a sample of ears of corn. You gather the following data:

662.2	595.4	639.8	593.4	481.6	945.6	390.1	614.7	593.3	613.4	589.3	377.1	415.4

    a. The 30th percentile =
    b. IQR =
    c. Apply the 1.5 IQR rule to search for outliers. Report the lower and upper cutoffs limits (fences).
    d. Find any outliers using the 1.5 IQR rule?

29. The following data set represents pulses of individuals measured at a blood drive. Identify any outliers and explain what they mean.

56	60	68	76
56	60	68	84
56	60	68	84
56	64	72	84
56	64	72	84
56	64	72	88
60	64	72	88
60	64	72	88
60	64	72	88
60	64	76	102

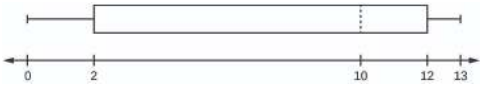
2.4: Identifying Outliers with Boxplots

1. Draw a quick sketch of a boxplot for the five-number summary below: {10, 20, 30, 40, 50}
2. In a boxplot, what does a long right whisker suggest about the data?
3. Choose a variable from your semester project dataset. List its five-number summary and indicate whether it has any outliers using the IQR method.
4. Describe one strength and one limitation of boxplots as visualization tools.
5. Create a box plot for the five number summary provided below.

Table of the five number summary

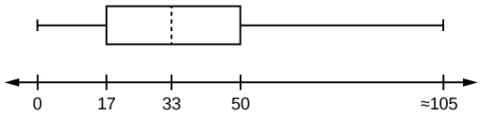
Min	Q1	Q2 (Median)	Q3	Max
57	72.5	78.5	82.5	94

6. What can you say about a data set when the “box” in the box plot is very wide but the “whiskers” do not go out very far from the box?
7. Give an example of a situation where an outlier might be removed from a dataset
8. Give an example of a situation where an outlier from a dataset needs to be emphasized.
9. Given the following box plot:



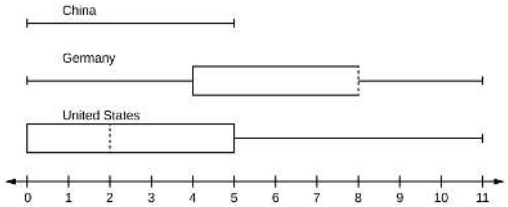
   a. Which quarter has the smallest spread of data? What is that spread?
   b. Which quarter has the largest spread of data? What is that spread?
   c. Find the interquartile range (*IQR*).
   d. Are there more data in the interval 5–10 or in the interval 10–13? How do you know this?
   e. Which interval has the fewest data in it? How do you know this?

10. The following box plot shows the age data for the U.S. population in 1990.



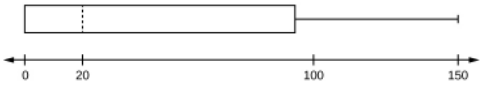
   a. Are there fewer or more children (age 17 and under) than senior citizens (age 65 and over)? How do you know?
   b. 12.6% are age 65 and over. Approximately what percentage of the population are working age adults (above age 17 to age 65)?

11. In a survey of 20-year-olds in China, Germany, and the United States, people were asked the number of countries they had visited in their lifetime. The following box plots display the results.



    a. In complete sentences, describe what the shape of each box plot implies about the distribution of the data collected.
    b. Have more Americans or more Germans surveyed been to over eight foreign countries?
    c. Compare the three box plots. What do they imply about the foreign travel of 20-year-old residents of the three countries when compared to each other?

12. Given the following box plot, answer the questions.



    a. Think of an example (in words) where the data might fit into the above box plot. In 2–5 sentences, write down the example.
    b. What does it mean to have the first and second quartiles so close together, while the second to third quartiles are far apart?

13. Twenty-five randomly selected students were asked the number of movies they watched the previous week. Construct a box plot of the data {0,0,0,0,0,1,1,1,1,1,1,1,1,1,2,2,2,2,2,2,3,3,3,3,4}

14. The percentage of people without health insurance for each state in the USA is given in the following table. Create a boxplot using spreadsheet software.

15.2	9.4	25.4	21.9	20.8	15.5	18.9	13.7	13.5	16.5
11.8	11.8	10.1	11.5	9.2	15.6	8.9	17.2	13.3	15.9
12.2	19.6	10.5	12.3	27.7	18.8	10.2	6.6	17.9	12.3
10.4	10.5	21.2	18.9	14.6	17.4	10	14	7.1	13.8
11.8	11.6	20.9	8.8	24.5	17.9	22.9	13.3	9.3	14

Practice problems include parts of:

Introductory Statistics 2e by OpenStax is licensed under CC BY 4.0

Access for free at https://openstax.org/books/introductory-statistics-2e/pages/1-introduction

and

OpenIntro Statistics by David Diez, Christopher Barr, & Mine Çetinkaya-Rundel is licensed under CC BY 3.0

2.6: Practice (Chapter 2) is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

- **Current page** has no license indicated.
- **1.E: Introduction to Data (Exercises)** by David Diez, Christopher Barr, & Mine Çetinkaya-Rundel is licensed CC BY-SA 3.0. Original source: https://www.openintro.org/book/os.
- **2.E: Descriptive Statistics (Exercises)** by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/introductory-statistics.
- **Licensing** by Mathematics Department has no license indicated.

3: Visualizing Data

Chapter 3: Visualizing Data

Once you've collected and summarized data, the next step is to **visualize** it. Graphical displays help you spot patterns, compare groups, identify outliers, and communicate your findings clearly. They often reveal insights that summary statistics alone can't show.

In this chapter, you'll learn how to create and interpret a variety of plots — starting with simple charts for categorical data, and moving to more detailed views for quantitative data. You'll also explore how frequency, distribution shape, and time trends can be seen more clearly with effective visuals.

Why Visuals Matter: A Motivating Example

Suppose two cities each report that the average monthly rent for a one-bedroom apartment is $1,200. But when you look at the data distributions, you find this:

- **City A**: Most apartments are clustered between $1,150 and $1,250 — rent is fairly consistent.
- **City B**: Half the apartments are under $1,000, and the rest are over $1,400 — rent is extremely variable.

Even with the same average, these cities paint two totally different pictures for a renter. A **graph** shows the whole story — helping you evaluate patterns, risks, and equity in a way a single number can't capture.

> **Coming Up:** You'll start by looking at how to graph categorical data, including bar charts and pie charts. Then we'll move on to histograms and other displays for quantitative variables, and explore how shape, variation, and time can be visualized.

This page titled 3: Visualizing Data is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Mathematics Department.

3.1: Graphical Displays for Categorical Data- Bar Charts, Pie Charts

Not all data represents numbers we can measure or order. In fact, some of the most common types of information we work with are *categories*: like restaurant type, customer satisfaction level, favorite color, or type of transportation. These variables are called **categorical variables**, and they require a different approach when we want to visualize them.

When working with **numerical data**, we spend time thinking about things like center, spread, or shape. But categorical data doesn't have a numeric “center” or follow a number line. Instead, we focus on **frequencies**: how many individuals fall into each category.

In this part of the chapter, we’ll introduce two powerful tools that help us see and compare category frequencies:

- **Bar Charts**: great for comparing categories side-by-side
- **Pie Charts**: good for showing how the categories make up a whole

Quick Review: What Is Categorical Data?

Categorical variables represent groups, labels, or classifications. They could be things like:

- Political party
- Type of payment (Cash, Credit, Debit)
- Brand preference
- Survey responses (Yes, No, Maybe)

We usually summarize these using **counts** or **relative frequencies** (percentages). And that's exactly what a bar chart or pie chart is designed to show.

Bar Charts: Compare Categories Clearly

A bar chart shows the frequency or proportion of each category using vertical or horizontal bars. The height or length of each bar represents the count (or percentage) in each category.

- Best for side-by-side comparisons
- Axes should be clearly labeled
- Include a title and consider sorting bars in chronological or frequency order

When to Use a Bar Chart

- You want to compare categories individually
- You want to see which category is largest or smallest
- You don’t need to focus on showing the “whole” (100%)

Example: Bar Chart — Preferred Study Environment

A college library surveyed 100 students, asking them where they most prefer to study on campus. The results were divided into five categories based on student responses.

Here is the frequency table from the survey:

Data table used for a bar chart.

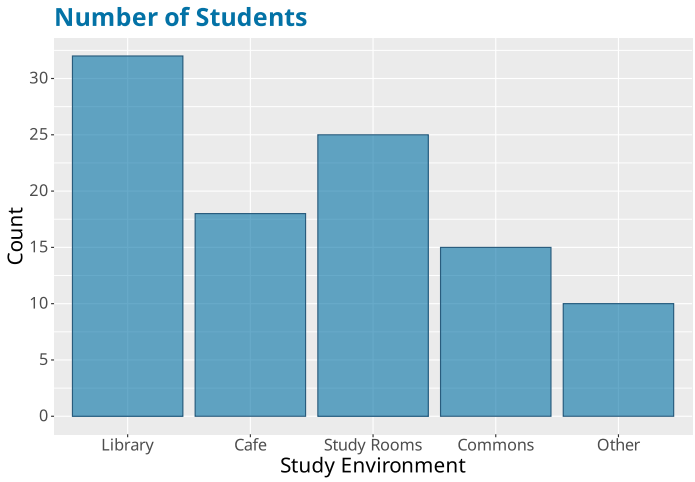
Study Environment	Number of Students
Library (Quiet Floor)	32
Café	18
Study Rooms	25
Commons	15
Other	10

Question: What does this distribution tell us about student study habits? Is one area clearly preferred, or are preferences spread evenly?

To visualize these results, we create a bar chart with the study environments on the x-axis and the number of students on the y-axis.

Interpretation

In the bar chart, it's immediately clear that the **Quiet Floor in the Library** is the preferred study environment, followed by **Group Study Rooms**. Only a few students prefer outdoor or alternate spaces. Bar charts like this make it easy to compare across categories and spot dominant choices.



Pie Charts: Highlight Part-to-Whole Relationships

A pie chart shows each category as a slice of a circular "pie." The angle of each slice represents that category's share of the total — usually displayed in percentages.

- Shows proportionality
- Works best with few categories
- Slices can be labeled with % or category name for clarity

Warning: Pie Chart Perception

Pie charts use **angles** to represent relative frequencies, but most people naturally interpret the **area** of each slice instead.

This can be misleading — especially when slices are similar in size or when visual embellishments exaggerate differences. Be cautious when **reading** or **designing** pie charts.

Example: Pie Chart: How Students Commute to Campus

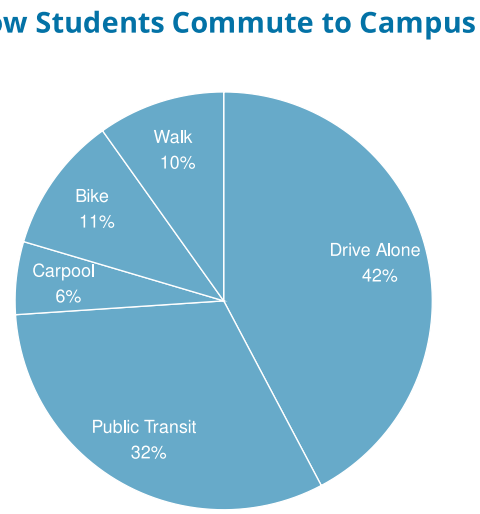
A student organization surveys 142 campus commuters to find out how students usually get to class. The goal is to understand what fraction of students use different transportation modes for possible sustainability planning.

Here is a breakdown from the survey:

Data table for pie chart.

Transportation Mode	Number of Students	Percentage
Drive Alone	60	42%
Public Transit	45	32%
Carpool	8	6%
Bike	15	11%
Walk	14	10%

Figure: Pie chart of student commuting choices as a percentage of all respondents. Without the percentages labeled, one would not be able to easily compare the 'Bike' and 'Walk categories.



Interpretation

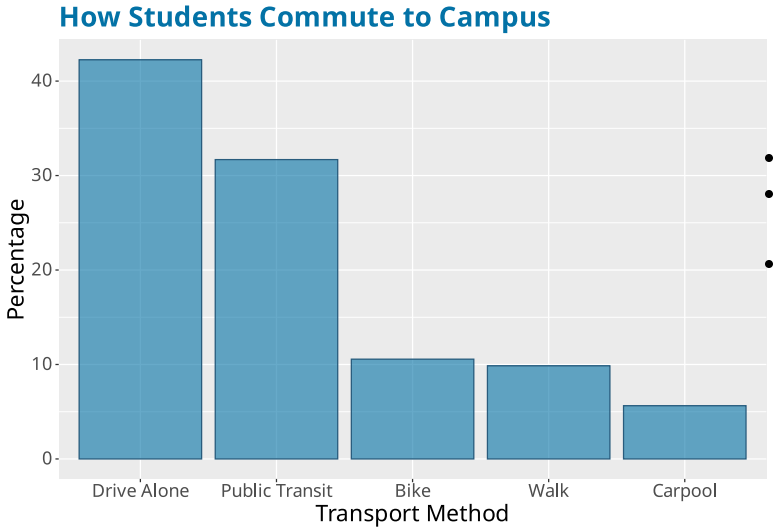
Driving alone is the most common commute method, used by 42% of students, but Public Transit is also heavily used (32%), and 27% of students are using more sustainable or shared modes (carpool, bike, or walk). This pie chart quickly shows how each mode contributes to the whole and helps support campus planning discussions. Note that rounding leads to the percentage total not adding to exactly 100%.

Why A Pie Chart is Admissible Here

- There are only 5 categories, all clearly distinct
- We want to emphasize percentages and part-of-whole relationships

Bar vs. Pie: Which Should You Use?

In general, one should primarily use bar charts. The data on student commute times is shown in a bar chart below. Although the emphasis is no longer on percentage out of the whole, the comparison between individual columns is much more apparent. As long as proper axis labels are used and show the full scale of the bars, the information conveyed in the bar chart will be easier to understand and interpret.

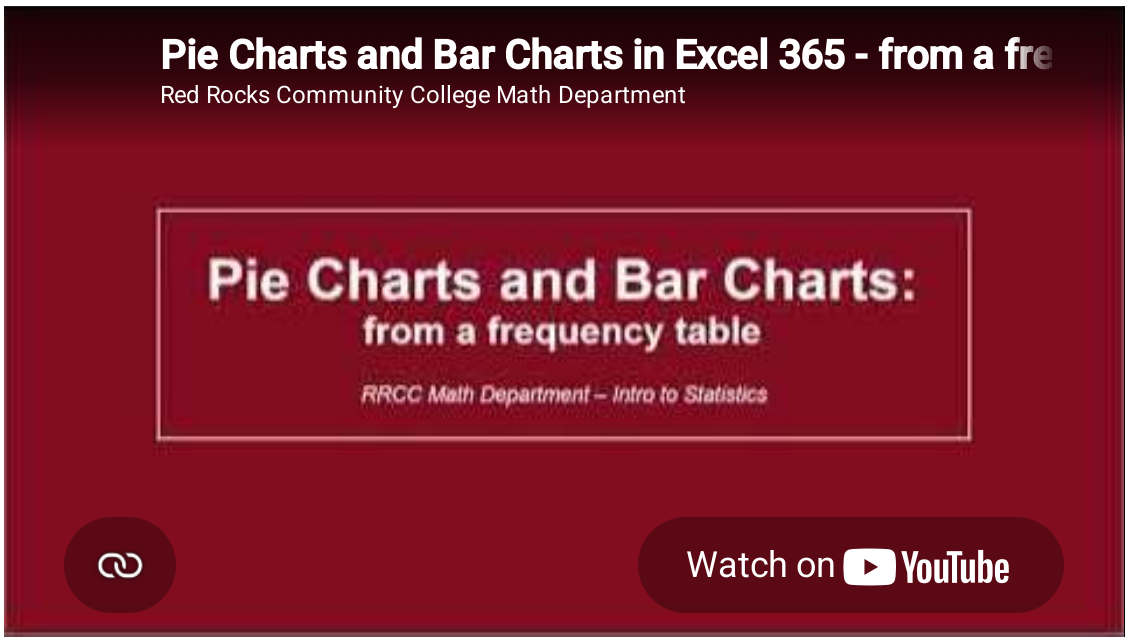


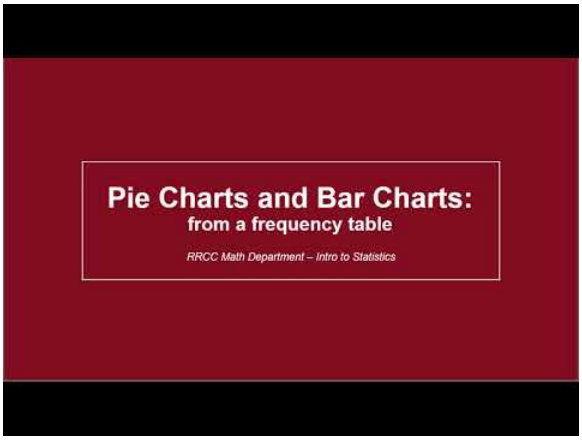
Check-In Prompt

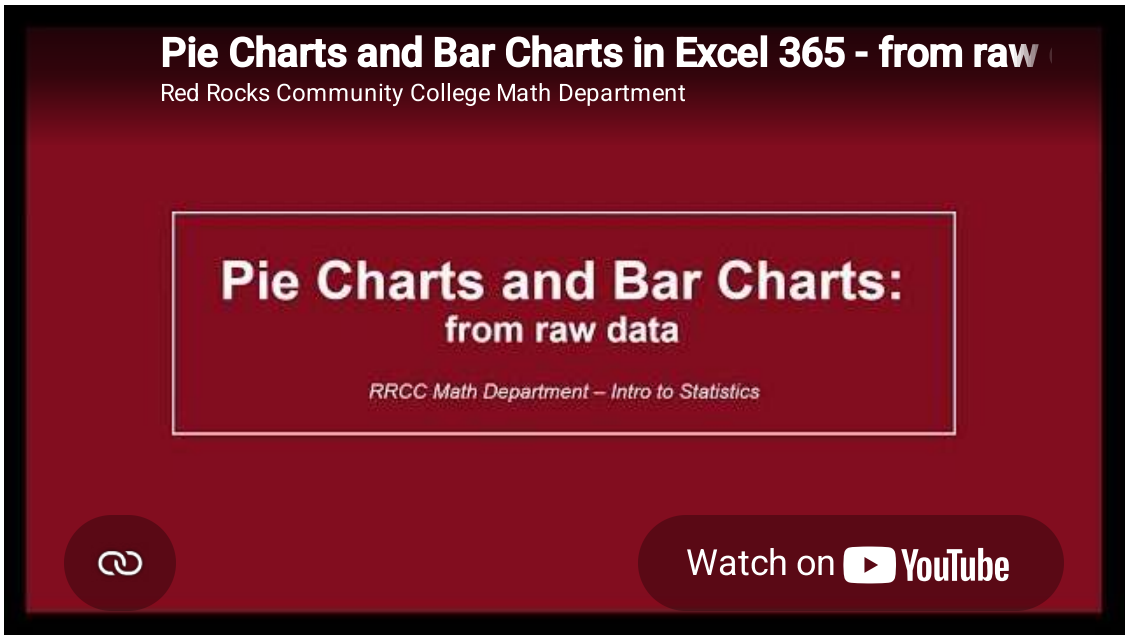
Find a categorical variable in your semester housing dataset (e.g., housing type, city, zip code range).

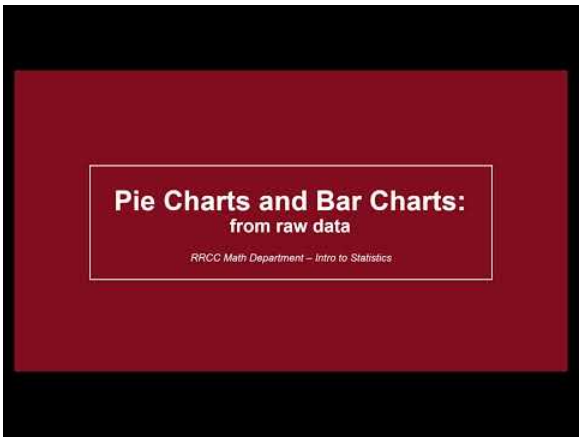
- How many categories are there?
- Would a bar chart or pie chart give the clearest picture of this variable?
- Create a rough sketch or build a quick chart using Excel or Google Sheets.

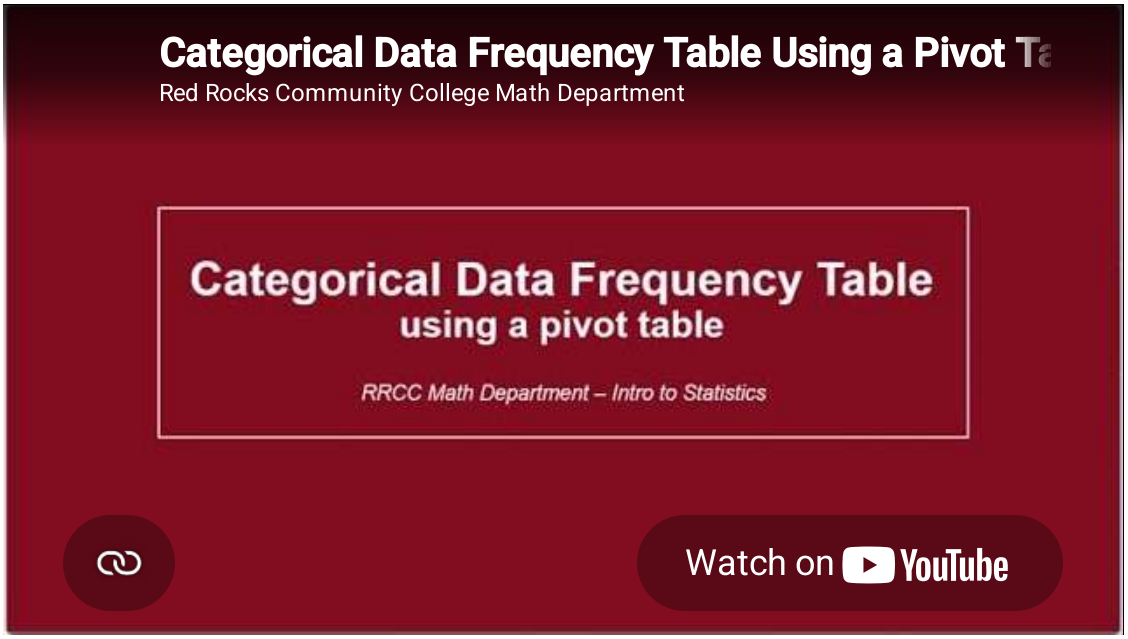
Related Videos

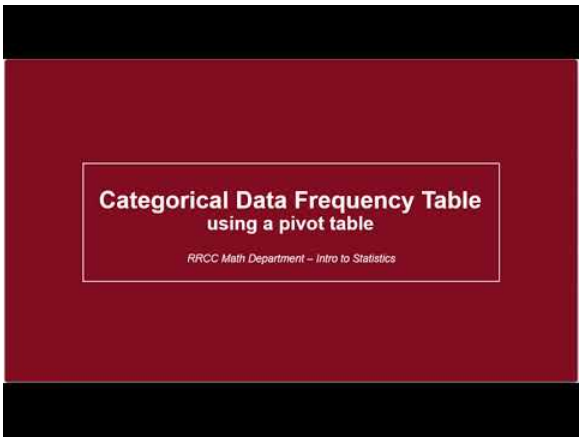












What's Next?

In the next section, we'll explore how to visualize **numerical (quantitative) data**, including some of the most important foundational plots in statistics: histograms, dot plots, and stemplots.

This page titled 3.1: Graphical Displays for Categorical Data- Bar Charts, Pie Charts is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Mathematics Department.

3.2: Graphical Displays for Quantitative Data- Dot Plots and Stemplots

When working with **quantitative variables** (like height, price, or temperature), we need visual tools that preserve the numerical nature of the data. Instead of comparing category labels (like with bar or pie charts), we're often interested in the overall **distribution**: How are the values spread out? Are they clustered? Symmetrical? Skewed?

In this chapter, we'll look at three powerful tools that help us graph quantitative data:

- **Histograms**: show overall shape and spread of a distribution
- **Dot Plots**: show individual data points with frequencies
- **Stemplots (Stem-and-Leaf Plots)**: preserve exact values while showing the distribution

Here we are going to take a look at Stemplots and Dot Plots and we will take a deeper look at histograms in a future section.

Dot Plot

A **dot plot** places one dot above a number line for each data point. When multiple observations share the same value, the dots stack upward. It's an effective way to show exact data values and frequencies, especially with smaller datasets.

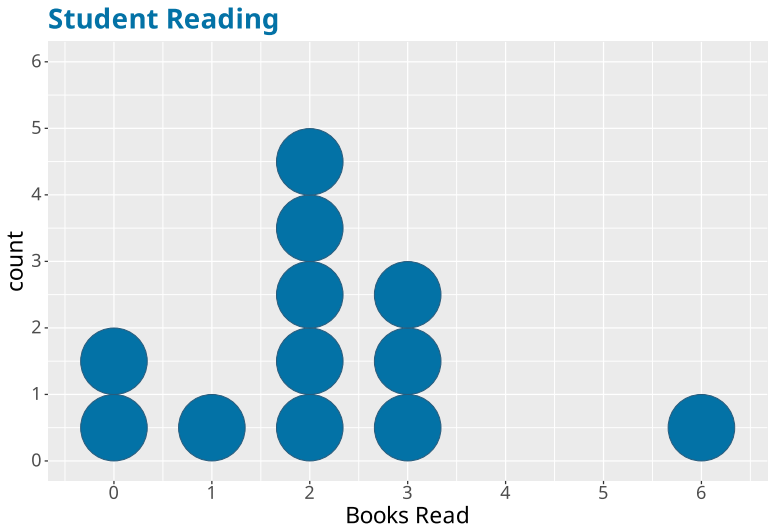
Also called: Line dot plot, stacked dot chart

Use when: You have a relatively small dataset and want an exact picture of the values and clusters.

Example: An English teacher asked 12 students how many books they read during a semester. The responses were:

$$[2, 6, 2, 0, 3, 2, 1, 0, 2, 3, 2, 3].$$

Though this isn't a large dataset, it offers a nice variety of values. When arranged in a dot plot, it's easy to see the rough distribution of reading, but also allows for one to count the exact amount of dots for a particular value. As the data set gets larger, this advange is lost if the dots are too numerous.



Stemplot (Stem-and-Leaf Plot)

A **stemplot** or **stem-and-leaf plot** displays quantitative data by splitting each value into a "stem" (typically the leading digit[s]) and a "leaf" (typically the final digit). This method keeps all original data visible while showing the distribution shape.

Also called: Stem plot, stem chart

Use when: You want to show both individual values and how they're distributed. Ideal for small to moderate samples.

Example data:
[59, 62, 65, 65, 66, 67, 71, 72, 78, 81, 85, 87]

Stem-and-leaf display:

```
5 | 9
6 | 2 5 5 6 7
7 | 1 2 8
8 | 1 5 7
```

To understand the data in this plot, we can examine a single row, such as the second one. The 6 on the left of the bar indicates that all values in this row begin with 6; they are all in the sixties. We have a 2, 5, 6 and 7 to the right of the bar, so we pair these values with the 6. That tells us that 62, 65, 66, and 67 are in our data set. Notice that the 5 appears twice, as 65 also appears twice in the data.

What's Next?

Now that you've learned how to visualize individual quantitative variables, we'll explore the overall **shapes of distributions** — like symmetry vs. skew and how these visual patterns connect to what we've learned about center and spread.

This page titled 3.2: Graphical Displays for Quantitative Data- Dot Plots and Stemplots is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Mathematics Department.

3.3: Frequency Tables and Relative Frequency

One of the simplest and most powerful ways to start summarizing raw data is by creating a **frequency table** which shows how many values fall into specific groups. With categorical data, we simply count the number of members in each category. When working with quantitative data, we often group values into **bins** or **ranges** of values.

> Definition: Frequency
>
> In statistics, **frequency** refers to the number of times a particular value or group of values appears in a dataset.

Example: Soil pH Levels

Suppose a team of agricultural scientists took pH readings from 30 different soil samples collected in the same region. pH is a measurement of acidity, with 7.0 representing neutral, lower values representing more acidic, and hight values representing more basic (alkaline). These readings can be used to help determine how well different crops might grow in those areas.

Soil pH Readings (n = 30):

```
5.7, 5.9, 6.0, 6.1, 6.1, 6.2, 6.3, 6.3, 6.4, 6.4,
6.4, 6.5, 6.5, 6.6, 6.6, 6.6, 6.7, 6.7, 6.8, 6.9,
7.0, 7.0, 7.0, 7.1, 7.2, 7.3, 7.4, 7.6, 7.8, 8.0
```

While this list is detailed, it's hard to see patterns just by scanning the numbers. A frequency table helps organize the pH levels into intervals so we can spot concentrations, gaps, or outliers more easily.

Frequency Table: Soil pH Readings

Let's build a frequency table by grouping the data into pH ranges (also known as **intervals** or **bins**). Since we're working with quantitative values, we can't treat each exact number as its own category as there are just too many individual values.

Instead, we group values into ranges of values, which we refer to as bins. The choice of bin isn't standardized. However, we do require each bin has the same length. Note in the below case we label the bins with one more significant figure than the data. If you are more experienced with algebra, note that we can also describe bins using intervals such as $[5.5, 6)$

Grouped Frequency Distribution Table.

Soil pH Range	Frequency
5.50 – 5.99	2
6.00 – 6.49	9
6.50 – 6.99	9
7.00 – 7.49	7
7.50 – 7.99	2
8.00 – 8.49	1

Definition: Relative Frequency

Relative frequency is the proportion or percentage of the total observations that falls into each category or interval. It is calculated by dividing the frequency of a group by the total number of observations.

Extended Table: Frequency and Relative Frequency

To get a clearer picture of how these groups compare, we convert the counts in our frequency table into relative frequencies by dividing each by the total number of observations (30).

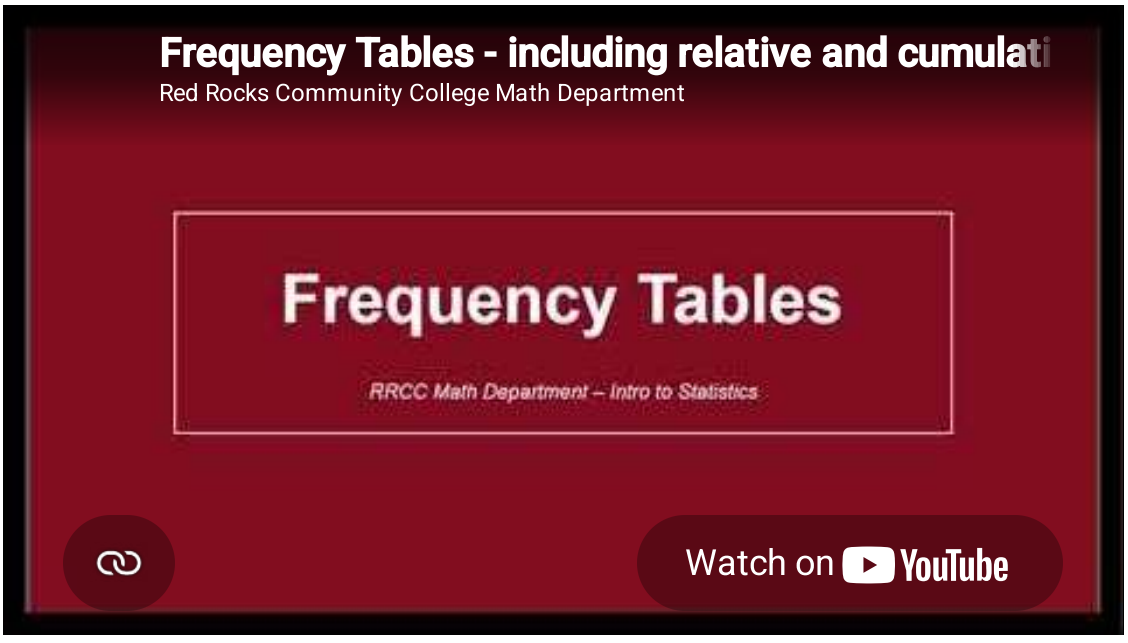
Grouped Relative Frequency Distribution Table.

Soil pH Range	Frequency	Relative Frequency	Percentage
5.50 – 5.99	2	2/30 ~ 0.067	6.7%
6.00 – 6.49	9	9/30 = 0.300	30.0%
6.50 – 6.99	9	9/30 = 0.300	30.0%
7.00 – 7.49	7	7/30 ~ 0.233	23.3%
7.50 – 7.99	2	2/30 ~ 0.067	6.7%
8.00 – 8.49	1	1/30 ~ 0.033	3.3%

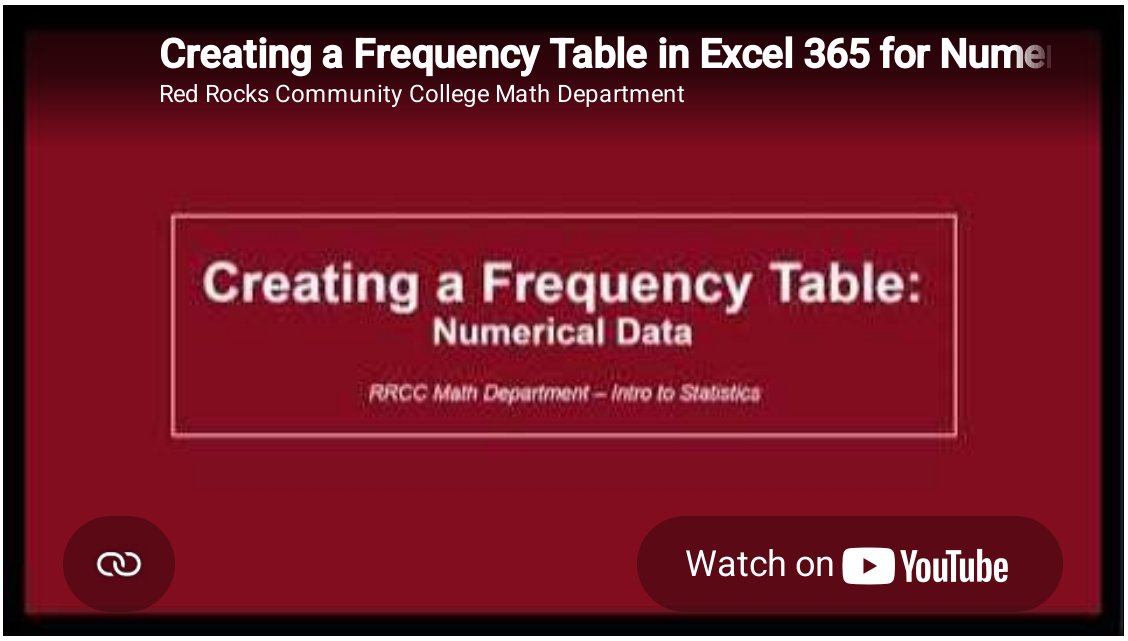
Interpretation

From the table, we notice that most soil pH readings fall between **6.00 and 6.99** which add up to 60% of all the data. This suggests that most of the tested fields have soil condition in the “slightly acidic to neutral” range, which is typically optimal for many crops. Only 1 measurement exceeds 8.0, indicating a very alkaline sample.

Related Videos - Frequency Tables









Histograms

The following graph is called a **histogram**. Each row in the table becomes a bar in a histogram, with the **x-axis** showing the interval (bin) and the **y-axis** showing either frequency or relative frequency as bar height. This gives us a visual way to convey the same information as the frequency table.

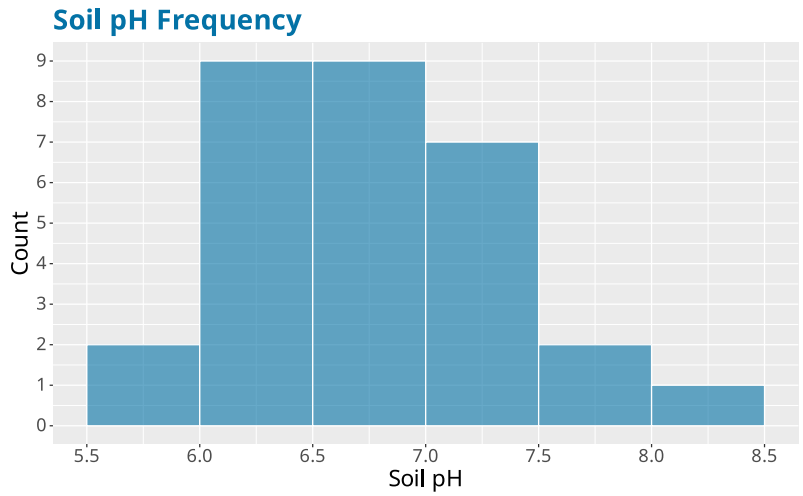


Figure: Histogram of soil pH readings based on grouped frequency intervals.

The rough shape of this chart, with the majority of the data in the middle and less of the data to the periphery, is a common occurrence.

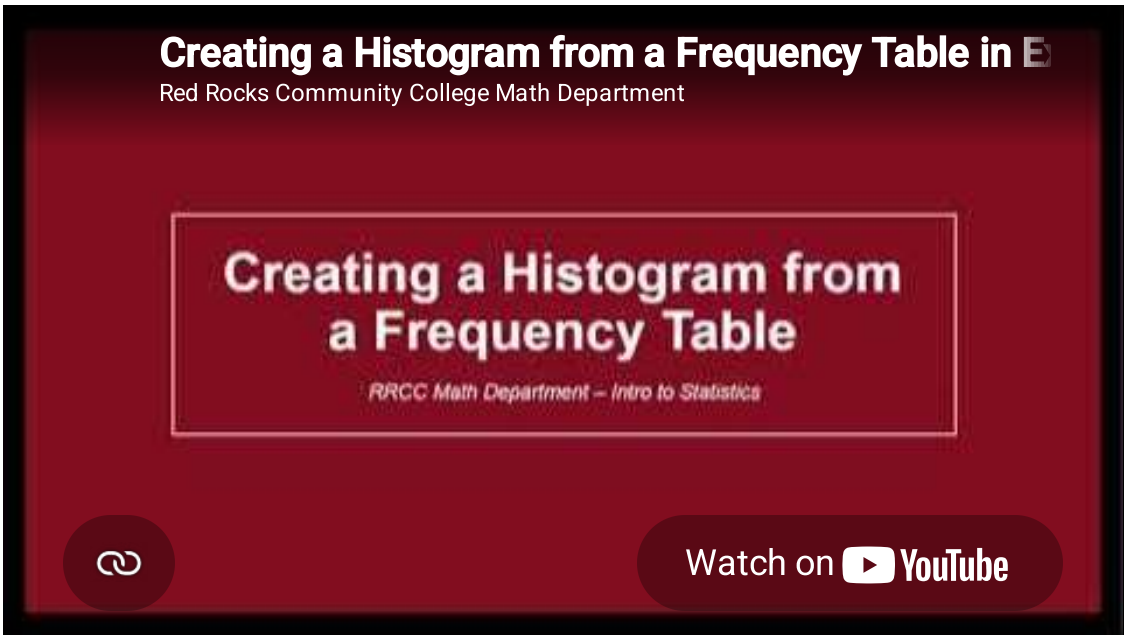
Step-by-Step: Building a Histogram

1. **Collect your data**
   Make sure you are working with quantitative (numerical data)
2. **Find the minimum and maximum values**
   This will help define where to group the values. In our soil samples, we had a minimum of 5.7 and a maximum of 8.0. Note that our data was sorted, however this may not be the case.
3. **Create your bins**
   In our case, we chose bins that were of width 0.5 and started at 5.5
   As this is a visualization tool, it is often best practice to make bins based on whole numbers or easy fractions.
   You can experiment with different bin widths to get the best picture of the data
   Very important: The bin widths must be the same length!
4. **Construct a frequency chart**
   Tally the number of values that fall into each bin
5. **Plot!**
   The height of each bar corresponds to the number of items in each bin
   Make sure the y-axis and x-axis tick marks are evenly spaced

Related Videos - Histograms









This page titled 3.3: Frequency Tables and Relative Frequency is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Mathematics Department.

3.4: Shapes of Distributions

When we visualize data — especially using histograms, dot plots, or boxplots — the **shape** of the distribution gives us important information about how the values are spread and where they tend to cluster. Recognizing the shape helps us choose the right summary statistics, spot unusual values, and prepare for more advanced ideas like probability and inference.

In this section, we'll explore the most common distribution shapes: symmetric, right-skewed, left-skewed, uniform, and bimodal.

> ### Side Note: Skewness and Kurtosis
>
> In more technical contexts, statisticians use formal formulas to describe the shape of a distribution using **skewness** (how lopsided the data is) and **kurtosis** (how "peaked" or "flat" it is). We won't calculate those here, but you can learn more on Wikipedia: Skewness(opens in new window) and Wikipedia: Kurtosis(opens in new window).

Symmetric Distribution

A distribution is **symmetric** if the left and right sides are roughly mirror images of each other. The most famous example is the bell-shaped normal distribution.

- The mean and median are approximately equal.
- The data is balanced; there's no lean to either side.
- This shape sets the foundation for many statistical models.

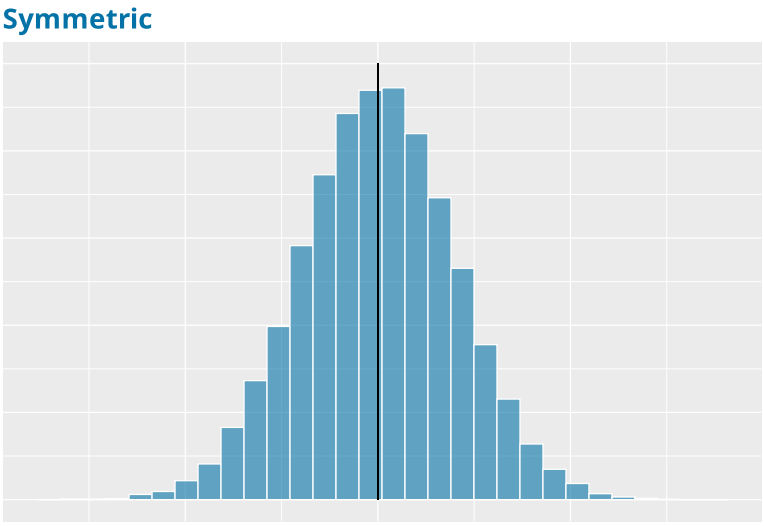


Figure: A symmetric (normal-like) distribution. Values are evenly spread around the center.

Right-Skewed (Positively Skewed) Distribution

A distribution is **right-skewed** when there's a long tail on the right. This often happens with data like income, where many people earn moderate amounts, but a few earn much higher than average.

- The **mean** is pulled right (higher) compared to the median.
- Most values cluster on the left side, with a few high values stretching the distribution.

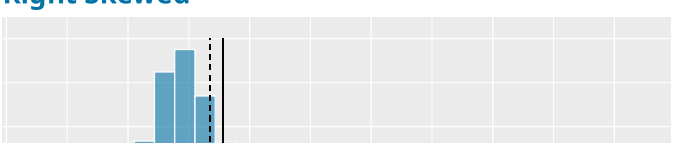


Figure: A right-skewed distribution. A few high values pull the mean upward.

Left-Skewed (Negatively Skewed) Distribution

A distribution is **left-skewed** when the tail stretches out to the left. This might occur with variables like exam scores when almost everyone does well, but a few students score much lower.

- The **mean** is pulled left (lower) compared to the median.
- Most values cluster on the higher end of the scale.

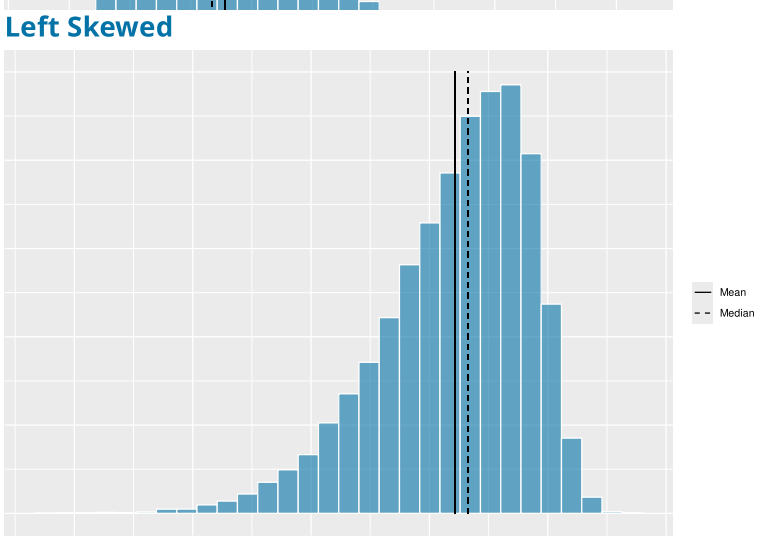


Figure: A left-skewed distribution. A few low values pull the mean down.

Uniform Distribution

A **uniform distribution** occurs when all values are roughly equally common. There are no obvious peaks, and each interval in the range has about the same frequency.

- Can indicate a random, equal probability across intervals.
- Looks “flat” — no shape or center stands out.

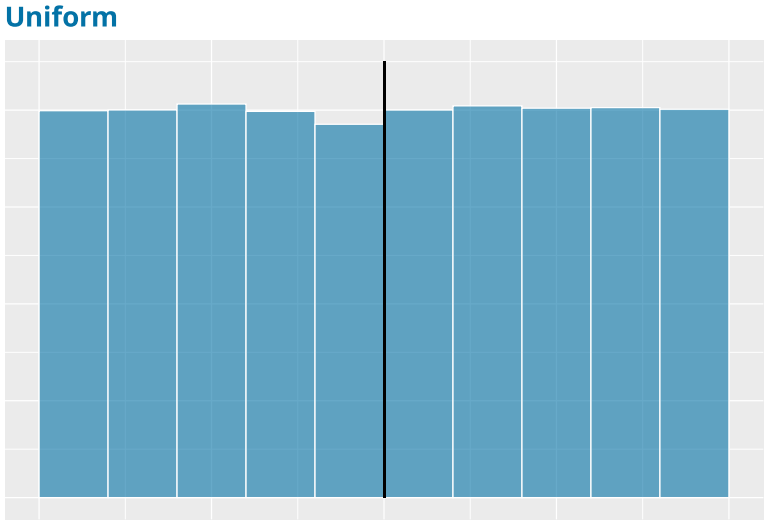


Figure: A uniform distribution. All bins are roughly equal in height.

Bimodal or Multimodal Distribution

A **bimodal distribution** has two clear peaks; a **multimodal distribution** has more than two. This may suggest that the data come from more than one group or process.

- Can result from combining groups (e.g., heights of children and adults)
- May signal the need to break the data into subgroups

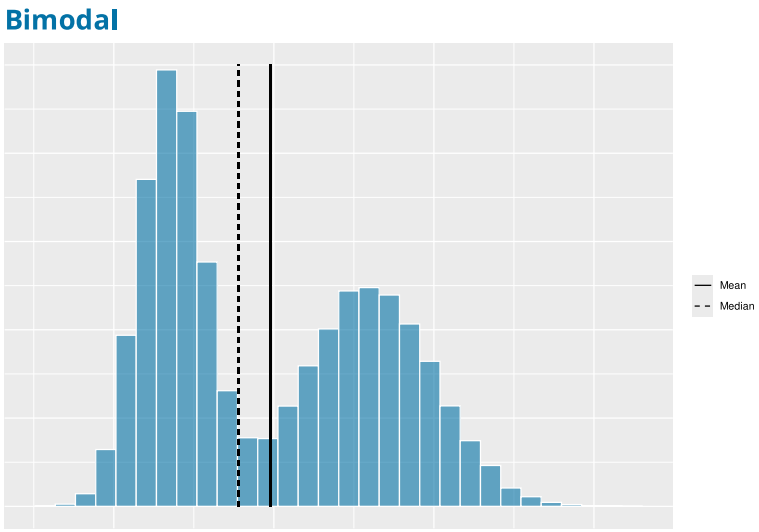


Figure: A bimodal distribution. Two distinct groups are present in the data.

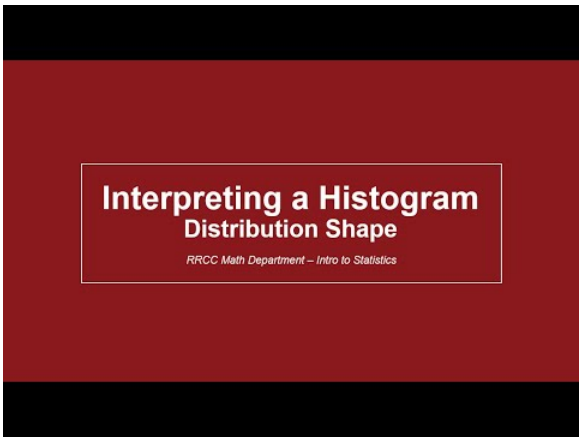
Summary

Understanding the shape of a distribution helps us choose the best descriptive statistics and prepares us for more advanced statistical tools.

- **Symmetric** → Consider using the mean and standard deviation.
- **Skewed** → The median and IQR might give a clearer picture.
- Visuals like histograms and boxplots are a key first step in shape recognition.

Related Videos





In the next section, we'll explore how data can change over **time** and how visual tools like time plots help us track shifts, trends, and patterns.

This page titled 3.4: Shapes of Distributions is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Mathematics Department.

3.5: Time Series Plots and Trends Over Time

Not all data lives in a moment but instead changes over time. The trend in this data can tell us what's rising, falling, or staying stable. These types of patterns show up everywhere: prices, temperatures, populations, interest rates, and even social media trends.

In this section, we'll explore how to visualize **time series data**, numerical observations tied to moments in time, using the simple and powerful **time series plot**.

What Is a Time Series Plot?

A **time series plot** (also called a time plot or time series graph) is a type of **scatter plot** where:

- The **x-axis** represents time, for example days, months, or years
- The **y-axis** represents a numerical variable that changes over time
- Points are often connected with lines to emphasize the direction or shape of change

Time plots help us capture patterns, cycles, and changes, and help us watch how data evolves so we can make better decisions in the future.

Example: U.S. Gas Prices Over the Last 12 Months

Suppose you collected weekly data on the average price of a gallon of regular gasoline in the United States over the last 12 months. Here are 12 monthly averages (for readability):

Gas Prices Dataset – Monthly Averages

Table of time series data

Month	Price (USD per gallon)
January	3.45
February	3.50
March	3.60
April	3.55
May	3.48
June	3.51
July	3.45
August	3.40
September	3.42
October	3.55
November	3.61
December	3.75

To analyze this data, we'll plot each month on the x-axis and the gas price (in dollars) on the y-axis to build a time plot.

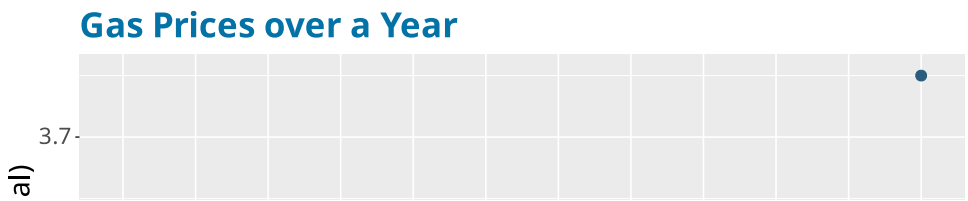
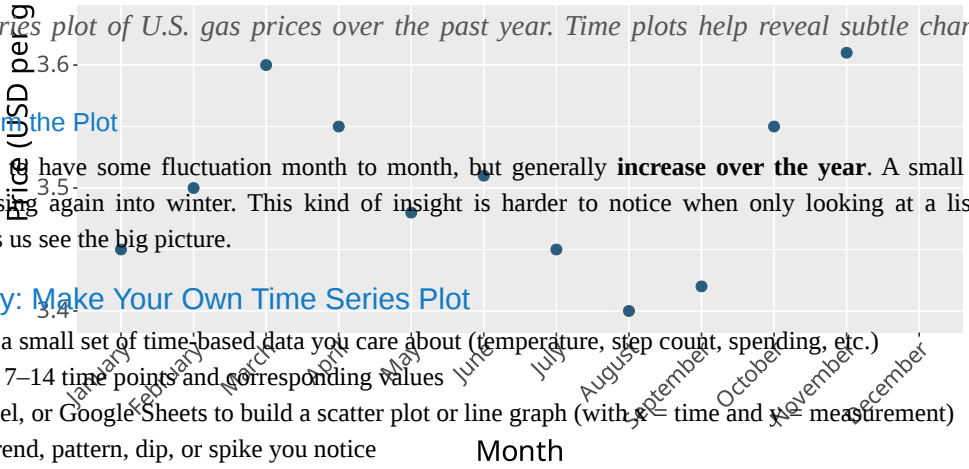


Figure: Time series plot of U.S. gas prices over the past year. Time plots help reveal subtle changes and overall direction.



Observations from the Plot

Gas prices appear to have some fluctuation month to month, but generally **increase over the year**. A small dip occurs in late summer before rising again into winter. This kind of insight is harder to notice when only looking at a list of numbers, but visualizing it helps us see the big picture.

Student Activity: Make Your Own Time Series Plot

1. Find or collect a small set of time-based data you care about (temperature, step count, spending, etc.)
2. Record at least 7–14 time points and corresponding values
3. Use paper, Excel, or Google Sheets to build a scatter plot or line graph (with x = time and y = measurement)
4. Describe any trend, pattern, dip, or spike you notice

How This Connects to Everything Before

Time plots are just one more way to describe data — and every visualization we've used so far helps us see different truths:

- **Bar charts** help compare quantities across named categories
- **Pie charts** help show parts of a whole
- **Histograms** reveal distributions and shape
- **Boxplots** show spread, center, and outliers
- **Time series plots** focus on change over time

This isn't the end of visualization — there are many more tools we can use, some of which we'll explore later in the course, like scatterplots between two variables, histograms with density overlays, trend lines, correlation matrices, and more.

Related Videos





Chapter Reflection

In this chapter, we've connected raw data to images that help people understand it. Data is difficult to interpret without mathematics and computations, but graphics allow for intuitive understanding.

The next time you see a chart, such as in other classes, the news, or on social media, consider: What kind of data is this? What is the visualization trying to show? How would I interpret it using what I've learned?

This page titled 3.5: Time Series Plots and Trends Over Time is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Mathematics Department.

3.5.1: Project Check-In 3

Project Check-In: Visualizing Your Housing Data

By this point in the semester, you've built a strong foundation for analyzing the housing dataset you've collected or been assigned:

1. **Data Overview:** You wrote a short summary describing where your data came from, what it represents, and what region or question you're exploring.
2. **Summary Statistics:** You calculated basic descriptive statistics (mean, median, standard deviation, five-number summary) for at least one continuous variable.
3. **Visual Representations:** Now it's time to deepen your understanding of that data by graphing it!

Goal of This Check-In

In this step, you'll create a few **appropriate visualizations** based on the variables in your housing dataset. This is your chance to turn numbers into a story — showing patterns, trends, clusters, and outliers that might influence how people understand prices in a neighborhood, square footage across home types, or rental fluctuations over time.

What to Include

1. **Create 2–3 Visualizations**
   Choose visuals appropriate for your data:
   - A **bar chart** or **pie chart** for a categorical variable (e.g., housing type, region, style)
   - A **histogram**, **boxplot**, or **dot plot** for a quantitative variable (e.g., price, square footage)
   - If collected over time, consider a **time plot**
2. **Label and Caption Each Chart**
   For each graph you create, include:
   - Clear title and labeled axes
   - A brief caption (1–2 sentences) explaining what it shows and why it's useful
   - Optional: annotate or highlight any interesting features (peaks, gaps, trends, etc.)
3. **Add These to Your Project Overview**
   Update your project write-up or slide deck so your visualizations follow your summary statistics section. You're starting to shape a narrative for your reader!

What to Consider as You Visualize

- Do my visuals reinforce or challenge what I saw in the summary statistics?
- Are there any trends, clusters, or potential outliers that are now more obvious?
- Does it look like my data is symmetric or skewed?
- What questions do these visualizations raise?

> #### Project Builder: Where You Should Be Now
>
> - Executive Summary (Project description + how data was collected)
> - Summary Statistics Table (for at least one continuous variable)
> - Visualizations added with captions + variable context
>
> You'll continue building and refining this throughout the next chapters — especially once we explore relationships between variables, begin comparing groups, and get into deeper analysis.

Optional Extension

If working in Google Sheets or Excel: try sorting your dataset by one variable before visualizing it. Does that sorting reveal new shapes or gaps that weren't as visible before?

This page titled 3.5.1: Project Check-In 3 is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Mathematics Department.

3.6: Vocabulary (Chapter 3)

Chapter 3 Vocabulary

Here is a list of key terms introduced throughout Chapter 3. These concepts help you visualize distributions, compare variables, and describe how data changes over time. Strong visualization skills will help you better analyze patterns, spot outliers, and prepare your data for deeper exploration in Chapters 4 and beyond.

Bar Chart

A graph that uses rectangular bars to represent the frequencies or proportions of different categories. The height (or length) of each bar corresponds to its value. Often used with categorical data. Bars do not touch.

Pie Chart

A circular chart divided into slices to represent parts of a whole. Each slice shows the proportion of a category relative to the total dataset. Best used with a small number of categories.

Histogram

A graph that displays the distribution of quantitative data by grouping values into equal-width intervals (bins). The height of each bar represents the frequency within each interval. Bars touch to show continuity.

Bin

An interval used when grouping numerical data for a histogram. Each bin includes a fixed range of values (e.g., 60–70), and all bins should be of equal width unless otherwise noted.

Dot Plot

A simple graph where each data point is shown as a dot placed above a number line. Repeated values are stacked vertically. Ideal for small datasets.

Stem-and-Leaf Plot

A tabular display where quantitative values are split into stems (typically the leading digits) and leaves (the final digit). This chart preserves the original dataset while showing distribution shape.

Frequency

The number of times a specific value or interval appears in a dataset. Frequencies are often shown in tables or bar charts to summarize distributions.

Relative Frequency

The proportion or percentage of total values that fall in a particular category or bin. Calculated as: frequency ÷ total sample size.

Distribution

The way data values are spread or clustered across a range. A distribution can be described by its center, spread, shape, and presence of outliers.

Symmetric Distribution

A distribution where the left and right sides are mirror images of each other. Often has a single peak near the center.

Skewed Distribution

Asymmetry in a dataset's shape. If the right tail (larger values) is longer, it’s right-skewed; if the left tail (smaller values) is longer, it’s left-skewed.

Uniform Distribution

A distribution where all values (or ranges) are approximately equally frequent. Appears flat with no clear peak.

Bimodal Distribution

A distribution with two clear peaks (modes) in different parts of the data. May suggest the combination of two separate groups.

Time Plot

A graph that shows how a numeric variable changes over time. Time is placed on the x-axis, and values are plotted on the y-axis, often with points connected by a line.

Scatter Plot

A graph that uses points to show the relationship between two numerical variables. Used in time plots (with time on the x-axis) or to display general associations between variables. Lines may or may not be included.

Axis (Plural: Axes)

The reference lines in a graph. The x-axis runs horizontally (typically showing categories or time), and the y-axis runs vertically (often showing numerical values or frequency).

Tip: You're now equipped with a whole toolkit of visualization types. Use these techniques to make your project data more clear, meaningful, and ready to analyze in Chapter 4 and beyond.

This page titled 3.6: Vocabulary (Chapter 3) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Mathematics Department.

3.7: Practice (Chapter 3)

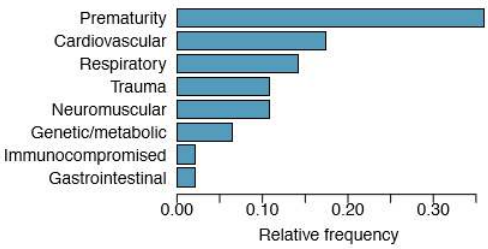
3.1: Graphical Displays for Categorical Data

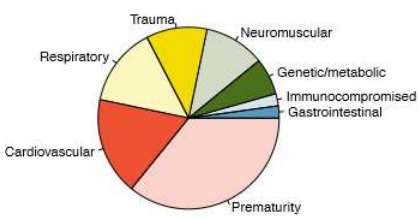
1. For each of the following, describe as categorical or numerical
   a. Hair color
   b. College GPA
   c. Humidity
   d. Character's level in a video game
   e. Subscribers to a content creator
   f. Score in a tennis match
   g. B-vitamins
   h. Round in the NCAA division tournament
   i. Musical notes in a melody
   j. Type of surround sound
   k. Speed limits
2. Give an example of a data set that would be best represented with a bar chart.
3. Give an example of a data set that would be best represented with a pie chart
4. Draw a bar chart to represent the following class survey data about animal companions: Cats (10), Dogs (8), Fish (4), Other (3)
5. Explain a limitation of pie charts when many categories are included.
6. Can a bar chart be used with relative frequencies instead of counts? Explain.
7. True or False: “There should be no space between bars in a bar chart.”
8. True or False: “There should be no space between slices in a pie chart.”
9. How do we decide if a pie chart or bar chart would be more appropriate?
10. The students in Ms. Ramirez’s math class have birthdays in each of the four seasons. The following table shows the four seasons, the number of students who have birthdays in each season, and the percentage (%) of students in each group. Construct a bar chart showing the number of students.

Birthday data for Math Class

Seasons	Number of students	Proportion of population
Spring	8	24%
Summer	9	26%
Autumn	11	32%
Winter	6	18%

11. Using the data from Ms. Ramirez’s math class supplied in the table above, construct a pie chart showing the percentages.
12. The bar plot and the pie chart below show the distribution of pre-existing medical conditions of children involved in a study on the optimal duration of antibiotic use in treatment of tracheitis, which is an upper respiratory infection.

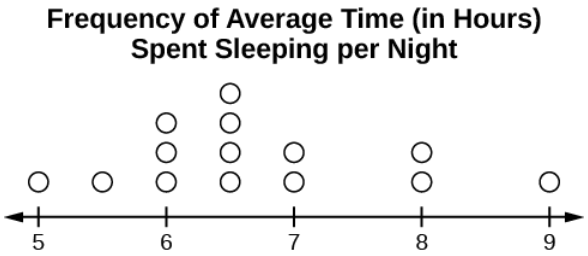




a. What features are apparent in the bar plot but not in the pie chart?
b. What features are apparent in the pie chart but not in the bar plot?
c. Which graph is better to use for this data and why?

3.2: Graphical Displays for Quantitative Data

1. What makes dot plots difficult to use for large sets of numerical data?
2. The following dot plot represents the average amount of sleep per night that students in a class report.



   a. What are the values in the original data set?
   b. If you created a similar plot with students in your class, do you think the results would be the same? Why or why not?
   c. Where do your data appear to cluster? How might you interpret the clustering?
3. Create dot plots for the following data sets and then use the dot plots to state how the means and standard deviations compare.
   a. set 1 {3, 5, 5, 5, 8, 11, 11, 11, 13}, set 2 {3, 5, 5, 5, 8, 11, 11, 11, 20}
   b. set 1 {-20, 0, 0, 0, 15, 25, 30, 30}, set 2 {-40, 0, 0, 0, 15, 25, 30, 30}
   c. set 1 {0, 2, 4, 6, 8, 10}, set 2 {20, 22, 24, 26, 28, 30}
   d. set 1 {100, 200, 300, 400, 500}, set 2 {0, 50, 300, 550, 600}
4. For Susan Dean's spring pre-calculus class, scores for the first exam were as follows (smallest to largest):

   33; 42; 49; 49; 53; 55; 55; 61; 63; 67; 68; 68; 69; 69; 72; 73; 74; 78; 80; 83; 88; 88; 88; 90; 92; 94; 94; 94; 94; 96; 100

Stem-and-Leaf Graph

Stem	Leaf
3	3
4	2 9 9
5	3 5 5
6	1 3 7 8 8 9 9
7	2 3 4 8
8	0 3 8 8 8
9	0 2 4 4 4 4 6
10	0

   What does the Stemplot tell you about the class's performance on the exam?
5. Consider a data set and Stemplot of the distances (in kilometers) from a home to local supermarkets.

   1.1; 1.5; 2.3; 2.5; 2.7; 3.2; 3.3; 3.3; 3.5; 3.8; 4.0; 4.2; 4.5; 4.5; 4.7; 4.8; 5.5; 5.6; 6.5; 6.7; 12.3

Stem	Leaf
1	1 5
2	3 5 7
3	2 3 3 5 8
4	0 2 5 5 7 8
5	5 6
6	5 7
7	
8	
9	
10	
11	
12	3

a. Do the data seem to have any concentration of values?

b. What does the bottom row of the Stemplot tell you?

6. Consider the data set represented by the ordered stem and leaf diagram:

10	0 0
9	1 1 1 1 2 3
8	0 1 1 2 2 3 4 5 7 8 8 9
7	0 0 1 1 2 4 4 5 6 6 6 7 7 7 8 8 9
6	0 1 2 2 2 3 4 4 5 7 7 7 7 8 8
5	0 2 3 3 4 4 6 7 7 8 9
4	2 5 6 8 8
3	9 9

a. Find the three quartiles.

b. Give the five-number summary of the data.

c. Find the range and the IQR.

7. For the following stem and leaf diagram the units on the stems are thousands and the units on the leaves are hundreds, so that

for example the largest observation is 3,800.

3	5 6 8
3	0 0 1 1 2 4
2	5 6 6 7 7 8 8 9 9
2	0 0 0 0 1 2 2 4
1	5 5 5 6 6 7 7 7 8 8 9
1	0 0 1 3 4 4 4
0	5 6 8 8
0	4

a. Find the percentile rank of 800.
b. Find the percentile rank of 3,200.

8. Construct a stem-and-leaf plot for: {34, 36, 37, 41, 42, 42, 43, 46}

9. Construct a Stemplot using the data in the table below

Table : Presidential Ages at Inauguration

President	Ageat Inauguration	President	Age	President	Age
Pierce	48	Harding	55	Obama	47
Polk	49	T. Roosevelt	42	G.H.W. Bush	64
Fillmore	50	Wilson	56	G. W. Bush	54
Tyler	51	McKinley	54	Reagan	69
Van Buren	54	B. Harrison	55	Ford	61
Washington	57	Lincoln	52	Hoover	54
Jefferson	57	Grant	46	Truman	60
Madison	57	Hayes	54	Eisenhower	62
J. Q. Adams	57	Arthur	51	L. Johnson	55
Monroe	58	Garfield	49	Kennedy	43
J. Adams	61	A. Johnson	56	F. Roosevelt	51
Jackson	61	Cleveland	47	Nixon	56
Taylor	64	Taft	51	Clinton	47
Buchanan	65	Coolidge	51	Trump	70
W. H. Harrison	68	Cleveland	55	Carter	52

10. Forty randomly selected students were asked the number of pairs of sneakers they owned. Let X = the number of pairs of sneakers owned. Use the data in the table to create a stem plot and find the sample mean.

Data table of number of shoes students own.

X	Frequency
1	2
2	5
3	8
4	12
5	12
6	0
7	1

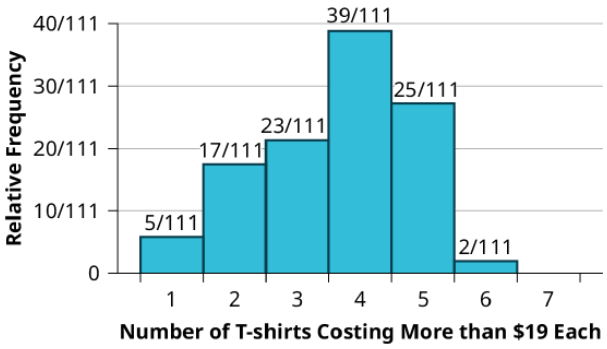
3.3: Frequency Tables and Relative Frequency

1. Fill in the missing frequency for a data set with sample size 40: A = 8, B = 12, C = ___, D = 4
2. Why is it useful to create a frequency table before creating a histogram?
3. Does a frequency table preserve original data values? Why does that matter?
4. What does the frequency column in a table sum to? Why?
5. What does the relative frequency column in table sum to? Why?
6. Explain the difference between frequency and relative frequency.
7. What is the difference between cumulative relative frequency and relative frequency for each data value?
8. What is the key difference between a bar chart and a histogram?
9. What does the height of a bar in a histogram represent?
10. Describe one advantage of a stem plot over a frequency histogram.
11. Choose a quantitative variable from your dataset. Would you rather display it with a histogram, dot plot, or stem plot? Why?
12. Construct a frequency table for: {2, 2, 3, 3, 3, 4, 5, 5}
13. Draw a histogram for the values: {2, 2, 3, 5, 5, 5, 6, 7, 8, 10}
14. Sixty-five randomly selected car salespersons were asked the number of cars they generally sell in one week. Fourteen people answered that they generally sell three cars; nineteen generally sell four cars; twelve generally sell five cars; nine generally sell six cars; eleven generally sell seven cars. Complete the table.

Empty data table to complete

Data Value (# cars)	Frequency	Relative Frequency	Cumulative Relative Frequency

15. Suppose 111 people who shopped in a special T-shirt store were asked the number of T-shirts they own costing more than $19 each.



a. What is the percentage of people who own at most three T-shirts costing more than $19 each?
b. If the data were collected by asking the first 111 people who entered the store, then what type of sampling was used?

16. Six hundred adult Americans were asked by telephone poll, "What do you think constitutes a middle-class income?" The results are in the table below.

Table of income Americans consider "middle-class"

Salary ($)	Relative Frequency
< 20,000	0.02
20,000–25,000	0.09
25,000–30,000	0.19
30,000–40,000	0.26
40,000–50,000	0.18
50,000–75,000	0.17
75,000–99,999	0.02
100,000+	0.01

a. What percentage of the survey answered "not sure"?
b. What percentage think that middle-class is from $25,000 to $50,000?
c. Construct a histogram of the data.
   i. Should all bars have the same width, based on the data? Why or why not?
   ii. How should the <20,000 and the 100,000+ intervals be handled? Why?
d. Find the 40th and 80th percentiles

17. List all the measurements for the data set represented by the following data frequency table.

X	31	32	33	34	35
frequency	1	5	6	4	2

18. Construct the data frequency table for the following data set.

22	25	22	27	24	23	26	24	22	24	26

19. The IQ scores of ten students randomly selected from an elementary school are given. {108, 100, 99, 125, 87, 105, 107, 105, 119, 118}. Grouping the measures in the 80's, the 90's, and so on, construct a stem and leaf diagram, a frequency histogram, and a relative frequency histogram.

20. The IQ scores of ten students randomly selected from an elementary school for academically gifted students are given. {133, 140, 152, 142, 137, 145, 160, 138, 139, 138}. Grouping the measures by their common hundreds and tens digits, construct a stem and leaf diagram, a frequency histogram, and a relative frequency histogram.
21. During a one-day blood drive 300 people donated blood at a mobile donation center. The blood types of these 300 donors are summarized in the table. Construct a relative frequency histogram for the data set.

Blood Type	O	A	B	AB
Frequency	136	120	32	12

22. In a particular kitchen appliance store an electric automatic rice cooker is a popular item. The weekly sales for the last 20 weeks are shown. Construct a frequency histogram with a "nice" class width.

20	15	14	14	18	17	16	16	18
15	16	19	9	15	13	12	15	15

23. Fifty part-time students were asked how many courses they were taking this term. The (incomplete) results are shown below:

Part-time Student Course Loads

# of Courses	Frequency	Relative Frequency	Cumulative Relative Frequency
1	30	0.6	
2	15		
3			

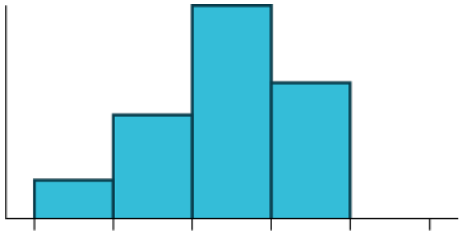
a. Fill in the blanks in the table.
b. What percent of students take exactly two courses?
c. What percent of students take one or two courses?

24. Javier and Ercilia are supervisors at a shopping mall. Each was given the task of estimating the mean distance that shoppers live from the mall. They each randomly surveyed 100 shoppers. The samples yielded the following information.

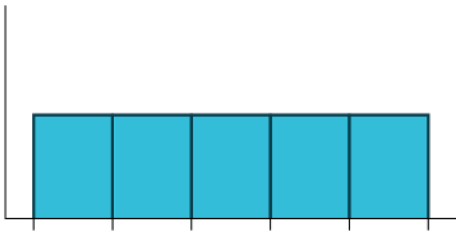
Data table of distance shoppers live from mall.

	Javier	Ercilia
mean	6.0 miles	6.0 miles
s	4.0 miles	7.0 miles

a. How can you determine which survey was correct ?
b. Explain what the difference in the results of the surveys implies about the data.
c. If the two histograms depict the distribution of values for each supervisor, which one depicts Ercilia's sample? How do you know?



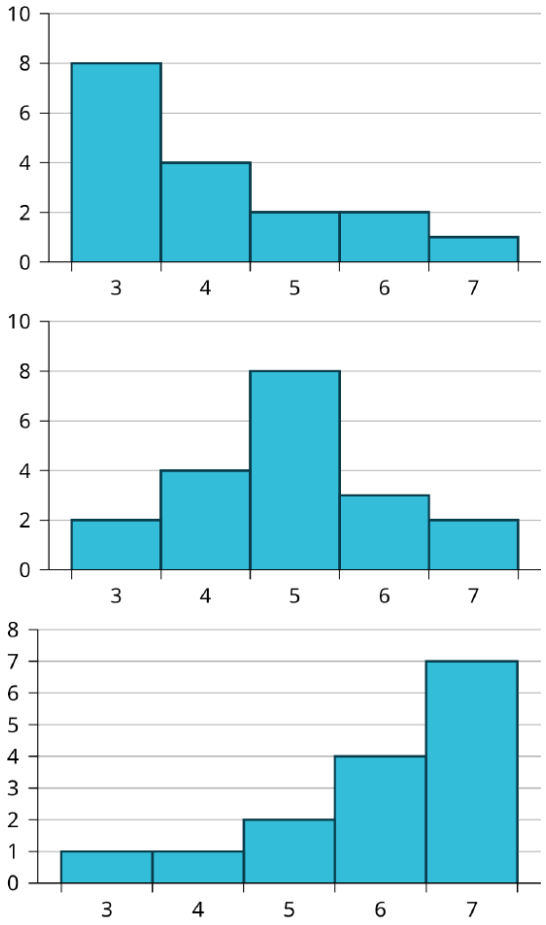
(a)



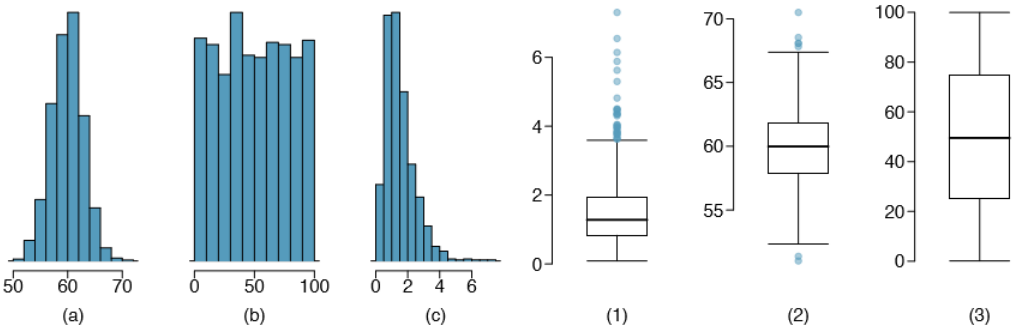
(b)

3.4: Shapes of Distributions

1. Describe how you can identify whether a histogram is right-skewed, left-skewed, or symmetric.
2. Which measure of center is better to use with skewed data: mean or median? Why?
3. Draw a rough histogram with a left-skewed distribution. Label axes.
4. How would a boxplot help verify the shape of a distribution?
5. What are the similarities and differences between unimodal and bimodal?
6. When the data are skewed left, what is the typical relationship between the mean and median?
7. When the data are symmetrical, what is the typical relationship between the mean and median?
8. True/False: A bimodal distribution has two modes and two medians.
9. True/False: The best way to describe a skewed distribution is to report the mean.
10. If the mean time to respond to a stimulus is much higher than the median time to respond, what can you say about the shape of the distribution of response times?
11. Provide a real-world example of a variable that likely has a right-skewed distribution.
12. For each of the following graphs, describe the shape of the distribution.



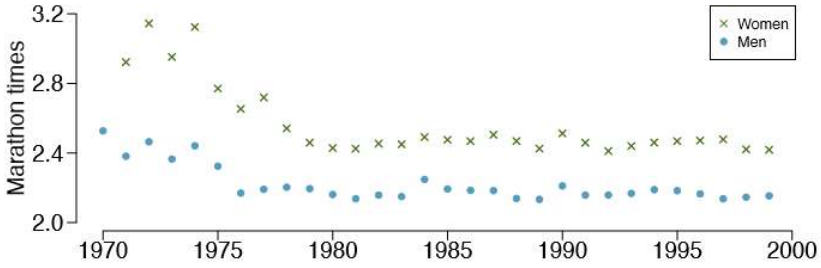
13. Describe the shape of the distributions in the histograms below and match them to the box plots.



14. For each of the following, describe whether you expect the distribution to be symmetric, right skewed, or left skewed. Also specify whether the mean or median would best represent a typical observation in the data, and whether the variability of observations would be best represented using the standard deviation or IQR.
    a. Housing prices in a country where 25% of the houses cost below $350,000, 50% of the houses cost below $450,000, 75% of the houses cost below $1,000,000 and there are a meaningful number of houses that cost more than $6,000,000.
    b. Housing prices in a country where 25% of the houses cost below $300,000, 50% of the houses cost below $600,000, 75% of the houses cost below $900,000 and very few houses that cost more than $1,200,000.
    c. Number of alcoholic drinks consumed by college students in a given week.
    d. Annual salaries of the employees at a Fortune 500 company.

3.5: Time Series Plots and Trends Over Time

1. Identify two types of data (outside this class) that would make good candidates for a time plot.
2. What does it mean for a variable to "trend upward" over time?
3. Can a time plot use bars instead of points and lines? Why or why not?
4. Look at your project data. If it includes any time-stamped data or dates, could you use a time plot? What would it show?
5. How would a time plot be different than a histogram for the same data?
6. The time series plot below shows the finishing times for male and female winners of the New York Marathon between 1980 and 1999.



    a. What points in time are significant for this data?
    b. What happened at those points in time?
7. Create a time plot for this data: Months: Jan–Jun; Visitors: 204, 228, 305, 311, 298, 287
8. Construct a times series graph for the total number of births in Scotland.

Table of the number of births in Scotland

Year	1855	1856	1857	1858	1859	1860	1861
Total Births	93,349	101,821	103,415	104,018	106,543	105,629	107,009

Number of births in Scotland (continued)

Year	1862	1863	1864	1865	1866	1867	1868	1869
Total Births	107,069	109,341	112,333	113,070	113,667	114,044	115,514	113,354

Number of births in Scotland (continued)

Year	1871	1870	1872	1871	1872	1827	1874	1875
Total Births	116,128	115,390	118,765	116,128	118,765	119,700	123,711	123,578

9. The following data sets list full time police per 100,000 citizens along with homicides per 100,000 citizens for the city of Detroit, Michigan during the period from 1961 to 1973.

Data table of Police and Homicides over time

Year	1961	1962	1963	1964	1965	1966	1967
Police	260.35	269.8	272.04	272.96	272.51	261.34	268.89
Homicides	8.6	8.9	8.52	8.89	13.07	14.57	21.36

Data table of Police and Homicides (continued)

Year	1968	1969	1970	1971	1972	1973
Police	295.99	319.87	341.43	356.59	376.69	390.19
Homicides	28.03	31.49	37.39	46.26	47.24	52.33

a. Construct a double time series graph using a common *x*-axis for both sets of data.
b. Which variable increased the fastest? Explain.
c. Did Detroit's increase in police officers have an impact on the murder rate? Explain.

10. The following data shows the Annual Consumer Price Index for ten years. Construct a time series graph for the Annual Consumer Price Index.

Year	Annual
2003	184.0
2004	188.9
2005	195.3
2006	201.6
2007	207.342
2008	215.303
2009	214.537
2010	218.056
2011	224.939
2012	229.594

11. The following table is a portion of a data set from www.worldbank.org. Use the table to construct a time series graph for CO_2 emissions.

CO_2 Emissions

	Ukraine	United Kingdom	United States
2003	352,259	540,640	5,681,664
2004	343,121	540,409	5,790,761
2005	339,029	541,990	5,826,394
2006	327,797	542,045	5,737,615
2007	328,357	528,631	5,828,697
2008	323,657	522,247	5,656,839
2009	272,176	474,579	5,299,563

12. The Gross Domestic Product Purchasing Power Parity (GDP PPP) is an indication of a country's currency value compared to another country. The table below shows the GDP PPP of Cuba as compared to US dollars. Construct a time series plot of the data.

Year	Cuba's PPP	Year	Cuba's PPP
1999	1,700	2006	4,000
2000	1,700	2007	11,000
2002	2,300	2008	9,500
2003	2,900	2009	9,700
2004	3,000	2010	9,900
2005	3,500		

13. For the following situations, determine which types of graphs are appropriate ways of displaying the data. Then choose which graph type would be best and explain why.
    a. A study on whether IQ level is related to birth order (oldest, middle, youngest)
    b. A study on the relationship between the annual income of a family and the amount of money the family spends on entertainment
    c. A store asked 250 of its customers whether they were satisfied with the service or not
    d. The median US household cost for housing each year since 1989
    e. Peoples' favorite television show
    f. The average length of time spent commuting to work
    g. The average cost of bread each day over the course of a month

Practice problems include parts of:

Introductory Statistics 2e by OpenStax is licensed under CC BY 4.0

Access for free at https://openstax.org/books/introductory-statistics-2e/pages/1-introduction

and

OpenIntro Statistics by David Diez, Christopher Barr, & Mine Çetinkaya-Rundel is licensed under CC BY 3.0

3.7: Practice (Chapter 3) is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

- **Current page** has no license indicated.
- **1.8: Practice (Chapter 1)** by Mathematics Department is licensed CC BY 4.0.
- **1.E: Introduction to Data (Exercises)** by David Diez, Christopher Barr, & Mine Çetinkaya-Rundel is licensed CC BY-SA 3.0. Original source: https://www.openintro.org/book/os.
- **2.E: Descriptive Statistics (Exercises)** has no license indicated.
- **1.E: Sampling and Data (Exercises)** by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/introductory-statistics.
- **2.E: Descriptive Statistics (Exercises)** by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/introductory-statistics.
- **Licensing** by Mathematics Department has no license indicated.

4: Introduction to Probability

One of the most fundamental ideas in statistics is **chance**: How likely is something to happen? How often should we expect an event to occur, and how different is that from what *actually* happens when we try it in real life?

Welcome to the world of **probability**, where we explore uncertainty, randomness, and patterns that emerge across large numbers of trials.

A Dice Problem: Is This Set Fair?

Imagine your friend hands you a brand-new set of dice and says, “I think one of these might be unfair.” You roll them once; no problem. You roll them a few more times, nothing obvious. But what if you rolled each die *100 times*? Or *1,000 times*? What patterns would you see? Would certain numbers show up too often?

If the dice behave the way you expect, each number appearing about the same number of times, you might feel confident that the set is fair. But if one die keeps landing on 6 more often than any other number, you’d start to wonder.

> **This is the essence of probability:** When we can’t be sure what will happen on the next roll, we use tools that help us understand *long-term behavior*.

In this chapter, we’ll look at how probability helps us:

- Make sense of randomness
- Recognize long-term patterns in repeated trials
- Disentangle expected results from surprising streaks
- Lay the foundation for inference and predictions later in the course

As we dive into coin flips, rolling dice, drawing cards, and running simulations, we’ll start to build a language for chance — and a framework for interpreting uncertainty, both in games and in real-world experiments.

Ready to roll? Let’s begin.

This page titled 4: Introduction to Probability is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Mathematics Department.

4.1: What is Probability?

In our everyday lives, we're constantly working with uncertainty. Will it rain today? What are the chances of getting in a car accident? How confident can we be that a medicine works? This is where probability comes in.

Probability is the critical for developing statistics. It gives us a language and a set of tools for understanding, describing, and modeling randomness. It allows us to measure uncertainty and reason logically about outcomes that aren't predictable with certainty.

> Definition: Probability
>
> **Probability** is the *long-run relative frequency* of an event. This is the proportion of times an outcome would occur if we repeated a random process over and over, under the same conditions.

A Simple Example: Coin Flipping

Think about flipping a coin. If you flip it once, you might get heads. If you flip it twice, maybe tails. But if you flip the coin 100 times, or 1,000 times, you'll notice a pattern emerge: you get heads about half the time.

That's the idea behind probability. In the **long run**, there is a predictable pattern in a repeating random process, even though you can't predict what will happen in the **short run**.

Important Concept: Probability Is Not Prediction

Probability doesn't guarantee an outcome, as it only describes what *tends to happen* in many repetitions. Even though flipping a coin has a 50% chance of heads, we might still get 7 heads out of 10 flips, or even 4 heads out of 10. That's not unusual in the short run. It is impossible for us to use probability to predict the next flip, only know how what we predict the statistics from many many flips to be.

As we gather more data or observe more repetitions, the pattern in the coin stabilizes. That's why probability is often described as measuring **long-term relative frequency**.

Explore It Yourself: Probability in the Long Run

To better understand how probability behaves in the long run, try the interactive applet below.

In this applet:

- You can set the value of the **theoretical probability** $P_{success}$ (e.g., 0.5 for a fair coin).
- The word trials refers to repeated investigations of a probabilistic idea. We represent the trials as k and you can see the effect of it by pressing play.
- As you increase k, the graph will show the proportion of successes over trials based on random outcomes using that value of P.

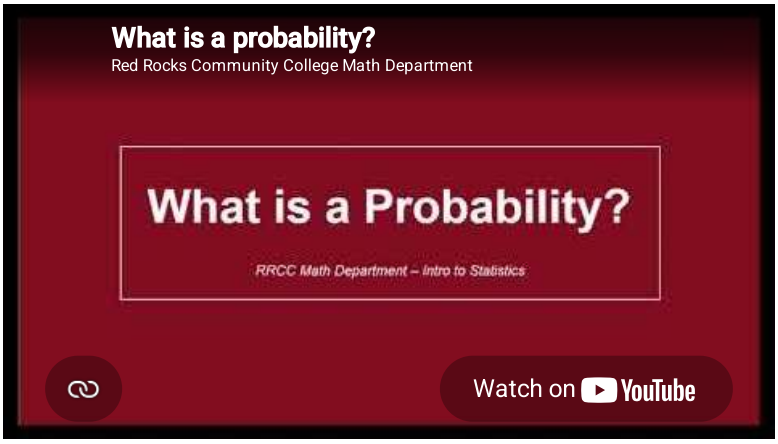
When we have a small number of trials, small deviations from a 50/50 ratio stand out more. For example, if we flip 4 coins and get 3 heads, that is 75% of our data set, however it is only one value off from 50/50. If we flip 100 coins, being off by one would be 51 heads to 49 tails. This is very close to 50/50!

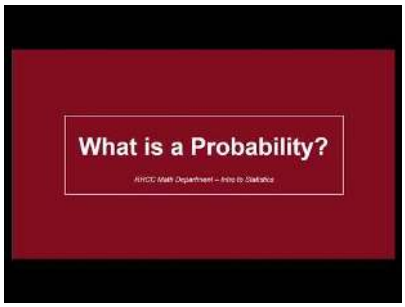
Think About It:

Try running the app with $P = 0.5$, $k = 50$, and then again with $k = 1000$.

What does the graph look like when trials are small vs. large? How does this demonstrate the definition of probability as a long-run relative frequency?

Related Videos





What's Coming Next

In the next sections, we'll explore tools that help us count and calculate probabilities. We'll look at defining **sample spaces**, building rules for combining probabilities, and simulating some events to get a better feel for what random processes look like.

For now, remember: probability is not a prediction engine; it's a map of what we expect over time.

This page titled 4.1: What is Probability? is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Mathematics Department.

4.2: Introduction to Counting

Probability relies on understanding what outcomes are **possible** and how many of them there are. But before we can study chance, we need to learn how to **count** correctly. This field is called **combinatorics**, and it helps us answer questions like:

- How many different ways can something happen?
- What changes when order *does* or *does not* matter?
- What do we do when we can't repeat selections?

> ### Counting is not Probability
>
> Counting doesn't tell us the likelihood of an event but only gives us the pieces we need to figure that out. Once we know how many outcomes are possible (and favorable), we can compute probability (we will see the approach in the next sections)

The multiplication rule of counting

Suppose we are curious about the total number of outcomes from flipping three coins in a row, while keeping track of the order. We list all possibilities below:

$$\{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\} \tag{4.2.1}$$

Listing these out gives us a total of 8 possibilities. We figured this number from *literal counting*. However, what if we were interested in how many ways there are to flip 100 coins? That would take an extremely long time to count. So we devise mathematical methods to calculate counts without needing to sit and tally. Let's consider the 3-coin scenario again. Instead of listing all possible outcomes we can consider one flip at a time.

For the first flip, we have two possibilities; a Heads or a Tails. After one of those occurs, there are now two possibilities for the second flip. This means, for each of the two possibilities of the first flip, there are two possibilities for the second. We can easily see this when we write them out: HH, HT, TH and TT. We have a total of $2 \times 2 = 4$. However think about the third flip. For each of the 4 possible outcomes of the first two, there will be 2 more possibilities on the third flip. Therefore we can take $4 \times 2 = 8$ which matches out result from counting above.

This is called the **multiplication rule of counting**. It says that if we have n ways of one thing occurring, and m ways of another, there are $n \times m$ ways they can happen together, assuming that the results of one way do not influence the other.

Now suppose we flip 100 coins. For each flip we multiply by 2. This leads to us needing to multiply by 2 over and over 100 times! However, we have notation for this, called **exponents** or **powers.**

$$\text{number of outcomes } = 2 \times 2 \times 2 \times \cdots \times 2 = 2^{100} \tag{4.2.2}$$

We can calculate that $2^{100} = 1267650600228229401496703205376$. That's the total number of ways to flip 100 coins! Imagine counting to a number that large.

How secure is your PIN?

A more practical computation involves checking security of passwords and PINs (Personal Identification Numbers). A PIN consists of 4 digits which can take the 10 values between 0 and 9 (check by counting that there are 10 digits including 0)

In this case, instead of a coin flip, we have 10 possibilities for each digit. Our total number of possibilities is $10 \times 10 \times 10 \times 10 = 10^4 = 10000$

This is the total number of possible PINs, which means randomly guessing one would be very unlikely.

What if repetition is not allowed?

Suppose we wanted to assign labels to 5 objects as a through e. The number of ways to do this is different from the previous examples. Let's consider the first item. We can label it a, b, c, d or e for a total of 5 possibilities. Let's say we decide to call it c. We move on to the second object. However, this time, there are only 4 possibilities remaining: a, b, d or e since we've already used the label c. Suppose we decide to call the second object a. Moving on to the third, we now have 3 choices: b, d or e. We call it e. The fourth object can be labeled in two ways, either b or d and we pick d. This means there is only 1 way to label the fifth object; it must be labeled b.

We end up with the specific sequence c, a, e, d, b. This is a particular way we could have labeled the objects, but how many total ways are there? Well, we appeal to the multiplication rule, but now have to consider that the number of ways to label decreases after each choice. There were 5 ways to label the first object, but once it is labeled there are only 4 ways for the second. This means there must be $5 \times 4 = 20$ ways just to label those first two. Continuing the pattern yields

$$\text{number of ways to label 5 items} = 5 \times 4 \times 3 \times 2 \times 1 = 120 \tag{4.2.3}$$

This type of pattern, where we start with a number and multiply by all whole numbers descending from it, is so common it has a specific name and symbol. We call it a **factorial** and write it as $5!$ for example. We would pronounce that as "five factorial" and it would be equal to 120. Many scientific calculators are equipped with a factorial which can be typed with the exclamation point. In Microsoft Excel, factorials are done via the command **=FACT()**

> ### Definition: Factorial
>
> The factorial of a positive integer n, written $n!$, is the product of all positive integers less than or equal to n.
> Example: $6! = 6 \times 5 \times 4 \times 3 \times 2 \times 1 = 720$

How many ways are there to shuffle a deck of cards?

A deck of playing cards consists of 52 standard cards and 2 jokers, usually omitted from play. If we shuffle the 52 standard cards, it is like randomly assigning them to an order. There are 52 possibilities for the top card, 51 for the card below it, 50 for the third card, and so on. We would have a total of $52!$ possibilities. Evaluating this in a calculator gives an answer of 8.0658×10^{67}. This number is so large that it requires a special type of notation called **scientific notation** to write it out. In the number to the left, you can see we have a value 8.0658 which is being multiplied by 10^{67}. As we previously saw, this means 10 multiplied by itself 67 times. Each time we multiply by 10 we move the **place value** over by 1, changing where the decimal point is:

$$8.0658 \times 10 = 80.658 \tag{4.2.4}$$

As we do this more times we can simply count the decimal change:

$$8.0658 \times 10^3 = 8065.8 \tag{4.2.5}$$

If we have more multiplications than digits we simply add trailing zeros.

$$8.0658 \times 10^6 = 8065800 \tag{4.2.6}$$

If we do this for the deck of cards we do this 67 times; can you see why we prefer scientific notation for large numbers?

$$\begin{aligned} &8.0658 \times 10^{67} \\ &= 80658000000000000000000000000000000000000000000000000000000000000000000 \end{aligned} \tag{4.2.7}$$

Let this number sink in for a minute; it is the number of possible ways to shuffle a deck of cards. If we had every single person on early shuffle a deck of cards that would be about 8000000000 shuffles. That's not even going to make a dent in the total number of ways! In fact, there are so many ways to shuffle a deck, that if you do a perfect shuffle it is overwhelmingly likely that your specific card order has *never been seen in the history of cards.*

What if we use only part of the digits?

Let's say we want to create a 3-digit ID from the numbers 0 through 5, with no repeats. We'd count:

$$5 \times 4 \times 3 = \frac{5\times4\times3\times2\times1}{2\times1} = \frac{5!}{2!}$$

This is called a **permutation**, which is an arrangement of items where **order matters**.

Definition: Permutation

A permutation is an arrangement of objects where **order matters**. The number of permutations of r items from a group of n is:

$$P(n,r) = \frac{n!}{(n-r)!}$$

What if order doesn't matter?

Now imagine forming a 3-student committee from a class of 10 students. It doesn't matter if Alex, Lee, and Jada are listed in a different order; it's the same group.

This is where we use a **combination**. We calculate:

$$\frac{10\times9\times8}{3\times2\times1} = \frac{10!}{3!(10-3)!} = \binom{10}{3}$$

Definition: Combination

A combination is a selection of items **where order does not matter**. The number of combinations of r items from a group of n is:

$$C(n,r) = \binom{n}{r} = \frac{n!}{r!(n-r)!}$$

More examples

Let's look at another example of each case.

Case 1: Digits Can Repeat

A United States Social Security Number (SSN) has 10 positions. Each position can be filled by any digit from 0–9. That gives us 10 choices per digit:

$$10 \times 10 \times 10 \times 10 \times 10 \times 10 \times 10 \times 10 \times 10 \times 10 = \mathbf{10^{10} = 10,000,000,000}$$

This represents the number of SSNs possible if any digit can appear any number of times. This is a straightforward use of the **multiplication rule**.

Case 2: No Repeating Digits

Suppose we wish to add 10 items to a list. Each item will receive a number in the list, but they cannot be repeated.

We still have 10 positions, but one fewer option at each step:

$$10 \times 9 \times 8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1 = \mathbf{10! = 3,628,800}$$

This use of **factorials** gives the number of ways to arrange 10 unique digits.

Case 3: Using Fewer Digits without Repeats

What if we only wanted to list out the top 4 items from 10 choices? There would only be 4 positions to fill, but 10 possibilities for the first one.

$$10 \times 9 \times 8 \times 7 = \frac{10!}{(10-4)!} = \frac{10!}{6!} = 5040$$

This example was a **permutation**, a partial selection **where order matters**.

Case 4: Order Doesn't Matter

Finally, suppose you're drawing a 4 raffle numbers and only care what the values are and not the order. This is a combination.

Using the use the **combinations formula**:

$$\binom{10}{4} = \frac{10!}{4!(10-4)!} = \frac{10!}{4!6!} = 210$$

So out of 10 unique digits, there are 210 different ways to choose 4 of them when order doesn't matter. That's many fewer than with order!

> **Next Idea:** Many problems in probability start by asking "What is the total number of possible outcomes?" We then divide the number of specific outcomes we are interested in to obtain their likelihood.

Summary

- **Multiplication Rule:** Use when steps are independent (e.g., 10×10 for repeated digits)
- **Factorials:** Represent total arrangements of n unique items
- **Permutations:** Order matters → $\frac{n!}{(n-r)!}$
- **Combinations:** Order doesn't matter → $\binom{n}{r}$

This page titled 4.2: Introduction to Counting is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Mathematics Department.

4.3: Foundational Probability Rules

Now that we've built an idea of probability as long-run behavior, we're ready to learn the tools and terminology that help us compute it. That starts with identifying the building blocks behind any chance-based scenario.

Definition: Experiment

An **experiment** is a process or action that produces a result, called an **outcome**, under conditions that involve randomness or unpredictability.

In probability, an experiment is something we can do repeatedly, where the exact result cannot be predicted in advance, but the set of all possible outcomes is known. Many of our examples will be simple for ease of learning, but realistic scenarios can also be thought of as experiments.

Examples:

- Flipping a coin
- Rolling a die
- Drawing a card from a shuffled deck of playing cards
- Collecting daily rainfall data
- Surveying a randomly selected person about their favorite food
- Administering a new drug to a patient and recording if they encounter side-effects
- Investing in a stock and seeing the resulting profit
- Giving an exam to a student and grading their results

In probability, everything begins with a clearly defined **experiment**. Each time we run the experiment, we observe one specific **outcome**. We may not know in advance which outcome will occur, but we can describe all the possible outcomes that could happen. Defining the experiment and the possible outcomes first helps us stay clear about what we're measuring and what each result means.

Definitions: Outcome & Sample Space

An **outcome** is a single possible result of a defined experiment. Every time we run a probability experiment, one specific result occurs. That result is the outcome.

For example, if the experiment is rolling a six-sided die, one possible outcome might be **4**. Order generally matters with outcomes. Rolling two dice and getting a **4** and a **2** is a different outcome than a **2** and a **4.**

The **sample space**, denoted by S, is the set of *all possible outcomes* of a probability experiment.

Using the die example, the sample space would be:
$S = \{1, 2, 3, 4, 5, 6\}$

In summary: An experiment produces one outcome, and the sample space lists *every* outcome that could possibly happen.

A few examples

Experiment: Flip two fair coins.
Sample space: $\{(H, H), (H, T), (T, H), (T, T)\}$

Experiment: Give a student an exam.
Sample space: $\{A, B, C, D, F\}$

Experiment: Recording daily rainfall
Sample space: All positive real numbers

Note that our sample space can be **discrete** or **continuous** just like the data from chapter 2. Additionally, there may be different ways to construct the sample space. If we wanted the percentage of the exam, our sample space would be all percentages between 0% and 100% instead of letter grades.

> Definition: Event
>
> An **event** is a *set of outcomes* from a random experiment (a subset of the sample space).

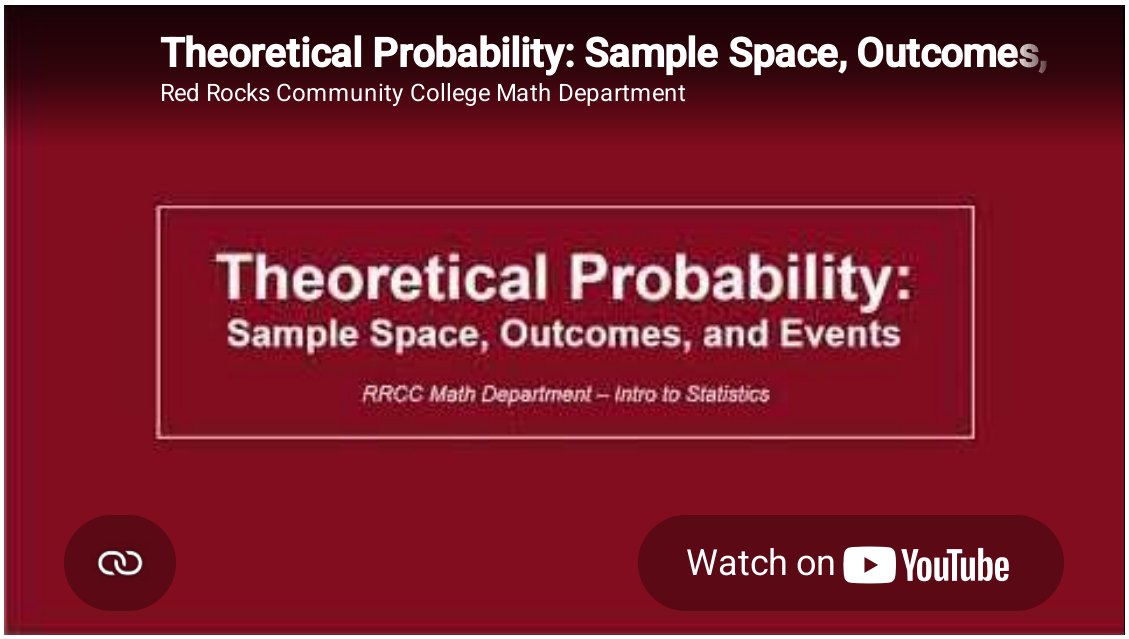
This is best illustrated with a few examples

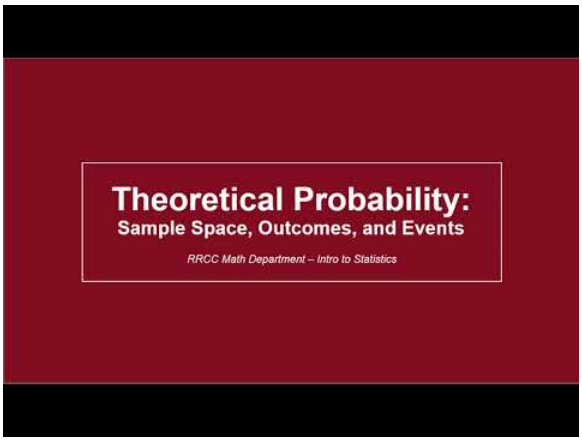
Experiment: Flip two coins
Event description: Get at least one Heads
Event: $\{(H, H), (H, T), (T, H)\}$

Experiment: Give a student an exam
Event description: Student passes
Event: $\{A, B, C\}$

Experiment: Recording daily rainfall
Event description: More than 2 inches of rain
Event: All real numbers greater than 2. We can write this as $R > 2$ where R is the number of inches of rain.

Related Video - Sample Space and Events





When Events Overlap or Not

Definition: Mutually Exclusive Events

Two events are **mutually exclusive** if they **cannot happen at the same time**. Their outcomes do not overlap. We also call these events **disjoint**.

Example:
Event A = "Both coins show heads" → {(H, H)}
Event B = "Both coins show tails" → {(T, T)}
A and B are **mutually exclusive** — they cannot both be true on the same trial.

Counterexample:
Event C = "At least one head" → {(H, H), (H, T), (T, H)}
Event D = "First coin is heads" → {(H, H), (H, T)}
C and D are **not mutually exclusive** — they both include (H, H).

Check Your Understanding: Are These Events Mutually Exclusive?

Question 1:
Event A: "Roll a 4 on a die"
Event B: "Roll an even number on a die"

Select one:

— Select an answer —

Question 2:
Event A: "Draw a red card from a deck"
Event B: "Draw a heart from a deck"

Select one:

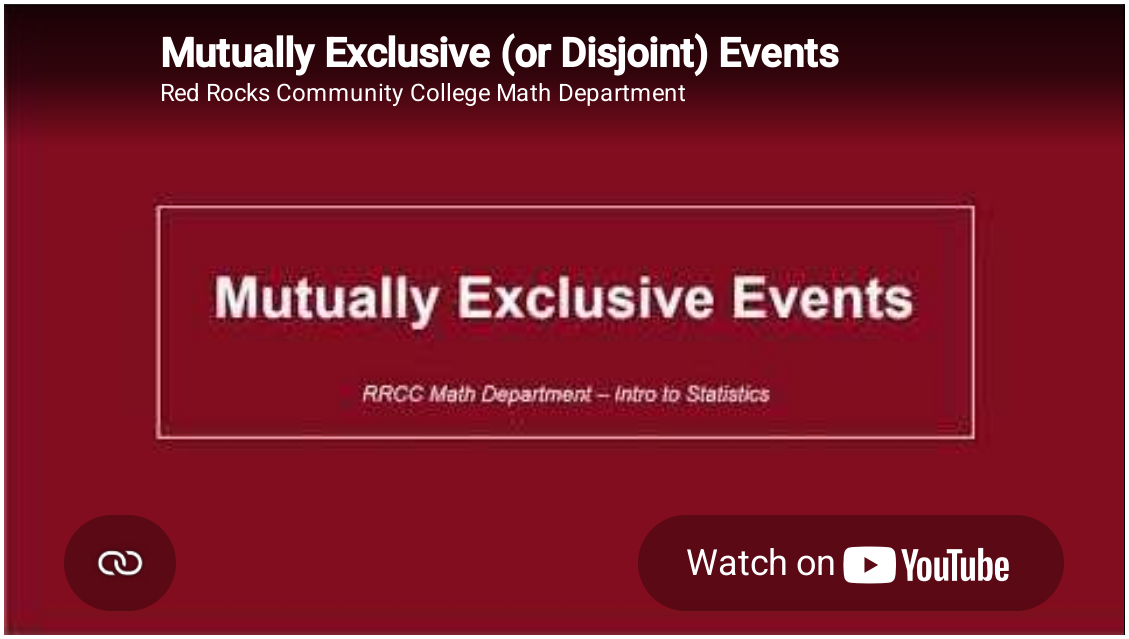
— Select an answer —

Question 3:
Event A: "Student is in 10th grade"
Event B: "Student is in 11th grade"

Select one:

— Select an answer —

Related Video - Mutually Exclusive





Probability as a Value

Definition revisit: Probability

We defined probability in 4.1 as the following:

Probability is the *long-run relative frequency* of an event. This is the proportion of times an outcome would occur if we repeated a random process over and over, under the same conditions.

Now that we have a definition of an event we can meaningfully use this idea. To make it more concrete we add the following conditions to the probability of an event:

Probabilities must be recorded as a real number between 0 and 1. A probability of 0 means that an event cannot occur and a probability of 1 indicates that an event is guaranteed to occur.

If we have denoted an event as A, we write probabilities as: $P(A)$, which means "the probability that event A occurs".

Basic Examples

Experiment: We flip two coins
Sample space: $\{(H,H),(H,T),(T,H),(T,T)\}$
Event: We get two heads; denoted as $H2$
Probability: Looking at the sample space we see there are 4 possible outcomes. However, only one of these outcomes yields two heads, when we have (H,H). Since all outcomes are equally likely, we have a 1 in 4 change of this event occurring. We would say $P(H2) = 1/4 = 0.25$. It is okay to represent probabilities as either fractions or decimals, although one notation may be preferable for communication.

Experiment: We roll a die
Sample space: $\{1,2,3,4,5,6\}$
Event: We get a roll of 5 or higher. We can denote this by $R \geq 5$ where R is our roll.
Probability: Since there are two possible rolls in this event, 5 or 6, out of a total of 6 outcomes in the sample space, we have $P(R \geq 5) = 2/6 = 0.3333$

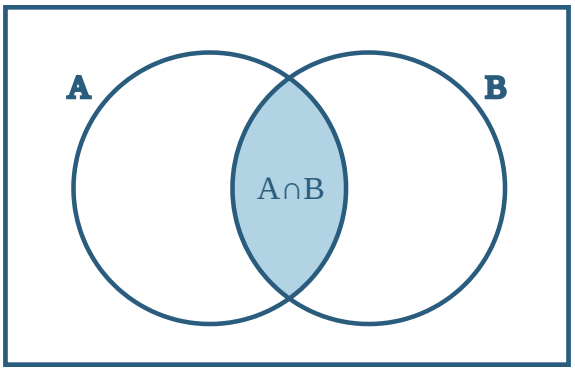
Building Events from Events

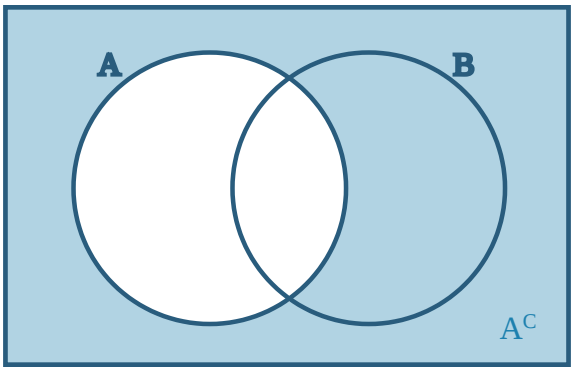
Sometimes, we define new events by combining existing ones. These compound events use the logic of **“or”**, **"and"**, and **“not”**.

- **"A or B"** (written as $A \cup B$) means event A happens, or event B happens, or both. It can be considered "at least one of A or B happens.
- **"A and B"** (written as $A \cap B$) means event A and B both occur. Notice that if A and B are disjoint, then $P(A \cap B) = 0$ since they cannot simultaneously happen.
- **"Not A"** (written as A^c) means A does not occur. This is also called the **complement** of A.

> ### Definition: Complement
>
> The **complement** of event A (written A^c) is the set of outcomes in the sample space that are *not* in A.

To show these concepts visually, we often use **Venn diagrams**. They’re helpful models for compound events.





The Addition Rule

The addition rule helps calculate the probability that event A *or* event B occurs:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \tag{4.3.1}$$

If A and B are mutually exclusive → $P(A \cap B) = 0$, so:

$$P(A \cup B) = P(A) + P(B) \tag{4.3.2}$$

> Think About It:
>
> Why do we subtract $P(A \cap B)$ in the general addition rule? What would happen if we didn't?

Examples Using the Addition Rule

Example 1: Events with Overlap (Not Mutually Exclusive)

In a group of 100 students:

- 40 have taken Algebra (Event A)
- 30 have taken Geometry (Event B)
- 15 have taken both Algebra and Geometry

What is the probability that a randomly selected student has taken **Algebra or Geometry**?

Using the addition rule:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \tag{4.3.3}$$

$$= \frac{40}{100} + \frac{30}{100} - \frac{15}{100} = \frac{55}{100} = 0.55 \tag{4.3.4}$$

Answer 55% of students have taken Algebra or Geometry.

We can analyze this piece by piece too. There are 40 students who have taken Algebra, but 15 out of that forty who have taken Geometry too. That leaves 25 who have only taken Algebra. Looking at the 30 students with Geometry, there are 15 who have also taken Algebra and 15 who haven't. This gives us now three disjoint groups. 25 students who have taken Algebra only, 15 who have taken Geometry only, and 15 who have taken both. You can check that adding up these values gives us 55 students out of the 100. This is an effective technique, but more time consuming and conceptual than the addition formula.

Example 2: Mutually Exclusive Events

When rolling a fair die:

- Let A = "roll a 1" → $P(A) = \frac{1}{6}$
- Let B = "roll a 5" → $P(B) = \frac{1}{6}$

A and B are mutually exclusive (you can't roll both at once), so:

$$P(A \cup B) = P(A) + P(B) = \frac{1}{6} + \frac{1}{6} = \frac{2}{6} \tag{4.3.5}$$

Answer Probability of rolling a 1 or a 5 is $\frac{2}{6} = 0.333$

Try These: Addition Rule

Question 1: In a class of 50 students:

- 30 like movies (Event A)
- 20 like sports (Event B)
- 10 like both

What is the probability a student likes movies or sports?
Select one:

— Select an answer —

Question 2: You flip a fair coin and roll a die:

- A = "get heads" → $P = 0.5$
- B = "roll a 6" → $P = \frac{1}{6}$

What's the probability of getting heads or rolling a 6?
Select one:

— Select an answer —

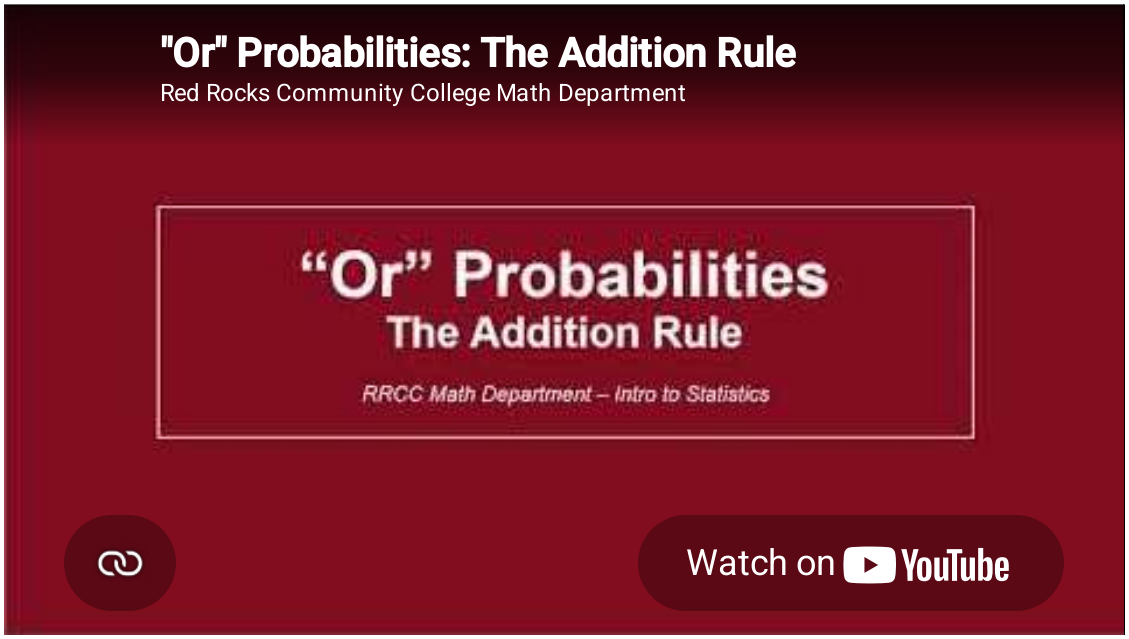
Question 3: Rolling a die, let:

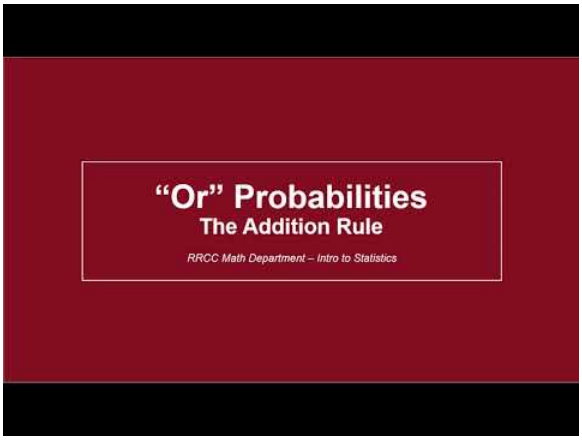
- A = "even number" → \{2, 4, 6\}
- B = "number less than 4" → \{1, 2, 3\}

They overlap on 2.
What is the probability of A or B?
Select one:

— Select an answer —

Related Video - The Addition Rule





Probability Rules So Far

- $P(A) \in [0,1]$
- $P(A^c) = 1 - P(A)$
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- If A and B are mutually exclusive, then $P(A \cap B) = 0$

Now that you've built a solid foundation with basic probability rules, you're ready to go further. In the next sections, we'll explore one of the most important ideas in statistics and probability: **independence**, or what it means for one event to have no effect on the chance of another. These concepts allow us to calculate more complex probabilities and model real-world scenarios more accurately.

This page titled 4.3: Foundational Probability Rules is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Mathematics Department.

4.4: Compound Events, Independence and Conditional Probability

So far, we've worked with probabilities where we are computing the likelihood of one event happening on its own. But in many real-world situations, we want to find the probability that **one event happens given that we already know another has occurred**. When one event influences or contains information about another, we call these **compound events**.

Understanding compound events leads us to a core concept in probability: **conditional probability**. This is the probability of one event happening *given that* another event has already occurred.

Definition: Conditional Probability

The **conditional probability** of event **A given B** is the probability that A occurs *assuming* that event B has already occurred. It is written as:

$$P(A \mid B) \tag{4.4.1}$$

The conditional probability can be computed with the following formula; though, in most cases, the formula is not necessary:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} \quad \text{(assuming } P(B) > 0) \tag{4.4.2}$$

Real-World Example: Flu Testing

Imagine that 20% of people in a population have the flu. A rapid test is administered. Among those who have the flu, 90% test positive. Among those who do *not* have the flu, 10% still test positive (false positive).

If someone tests positive, what is the probability that they actually have the flu? That's a **conditional probability**. It represents the interaction between:

- A = person has the flu
- B = person tests positive

We want to find: $P(A \mid B)$, which asks: "Among people who tested positive, how many actually have the flu?"

These types of questions appear everywhere from healthcare to quality control to legal trials, and accurate answers often depend on understanding if and how **events relate to each other**.

Statistical Independence

Sometimes two events have no influence on each other. In this case, knowing one occurs doesn't change the probability of the other. We call this **statistical independence**.

Definition: Independent Events

Events A and B are **independent** if:

$$P(A \mid B) = P(A) \tag{4.4.3}$$

$$\text{or equivalently:} \quad P(A \cap B) = P(A) \cdot P(B) \tag{4.4.4}$$

In other words: knowing that B happens does *not* change the probability of A happening.

Intuitively you can ask "**Does one event happening impact the probability that another event happens?**" If no, the events are independent.

Check Your Understanding: Dependent or Independent?

For each of the scenarios below, decide whether the two events are **independent** or **dependent**. Select your answer and check the feedback!

Question 1: You randomly select a marble from a bag, and then (without putting it back) select a second marble.
Are these events dependent or independent?

Select one:

— Select an answer —

Question 2: You roll a die, and then flip a coin.
Are these events dependent or independent?

Select one:

— Select an answer —

Question 3: You draw a card from a deck, note its suit, and then put it back. Later, you draw another card.
Are these events dependent or independent?

Select one:

— Select an answer —

Question 4: You take a reading quiz, and then use your score to decide whether to go to tutoring.
Are these events dependent or independent?

Select one:

— Select an answer —

Dice Example: Dependent Events

Suppose you roll two standard 6-sided dice (let's call them Die 1 and Die 2). Each die will have a **sample space** of: {1, 2, 3, 4, 5, 6}. Let's consider the two possible events below:

- Event A: "The sum of the two dice is 10"
- Event B: "The first die shows a 4"

To determine if these events are independent we can compare if $P(A \mid B) = P(A)$. You may already intuitively see if this true.

Let's first find $P(A)$. This is a little more tricky than finding the probability associated with a single die. There are actually 36 unique outcomes from the two dice rolls, using the **multiplication rule** from counting. There are exactly three ways for the sum of the two rolls to total 10. We can either have the first die roll a 4 and the second a 6, which we can denote as $(4, 6)$, we can have both dice roll 5, denoted as $(5, 5)$, or we can have the first a 6 and the second a 4 denoted as $(6, 4)$. Note here we do care about order, as we are interested in the specifics of the first roll. Since there are three different ways to sum to 10, we have $P(A) = 3/36 = 1/12$.

Let's next find $P(A \mid B)$. This is asking the probability of have the dice sum to 10 assuming that the first die was a 4. If we make the assumption that die 1 rolled a 4, then there is only a single possible roll of die 2 that can lead to a sum of 10: the second die *must* be a 6. This is the scenario seen above $(4, 6)$. Since there is a 1 in 6 chance of this happening with the second die, the probability is 1/6.

Since $P(A) \neq P(A \mid B)$, these events are **not independent**! Event A is influenced by knowing if B has occurred or not!

While we have the dice out, let's investigate another question. Is $P(A \mid B) = P(B \mid A)$? Let's calculate the second conditional probability below.

Assume that we know A has occurred. This means we know the two dice have been rolled and summed to 10. That means there are only three possibilities as listed above: $(4, 6)$, $(5, 5)$ and $(6, 4)$. Recall, these are the only possibilities because we are taking event A

as know to have happened! If we look at these three possibilities, only one of them fulfills event B. Therefore $P(B \mid A) = 1/3$. This is not equal to $P(A \mid B)$!

Summary

- **Conditional probability:** "What is the probability something happens *given* something else already did?"
  - $P(A|B) = P(A \text{ occuring assuming } B \text{ has already occurred})$
- **Independence:** One event having occurred does not influence the probability of a different event occurring.
  - If independent, $P(A \mid B) = P(A)$ and $P(B \mid A) = P(B)$
- **Multiplication rule:**

$$P(A \cap B) = P(A \mid B) \cdot P(B) = P(B \mid A) \cdot P(A) \tag{4.4.5}$$

  which becomes $P(A) \cdot P(B)$ *only if A and B are independent.*

Summary of Probability Rules (So Far)

Here's a growing list of the foundational probability rules and relationships we've learned so far. You'll continue to use these rules throughout your study of probability and statistics.

1. **Range of Probability:**
   The probability of any event A is a number between 0 and 1:

$$0 \leq P(A) \leq 1 \tag{4.4.6}$$

2. **Complement Rule:**
   The probability that event A does not occur is:

$$P(A^c) = 1 - P(A) \tag{4.4.7}$$

3. **Addition Rule (General):**
   For any two events A and B:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \tag{4.4.8}$$

4. **Addition Rule (Mutually Exclusive Events):**
   If A and B are mutually exclusive (disjoint):

$$P(A \cup B) = P(A) + P(B) \tag{4.4.9}$$

5. **Conditional Probability:**
   The probability of A given B:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} \quad (P(B) > 0) \tag{4.4.10}$$

6. **Multiplication Rule (General):**
   For any two events A and B:

$$P(A \cap B) = P(A) \cdot P(B \mid A) \tag{4.4.11}$$

7. **Multiplication Rule (Independent Events):**
   If A and B are independent:

$$P(A \cap B) = P(A) \cdot P(B) \tag{4.4.12}$$

8. **Definition of Independence:**
   Events A and B are independent if:

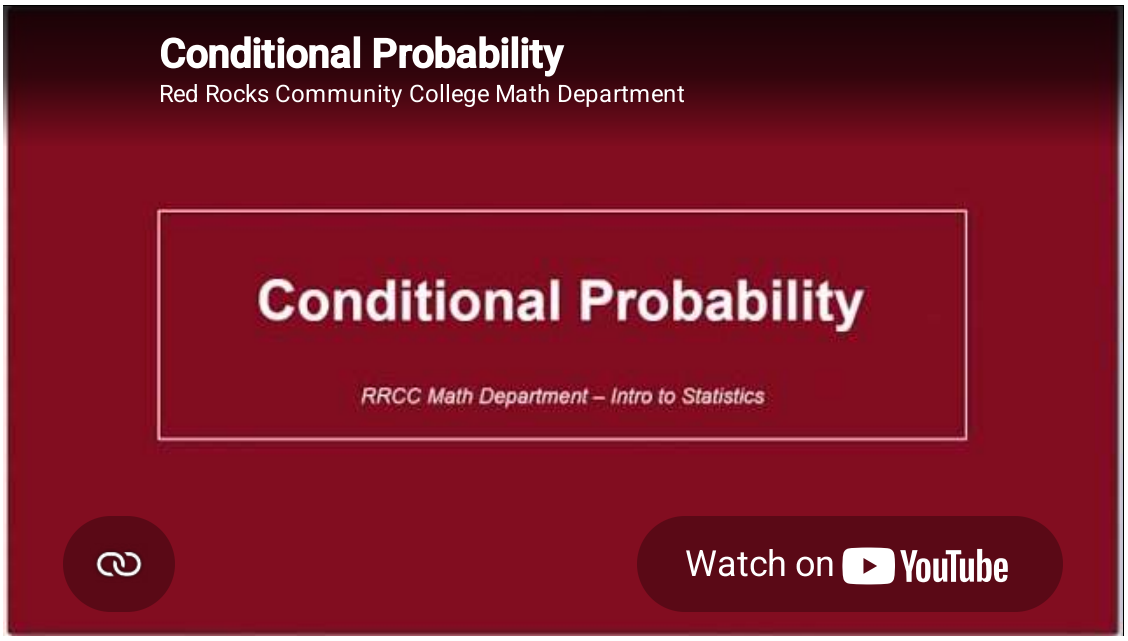
$$P(A \mid B) = P(A) \quad \text{or} \quad P(B \mid A) = P(B) \tag{4.4.13}$$

Tip: Keep this list handy! You'll use these rules to simplify and solve problems involving multiple steps, conditional outcomes, and dependent/independent scenarios.

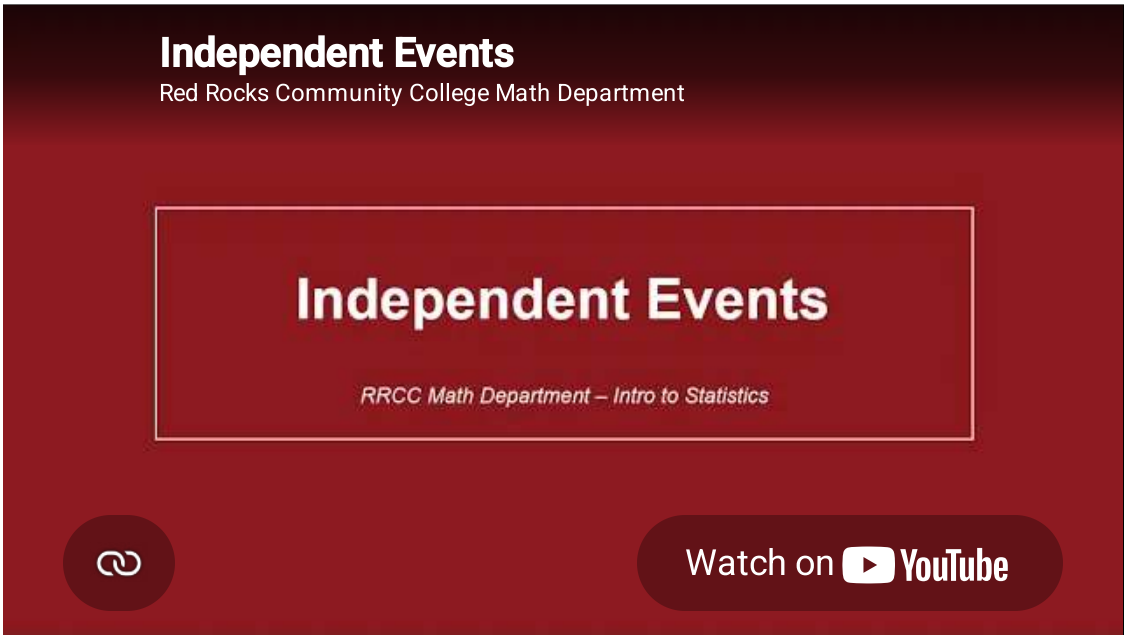
Think About It

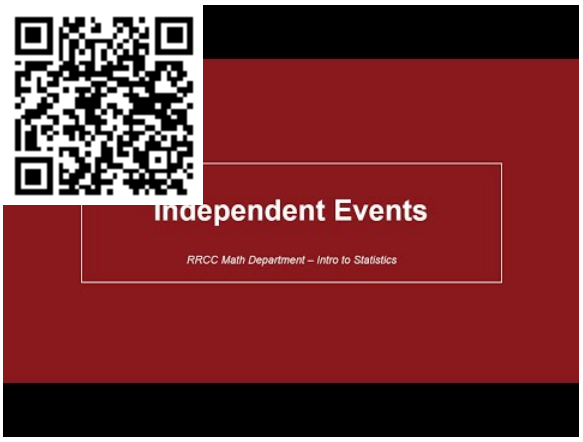
Can you think of two events from your life or your studies that would be considered dependent? What about two that are truly independent?

Related Videos






Independent Events
Red Rocks Community College Math Department
Independent Events
RRCC Math Department – Intro to Statistics
Watch on YouTube


Independent Events
RRCC Math Department – Intro to Statistics


"And" Probabilities: The Multiplication Rule
Red Rocks Community College Math Department
"And" Probabilities
The Multiplication Rule
RRCC Math Department – Intro to Statistics
Watch on YouTube



This page titled 4.4: Compound Events, Independence and Conditional Probability is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Mathematics Department.

4.5: Vocabulary (Chapter 4)

Chapter 4 Vocabulary

This list includes key terms introduced throughout Chapter 4. These terms form the foundation of probability, helping us describe, calculate, and interpret randomness in structured ways.

Probability

The long-run relative frequency of an event. Describes the proportion of times an event would occur if a random experiment were repeated many times under the same conditions. Values range from 0 to 1.

Relative Frequency

The proportion of times an event occurs out of the total number of trials in a simulation or experiment. Approximates probability in practice.

Experiment

An activity or process with an observable result. In probability, experiments are repeatable under identical conditions but have uncertain outcomes.

Sample Space

The set of all possible outcomes of a random experiment. Usually denoted by the symbol S or Ω.

Event

A set of outcomes (subset of the sample space) that satisfies a particular condition. Events can consist of one or more outcomes.

Complement

The set of all outcomes in the sample space that are *not* part of a given event. Denoted A^c. Represented as "not A".

Mutually Exclusive Events

Two events that cannot occur at the same time. The intersection of mutually exclusive events is the empty set; $P(A \cap B) = 0$.

Addition Rule

A formula to find the probability that at least one of two events occurs:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \tag{4.5.1}$$

When A and B are mutually exclusive:

$$P(A \cup B) = P(A) + P(B) \tag{4.5.2}$$

Conditional Probability

The probability that an event A occurs, given that another event B has occurred. Notated as $P(A \mid B)$ and calculated as:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} \quad \text{provided } P(B) > 0 \tag{4.5.3}$$

Independent Events

Two events A and B are independent if knowing that one occurred does not affect the probability of the other. Mathematically:

$$P(A \cap B) = P(A) \cdot P(B) \quad \text{or equivalently} \quad P(A \mid B) = P(A) \tag{4.5.4}$$

Dependent Events

Events for which the outcome or occurrence of one affects the probability of the other. For dependent events, $P(A \mid B) \neq P(A)$.

Counting Rule (Multiplication Rule of Counting)

If one task can be done in m ways and another in n ways, then the total number of ways to complete both tasks is $m \times n$.

Factorial

The product of all positive integers less than or equal to a given number n. Denoted $n!$. For example, $4! = 4 \times 3 \times 2 \times 1 = 24$.

Permutation

An arrangement of items where order matters. The number of ways to arrange r items from n is:

$$P(n,r) = \frac{n!}{(n-r)!} \tag{4.5.5}$$

Combination

A selection of items where order does *not* matter. The number of combinations of r items chosen from n is:

$$C(n,r) = \binom{n}{r} = \frac{n!}{r!(n-r)!} \tag{4.5.6}$$

Tip: Use these definitions as a reference while solving probability problems. Many of these ideas will return again when we explore inference, decision-making, and probability distributions later in the course.

This page titled 4.5: Vocabulary (Chapter 4) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Mathematics Department.

4.6: Practice (Chapter 4)

4.1: What is Probability?

1. Define probability in your own words.
2. Why do we refer to probability as a “long-run” relative frequency?
3. What does the Law of Large Numbers say about long-run probabilities as the number of trials increases?
4. Why is randomness important in interpreting statistical results?
5. Describe a real-life context where probability helps people make better decisions.
6. For the following problems consider flipping a fair coin with heads on one side and tails on the other.
    a. What is the theoretical probability of the coin landing heads up?
    b. Flip a fair coin 20 times and record your results. What proportion of the flips were heads?
    c. Simulate 100 coin flips (you can use an applet or spreadsheet). How did the proportion of heads compare to the theoretical probability?
    d. If a fair coin is tossed many times and the last eight tosses are all heads, then is the chance that the next toss will be heads is somewhat less than 50%? Explain
    e. How many times could we get the same number from rolling a die before we start to think it is unfair?
7. For the following problems consider rolling a fair six sided die numbered one through six.
    a. What is the probability of rolling a 5?
    b. What is the probability of rolling an even number?
    c. What is the probability of rolling a prime number?
    d. Roll a die 20 times and record your results. What proportion of the rolls were 5's?
    e. Simulate 100 die rolls (you can use an applet or spreadsheet). How did the proportion of 5's compare to the theoretical probability?
    f. If a fair die is rolled many times and the last ten rolls are all 5's, then is the chance that the next roll will be a 5 somewhat less than 16.6%? Explain
8. For the following problems consider drawing a card from a standard deck of playing cards.
    a. What is the theoretical probability of drawing a red card?
    b. What is the theoretical probability of drawing a club?
    c. What is the theoretical probability of drawing a face card (jack, queen, or king)?
    d. How does shuffling the deck of cards affect the probability of drawing a specific card?
9. Explain what is wrong with the following statements. Use complete sentences.
    a. If there is a 60% chance of rain on Saturday and a 70% chance of rain on Sunday, then there is a 130% chance of rain over the weekend.
    b. The probability that a baseball player hits a home run is greater than the probability that he gets a successful hit.

4.2: Introduction to Counting

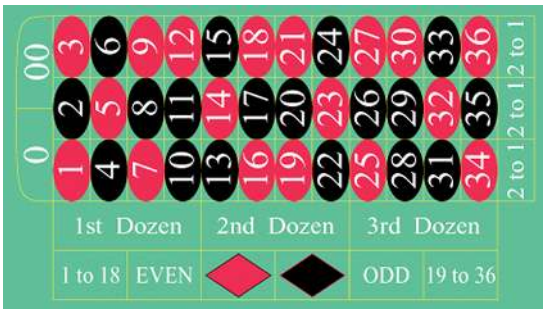
1. How many different outcomes are there for flipping a coin 23 times?
2. How many different outcomes are there for rolling a die 100 times?
3. How many different outcomes are there for drawing five cards from a shuffled deck of playing cards?
4. 5 different color dice are rolled, and the numbers showing are recorded. How many different outcomes are possible?
5. You’re choosing an outfit from 3 shirts, 2 pants, and 2 pairs of shoes. How many different outfits are possible?
6. Suppose a designer has a palette of 6 colors to work with, and wants to design a flag with 4 vertical stripes, all of different colors. How many different flags can be designed?
7. A restaurant offers 5 appetizers and 4 entrees. If you order one of each, how many meal combinations are available?
8. Taco del Tierra sells only tacos. They offer four types of tortillas, two types of beans, four types of vegetables, and four types of salsa. If you must choose exactly one of each for your taco, how many different kinds of tacos can you order at Taco del Tierra?
9. Compute 5!
10. Write the following in factorial notation 1x2x3x...x22x23

11. Simplify and then compute 5!/3!
12. Simplify and then compute 11!/(9!2!)
13. Write the following using factorial notation 7x8x9X10
14. The Colorado Lottery Mega Millions has players choose five numbers from 1 to 70 and one number from 1 to 24. How many different lottery tickets are there?
15. How many 3-letter codes can be made from A–Z with no repetition?
16. How many different ways can the 7 main Harry Potter books be arranged on a shelf?
17. The standard Colorado license plate includes seven characters. These can be: letters (A-Z), digits (1-9), blank space, dash, and period. How many different CO license plates are there?
18. In how many ways can first, second, and third prizes be awarded in a contest with 590 contestants?
19. A committee consists of 13 people. How many different subcommittees of four people can be formed?
20. You are selecting 3 students from a class of 10 to represent the class. How many different combinations of students are possible?
21. Joni is reorganizing their video game collection and is having trouble determining how to arrange them on the shelf. The collection consists of 39 games. Which method of counting should be used to determine the number of different ways to arrange the games?
22. A vending machine has 8 different soft drink options. How many pairs of drinks can you buy?
23. Explain the difference between permutations and combinations. Give an example of when each would be appropriate.

4.3: Foundational Probability Rules

1. An experiment consists of first rolling a die and then tossing a coin.
    a. List the sample space.
    b. List the outcomes for the event described by rolling a number greater than 4
2. An experiment consists of measuring the weight of a dog in a shelter. Describe the sample space.
3. What is the event set for rolling a pair of dice and getting the same number for both?
4. Two Queen cards are dealt from a standard deck of playing cards. Write out the event set.
5. What is the relationship between experiments and events?
6. What is the relationship between outcomes and sample space?
7. When is the sample space the same as the event space?
8. There are 23 countries in North America, 12 countries in South America, 47 countries in Europe, 44 countries in Asia, 54 countries in Africa, and 14 in Oceania (Pacific Ocean region). Let A = the event that a country is in Asia, let E = the event that a country is in Europe, let F = the event that a country is in Africa, let N = the event that a country is in North America, let O = the event that a country is in Oceania, and let S = the event that a country is in South America.
    a. Find $P(A)$
    b. Find $P(E)$
    c. Find $P(F)$
    d. Find $P(N)$
    e. Find $P(O)$
    f. Find $P(S)$
9. A special deck of cards has ten cards. Four are green, three are blue, and three are red. When a card is picked, its color of it is recorded. An experiment consists of first picking a card and then tossing a coin.
    a. List the sample space.
    b. Let A be the event that a blue card is picked first, followed by landing a head on the coin toss. Find P(A).
    c. Let B be the event that a red or green is picked, followed by landing a head on the coin toss. Are the events A and B mutually exclusive? Explain your answer in one to three complete sentences, including numerical justification.
    d. Let C be the event that a red or blue is picked, followed by landing a head on the coin toss. Are the events A and C mutually exclusive? Explain your answer in one to three complete sentences, including numerical justification.
10. When rolling a pair of dice, is getting a total of 7 mutually exclusive with rolling a pair of the same number? Explain
11. Are rain and sunny weather mutually exclusive? Explain

12. Are drawing a face card (jack, queen, or king) and drawing a red card from a full deck of playing cards mutually exclusive events? Explain
13. Are drawing a face card and drawing an ace from a full deck of playing cards mutually exclusive events? Explain
14. What does the complement rule say about $P(A^C)$? Use it to calculate the probability that a randomly selected person does not own a pet if 72% of people do.
15. Make a statement in ordinary English that describes the complement of each event (do not simply insert the word "not").
    a. In the roll of a die: "five or more."
    b. In a roll of a die: "an even number."
    c. In two tosses of a coin: "at least one heads."
    d. In the random selection of a college student: "Not a first-year student."
16. Use the complement rule to show that the probability of flipping a coin and **not** getting heads is 0.5.
17. Draw a Venn diagram showing two overlapping events A and B. Shade in the area that corresponds to the B^C.
18. Write the Addition Rule for any two events A and B.
19. If P(A) = 0.4, P(B) = 0.5, and A and B are mutually exclusive, what is $P(A \cup B)$?
20. Use the addition rule to show that the probability of flipping a coin and getting heads or tails is 1.
21. Compute the following probabilities for rolling a fair die:
    a. $P(5^C)$
    b. P(even or 2)
    c. P(prime or odd)
    d. P(<2 or >4)
22. Compute the following probabilities for drawing a card from a standard deck of playing cards:
    a. $P(\text{heart}^C)$
    b. P(red or king)
    c. P(ace or club)
    d. P(diamond or face card)
    e. P(black or spade)
    f. P(J or Q or K)
23. The casino game, roulette, allows the gambler to bet on the probability of a ball, which spins in the roulette wheel, landing on a particular color, number, or range of numbers. The table used to place bets contains of 38 numbers, and each number is assigned to a color and a range.



    a. List the sample space of the 38 possible outcomes in roulette.
    b. Find P(red)
    c. Find P(1st Dozen)
    d. Find P(even or 1 to 18)
    e. Find $P((0 \text{ or } 00)^C)$
    f. Is getting an odd number the complement of getting an even number? Why?
    g. Find two mutually exclusive events.
24. The graph shown is based on more than 170,000 interviews done by Gallup that took place from January through December 2012. The sample consists of employed Americans 18 years of age or older. The Emotional Health Index Scores are the sample

space. We randomly sample one Emotional Health Index Score.



Figure 3.3.1.

a. Find the probability that an Emotional Health Index Score is 82.7.
b. Find the probability that an Emotional Health Index Score is 81.0.
c. Find the probability that an Emotional Health Index Score is more than 81?
d. Find the probability that an Emotional Health Index Score is between 80.5 and 82?
e. If we know an Emotional Health Index Score is 81.5 or more, what is the probability that it is 82.7?
f. What is the probability that an Emotional Health Index Score is 80.7 or 82.7?
g. What is the probability that an Emotional Health Index Score is less than 80.2 given that it is already less than 81.
h. What occupation has the highest emotional index score?
i. What occupation has the lowest emotional index score?
j. What is the range of the data?
k. Compute the average EHIS.
l. If all occupations are equally likely for a certain individual, what is the probability that they will have an occupation with lower than average EHIS?

4.4: Compound Events, Independence, and Conditional Probability

1. What is a compound event? Provide one example from daily life.
2. Define conditional probability and explain how it differs from basic probability.
3. A die is rolled and then a coin is flipped. Are these events independent? Why?
4. Differentiate between dependent and independent events. Provide one example of each.
5. In your own words, explain the difference between mutually exclusive and independent events.
6. You randomly draw two marbles from a jar with 3 red and 2 blue marbles. Compute the probability that both are red if you do not replace the first marble.
7. Repeat the previous question, but assume you replace the first marble before drawing the second. How does the probability change?
8. Suppose you roll two standard 6-sided dice. Event A is the total of the two dice is 7. Event B is that the first die is even. Are events A and B independent?
9. When is drawing a second card from a deck independent of the first?
10. In a group of 100 people, 30 have the flu, and 24 of those tested positive. Write and calculate P(Test Positive | Has Flu).
11. A survey shows 60% of students have laptops. Of those, 40% also have tablets. What is the probability that a randomly selected student has both?
12. Suppose 80% of people like peanut butter, 89% like jelly, and 78% like both. Given that a randomly sampled person likes peanut butter, what's the probability that they also like jelly?
13. If P(A) = 0.6 and P(B) = 0.3, and A and B are independent, calculate P(A ∩ B).
14. Let P(A ∩ B) = 0.1 and P(B) = 0.2. Find P(A | B).
15. Using the formula for conditional probability, explain what each component of P(A | B) = P(A ∩ B)/P(B) represents in words.
16. If two events A and B are independent, how is conditional probability simplified?

17. Use a tree diagram or two-way table to organize outcomes for students who pass/fail a quiz based on whether they studied. Use hypothetical values and calculate P(Pass | Studied).
18. Suppose we roll two fair dice. Each die has six faces.
    a. List the sample space.
    b. Let A be the event that either a three or four is rolled first, followed by an even number. Find P(A).
    c. Let B be the event that the sum of the two rolls is at most seven. Find P(B).
    d. Find P(A ∩ B).
    e. In words, explain what P(A|B) represents. Find P(A|B).
    f. Are A and B mutually exclusive events? Explain your answer in one to three complete sentences, including numerical justification.
    g. Are A and B independent events? Explain your answer in one to three complete sentences, including numerical justification.
19. A 2012 Pew Research survey asked 2,373 randomly sampled registered voters their political affiliation (Republican, Democrat, or Independent) and whether or not they identify as swing voters. 35% of respondents identified as Independent, 23% identified as swing voters, and 11% identified as both.
    a. Are being Independent and being a swing voter disjoint, i.e. mutually exclusive?
    b. Draw a Venn diagram summarizing the variables and their associated probabilities.
    c. What percent of voters are Independent but not swing voters?
    d. What percent of voters are Independent or swing voters?
    e. What percent of voters are neither Independent nor swing voters?
    f. What percent of voters are Independent and swing voters?
    g. Is the event that someone is a swing voter independent of the event that someone is a political Independent?
20. The American Community Survey is an ongoing survey that provides data every year to give communities the current information they need to plan investments and services. The 2010 American Community Survey estimates that 14.6% of Americans live below the poverty line, 20.7% speak a language other that English at home, and 4.2% fall into both categories.
    a. Are living below the poverty line and speaking a language other than English at home disjoint?
    b. Draw a Venn diagram summarizing the variables and their associated probabilities.
    c. What percent of Americans live below the poverty line and only speak English at home?
    d. What percent of Americans live below the poverty line or speak a language other than English at home?
    e. What percent of Americans live above the poverty line and only speak English at home?
    f. Is the event that someone lives below the poverty line independent of the event that the person speaks a language other than English at home?
21. A genetic test is used to determine if people have a predisposition for thrombosis, which is the formation of a blood clot inside a blood vessel that obstructs the flow of blood through the circulatory system. It is believed that 3% of people actually have this predisposition. The genetic test is 99% accurate if a person actually has the predisposition, meaning that the probability of a positive test result when a person actually has the predisposition is 0.99. The test is 98% accurate if a person does not have the predisposition. What is the probability that a randomly selected person who tests positive for the predisposition by the test actually has the predisposition?

Practice problems include parts of:

Introductory Statistics 2e by OpenStax is licensed under CC BY 4.0

Access for free at https://openstax.org/books/introductory-statistics-2e/pages/1-introduction

and

OpenIntro Statistics by David Diez, Christopher Barr, & Mine Çetinkaya-Rundel is licensed under CC BY 3.0

4.6: Practice (Chapter 4) is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

- **Current page** has no license indicated.
- **2.E: Probability (Exercises)** by David Diez, Christopher Barr, & Mine Çetinkaya-Rundel has no license indicated. Original source: https://www.openintro.org/book/os.

- **Licensing** by Mathematics Department has no license indicated.

5: Probability Distributions

Have you ever considered the amount of time you wait for a bus on a given day? Or how much snow you will get in a storm? What about the mechanics of a game that involves rolling different types of dice? All of these are examples of randomness in our lives that exhibit certain patterns. We can predict these patterns using **probability distributions.**

Probability distributions are used to calculate insurance rates, determine lottery winnings, predict weather, and estimate delivery times. Companies and government agencies utilize massive data sets and hugely complicated mathematical models to do these, but we will take a look at simpler examples. We will encounter two different types of distributions:

- **Discrete distributions** – where outcomes are countable (like rolling dice or flipping coins)
- **Continuous distributions** – where outcomes are measured (like waiting times or test scores)

We'll explore commonly used distributions including: Bernoulli, binomial, and normal, and connect them to real-world scenarios. You'll learn how to calculate **expected values**, measure **variability**, and visualize how outcomes are shaped.

Most importantly, understanding these distributions will prepare us for one of the most powerful tools in statistics: the normal distribution and its role in inference.

Let's dive in and see how structured randomness can be.

This page titled 5: Probability Distributions is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Mathematics Department.

5.1: Random Variables- Discrete and Continuous

In earlier chapters, we explored types of data, in particular, **categorical vs. quantitative**, and **discrete vs. continuous** variables. We will connect these qualities of variables with notions of probability in this chapter.

Definition: Random Variable

A **random variable** is a quantitative variable whose value is determined by a random event.

We usually represent random variables with an uppercase letter like **X** or **Y**. For example:

- Let X = number of heads after flipping 3 coins
- Let Y = the waiting time (in minutes) until a bus arrives

The word “random” here just means that we don’t know what value the variable will take *before* we observe the outcome, it depends on chance. Other terminology you may encounter to describe this is **non-deterministic** or **stochastic**.

Some non-examples of random variables could include:

- A person's favorite food. Although we cannot predict this, it is not a quantitative variable.
- The freezing temperature of pure water at sea-level (0° C). This is a measurable and quantitative variable, but is not due to chance. It's a fixed value that is a property of nature. In fact the celsius temperature scale was originally based on this measurable temperature!
- Population parameters. For example, the average height of *every single person on earth*. This is a fixed value that we could theoretically calculate. However, it's feasibly impossible to do so, so any attempt to measure it ultimately results in requiring a sample, which introduces randomness.

When we work with random variables, we’re combining the power of probability with data. We’re no longer just talking about events like rolling a die, but about the *numerical outcome* we could observe from that process.

Two Types of Random Variables

Just like with data earlier, we can divide random variables into two major types:

Discrete Random Variables

Take on separate, distinct values (usually whole numbers).

- Examples:
  - X = number of students absent on a given day (0, 1, 2, …)
  - Y = number of correct answers on a quiz
  - Z = number of Atlantic hurricanes that make landfall in a season
- Usually represent things you can “count”

Continuous Random Variables

Can take on any value in an interval (there are infinite possibilities between any two values).

- Examples:

- X = the time (in seconds) it takes an athlete to run 100 meters
- Y = the height (in meters) of a randomly selected medical patient
- Z = the cost (in dollars) associated with a traffic accident

- Usually represent things you "measure"

Distributions

Every random variable has a certain set of outcomes it can take, and each of those outcomes comes with a probability. This is why we are interested in random variables. They allow for us to categorize and calculate probabilities associated with certain scenarios. We design a way to organize this information and call it a **probability distribution.**

Example

Suppose we flip two fair coins and count the number of heads. Let the random variable X denote this. There are 4 different outcomes to this experiment, listed below:

(H, H) : two heads

(H, T) : first coin heads, second coin tails (1 head)

(T, H) : first coin tails, second coin heads (1 head)

(T, T) : two tails (no heads)

If we imagine the possibilities for X, there are only 3. X can either be equal to 0, 1, or 2.

If we wanted to tabulate the probabilities of these occurrences, we only need count the number of events corresponding to each value that X can be. For example, there is only one way for X to be equal to 0. It must be the case that we flipped two tails, which has a 1 in 4 change. In contrast, there are two ways that X could end up equal to 1. Either (H, T) or (T, H). This gives us a 2 in 4 chance. The summary is given below using our notation from chapter 4. When we write $P(X = x)$, this means the probability that our random variable has taken on a value of x. If it is completely clear which random variable and context we are using, we might use the shorter notation $P(x)$.

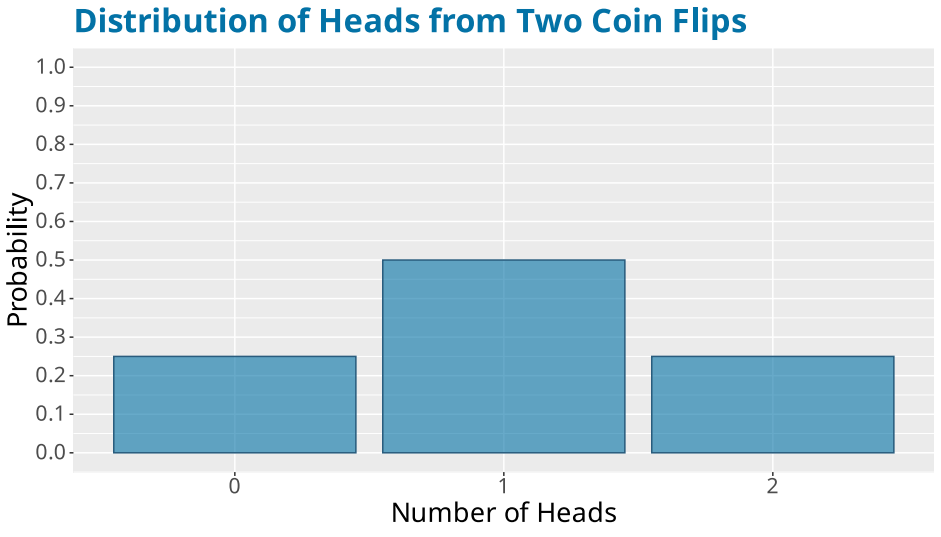
- $P(X = 0) = 0.25$
- $P(X = 1) = 0.5$
- $P(X = 2) = 0.25$

We can alternatively summarize this information in a table. We call this relationship that tells us the probability of each value the **Probability Mass Function** (PMF)

Probability Mass Function values

x	$P(X = x)$
0	0.25
1	0.5
2	0.25

Each column of the table stores the possible value for X in the top row, and its corresponding probability in the bottom row. Finally, we can also organize this information in a bar graph.



Check Your Understanding

Take a moment and sort the examples below as **Discrete** or **Continuous** random variables:

- **1.** Number of books you check out at the library
- **2.** The amount of rainfall in a day (in inches)
- **3.** Number of people who show up for your group meeting
- **4.** Temperature in degrees Fahrenheit at noon today

Discuss: Can we define a probability distribution for each of these? What would it look like?

Why This Matters

As we move forward into probability distributions, we'll focus on two types:

- **Discrete:** starting with the binomial distribution
- **Continuous:** using the normal distribution

These tools let us assign probabilities to outcomes not just with logic or counting, but with a defined distribution that follows real-world patterns. And that's powerful.

Project tie in

Choose one variable from your housing dataset. Could it be modeled as a random variable? If it's discrete, what values can it take on? If it's continuous, what units would you use?

Probability Distribution Definition

A **probability distribution** outlines how probabilities are distributed across all possible values of a random variable. If our random variable describes discrete data, we refer to its distribution as a **discrete probability distribution.** We saw an example of this above in the coin flips.

To be a valid discrete probability distribution, your table **must** meet these two rules:

1. Each probability must be between 0 and 1 (inclusive):
   → $0 \leq P(X = x) \leq 1$

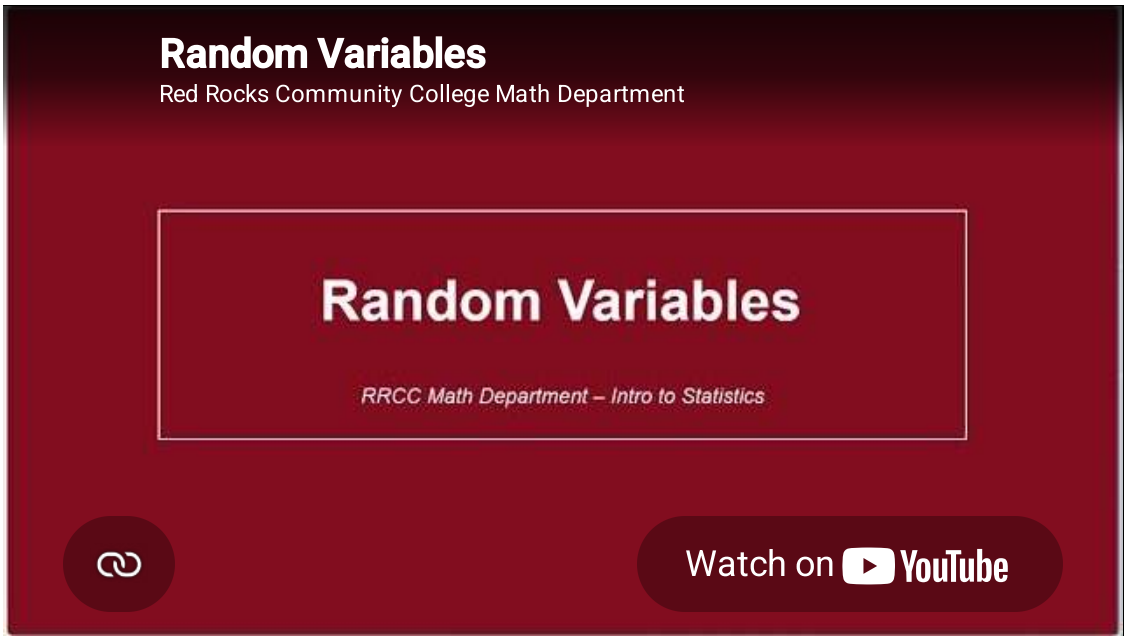
2. The sum of all probabilities must equal 1 (in the below example we have three values)
   → $P(X=0)+P(X=1)+P(X=2)=1$

Take a second to verify these rules check out for the coin flip example above.

Realistic Examples

We have only discussed a simple example that is calculated from basic probability. But in reality, when we want to investigate meaningful and important ideas, our distributions can be extremely complicated! For example, how could we build a distribution that tells us the probability of the number of patients that an emergency room sees on a given day? Or the amount of produce a grocery store sells and therefore needs to order? We will start small and then work our way up, but there are two techniques we will investigate. The first is how we can use probabilistic models (the binomal distribution) to create more complicated ideas. The second goes back to our topic in the beginning of the course, that we can utilize descriptive statistics and data to inform probabilities.

Related Video





This page titled 5.1: Random Variables- Discrete and Continuous is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Mathematics Department.

5.2: Expected Value of Discrete Distributions

In this section, we begin putting probability distributions to work in a new way by calculating what we expect to happen on average. The **expected value** tells us about the *typical* value or the long term average of repeating an experiment leading to the random variable.

Definition: Expected Value (Mean)

The **expected value** of a discrete random variable X, denoted $E(X)$ or μ, is the theoretical average outcome of the distribution.

$$E(X) = \sum x \cdot P(x) \tag{5.2.1}$$

In words: multiply each value x by its probability, then add them all up.

Example: Lottery Scratch Ticket

A scratch ticket game costs \$5 to play. The possible prizes and their probabilities are shown below. To analyze the game, we'll compute the **net gain** (prize minus the cost to play).

Calculating expected value

Prize	Net Gain (x)	Probability P(x)	$x \cdot P(x)$
\$0	–5	0.849	–4.245
\$5	0	0.08	0.00
\$20	15	0.05	0.75
\$100	95	0.02	1.90
\$500	495	0.001	0.495
		Total Expected Value:	**-1.10**

The value of -1.10 is our expected value, and takes the same units as the random variable. In this case it represents a loss of \$1.10. It is impossible for any individual player to lose this amount, but if a large number of individuals buy tickets, and we calculate the average of all their gains and losses, the average will be -1.10. This value is negative because the majority of players lose a little money, whereas very few players will win a considerable amount. The fact that the expected value is negative is important for the success of the lottery; if the expected value was positive, then the lottery would on average pay out money and lose their funds. With a negative expected value, the lottery will bring in more money than they pay out, which keeps the lottery running and may be used to raise funds for public services such as education or conservation.

Measuring Variation: Variance and Standard Deviation

The expected value, or mean, describes the center of the data, but it doesn't tell us how much values vary around that center. For that, we need to revisit measures of spread.

The **variance** of a discrete random variable X, denoted σ^2, measures the average squared distance from the mean:

$$\text{Var}(X) = \sigma^2 = \sum (x - \mu)^2 \cdot P(x) \tag{5.2.2}$$

The **standard deviation** σ is the square root of the variance. It measures typical distance from the mean, in the same units as the data:

$$\sigma = \sqrt{\sum (x - \mu)^2 \cdot P(x)} \tag{5.2.3}$$

The standard deviation gives us a measure of how the data, on average, differs from the mean. Let's revisit our lottery example to calculate its variance and standard deviation. We found before that $\mu = -1.10$. We can build a table to compute the variance in a similar manner as we did the expected value; you can verify these results yourself using a spreadsheet tool such as Microsoft Excel or Google Sheets.

Variance calculations

Prize	Net Gain (x)	Probability P(X = x)	x - μ	$(x - \mu)^2$	$(x - \mu)^2 \cdot P(x)$
$0	–5	0.849	-3.9	15.21	12.91329
$5	0	0.08	1.1	1.21	0.09605
$20	15	0.05	16.1	259.21	12.9605
$100	95	0.02	96.1	9235.21	184.7042
$500	495	0.001	496.1	246115.2	246.11521
				Variance:	**456.79**

Taking the square root of this value for the variance, we obtain $\sigma = \sqrt{456.79} \approx 21.37$. The larger this value of σ, the more spread out the data is. In this case, our data is relatively spread out!

Example: Rolling a Loaded Die

Suppose a six-sided die has been tampered with so that it produces the following outcomes:

Distribution of a loaded die

Outcome (x)	Probability P(x)
1	0.10
2	0.15
3	0.20
4	0.20
5	0.25
6	0.10

Step 0: Verify this is a random variable.

- All probabilities are between 0 and 1
- All probabilities add to 1

Step 1: Compute the expected value

$$E(X) = 1(0.10) + 2(0.15) + 3(0.20) + 4(0.20) + 5(0.25) + 6(0.10) = 3.65 \tag{5.2.4}$$

Step 2: Compute the variance

$$\sigma^2 = \sum (x - 3.65)^2 \cdot P(x) = (1 - 3.65)^2(0.10) + \cdots + (6 - 3.65)^2(0.10) \tag{5.2.5}$$

You can use the formula, but it is easier to keep track of with a table. These values were actually calculated in Microsoft Excel

Calculating variance

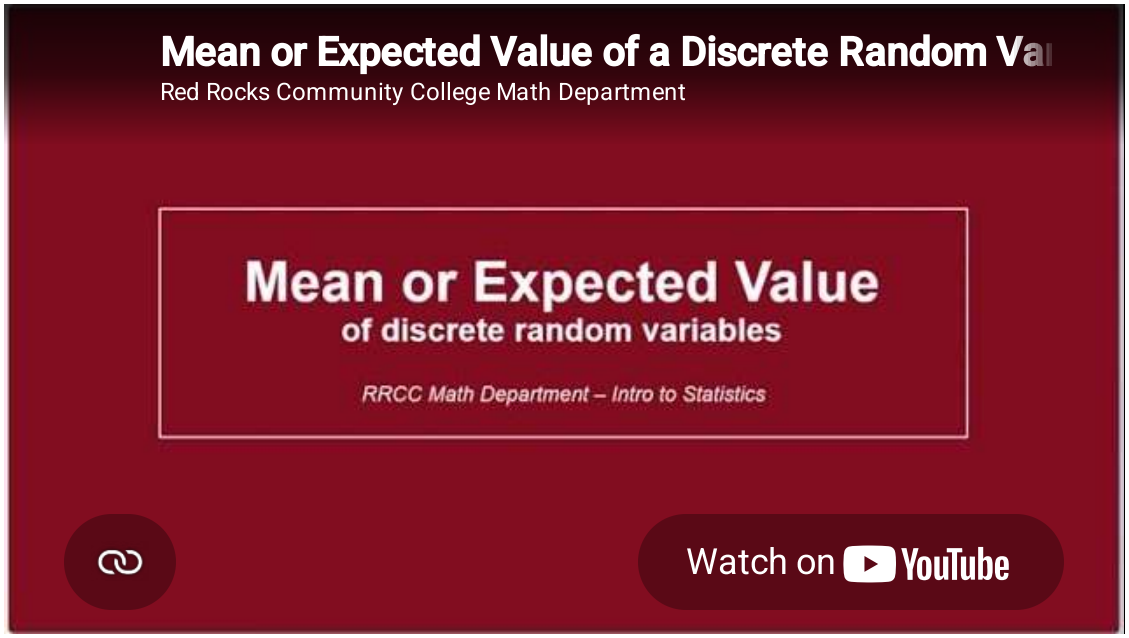
x	P(x)	(x – μ)	(x – μ)²	(x – μ)² · P(x)
1	0.10	–2.65	7.0225	0.70225
2	0.15	–1.65	2.7225	0.408375
3	0.20	–0.65	0.4225	0.0845
4	0.20	0.35	0.1225	0.0245
5	0.25	1.35	1.8225	0.455625
6	0.10	2.35	5.5225	0.55225
			Total Variance:	**2.2275**

Standard deviation: $\sigma = \sqrt{2.2275} \approx 1.4925$

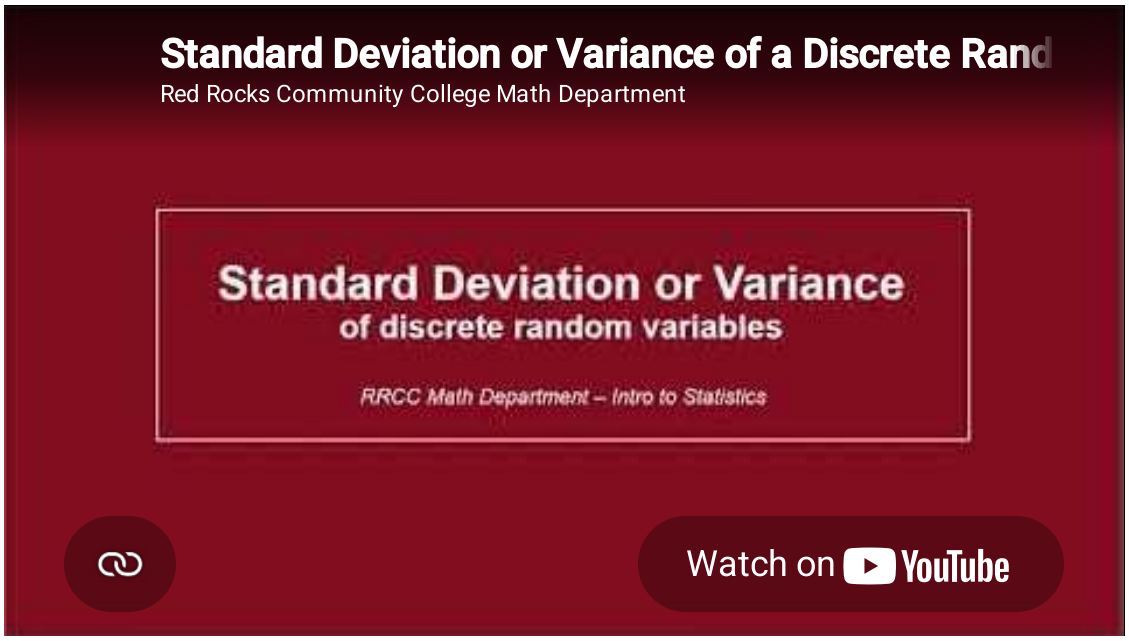
Summary

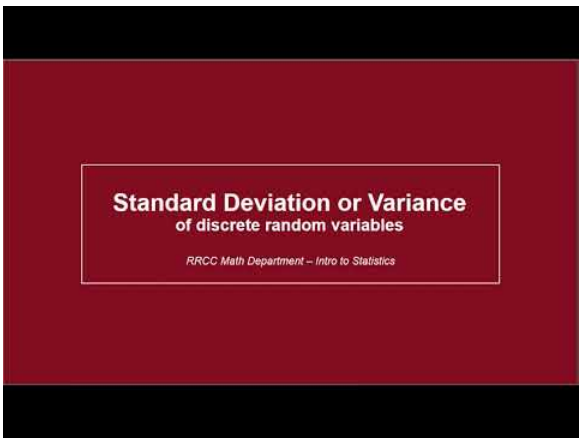
- $E(X)$ tells us what to expect on average (long term).
- $\mathrm{Var}(X)$ and σ tell us how spread out the values are around the mean.
- Just like with data summaries: mean for center, standard deviation for spread!

Related Video









This page titled 5.2: Expected Value of Discrete Distributions is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Mathematics Department.

5.3: Discrete Distributions- Bernoulli and Binomial

The Sock Drawer Scenario

You're late to class and grab a sock, just one, from a messy drawer. The drawer has a mix of black and white socks, but you don't know the exact count.

> *"Is this sock black?"*

That's a yes/no question. If you define **"black sock = success"**, then pulling a sock is a simple **Bernoulli trial**.

There are just two outcomes:

- **Success** (it's black), with some probability p
- **Failure** (not black), with probability $1-p$

Whether you realize it or not, this super simple setup — one random trial, two possible outcomes — is the starting point for some of the most powerful models in statistics. And it all starts with a sock.

Let's break down what the Bernoulli distribution really is...

Definition: Bernoulli Distribution

The **Bernoulli Distribution**, named after Jacob Bernoulli, is the simplest kind of *discrete probability distribution*. It models the outcomes of a single experiment (called a **Bernoulli trial**) that has only two possible outcomes:

- **Success**, with probability p
- **Failure**, with probability $1-p$

Note that success and failure do not have to carry any specific meanings. We often use terminology such as "**yes, true, success**" versus "**no, false, failure**", but it could also be the difference between heads and tails of a coin flip or any other binary. We may also use the term 'success' to denote something with a negative connotation, such as developing side-effects from a medication, or an extreme weather event successfully materializing.

If we define a random variable X such that:

- $X = 1$ if the outcome is a success
- $X = 0$ if the outcome is a failure

Then we say:

$$X \sim \text{Bernoulli}(p)$$

For the Bernoulli distribution, we can find a formula for the probability mass function (PMF). Hopefully this illustrates the reason why we have different terminology. The distribution represents the overall distributions of probabilities and can tell us about complicated events for multiple outcomes, whereas the PMF gives specifically the probability of a single outcome. The PMF of the Bernoulli distribution can be summarized in the following way:

$$\begin{aligned} P(X=1) &= p \\ P(X=0) &= 1-p \end{aligned} \tag{5.3.1}$$

For more complicated distributions we will need an algebraic expression. For the Bernoulli, it is given by:

$$P(X=x) = p^x(1-p)^{1-x} \text{ for } x=0 \text{ or } x=1$$

Where:

- p = probability of success (a number between 0 and 1)

- X takes on only two values: 0 or 1

If $X = 1$ then $P(X = 1) = p^1 \cdot (1-p)^{1-1} = p$ and if $X = 0$, then $P(X = 0) = p^0 \cdot (1-p)^{1-0} = 1-p$

The expected value (mean) of a Bernoulli random variable is:

$$\mu = E(X) = p$$

Example: Guessing a Multiple Choice Question

Suppose that you take a guess on a multiple choice test question with 4 answers. There is a 1 in 4 chance of guessing the correct answer. This can be considered a Bernoulli trial with $p = 0.25$. Furthermore, in this scenario, X also represents the points allotted for this question. You get 1 point if you guess correctly, and 0 if you do not.

Guessing on a multiple choice

Outcome	Random Variable X	$P(X)$
Guess incorrect	0	0.75
Guess correct	1	0.25

There is not a lot of investigation that can be performed with the Bernoulli distribution. Instead we move on to the binomial distribution, which asks the question about *a series* of Bernoulli trials.

The Binomial Distribution

Definition: Binomial Distribution

The **Binomial Distribution** is a *discrete probability distribution* that models the number of **successes** in a fixed number of repeated, independent Bernoulli trials, where each trial has only two possible outcomes: **success** or **failure**.

You can use the Binomial distribution when:

- The number of trials, n, is fixed.
- Each trial is **independent** of the others (the results of one trial do not affect subsequent trials)
- Each trial results in either **success** or **failure** (a Bernoulli trial).
- The probability of success, p, is the same for every trial.

We sometimes call these trials IID or *independent and identically distributed.*

If a random variable X counts the number of successes, we write:

$$X \sim \text{Binomial}(n, p)$$

The probability of getting exactly x successes in n trials is given by the **Binomial probability formula**:

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$$

Where:

- n = number of trials
- x = number of successes
- p = probability of success on a single trial
- $\binom{n}{x}$ = number of ways to choose x successes from n trials (combination from chapter 4)

After seeing the PMF, the expected value of a Binomial distribution is surprisingly simple! It is just given by the formula:

$$\mu = n \cdot p \tag{5.3.2}$$

We can also compute the standard deviation of a Binomial random variable with:

$$\sigma = \sqrt{n \cdot p \cdot (1-p)} \tag{5.3.3}$$

Example: Multiple Choice Test

Let's revisit a multiple choice test, this time with 5 questions each with 4 possible choices. Since there is a 1 in 4 chance of guessing each question correct, we have:

- $p = 0.25$
- $n = 5$
- X: number of correctly guessed answers out of 5.

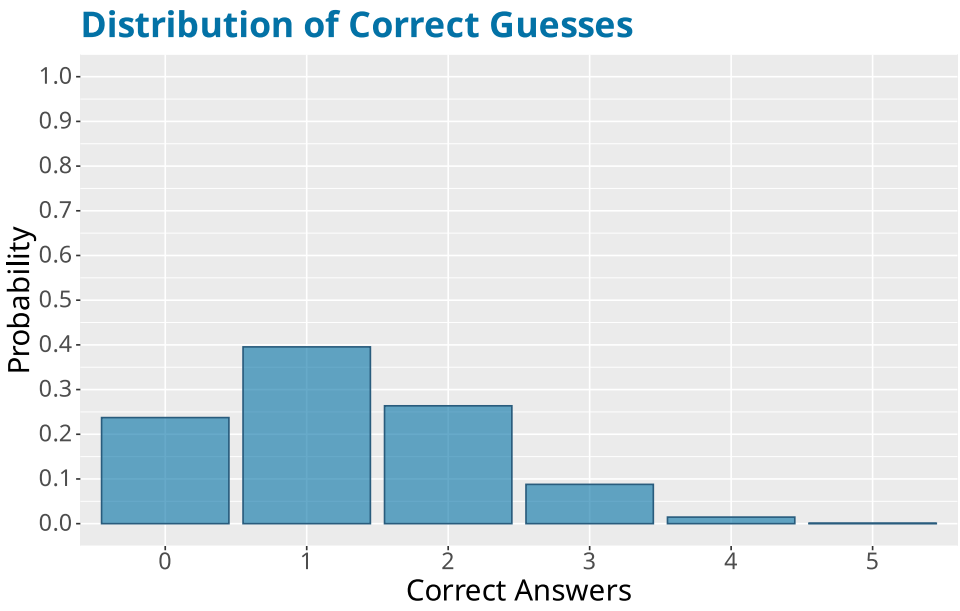
We can use the PMF for the Binomial distribution to calculate the probabilities of getting a certain number of correct questions on the test. The following values were actually calculated with Excel, using the formula: **=BINOM.DIST(x, n, p, FALSE)**. The "FALSE" input chooses the non-cumulative probability function in Excel. If you used "TRUE", we get the probability of x *or fewer* successes. Ex: If n=10 and p=0.3, P(x=3)=BINOM.DIST(3, 10, 0.3, FALSE) and P(x≤ 3)=BINOM.DIST(3, 10, 0.3, TRUE). We only need the "FALSE" version for creating a probability distribution table like below.

Guessing on a multiple choice test

Number of Correct Answers (X)	Probability $P(X)$
0	$P(X=0) = \binom{5}{0}(0.25)^0(0.75)^5 = 1 \cdot 1 \cdot 0.2373 = \mathbf{0.23731}$ *or* BINOM.DIST(0, 5, 0.25, FALSE)
1	$P(X=1) = \binom{5}{1}(0.25)^1(0.75)^4 = 5 \cdot 0.25 \cdot 0.3164 = \mathbf{0.39551}$ *or* BINOM.DIST(1, 5, 0.25, FALSE)
2	$P(X=2) = \binom{5}{2}(0.25)^2(0.75)^3 = 10 \cdot 0.0625 \cdot 0.4219 = \mathbf{0.26367}$ *or* BINOM.DIST(2, 5, 0.25, FALSE)
3	$P(X=3) = \binom{5}{3}(0.25)^3(0.75)^2 = 10 \cdot 0.0156 \cdot 0.5625 = \mathbf{0.08789}$ *or* BINOM.DIST(3, 5, 0.25, FALSE)
4	$P(X=4) = \binom{5}{4}(0.25)^4(0.75)^1 = 5 \cdot 0.0039 \cdot 0.75 = \mathbf{0.01465}$ *or* BINOM.DIST(4, 5, 0.25, FALSE)
5	$P(X=5) = \binom{5}{5}(0.25)^5(0.75)^0 = 1 \cdot 0.00098 \cdot 1 = \mathbf{0.00098}$ *or* BINOM.DIST(5, 5, 0.25, FALSE)

This is a probability distribution for this scenario! Feel free to double check that all the probability values sum to 1. We can answer some questions using this distribution:

- The most likely number of correct answers to obtain from guessing is 1, with a probability of 0.3955
- There is only a 0.00098 chance of getting them all correct! To put this into perspective, this is a little less than a 1 in 1000 chance.
- If a passing grade is getting at least 4 out of 5 correct, then there is a 0.01465 + 0.00098 = 0.01563 chance of passing by guessing.
- The expected value is: $\mu 5 \cdot 0.25 = 1.25$. It is impossible to get this exact value; it is the average number of correct answers that many people guessing would end up with. You can compare to the distribution below and note that 1.25 would be roughly in the middle of the distribution.
- Finally, note that this distribution is skewed right!



Example: Manufacturing

A common use of statistics is in quality control for engineering and manufacturing. During the process of manufacturing, a combination of sloppy tolerances, human error, environmental conditions, and luck all contribution to the potential for faulty manufacturing. Companies do a significant amount of calculations to determine how likely manufacturing errors are to occur and then factor that into the price and warranty for objects. When we have particularly high rates of manufacturing errors, especially if these can lead to injury, companies may be forced to recall products. Recalls are a good thing for the public versus letting a defect go unfixed, but companies can do their best to minimize the chance of these defects to avoid facing these costly recalls, warranties, and returns.

Let's consider the manufacture of a cheap LED light, where it is calculated 1 in every 10000 bulbs is defective, and labeled a failure. We have:

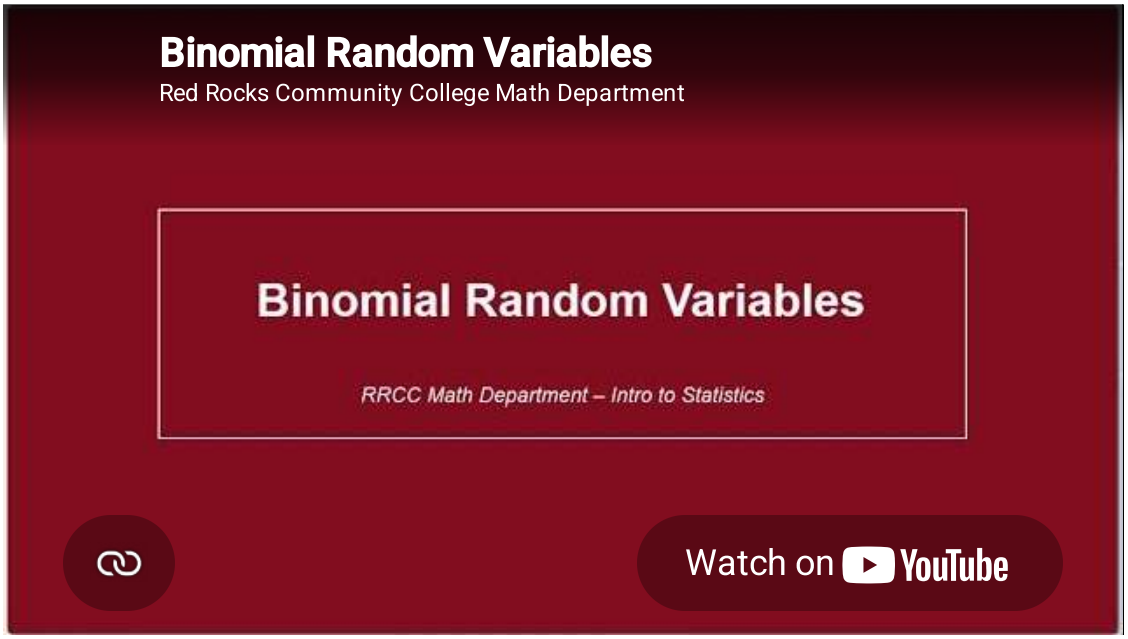
- $P(\text{Success}) = 0.9999$
- $P(\text{Failure}) = 0.0001$

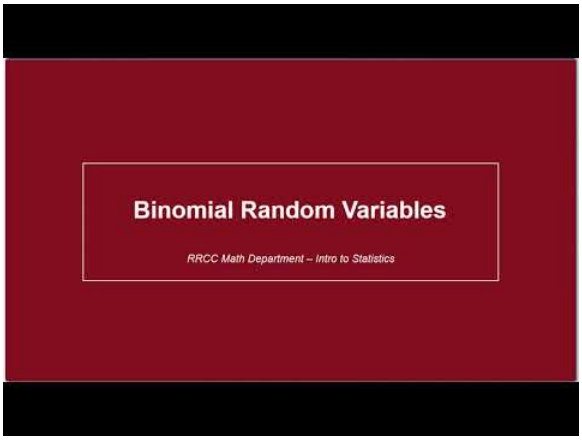
Suppose that we have a batch of 100 bulbs shipping out. What is the probability that we have 1 defective bulb in the batch (which means 99 successes)? We use the binomial PMF to calculate in Excel: **=BINOM.DIST(99, 100, 0.9999, FALSE)**

$$P(X = 99) = \binom{100}{99}(0.9999)^{99} \cdot (0.0001)^{1} \approx 0.0099 \tag{5.3.4}$$

This is a small chance, but still big enough to be worrisome!

Related Video





This page titled 5.3: Discrete Distributions- Bernoulli and Binomial is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Mathematics Department.

5.4: The Normal Distribution

The **Normal Distribution** is one of the most important and widely used probability distributions in statistics. Also known as the *bell curve*, it shows up naturally in many real-world processes from test scores to measurement errors, heights to IQs, and much more.

Continuous Distributions: A Broader View

Before we zoom in on the **Normal Distribution**, it's helpful to remember that it's just one of many types of **continuous probability distributions** used in statistics.

Unlike discrete distributions (which list clear, separate outcomes like rolling a die), continuous distributions model variables that can take on *any value* in an interval or range. We do not assign probabilities to individual values instead, we talk about the **probability of falling within an interval**.

- A continuous probability distribution is represented by a **density curve**.
- The **total area under the curve equals 1**.
- The probability of landing exactly on one number (like $P(X = 5)$) is always **zero**. We only deal with ranges, like $P(4 \leq X \leq 6)$.

There are many different continuous distributions beyond the Normal! Think of these models as tools for different situations:

- **Normal distribution** — symmetric, bell-shaped
- **Student t-distribution** — symmetric, bell-shaped
- **Uniform distribution** — all intervals of equal width are equally likely
- **Exponential distribution** — often used for waiting times
- **Beta & Gamma distributions** — useful in Bayesian statistics and modeling skew

Now let's dive deeper into the most famous of these: the **Normal Distribution**.

Definition: Normal Distribution

The **Normal Distribution** is a *symmetric, bell-shaped, continuous probability distribution* defined by two parameters:

- **μ (mu):** the mean or center of the distribution
- **σ (sigma):** the standard deviation, which controls the spread

We write this as:

$$X \sim N(\mu, \sigma)$$

The probability density function (PDF) for a normal distribution is:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \tag{5.4.1}$$

And for the **standard normal distribution**, where μ = 0 and σ = 1:

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} \tag{5.4.2}$$

Note: You won't be computing probabilities directly using this formula. Finding the area under this curve (which gives us probability) requires technology, integration, or a Z-table. But knowing the form of the function helps us understand how the curve behaves, especially how it centers around μ and falls off with distance from μ.

Key Features of the Normal Distribution:

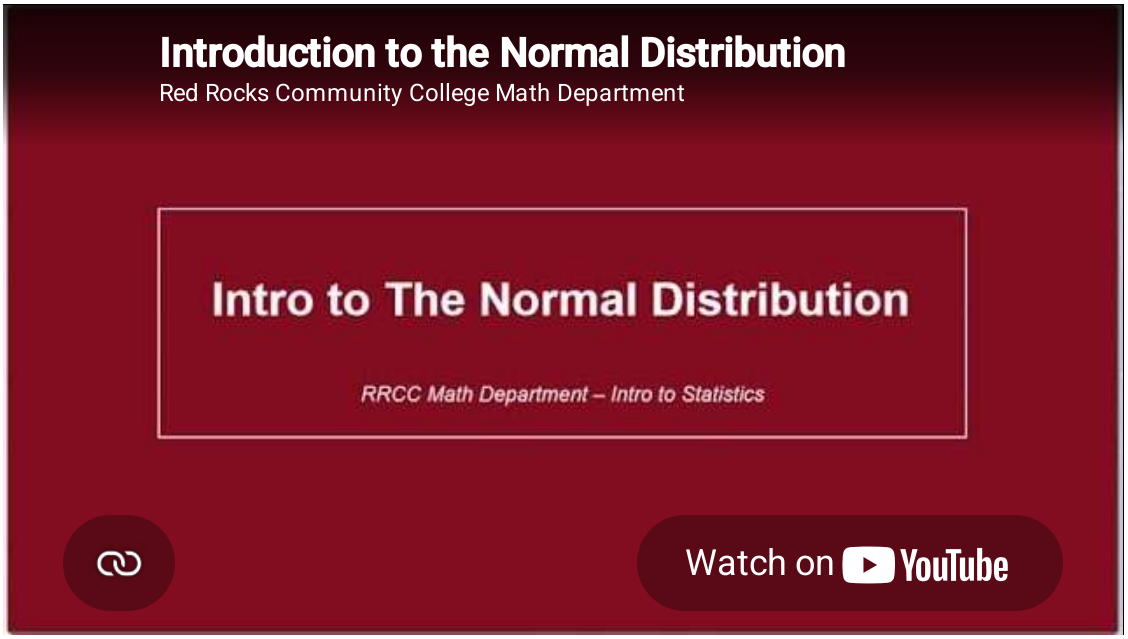
- It is **symmetric** around the mean.
- The **mean** = **median** = **mode**.
- It follows a predictable shape centered on μ with most of the area near the center and less in the tails.
- The curve never touches the x-axis (it is asymptotic).
- About **68%** of the data fall within 1 standard deviation of the mean, **95%** within 2, and **99.7%** within 3 (we'll formalize this soon as the **Empirical Rule**).

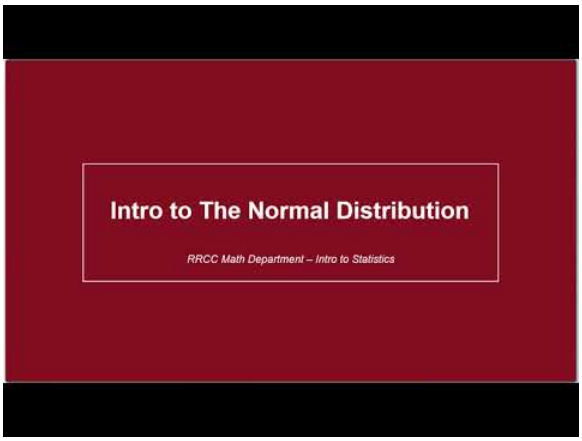
Think About It

Have you ever heard of "grading on a curve"? Or taken a standardized test? The logic behind those systems often assumes that results follow a normal distribution.

Can you think of a situation in which a bell-shaped curve might **not** be a good fit for data? What might make a distribution not normal?

Related Videos





This page titled 5.4: The Normal Distribution is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Mathematics Department.

5.5: The Empirical Rule and Standard Normal (Z) Distribution

Imagine you just received your exam score back from a national standardized test. You scored a 720, but you're not sure if that's good, average, or below expectations. The raw score alone doesn't tell you much. What really matters is how your score compares to *everyone else's*, and whether it's above or below average. Was your score typical, or was it unusually high?

This is where **z-scores** and the **standard normal distribution** come in. They let us translate raw scores into a common frame of reference, showing exactly how far a score is from the average (mean), measured in standard deviations. In this section, we'll learn how the **Empirical Rule** and **z-scores** help us reframe raw data in terms we can easily interpret, compare, and explain.

Definition: Z-score

A **z-score** tells you how far a data point is from the mean in standard deviation units.

$$z = \frac{x - \mu}{\sigma} \tag{5.5.1}$$

Where:

- x = your data value
- μ = the dataset mean
- σ = the standard deviation

A positive z-score means the value is above the mean, and a negative z-score means it is below the mean.

Computing a z-score **normalizes** a value in a distribution. In the standard normal distribion ($\mathcal{N}(0,1)$), a value is equal to its z-score. (Check this!)

Example: Converting to a Z-score

Suppose we find out that the test you scored a 720 on has a mean of 700 and a standard deviation of 40.

We calculate the z-score:

$$z = \frac{720 - 700}{40} = \frac{20}{40} = 0.5$$

This means you scored one half of a standard deviation above the average.

While z-scores are most commonly associated with the **standard normal distribution**, they can still be useful for comparing individual data points across different data sets, even when those distributions are not normal.

This works because a z-score doesn't rely on the shape of the distribution, it simply tells us how many standard deviations away a value is from the mean of *its own* distribution.

Note: Z-scores are useful to compare between different normal distributions!

You can use z-scores to compare values from **different distributions**, like comparing a math test score to a science test score, even if those datasets have different units, means, and standard deviations.

Why? Because z-scores *standardize* the values. They translate everything into "how far from the mean, in standard deviations", which is a unitless scale.

Example: Comparing Across Tests

Suppose you score an 82 on a math test with a class mean of $\mu = 75$ and $\sigma = 5$, and you score a 87 on your history test, where the class mean is 85 and $\sigma = 4$.

Good job on your tests! Note that your history score is higher than the math score, but the history class average is also higher than the math class average! Rather than comparing the scores directly, we can compare where they fall in the distribution, which tells us if you're doing better in each class relative to your peers, although one should not necessarily compare themselves to others!

Let's compute the z-scores:

- Math: $z = \frac{82-75}{5} = 1.4$
- History: $z = \frac{87-85}{4} = 0.5$

We can see from the z-scores that your math performance, relative to the class, was better than your history performance, regardless of the fact that your history score was higher.

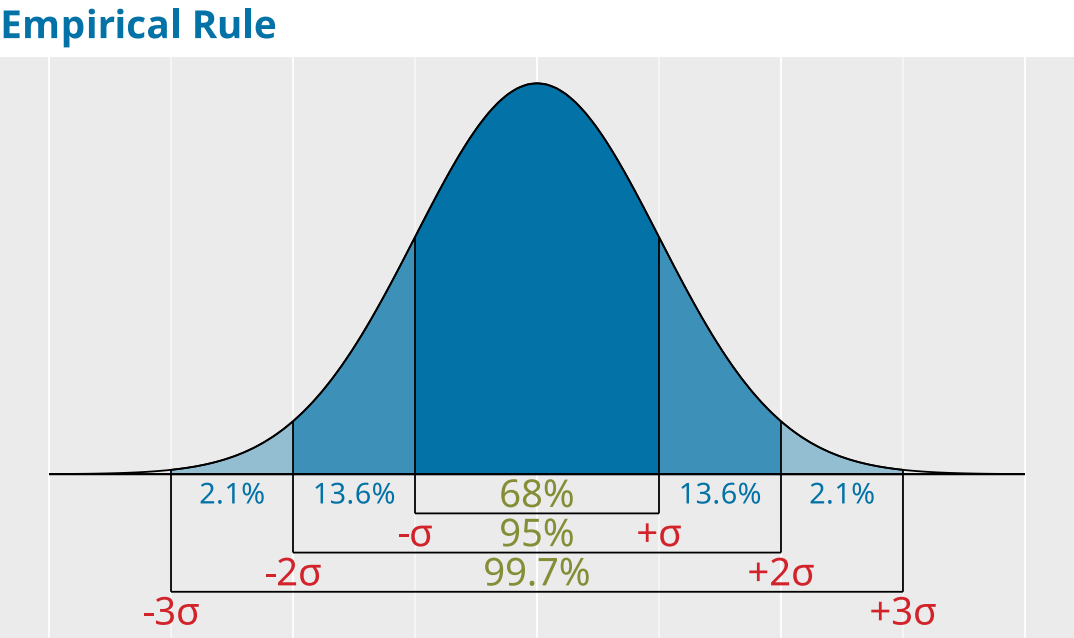
We have calculate standard deviations, and described values relative to them using z-scores, but we have a tool to build a better intuitive idea of what being "1 standard deviation above the mean" actually implies.

The Empirical Rule (68–95–99.7 Rule)

The **Empirical Rule** tells us how data is distributed in a normal distribution when we know the mean and standard deviation. It estimates the proportion of data within 1, 2, and 3 standard deviations of the mean.

Here's how it breaks down:

- About **68%** of the data falls within **1 standard deviation** of the mean ($\mu \pm \sigma$)
- About **95%** falls within **2 standard deviations** ($\mu \pm 2\sigma$)
- About **99.7%** falls within **3 standard deviations** ($\mu \pm 3\sigma$)



Shown are the various components of the distribution up to 3 standard deviations. All values are approximate.

The Empirical Rule gives us a quick way to estimate the spread of the data when we know it's approximately normal. It helps us identify what range is "typical" or "unusual."

Using a Z-Table (Standard Normal Table)

The **Z-table** allows us to find the proportion of values that fall below (to the left of) a given z-score in a standard normal distribution. This area represents a cumulative probability.

Excerpt from Z-table (Left-Tail Cumulative Area)

z-scores and probabilities

Z	0.00	0.01	0.02	0.03	0.04
0.0	0.5000	0.5040	0.5080	0.5120	0.5160
0.1	0.5398	0.5438	0.5478	0.5517	0.5557
0.2	0.5793	0.5832	0.5871	0.5910	0.5948

For example, a z-score of 0.20 corresponds to an area of approximately **0.5793**, or 57.93% of values falling below that score.

Worked Example: Exam Scores and Z-scores

Suppose exam scores for a national standardized math test are approximately normally distributed with a mean of 500 and a standard deviation of 100.

Javier scores a 630. You want to find out how well Javier did compared to everyone else, specifically, **what proportion of test takers scored below him**.

Step 1: Find the Z-score

Use the z-score formula:

$$z = \frac{x-\mu}{\sigma} = \frac{630-500}{100} = \frac{130}{100} = 1.30$$

This tells us that Javier scored **1.3 standard deviations above the mean**.

Step 2: Use the Standard Normal Table

Next, use the z-table to find the proportion of scores to the left of $z = 1.30$.

Look up 1.3 on the leftmost column and use the 0.00 column (since we want exactly 1.30):

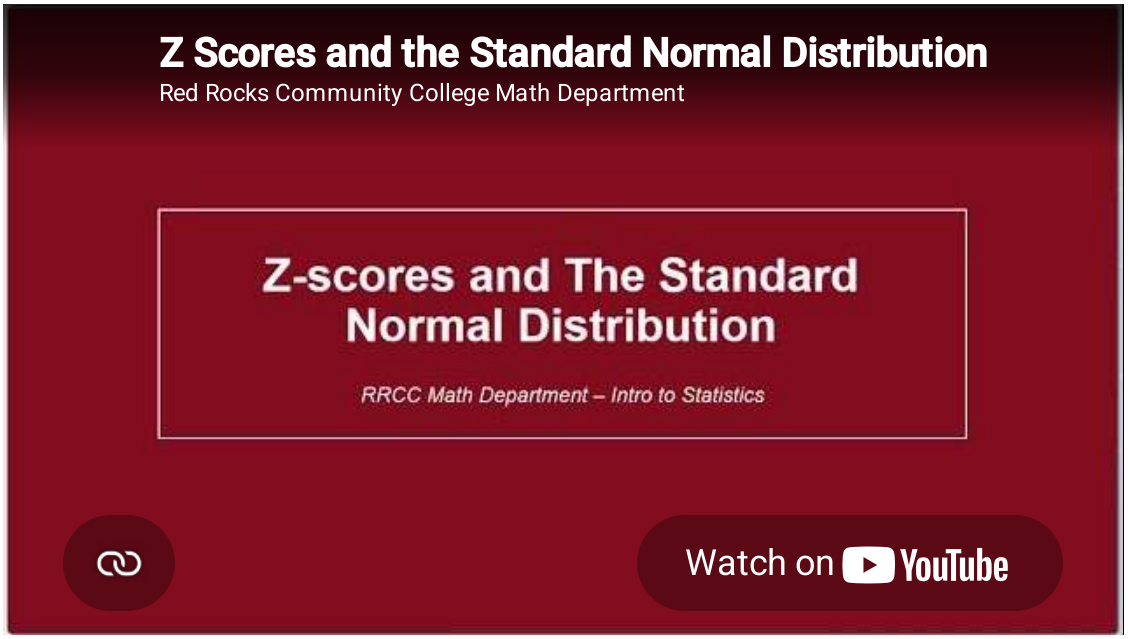
- $P(Z < 1.30) = 0.9032$

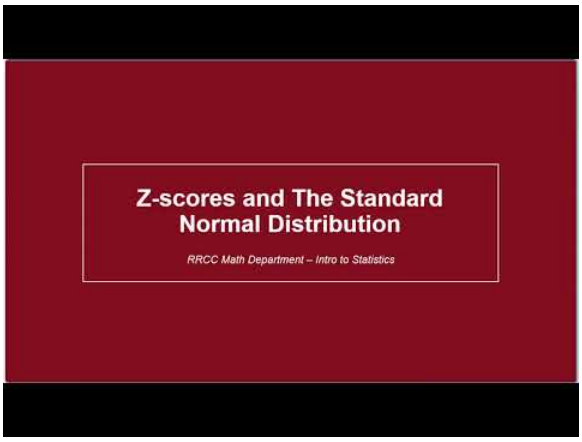
This means that approximately **90.32% of test takers scored below Javier**.

> **Final Answer**
>
> Javier's score of 630 places him in the **90th percentile** of all test takers.

> ### Think About It:
>
> If a student scores 2 standard deviations *below* the mean on a national exam, where would that person fall in the population? Use the Z-table to help you estimate.

Related Videos





This page titled 5.5: The Empirical Rule and Standard Normal (Z) Distribution is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Mathematics Department.

5.6: Application of the Normal Distribution

Now that we've explored the normal distribution and how z-scores standardize values, we're ready to apply these tools to real-world problems.

We'll work through five example scenarios that involve different types of probability questions using the normal distribution. You may use a z-table, calculator, statistical software, or graphing technology to answer these.

If you are using Excel, you will use these two functions:

Excel function	Output	Example and interpretation
=NORM.DIST(x, μ, σ, TRUE)	This gives us the **area under the curve to the left of x**	Ex: =NORM.DIST(3, 2, 1, TRUE) calculates to be 0.84134475. This means that the probability that a random value distributed normally with mean 2 and standard deviation 1 is *less than 3* is 0.84 or about 84%.
=NORM.INV(p, μ, σ)	(Inverse) This will find the x- value with *p* as the area under the curve to its left	Ex: =NORM.INV(0.25, 2, 1) calculates to be 1.3255. This means that there is a 25% probability of getting a value less than 1.3255 in the normal distribution with mean 2 and standard deviation 1

Reminder: For all these problems, assume the data is approximately normally distributed unless stated otherwise.

Example 1: Area to the Left (Lower Tail)

The heights of adult women in the U.S. are approximately normally distributed with a mean of 64 inches and a standard deviation of 3 inches.

Question: What proportion of adult women are shorter than 60 inches?

Using Technology: Calculate P(x<60) where $\mu = 64$ and $\sigma = 3$. This is the area to the left of 60 in this normal distribution

With Excel: =NORM.DIST(60, 64, 3, TRUE) = 0.091211

Using a z-table: Convert to a z-score: $z = \frac{60-64}{3} \approx -1.33$ then look up P(Z<1.33) in the table: 0.1020

Example 2: Area to the Right (Upper Tail)

IQ scores are normally distributed with a mean of 100 and a standard deviation of 15.

Question: What percentage of people have an IQ above 120?

1. Find the z-score for 120
2. Find the area *to the left* (from the table), then subtract from 1
3. $P(X > 120) = 1 - P(Z < z)$

Example 3: Area Between Two Values

A college class takes a final exam. The scores are normally distributed with a mean of 72 and standard deviation of 8.

Question: What proportion of students scored a C?

1. Find the z-scores for both endpoints (70 and 80)
2. Use the table to find $P(Z < z_2)$ and $P(Z < z_1)$
3. Subtract to find area between: $P(z_1 < Z < z_2) = P(Z < z_2) - P(Z < z_1)$

Example 4: Percentile to Value (Reverse Z)

The SAT Math test scores are normally distributed with a mean of 520 and standard deviation of 100.

Question: What score corresponds to the 90th percentile? Recall that a student in the 90th percentile would score *higher* than 90% of other students. In otherwords, we have a proportion of 0.90

1. Look up the area closest to 0.9000 in the z-table.
2. Take that z-score and plug into: $x = \mu + z\sigma$

Example 5: Comparing Two z-scores

Jordan earns an 88 on their statistics test. The class mean is 75 with a standard deviation of 6.

However, Jordan scores an 82 on their biology test, where the mean is 70 with a standard deviation of 8.

Question: In which class did Jordan score higher relative to their peers?

1. Calculate each class's z-score.
2. Interpret: the higher positive z-score is the one further above average.



Normal Distribution Check-In

Use the standard normal table or technology for these questions. Answers rounded to 4 decimal places.

1. Z-score

A student scores 84 on a science test. The mean is 75 and the standard deviation is 5. What is the z-score?

Check Answer

2. Area to the Left

What proportion of data falls below a z-score of 0.85?

Check Answer

3. Area to the Right

What proportion of values are greater than a z-score of 1.2?

Check Answer

4. Area Between

What proportion of values fall between z = -1.0 and z = 1.0?

Check Answer

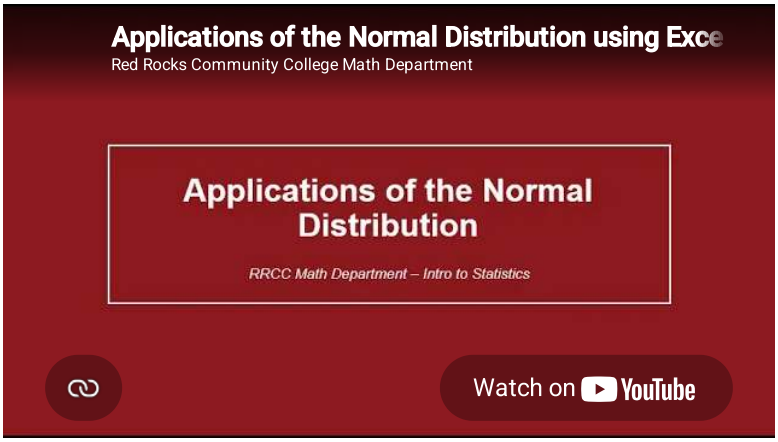
5. Value from Percentile

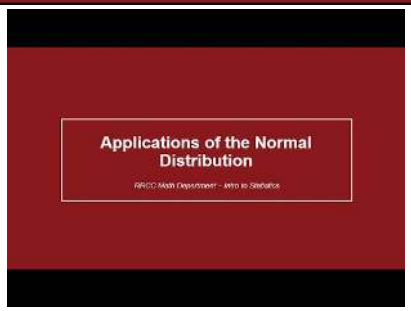
SAT Math scores are normally distributed with mean = 500 and standard deviation = 100. What score corresponds to the 84th percentile?

Check Answer

> **Tip for Practice**
>
> Use a z-table or software to determine areas under the standard normal curve. The goal here is to **understand the logic and patterns** — not just get numbers. These tools will be used again when we talk about **confidence intervals** and **hypothesis testing**.

Related Video





This page titled 5.6: Application of the Normal Distribution is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Mathematics Department.

5.7: Vocabulary (Chapter 5)

Chapter 5 Vocabulary

This chapter introduced several key terms and definitions related to probability distributions, random variables, and the normal distribution. Review the terms below and be prepared to use them in context when solving problems or justifying calculations.

Random Variable

A variable that takes numerical values determined by the outcome of a random process. Each value is associated with a probability.

Discrete Random Variable

A random variable with a finite or countable set of possible outcomes. Example: number of heads in 3 coin flips.

Continuous Random Variable

A random variable with an infinite number of possible values within a given interval. Example: time to complete a quiz.

Probability Distribution

A table, graph, or formula that assigns probabilities to each possible value of a random variable.

Bernoulli Trial

An experiment with only two possible outcomes: success or failure.

Bernoulli Distribution

A discrete probability distribution for a single Bernoulli trial. Only two values possible: 0 and 1.

Binomial Distribution

A probability distribution that models the number of successes in n independent Bernoulli trials with probability p.

Binomial Coefficient

The number of ways to choose x successes out of n trials: $\binom{n}{x}$.

Probability Mass Function (PMF)

A function that gives the probability that a discrete random variable is exactly equal to some value.

Expected Value

The theoretical mean of a random variable, computed as $E(X) = \sum x \cdot P(x)$.

Variance

The average of the squared differences from the mean: $\sigma^2 = \sum (x - \mu)^2 P(x)$.

Standard Deviation

The square root of the variance. It measures the typical distance between the data values and the mean.

Net Gain

The outcome of a situation after subtracting the initial cost or investment. Often used in lottery and game models.

Normal Distribution

A continuous, symmetric, bell-shaped distribution commonly found in natural and social processes. Defined by its mean and standard deviation.

Density Curve

A curve that represents the probability distribution of a continuous random variable. The area under the curve equals 1.

Symmetric Distribution

A distribution where the left and right halves are mirror images of each other.

Standard Normal Distribution

A normal distribution with mean $\mu = 0$ and standard deviation $\sigma = 1$.

Z-score

A standardized value that indicates how many standard deviations a data point is from the mean: $z = \frac{x - \mu}{\sigma}$.

Empirical Rule (68–95–99.7 Rule)

In a normal distribution: ~68% of data fall within 1 standard deviation, ~95% within 2, and ~99.7% within 3.

Percentile

The percentage of data values below a given value. Example: 90th percentile means 90% scored below that point.

Z-table (Standard Normal Table)

A table showing the cumulative probability up to a given z-score in the standard normal distribution.

Area under the curve

In a normal distribution, the area under the curve corresponds to probability. The total area is 1.

Left Tail

The area under the normal curve to the left of a z-score. Represents $P(Z < z)$.

Right Tail

The area to the right of a z-score: $P(Z > z) = 1 - P(Z < z)$.

Standardization

The process of converting raw data into z-scores so values can be compared across distributions with different scales.

Between Probability

The probability that a random variable falls between two values. Calculated as the difference between cumulative areas.

This page titled 5.7: Vocabulary (Chapter 5) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Mathematics Department.

5.8: Practice (Chapter 5)

5.1: Random Variables – Discrete and Continuous

1. Identify whether each of the following is a random variable or not:
   a. The amount of rain, in inches, that will fall next Friday in Denver, CO.
   b. The major of a randomly drawn student from Red Rocks Community College.
   c. The number of books purchased next year by your local library.
   d. Whether people prefer cheese pizza or pepperoni pizza.
   e. The number of credits a student takes in a semester.
   f. The average weight of a cat.
2. In your own words, what is a random variable?
3. Identify whether each variable below is discrete or continuous:
   a. Number of emails you receive in a day
   b. Temperature outside (in °F)
   c. Cost of a cup of coffee at a local shop
   d. Number of correct answers on a 10-question quiz
   e. Height of students in a class
4. Consider the following experiment. You are one of 100 people enlisted to take part in a study to determine the percent of nurses in America with an R.N. (registered nurse) degree. You ask nurses if they have an R.N. degree. The nurses answer "yes" or "no." You then calculate the percentage of nurses with an R.N. degree. You give that percentage to your supervisor.
   a. What part of the experiment will yield discrete data?
   b. What part of the experiment will yield continuous data?
5. When age is rounded to the nearest year, do the data stay continuous, or do they become discrete? Why?
6. What is the difference between a discrete and a continuous random variable? Include one real-world example of each.
7. Why can't we list individual probabilities for continuous random variables the way we do for discrete ones?
8. Determine whether or not the table is a valid probability distribution of a discrete random variable. Explain fully.

   a.

x	-2	0	2	4
$P(x)$	0.3	0.5	0.2	0.1

   (5.8.1)

   b.

x	0.5	0.25	0.25
$P(x)$	-0.4	0.6	0.8

   (5.8.2)

   c.

x	1.1	2.5	4.1	4.6	5.3
$P(x)$	0.16	0.14	0.11	0.27	0.22

   (5.8.3)

9. Determine whether or not the table is a valid probability distribution of a discrete random variable. Explain fully.

   a.

x	0	1	2	3	4
$P(x)$	-0.25	0.50	0.35	0.10	0.30

   (5.8.4)

   b.

x	1	2	3
$P(x)$	0.325	0.406	0.164

   (5.8.5)

   c.

x	25	26	27	28	29
$P(x)$	0.13	0.27	0.28	0.18	0.14

   (5.8.6)

10. What is a "probability distribution"? Describe it in your own words.
11. Create a discrete probability distribution for each of the following. (Make reasonable estimates where necessary.)
    a. The number of heads in two tosses of a coin.

b. The number of newborn babies born in a particular county in a month.
c. The number of 12-ounce cans of soft drink someone drinks.
d. The number of games in the next World Series.
e. The number of cards that match when three cards are drawn from a deck.

12. Suppose that the probability distribution for the number of years it takes to earn a Bachelor of Science (B.S.) degree is given in the following table.

a. In words, define the random variable X.
b. What does it mean that the values zero, one, and two are not included for X in the distribution?

X	P(X)
3	0.05
4	0.40
5	0.30
6	0.15
7	0.10

13. A physics professor wants to know what percent of physics majors will spend the next several years doing post-graduate research. She has the following probability distribution.

X	P(X)
1	0.35
2	0.20
3	0.15
4	
5	0.10
6	0.05

a. Define the random variable X
b. What should the column P(X) add up to and why?
c. Find the probability that a physics major will do post-graduate research for four years
d. Find the probability that a physics major will do post-graduate research for at least three years.
e. Find the probability that a physics major will no do research for one year.

14. A ballet instructor is interested in knowing what percent of each year's class will continue on to the next, so that they can plan what classes to offer. Over the years, they have established the following probability distribution.

X	P(X)
1	0.10
2	0.05
3	0.10
4	
5	0.30
6	0.20
7	0.10

a. In words, describe the random variable X.
b. Determine P(X=4)
c. Calculate and interpret P(x<4)
d. Calculate and interpret P(x\=7)

5.2: Expected Value of Discrete Distributions

1. In your own words, what is expected value? What does it represent?
2. Complete the following table and calculate the expected value.

X	P(X)	X P(X)
2	0.1	
4	0.3	4(0.3) = 1.2
6	0.4	6(0.4) = 2.4
8	0.2	8(0.2) = 1.6

3. Complete the following table and calculate the expected value.

Expected values

x	$P(x)$	$x * P(x)$
0	0.2	
1	0.2	
2	0.4	
3	0.2	

4. Identify the mistake in the probability distribution table.

Find the mistake

x	$P(x)$	$x * P(x)$
1	0.15	0.15
2	0.25	0.50
3	0.30	0.90
4	0.20	0.80
5	0.15	0.75

5. Suppose that the probability distribution for the number of years it takes to earn a Bachelor of Science (B.S.) degree is given as in the table.

B.S. completion times

x	$P(x)$
3	0.05
4	0.40
5	0.30
6	0.15
7	0.10

On average, how many years do you expect it to take for an individual to earn a B.S.?

6. You play a game where you win $10 with probability 0.2, win $5 with probability 0.3, and lose $2 with probability 0.5.
   a. Create a probability distribution table for this game.
   b. What is the expected value of your winnings?
   c. Would you recommend playing this game? Why or why not?
7. A lottery ticket costs $3. The possible prizes (net winnings) and associated probabilities are as follows: What is the expected value of this ticket (accounting for the cost to play)?
   a. $0 (prob = 0.85)
   b. $5 (prob = 0.10)
   c. $50 (prob = 0.04)
   d. $500 (prob = 0.01)
8. A theater group holds a fund-raiser. It sells 100 raffle tickets for $5 each. Suppose you purchase four tickets. The prize is two passes to a Broadway show, worth a total of $150.
   a. What are you interested in here?
   b. Define the random variable $\boldsymbol{X}$.
   c. List the values that $\boldsymbol{X}$ may take on.
   d. Construct a probability distribution table for this situation.
   e. If this fund-raiser is repeated often and you always purchase four tickets, what would be your expected average winnings per raffle?
9. A game involves selecting a card from a regular 52-card deck and tossing a coin. The coin is a fair coin and is equally likely to land on heads or tails.
   a. If the card is a face card, and the coin lands on Heads, you win $6
   b. If the card is a face card, and the coin lands on Tails, you win $2
   c. If the card is not a face card, you lose $2, no matter what the coin shows.
   d. Find the expected value for this game (expected net gain or loss).
   e. Explain what your calculations indicate about your long-term average profits and losses on this game.

f. Should you play this game to win money?

10. A venture capitalist, willing to invest $1,000,000, has three investments to choose from. The first investment, a social media company, has a 20% chance of returning $7,000,000 profit, a 30% chance of returning no profit, and a 50% chance of losing the million dollars. The second company, an advertising firm has a 10% chance of returning $3,000,000 profit, a 60% chance of returning a $2,000,000 profit, and a 30% chance of losing the million dollars. The third company, a chemical company has a 40% chance of returning $3,000,000 profit, a 50% chance of no profit, and a 10% chance of losing the million dollars.
    a. Construct a probability distribution table for each investment. In your table the X column is the net amount of profit/loss for the venture capitalist and the P(X) column is the likelihood given above.
    b. Find the expected value for each investment.
    c. Which investment has the highest expected return?
    d. Which is the safest investment and why?
    e. Which is the riskiest investment and why?
11. A company estimates that 3% of their products will fail after the original warranty period but within 2 years of the purchase, with a replacement cost of $250. If they want to offer a 2 year extended warranty, what price should they charge so that they'll break even (in other words, so the expected value will be 0).
12. Write an expression (but do not calculate) for the variance of the following random variable:
    a. x = {–1, 0, 5}, with P(x) = {0.2, 0.5, 0.3}
13. Find the standard deviation for the following probability distribution

X	P(X)	X P(X)	$(X-\mu)^2P(X)$
2	0.1	2(0.1) = 0.2	$(2–5.4)^2(0.1) = 1.156$
4	0.3	4(0.3) = 1.2	$(4–5.4)^2(0.3) = 0.588$
6	0.4	6(0.4) = 2.4	$(6–5.4)^2(0.4) = 0.144$
8	0.2	8(0.2) = 1.6	$(8–5.4)^2(0.2) = 1.352$

14. Red Rocks Community College has 14 statistics classes scheduled for its Fall 2026 term. One class has space available for 30 students, eight classes have space for 60 students, one class has space for 70 students, and four classes have space for 100 students.
    a. What is the average class size assuming each class is filled to capacity?
    b. Space is available for 980 students. Suppose that each class is filled to capacity and select a statistics student at random. Let the random variable X equal the size of the student's class. Create a probability distribution for X.
    c. Find the mean of X.
    d. Find the standard deviation of X.
15. Borachio works in an automotive tire factory. The number X of sound but blemished tires that he produces on a random day has the probability distribution

x	2	3	4	5
$P(x)$	0.48	0.36	0.12	0.04

(5.8.7)

    a. Find the probability that Borachio will produce more than three blemished tires tomorrow.
    b. Find the probability that Borachio will produce at most two blemished tires tomorrow.
    c. Compute the mean and standard deviation of X. Interpret them in the context of the problem.
16. The number X of days in the summer months that a construction crew cannot work because of the weather has the probability distribution

x	6	7	8	9	10
$P(x)$	0.03	0.08	0.15	0.20	0.19

(5.8.8)

x	11	12	13	14
$P(x)$	0.16	0.10	0.07	0.02

(5.8.9)

a. Find the probability that no more than ten days will be lost next summer.
b. Find the probability that from 8 to 12 days will be lost next summer.
c. Find the probability that no days at all will be lost next summer.
d. Compute the mean and standard deviation of X. Interpret them in the context of the problem.

17. Ice cream usually comes in 1.5 quart boxes (48 fluid ounces), and ice cream scoops hold about 2 ounces. However, there is some variability in the amount of ice cream in a box as well as the amount of ice cream scooped out. We represent the amount of ice cream in the box as X and the amount scooped out as Y. Suppose these random variables have the following means, standard deviations, and variances

	mean	SD	variance
X	48	1	1
Y	2	0.25	0.0625

a. An entire box of ice cream, plus 3 scoops from a second box is served at a party. How much ice cream do you expect to have been served at this party? What is the standard deviation of the amount of ice cream served?
b. How much ice cream would you expect to be left in the box after scooping out one scoop of ice cream? That is, find the expected value of X - Y . What is the standard deviation of the amount left in the box?
c. Using the context of this exercise, explain why we add variances when we subtract one random variable from another.

5.3: Discrete Distributions – Bernoulli and Binomial

1. Suppose you ask 40 individuals whether they ride their bicycle to work and 2 individuals say "yes". What proportion of the individuals in the sample responded "yes"? What proportion responded "no"?
2. Suppose you are a caterer and planning an event for 250 attendees. The event is held for a population that you know to be 15% vegan. How many attendees would you expect to be vegan?
3. Determine if each trial can be considered an independent Bernouilli trial for the following situations.
   a. Cards dealt in a hand of poker.
   b. Outcome of each roll of a die.
   c. The number of coins that match when three coins are flipped.
   d. The response from an individual to a yes/no question on a survey.
   e. The number of times you roll a die until you get a six.
4. Determine whether or not the random variable X is a binomial random variable. If so, give the values of n and p. If not, explain why not.
   a. X is the number of dots on the top face of fair die that is rolled.
   b. X is the number of hearts in a five-card hand drawn (without replacement) from a well-shuffled ordinary deck.
   c. X is the number of defective parts in a sample of ten randomly selected parts coming from a manufacturing process in which 0.02% of all parts are defective.
   d. X is the number of times the number of dots on the top face of a fair die is even in six rolls of the die.
   e. X is the number of dice that show an even number of dots on the top face when six dice are rolled at once.
5. Determine whether or not the random variable X is a binomial random variable. If so, give the values of n and p. If not, explain why not.
   a. X is the number of black marbles in a sample of 5 marbles drawn randomly and without replacement from a box that contains 25 white marbles and 15 black marbles.

b. X is the number of black marbles in a sample of 5 marbles drawn randomly and with replacement from a box that contains 25 white marbles and 15 black marbles.
c. X is the number of voters in favor of proposed law in a sample $1,200$ randomly selected voters drawn from the entire electorate of a country in which 35% of the voters favor the law.
d. X is the number of coins that match at least one other coin when four coins are tossed at once.

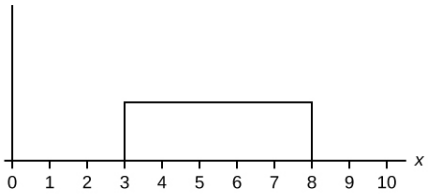
6. What are the requirements for a binomial experiment?
7. When dealing with the binomial distribution, why are the possible values for the random variable always 0,1,2,3,…,n where n is the number of trials or sample size? Why can't we use negative values, or fractions, or numbers greater than n?
8. Under what conditions is a binomial distribution symmetric? Skewed left? Skewed right? Why?
9. How do you know when to use the binomial distribution to model a situation?
10. How are the areas in the bars of a binomial histogram related to the probability of choosing those X values? (Hint: figure it out for a single bar)
11. Compare and contrast the Bernoulli and Binomial distributions.
12. A fair coin is flipped once. Define a random variable X to represent success = heads. Construct a table showing the probability distribution for this Bernoulli variable.
13. A basketball player makes 80% of their free throws. If they shoot 3 times, what is the probability they make exactly 2?
14. The Substance Abuse and Mental Health Services Administration estimated that 70% of 18-20 year olds consumed alcoholic beverages in 2008.
    a. Suppose a random sample of ten 18-20 year olds is taken. Is the use of the binomial distribution appropriate for calculating the probability that exactly six consumed alcoholic beverages? Explain.
    b. Calculate the probability that exactly 6 out of 10 randomly sampled 18-20 year olds consumed an alcoholic drink.
    c. What is the probability that exactly four out of the ten 18-20 year olds have not consumed an alcoholic beverage?
    d. What is the probability that at most 2 out of 5 randomly sampled 18-20 year olds have consumed alcoholic beverages?
    e. What is the probability that at least 1 out of 5 randomly sampled 18-20 year olds have consumed alcoholic beverages?
15. The National Vaccine Information Center estimates that 90% of Americans have had chickenpox by the time they reach adulthood.
    a. Suppose we take a random sample of 100 American adults. Is the use of the binomial distribution appropriate for calculating the probability that exactly 97 had chickenpox before they reached adulthood? Explain.
    b. Calculate the probability that exactly 97 out of 100 randomly sampled American adults had chickenpox during childhood.
    c. What is the probability that exactly 3 out of a new sample of 100 American adults have not had chickenpox in their childhood?
    d. What is the probability that at least 1 out of 10 randomly sampled American adults have had chickenpox?
    e. What is the probability that at most 3 out of 10 randomly sampled American adults have not had chickenpox?
16. A 2005 Gallup Poll found that that 7% of teenagers (ages 13 to 17) suffer from arachnophobia and are extremely afraid of spiders. At a summer camp there are 10 teenagers sleeping in each tent. Assume that these 10 teenagers are independent of each other.
    a. Calculate the probability that at least one of them suffers from arachnophobia.
    b. Calculate the probability that exactly 2 of them suffer from arachnophobia?
    c. Calculate the probability that at most 1 of them suffers from arachnophobia?
    d. If the camp counselor wants to make sure no more than 1 teenager in each tent is afraid of spiders, does it seem reasonable for him to randomly assign teenagers to tents?
17. About 24% of flights departing from New York's John F. Kennedy International Airport were delayed in 2009. Assuming that the chance of a flight being delayed has stayed constant at 24%, we are interested in finding the probability of 10 out of the next 100 departing flights being delayed. Noting that if one flight is delayed, the next flight is more likely to be delayed, which of the following statements is correct?
    a. We can use the binomial distribution with n = 100, k = 10, and p = 0.24 to calculate this probability.
    b. We can use the geometric distribution with n = 100, k = 10, and p = 0.24 to calculate this probability.

c. We cannot calculate this probability using the binomial distribution since whether or not one flight is delayed is not independent of another.
d. We can use the binomial distribution with n = 10, k = 100, and p = 0.24 to calculate this probability.

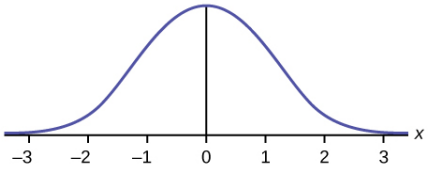
18. Write out the full binomial distribution table for $n = 4, p = 0.5$. Label columns clearly.
19. In a binomial experiment with $n = 6, p = 0.3$, what is the probability of 0 successes? One success? (Use binomial probabilities)
20. Suppose you are a lawyer representing a Hispanic individual in a criminal trial. Twelve jurors are randomly selected by the state from a population that you know to be 50% Hispanic.
    a. How many jurors would you expect to be Hispanic?
    b. The selection process produces a jury with 3 Hispanic individuals. What proportion of the jury is Hispanic?
    c. Use the binomial distribution and find the probability that 3 or fewer jurors would be Hispanic.
    d. Create a graph (using technology) of the distribution with the area corresponding to the previous answer highlighted
    e. Use the above results to write an argumentative essay to the court on behalf of your client.
21. A small regional carrier accepted 21 reservations for a particular flight with 19 seats. 17 reservations went to regular customers who will arrive for the flight. Each of the remaining passengers will arrive for the flight with a 41% chance, independently of each other.
    a. Find the probability that overbooking occurs.
    b. Find the probability that the flight has empty seats.
22. After being rejected for employment, Kim Kelly learns that the Bellevue Credit Company has hired only four women among the last 17 new employees. She also learns that the pool of applicants is very large, with an approximately equal number of qualified men as qualified women.
    a. Help her address the charge of gender discrimination by finding the probability of getting four or fewer women when 17 people are hired, assuming that there is no discrimination based on gender.
    b. Because this is a serious claim, we will use a stricter cutoff value for unusual events. We will use 0.5% as the cutoff value (1 in 200 chance of happening by chance). With this in mind, does the resulting probability really support such a charge?
23. Among patients claiming to have the flu, the chance that someone has the flu (and not just a nasty cold) is only about 4%. Of the next 25 patients calling in claiming to have the flu, we are interested in how many actually have the flu.
    a. Create a probability distribution table for this situation
    b. Find the probability that at least four of the 25 patients actually have the flu.
    c. On average, for every 25 patients calling in, how many do you expect to have the flu?
24. A student takes a ten-question true-false quiz, but did not study and randomly guesses each answer. Find the probability that the student passes the quiz with a grade of at least 70% of the questions correct.
25. According to The World Bank, only 9% of the population of Uganda had access to electricity as of 2009. Suppose we randomly sample 150 people in Uganda. Let X represent the number of people who have access to electricity.
    a. Create a probability distribution for X?
    b. Calculate the mean and standard deviation of X.
    c. Use technology to find the probability that 15 people in the sample have access to electricity.
    d. Find the probability that at most ten people in the sample have access to electricity.
    e. Find the probability that more than 25 people in the sample have access to electricity.
26. If you flip a fair coin 10 times, what is the probability of
    a. getting all tails?
    b. getting all heads?
    c. getting at least one tails?

5.4: The Normal Distribution

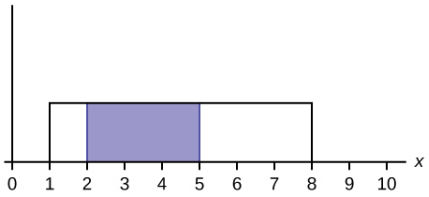
1. Which type of distribution does the graph illustrate?



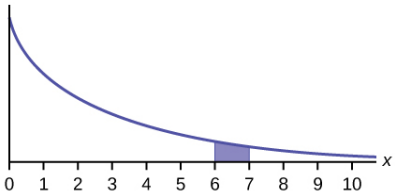
2. Which type of distribution does the graph illustrate?



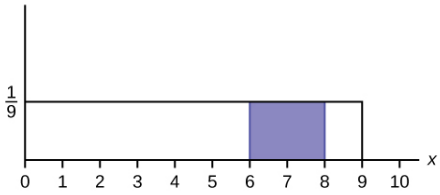
3. What does the shaded area represent?



4. What does the shaded area represent?



5. How is probability determined from a continuous distribution? Why is this easy for the uniform distribution and not so easy for the normal distribution?
6. Why is the probability that a continuous random variable is equal to a single number zero? (i.e. Why is P(X=a)=0 for any number a)
7. For a continuous probability distribution, 0<X<15.
    1. What is P(X>15)?
    2. What is P(X=7)?
8. Find the probability that falls in the shaded area.

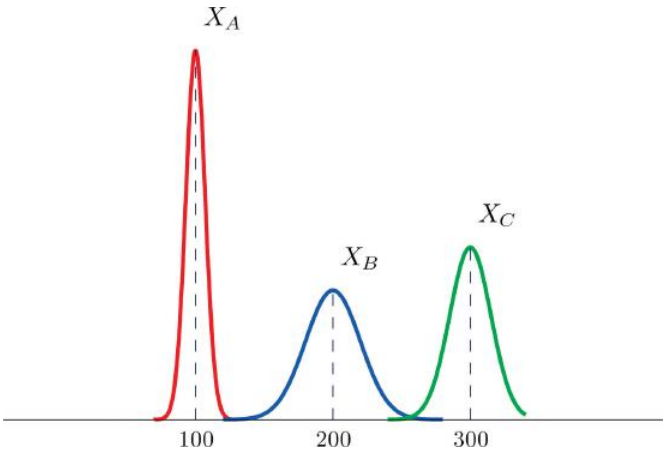


9. Births are approximately uniformly distributed between the 52 weeks of the year. They can be said to follow a uniform distribution from one to 53 (spread of 52 weeks).Graph the probability distribution.
    a. What is the mean?
    b. What is the standard deviation?
    c. Find the probability that a person is born at the exact moment week 19 starts.
    d. Find the probability that a person is born after week 40.

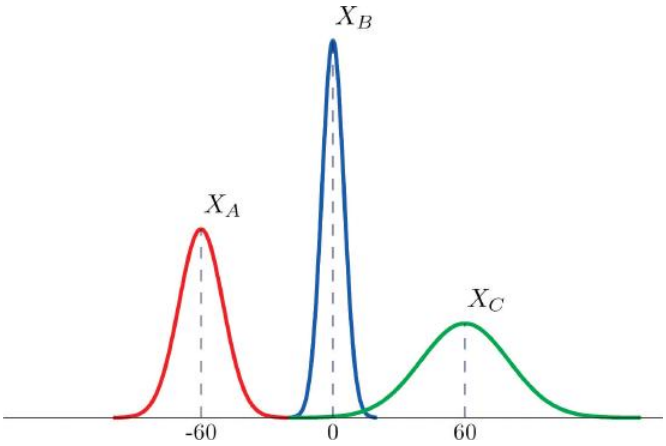
e. Find the 70th percentile.

f. Find the minimum for the upper quarter.

10. What two quantities do we need to fully describe a normal distribution?
11. Sketch and label a normal distribution. Indicate the mean and one, two, and three standard deviations on either side.
12. What does it mean for a distribution to be "normal"? List 3 key properties.
13. What does the symmetric bell shape of the normal curve imply about the distribution of individuals in a normal population?
14. The length of time to complete a survey is normally distributed with a mean of 10 minutes and standard deviation of 2 minutes.
    a. What proportion of people take less than 10 minutes?
    b. What proportion take between 8 and 12 minutes?
    c. What is the median completion time?
15. The heights of the 430 National Basketball Association players were listed on team rosters at the start of the 2005–2006 season. The heights of basketball players have an approximate normal distribution with mean, $\mu = 79$ inches and a standard deviation, $\sigma = 3.89$ inches. For each of the following heights, calculate the z-score and interpret it using complete sentences.
    a. 77 inches
    b. 85 inches
    c. If an NBA player reported his height had a z-score of 3.5, would you believe him? Explain your answer.
16. Compare a normal distribution to a skewed distribution. How might their histogram shapes differ?
17. Suppose you randomly select an individual from a population that is normally distributed and they are above average. When you find out the probability of randomly selecting that individual is very very small, what are some possible explanations? In other words what does this very very small probability suggest?
18. A continuous random variable X has a normal distribution with mean 100 and standard deviation 10. Sketch a qualitatively accurate graph of its density function.
19. A continuous random variable X has a normal distribution with mean 73 and standard deviation 2.5. Sketch a qualitatively accurate graph of its density function.
20. A continuous random variable X has a normal distribution with mean 73. The probability that X takes a value greater than 80 is 0.212. Use this information and the symmetry of the density function to find the probability that X takes a value less than 66. Sketch the density curve with relevant regions shaded to illustrate the computation.
21. A continuous random variable X has a normal distribution with mean 169. The probability that X takes a value greater than 180 is 0.17. Use this information and the symmetry of the density function to find the probability that X takes a value less than 158. Sketch the density curve with relevant regions shaded to illustrate the computation.
22. A continuous random variable X has a normal distribution with mean 50.5. The probability that X takes a value less than 54 is 0.76. Use this information and the symmetry of the density function to find the probability that X takes a value greater than 47. Sketch the density curve with relevant regions shaded to illustrate the computation.
23. A continuous random variable X has a normal distribution with mean 12.25. The probability that X takes a value less than 13 is 0.82. Use this information and the symmetry of the density function to find the probability that X takes a value greater than 11.50. Sketch the density curve with relevant regions shaded to illustrate the computation.
24. The figure provided shows the density curves of three normally distributed random variables X_A, X_B and X_C. Their standard deviations (in no particular order) are 15, 7, and 20. Use the figure to identify the values of the means μ_A, μ_B, and μ_C and standard deviations σ_A, σ_B, and σ_C of the three random variables.



25. The figure provided shows the density curves of three normally distributed random variables X_A, X_B and X_C. Their standard deviations (in no particular order) are 20, 5, and 10. Use the figure to identify the values of the means μ_A, μ_B, and μ_C and standard deviations σ_A, σ_B, and σ_C of the three random variables.



26. The amount X of orange juice in a randomly selected half-gallon container varies according to a normal distribution with mean 64 ounces and standard deviation 0.25 ounce.
    a. Sketch the graph of the density function for X.
    b. What proportion of all containers contain less than a half gallon (64 ounces)? Explain.
    c. What is the median amount of orange juice in such containers? Explain.
27. The weight X of grass seed in bags marked 50 lb varies according to a normal distribution with mean 50 lb and standard deviation 1 ounce (0.0625 lb).
    a. Sketch the graph of the density function for X.
    b. What proportion of all bags weigh less than 50 pounds? Explain.
    c. What is the median weight of such bags? Explain

5.5: The Empirical Rule and Standard Normal (Z) Distribution

1. Find the z-score for the following:
    a. x = 88, μ = 75, σ = 5
    b. x = 64, μ = 70, σ = 3
2. The patient recovery time from a particular surgical procedure is normally distributed with a mean of 5.3 days and a standard deviation of 2.1 days. What is the *z*-score for a patient who takes ten days to recover?
3. The widths of platinum samples manufactured at a factory are normally distributed, with a mean of 1.2 cm and a standard deviation of 0.5 cm. Find the z-scores that correspond to each of the following widths.
    a. 1.9 cm

b. 0.5 cm

4. The heights of the 430 National Basketball Association players were listed on team rosters at the start of the 2005–2006 season. The heights of basketball players have an approximate normal distribution with mean, μ = 79 inches and a standard deviation, σ = 3.89 inches. For each of the following heights, calculate the *z*-score and interpret it using complete sentences.
   a. 77 inches
   b. 85 inches
   c. If an NBA player reported his height had a *z*-score of 3.5, would you believe him? Explain your answer.
5. Two students take different exams:
   a. Adrian: Score = 87, μ = 75, σ = 6
   b. Blair: Score = 83, μ = 70, σ = 8
   c. Who performed better, relative to their class?
6. Kyle's doctor told him that the *z*-score for his systolic blood pressure is 1.75. Which of the following is the best interpretation of this standardized score?
7. Explain what a z-score of –2.1 tells us about a data value.
8. In general, what does a z-score tell you about a number in a data set?
9. Which of the following is false? Hint: It might be useful to sketch the distributions.
   a. Calculating percentiles based on the Z table is only appropriate for observations that come from a nearly normal distribution.
   b. The Z score for the median of a left skewed distribution is negative.
   c. The Z score for the median of a symmetric distribution is approximately 0.
   d. Z scores are defined for observations from distributions of any shape and skew.
10. According to the Empirical Rule, about what % of data fall within 1 standard deviation of the mean in a normal distribution?
11. If a dataset is normally distributed with mean = 60 and standard deviation = 4:
    a. What interval contains approximately 68% of the data?
    b. What interval contains approximately 95% of the data?
12. How can the empirical rule be restated in terms of z-scores and percentiles? Restate it for each of the seven z-scores.
13. Suppose the systolic blood pressure (in mm) of adults has an approximately normal distribution with mean 125 and standard deviation 14.
    a. Create an empirical rule graph for this situation. The use your graph to answer the following questions.
    b. About 99.7% of adults will have blood pressure between what amounts?
    c. What percentage of adults will have a systolic blood pressure outside the range 111 mm to 153 mm?
    d. Suppose you are a health practitioner and an adult patient has systolic blood pressure of 171 mm. Use the above results to explain the gravity of their situation.
14. We flip a coin 100 times ($n = 100$) and note that it only comes up heads 20% ($p = 0.20$) of the time. The mean and standard deviation for the number of times the coin lands on heads is $\mu = 20$ and $\sigma = 4$ (verify the mean and standard deviation). Solve the following:
    a. There is about a 68% chance that the number of heads will be somewhere between ___ and ___.
    b. There is about a ____chance that the number of heads will be somewhere between 12 and 28.
    c. There is about a ____ chance that the number of heads will be somewhere between eight and 32.
15. The maintenance department at the Lakewood campus of Red Rocks Community College receives daily requests to replace fluorecent lightbulbs. The distribution of the number of daily requests is bell-shaped and has a mean of 37 and a standard deviation of 8. Using the 68-95-99.7 rule, what is the approximate percentage of lightbulb replacement requests numbering between 37 and 53?
16. A NUMMI assembly line, which has been operating since 1984, has built an average of 6,000 cars and trucks a week. Generally, 10% of the cars were defective coming off the assembly line. Suppose we draw a random sample of n = 100 cars. Let X represent the number of defective cars in the sample. What can we say about X in regard to the 68-95-99.7 empirical rule (one standard deviation, two standard deviations and three standard deviations from the mean are being referred to)? Assume a normal distribution for the defective cars in the sample.

17. Thuy Dau, Ngoc Bui, Sam Su, and Lan Voung conducted a survey as to how long customers at Lucky claimed to wait in the checkout line until their turn. Let $X =$ time in line. The following table displays the ordered real data (in minutes):

Waiting data

0.50	4.25	5	6	7.25
1.75	4.25	5.25	6	7.25
2	4.25	5.25	6.25	7.25
2.25	4.25	5.5	6.25	7.75
2.25	4.5	5.5	6.5	8
2.5	4.75	5.5	6.5	8.25
2.75	4.75	5.75	6.5	9.5
3.25	4.75	5.75	6.75	9.5
3.75	5	6	6.75	9.75
3.75	5	6	6.75	10.75

a. Calculate the sample mean and the sample standard deviation.
b. Construct a histogram.
c. Draw a smooth curve through the midpoints of the tops of the bars.
d. In words, describe the shape of your histogram and smooth curve.
e. Let the sample mean approximate μ and the sample standard deviation approximate σ. The distribution of X can then be approximated by $X \sim$
f. Use the distribution in part e to calculate the probability that a person will wait fewer than 6.1 minutes.
g. Determine the cumulative relative frequency for waiting less than 6.1 minutes.
h. Why aren't the answers to part f and part g exactly the same?
i. Why are the answers to part f and part g as close as they are?
j. If only ten customers has been surveyed rather than 50, do you think the answers to part f and part g would have been closer together or farther apart? Explain your conclusion.

18. For normally distributed data, what proportion of observations have a z-score less than -0.14.
19. For normally distributed data, what proportion of observations have a z-score greater than 2.51.
20. What proportion of data fall between z = –1.5 and z = 2?
21. A z-score is given as 1.75. Use a z-table or calculator to approximate $P(Z > 1.75)$.
22. What percent of a standard normal distribution is found in each region?

a. Z < -1:35
b. Z > 1:48
c. -0:4 < Z < 1:5
d. |Z| > 2

5.6: Applications of the Normal Distribution

1. A weightlifting competition has normally distributed scores with μ = 100 and σ = 15. What proportion of athletes score above 120?
2. The average gas mileage for a hybrid car is 48 mpg (σ = 4). What is the probability that a randomly selected car gets less than 43 mpg?
3. What SAT Math score corresponds to the 95th percentile if μ = 520 and σ = 100?
4. The empirical rule says that 95% of the population is within 2 standard deviations of the mean, but when I find the z-scores that mark off the middle 95% of the standard normal distribution I calculate -1.96 and 1.96. Is this a contradiction? Why or why not? In other words why are the normal distribution calculators not agreeing with the empirical rule?

5. Suppose that you are an elementary school teacher and you are evaluating the reading levels of your students. You find an individual that reads 42.3 word per minute. You do some research and determine that the reading rates for their grade level are normally distributed with a mean of 85 words per minute and a standard deviation of 23 words per minute.
   a. At what percentile is the child's reading level.
   b. Create a graph with a normal curve that illustrates the problem.
   c. Write an essay to the parents of the child justifying the need for remediation.
6. Suppose that you are working for a chain restaurant and wish to design a promotion to disabuse the public of notions that the service is slow. You decide to institute a policy that any customer that waits too long will receive their meal for free. You know that the wait times for customers are normally distributed with a mean of 19 minutes and a standard deviation of 3.4 minutes. Use statistics to decide the maximum wait time you would advertise to customers so that you only give away free meals to at most 1% of the customers.
   a. Determine an estimate of an advertised maximum wait time so that 1% of the customers would receive a free meal.
   b. Create a graph illustrating the solution.
   c. Write a response to the vice president explaining your prescribed maximum wait time.
7. The amount X of beverage in a can labeled 12 ounces is normally distributed with mean 12.1 ounces and standard deviation 0.05 ounce. A can is selected at random.
   a. Find the probability that the can contains at least 12 ounces.
   b. Find the probability that the can contains between 11.9 and 12.1 ounces.
8. The systolic blood pressure X of adults in a region is normally distributed with mean 112 mm Hg and standard deviation 15 mm Hg. A person is considered “pre-hypertensive” if his systolic blood pressure is between 120 and 130 mm Hg. Find the probability that the blood pressure of a randomly selected person is pre-hypertensive.
9. Heights X of adult women are normally distributed with mean 63.7 inches and standard deviation 2.71 inches. Juliet, who is 69.25 inches tall, wishes to date only women who are shorter than she is, but within 4 inches of her height. Find the probability that the next woman she meets will have such a height.
10. Heights X of adult men are normally distributed with mean 69.1 inches and standard deviation 2.92 inches. Romeo, who is 63.25 inches tall, wishes to date only men who are taller than he is but within 6 inches of his height. Find the probability that the next man he meets will have such a height.
11. A regulation hockey puck must weigh between 5.5 and 6 ounces. The weights X of pucks made by a particular process are normally distributed with mean 5.75 ounces and standard deviation 0.11 ounce.
    a. Find the probability that a puck made by this process will meet the weight standard.
    b. Find the maximum allowable standard deviation so that at most 0.005 of all pucks will fail to meet the weight standard. (Hint: The distribution is symmetric and is centered at the middle of the interval of acceptable weights.)
12. A regulation golf ball may not weigh more than 1.620 ounces. The weights X of golf balls made by a particular process are normally distributed with mean 1.361 ounces and standard deviation 0.09 ounce. Find the probability that a golf ball made by this process will meet the weight standard.
13. The length of time that the battery in Hippolyta's cell phone will hold enough charge to operate acceptably is normally distributed with mean 25.6 hours and standard deviation 0.32 hour. Hippolyta forgot to charge their phone yesterday, so that at the moment they first want to use it today it has been 26 hours 18 minutes since the phone was last fully charged. Find the probability that the phone will operate properly.
14. The amount of non-mortgage debt per household for households in a particular income bracket in one part of the country is normally distributed with mean $\$28,350$ and standard deviation $\$3,425$. Find the probability that a randomly selected such household has between $\$20,000$ and $\$30,000$ in non-mortgage debt.
15. Birth weights of full-term babies in a certain region are normally distributed with mean 7.125 lb and standard deviation 1.290 lb. Find the probability that a randomly selected newborn will weigh less than 5.5 lb, the historic definition of prematurity.
16. The distance from the seat back to the front of the knees of seated adults is normally distributed with mean 23.8 inches and standard deviation 1.22 inches. The distance from the seat back to the back of the next seat forward in all seats on aircraft flown by a budget airline is 26 inches. Find the proportion of adults flying with this airline whose knees will touch the back of the seat in front of them.

17. The distance from the seat to the top of the head of seated adults is normally distributed with mean 36.5 inches and standard deviation 1.39 inches. The distance from the seat to the roof of a particular make and model car is 40.5 inches. Find the proportion of adults who when sitting in this car will have at least one inch of headroom (distance from the top of the head to the roof).
18. The useful life of a particular make and type of automotive tire is normally distributed with mean $57,500$ miles and standard deviation 950 miles.
    a. Find the probability that such a tire will have a useful life of between $57,000$ and $58,000$ miles.
    b. Hamnet buys four such tires. Assuming that their lifetimes are independent, find the probability that all four will last between $57,000$ and $58,000$ miles. (If so, the best tire will have no more than $1,000$ miles left on it when the first tire fails.) Hint: There is a binomial random variable here, whose value of p comes from part a.
19. A machine produces large fasteners whose length must be within 0.5 inch of 22 inches. The lengths are normally distributed with mean 22.0 inches and standard deviation 0.17 inch.
    a. Find the probability that a randomly selected fastener produced by the machine will have an acceptable length.
    b. Between what two sizes are the middle 50% of all fasteners?
20. The lengths of time taken by students on an algebra proficiency exam (if not forced to stop before completing it) are normally distributed with mean 28 minutes and standard deviation 1.5 minutes.
    a. Find the proportion of students who will finish the exam if a 30-minute time limit is set.
    b. Between what two scores are the middle 70% of all students?
21. The amount of gasoline X delivered by a metered pump when it registers 5 gallons is a normally distributed random variable. The standard deviation σ of Xmeasures the precision of the pump; the smaller σ is the smaller the variation from delivery to delivery. A typical standard for pumps is that when they show that 5 gallons of fuel has been delivered the actual amount must be between 4.97 and 5.03 gallons (which corresponds to being off by at most about half a cup). Supposing that the mean of X is 5, find the largest that σ can be so that $P(4.97 < X < 5.03)$ is 1.0000 to four decimal places. (Hint: The z-score of 5.03 will be the smallest value of Z so that (P(Z<z)=1.0000\)).
22. Scores on a national exam are normally distributed with mean 382 and standard deviation 26.
    a. Find the score that is the 50th percentile.
    b. Find the score that is the 90th percentile.
23. Heights of adults are normally distributed with mean 63.7 inches and standard deviation 2.47 inches.
    a. Find the height that is the 10th percentile.
    b. Find the height that is the 80th percentile.
24. The monthly amount of water used per household in a small community is normally distributed with mean $7,069$ gallons and standard deviation 58 gallons. Find the three quartiles for the amount of water used.
25. The quantity of gasoline purchased in a single sale at a chain of filling stations in a certain region is normally distributed with mean 11.6 gallons and standard deviation 2.78 gallons. Find the three quartiles for the quantity of gasoline purchased in a single sale.
26. Scores on the common final exam given in a large enrollment multiple section course were normally distributed with mean 69.35 and standard deviation 12.93. The department has the rule that in order to receive an A in the course his score must be in the top 10% of all exam scores. Find the minimum exam score that meets this requirement.
27. The average finishing time among all high school boys in a particular track event in a certain state is 5 minutes 17 seconds. Times are normally distributed with standard deviation 12 seconds.
    a. The qualifying time in this event for participation in the state meet is to be set so that only the fastest 5% of all runners qualify. Find the qualifying time. (Hint: Convert seconds to minutes.)
    b. In the western region of the state the times of all boys running in this event are normally distributed with standard deviation 12 seconds, but with mean 5 minutes 22 seconds. Find the proportion of boys from this region who qualify to run in this event in the state meet.
28. Tests of a new tire developed by a tire manufacturer led to an estimated mean tread life of $67,350$ miles and standard deviation of $1,120$ miles. The manufacturer will advertise the lifetime of the tire (for example, a "$50,000$ mile tire") using the largest value

for which it is expected that 98% of the tires will last at least that long. Assuming tire life is normally distributed, find that advertised value.

29. Tests of a new light led to an estimated mean life of $1,321$ hours and standard deviation of 106 hours. The manufacturer will advertise the lifetime of the bulb using the largest value for which it is expected that 90% of the bulbs will last at least that long. Assuming bulb life is normally distributed, find that advertised value.
30. The lengths X of hardwood flooring strips are normally distributed with mean 28.9 inches and standard deviation 6.12 inches. Strips whose lengths lie in the middle 80% of the distribution of lengths of all strips are classified as “average-length strips.” Find the maximum and minimum lengths of such strips. (These lengths are endpoints of an interval that is symmetric about the mean and in which the lengths of 80% of the hardwood strips lie.)
31. All students in a large enrollment multiple section course take common in-class exams and a common final, and submit common homework assignments. Course grades are assigned based on students' final overall scores, which are approximately normally distributed. The department assigns a C to students whose scores constitute the middle $2/3$ of all scores. If scores this semester had mean 72.5 and standard deviation 6.14, find the interval of scores that will be assigned a C.
32. Researchers wish to investigate the overall health of individuals with abnormally high or low levels of glucose in the blood stream. Suppose glucose levels are normally distributed with mean 96 and standard deviation 8.5 mg/dl, and that “normal” is defined as the middle 90% of the population. Find the interval of normal glucose levels, that is, the interval centered at 96 that contains 90% of all glucose levels in the population.
33. A machine for filling 2-liter bottles of soft drink delivers an amount to each bottle that varies from bottle to bottle according to a normal distribution with standard deviation 0.002 liter and mean whatever amount the machine is set to deliver.
    a. If the machine is set to deliver 2 liters (so the mean amount delivered is 2 liters) what proportion of the bottles will contain at least 2 liters of soft drink?
    b. Find the minimum setting of the mean amount delivered by the machine so that at least 99% of all bottles will contain at least 2 liters.
34. A nursery has observed that the mean number of days it must darken the environment of a species poinsettia plant daily in order to have it ready for market is 71 days. Suppose the lengths of such periods of darkening are normally distributed with standard deviation 2 days. Find the number of days in advance of the projected delivery dates of the plants to market that the nursery must begin the daily darkening process in order that at least 95% of the plants will be ready on time. (Poinsettias are so long-lived that once ready for market the plant remains salable indefinitely.)

Practice problems include parts of:

Introductory Statistics 2e by OpenStax is licensed under CC BY 4.0

Access for free at https://openstax.org/books/introductory-statistics-2e/pages/1-introduction

and

OpenIntro Statistics by David Diez, Christopher Barr, & Mine Çetinkaya-Rundel is licensed under CC BY 3.0

5.8: Practice (Chapter 5) is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

- **Current page** has no license indicated.
- **4.E: Discrete Random Variables (Exercises)** has no license indicated.
- **4.E: Discrete Random Variables (Exercises)** by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/introductory-statistics.
- **6.E: The Normal Distribution (Exercises)** by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/introductory-statistics.
- **5.E: Continuous Random Variables (Exercises)** has no license indicated.
- **Licensing** by Mathematics Department has no license indicated.

6: Sampling Distributions

Pharmaceutical companies are interested in knowing the rates of side effects among patients using medicines. It would be impossible to check if every single individual who would ever take that medicine develops side effects, but we can look at a smaller group of subjects and develop an idea of the frequency in this sample.

But every sample, even if collected fairly, comes with some **variation**. If we sampled again, we'd likely get a different result. So how do we make sense of those differences? That's what **sampling distributions** are designed to explain.

In this chapter, we shift to thinking not just about data, but about *statistics themselves as data*: the mean from a sample, the proportion who responded, even the spread. If we took lots of samples from the same population, what kind of pattern would the statistics follow? That pattern **is a distribution too**, we call it the **sampling distribution**.

Understanding sampling distributions helps us separate signal from noise. It gives us a way to measure how far off a sample stat might be from the true population parameter we *really* care about. And it leads us directly to one of the most important ideas in all of statistics: the **Central Limit Theorem**.

Over the next few sections, you'll:

- Understand the difference between a population, a sample, and a *sampling distribution*
- See how variation across samples leads to a new measure: the **standard error**
- Investigate what happens when we take **many** samples and how their means behave
- Discover how sample sizes impact accuracy and reliability
- Arrive at the **Central Limit Theorem**, the foundation behind confidence intervals and hypothesis testing

Sampling distributions are where the practice of statistics becomes the power of inference. By building up our understanding here, we'll set the stage for estimation, decision-making, and statistical reasoning throughout the rest of the course.

Let's zoom out and start thinking about statistics as data.

This page titled 6: Sampling Distributions is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Mathematics Department.

6.1: Sampling Distributions

So far, we've worked with samples to estimate characteristics of a population. And we've seen that statistics (like averages or proportions) can vary from sample to sample. How do we know if our sample is representative of the population? We can express a distribution of possible samples, which is fittingly called the **Sample distribution.**

Review: A Few Key Terms

- **Population:** The entire group we care about (e.g., all voters, all students at a college).
- **Sample:** A subset of the population that we actually measure (e.g., 100 randomly chosen voters).
- **Statistic:** A summary measure calculated from a sample (e.g., the sample mean $\bar{x}$, or sample proportion $\hat{p}$).
- **Parameter:** A summary measure from a population (e.g., the population mean μ, or population proportion p).

We use sample statistics to make guesses about underlying population parameters.

What is a Distribution of Samples?

Definition: Sampling Distribution

The **sampling distribution** of a statistic is the distribution of values that the statistic takes on in repeated random samples of the same size from the same population.

You can imagine if we repeated took a new random sample and recorded the statistic obtained, we would build up a histogram showing where the most likely statistics fall.

Sampling Distribution Simulation (n = 5)

This simulation draws random samples from a fixed population (sample size = 5). Samples are shown in the left. The distribution of sample means builds on the right.

Draw 1 Sample | Draw 10 Samples | Reset

Why Does This Matter?

Sampling distributions allow us to study the long-run behavior of sample statistics. They help us answer questions like:

- How much do sample means typically vary?
- How close is a sample proportion likely to be to the true population proportion?
- What's the probability of getting a certain sample result if nothing unusual is happening?

> **Think About It:** Imagine 10 different people take random samples of 25 students to estimate the average number of hours of sleep students at your school get per night. Do you expect them to get the same result? What do you think the distribution of their results would look like?

In the next section, we'll study the sampling distribution of the sample mean one of the most important and widely used statistics in all of data analysis.

This page titled 6.1: Sampling Distributions is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Mathematics Department.

6.2: Mean and Standard Deviation of Sampling Distributions

In the last section, we saw how we can build a **sampling distribution**: take many samples from the same population, calculate a statistic (like the mean) for each sample, and plot those values.

Now we focus more carefully on two key questions:

- What is the **center** of that sampling distribution?
- How much do the sample statistics typically **vary**? That is, how wide is the distribution?

To answer these, we revisit two ideas from earlier chapters: **mean** and **standard deviation**. But this time, we're not looking at raw data, we're looking at a *distribution of statistics* from many samples. And that leads us to two new terms: **unbiased estimators** and **standard error**.

Review: Mean and Standard Deviation

- **Mean:** The average of a set of values; center of a distribution
- **Standard Deviation (SD):** A measure of how spread out values are around the mean on average

In a sampling distribution, the values we're working with are not individual data points, they're **sample statistics**, like sample means from different samples. So when we talk about "the mean of all the sample means" or their spread, we're talking about the *sampling distribution of the statistic*.

Definition: Mean of the Sampling Distribution

The mean of the sampling distribution is the mean of all of the sample statistics from all possible samples.

For sample means, the mean of the sampling distribution is equal to the population mean:

$$\mu_{\bar{x}} = \mu \tag{6.2.1}$$

For sample proportions, the mean of the sampling distribution is equal to the population proportion:

$$\mu_{\hat{p}} = p \tag{6.2.2}$$

The sample mean and sample proportion are both called **unbiased estimators** since the mean of all the sample means matches the population mean and the mean of all the sample proportions matches the population proportion. The statistics "target" the population parameter. This means that *on average* the sample mean (and proportion) is correct, even though individual samples may vary. This is NOT true of other sample statistics like the range or median. The mean of all sample ranges may NOT match the population range.

This concept, known as **unbiasedness**, is important: it means that even though sample statistics fluctuate, the process doesn't consistently overestimate or underestimate the population value. Remember, not all statistics are unbiased!

Definition: Standard Error (SE)

The **standard error** is the standard deviation of a sampling distribution. It measures how much the sample statistic varies from sample to sample. We often abbreviate standard error as **SE**. However, it may also be written like $\sigma_{\bar{x}}$ for sample means, $\sigma_{\hat{p}}$ for sample proportions, or σ_s for sample standard deviations.

As an example, for sample means the standard error is:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \tag{6.2.3}$$

where n is the sample size and σ is the population standard deviation.

Note the n in the denominator of the fraction. As sample size increases, the standard error shrinks. Larger samples give more stable sample means. This is generally true for all sampling distributions, not just sample means, but this particular formula $\frac{\sigma}{\sqrt{n}}$ is specific to sample means. You can see an example of this plotted below.

Interpreting the Horizontal Axis

When we draw or simulate a sampling distribution (like we did in Section 6.1), the **x-axis now shows sample statistics**, not individual data points.

The **center of the distribution is** $\mu_{\bar{x}}$ or $\mu_{\hat{p}}$ or whichever statistic you're using.
The **spread of the distribution is** $\sigma_{\bar{x}}$, the standard error.

Nightly Sleep Amounts

Suppose that the average college student gets 6.8 hours of sleep per night with a population standard deviation is 1.1 hours. If we were studying student sleep habits, we would not know these actual population values! Instead, we would take a sample to determine an estimate.

We take a random sample of 25 students and compute this sample mean. What would the center and spread of the sampling distribution of sample means look like?

Mean: $\mu_{\bar{x}} = 6.8$ hours. The average sample mean is just the pre-existing population mean.

Standard Error: $\sigma_{\bar{x}} = \frac{1.1}{\sqrt{25}} = \frac{1.1}{5} = 0.22$ hours. We have to calculate the standard error by dividing by the sample size. This is because as the sample size goes up, the likelihood of having a sample mean further from the population mean decreases.

Conclusion: Most of the sample means (not individual sleep values!) would fall within 0.22 hours of 6.8 on average. Using our rule of thumb, we can say that with perfectly random sampling, there is a 68% of getting a sample within one standard deviation, or between 6.58 and 7.02 hours.

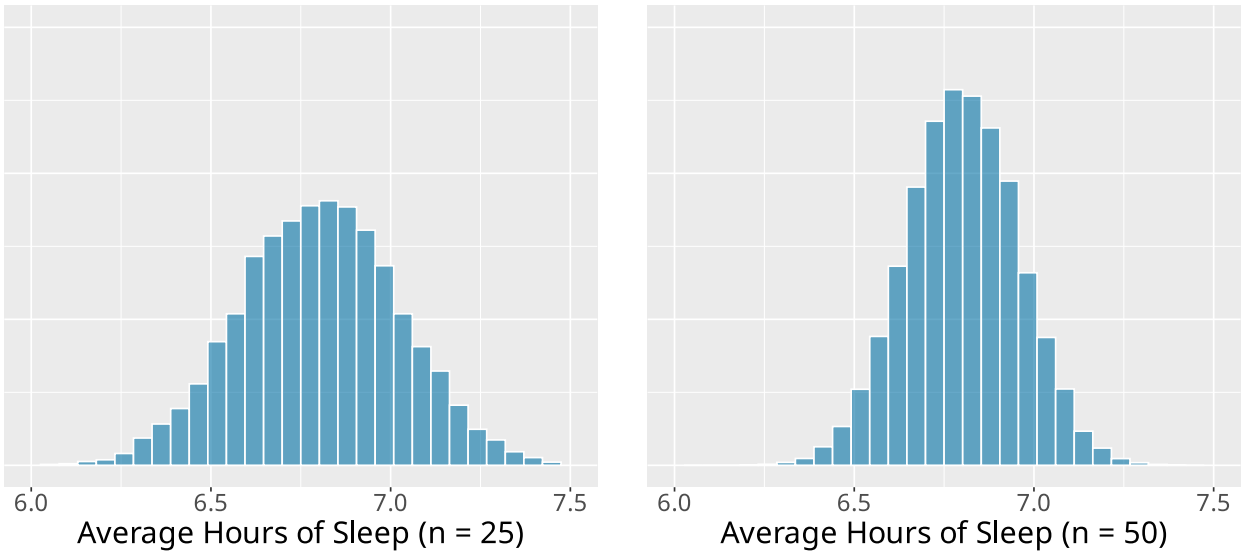
A Larger Sample

Let's suppose we change our minds and double our sample size for our sleep study. We now have $n = 50$ and can recompute the standard error.

$\sigma_{\bar{x}} = \frac{1.1}{\sqrt{50}} = \frac{1.1}{7.07} = 0.156$

Note that with a doubling of sample size, the standard error does **not** get twice as small. This is still helpful however, as our 68% area would now span between 6.64 and 6.96, which is closer to the true mean.

Distribution of Sample Means for Different Sample Sizes



The plotted histograms show the theoretical results from sampling 25 or 50 students repeatedly and tracking what the average amount of sleep for the sample is. To generate this image, we used a computer to randomly sample 20,000 times. This means for the image on the left, there are 20,000 samples, each with 25 students for a total of 500,000 data points! In reality, we would take one sample, of 25 students, and have no idea where it lies on the distribution. Note that it is more likely to like close to the real mean, and with a larger sample that probability increases.

Connecting to Chapter 5

Back in Chapter 5, we plotted raw data and observed shapes and spread. That was the **distribution of the data**.

Here in Chapter 6, we are plotting **statistic values from many samples**. It's a shift in thinking:

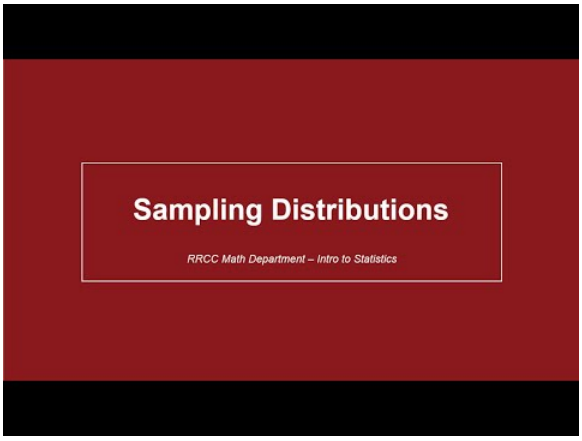
- Before: Each point = one person's data value
- Now: Each point = one sample's *mean* (or another statistic)

There are two major points to consider however:

1. In reality, we don't know the population values and therefore cannot calculate our sample distribution. Instead, we have a sample that came from that distribution.
2. We don't take multiple samples to build the distribution. It's a tool that is used to assess a single sample that comes from the distribution.

Related Video





This page titled 6.2: Mean and Standard Deviation of Sampling Distributions is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Mathematics Department.

6.3: Central Limit Theorem- Meaning and Implications

The **Central Limit Theorem (CLT)** is one of the most powerful and surprising results in all of statistics. It explains why sampling distributions are often bell-shaped even when the original population isn't.

Definition: Central Limit Theorem

The **Central Limit Theorem** (CLT) states the following:

If we take many random samples of size n from any population (with mean μ and standard deviation σ), then:

The sampling distribution of the sample mean $\bar{x}$ will follow an approximately normal distribution with mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$,
getting close to normal as the sample size n increases.

A few important notes:

- This works **regardless** of the original population distribution.
- The bigger the sample size, the more normal the sampling distribution becomes.

What Does This Do?

The CLT tells us that the average of the sample we take comes from a known distribution. We can then try to theoretically backtrack to the population

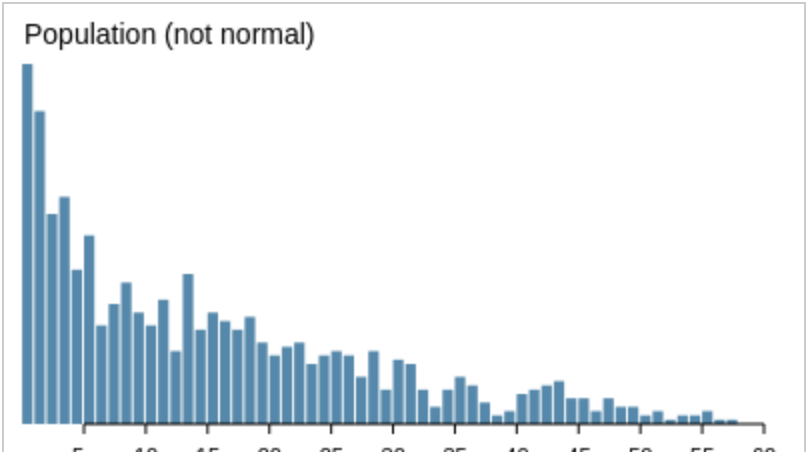
If we want to estimate a population mean (like average home price or GPA), the CLT tells us that our *sample* means have a predictable shape and center:

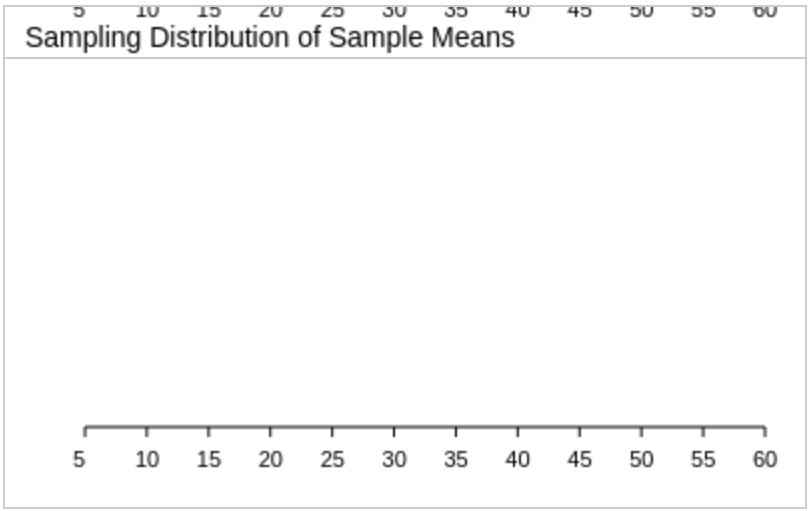
- Center: $\mu_{\bar{x}} = \mu$
- Spread: $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$
- Shape: approximately normal (if n is large enough)

This gives us the foundation to compute probabilities, find margins of error, and test whether sample results could reasonably occur under specific assumptions.

Central Limit Theorem: Simulation (n = 30)

This simulation draws samples from a right-skewed population and tracks the sample means. The histogram below builds into a bell shape, even though the original population is not normal.





Draw 1 Sample | Draw 10 Samples | Reset

Exploration: What Did You See?

Use the simulation above to explore how the Central Limit Theorem works in practice. As you interact with it, reflect on these questions:

- What does the **population distribution** look like? Is it symmetric? Skewed?
- After 10, 20, then 50 samples, what does the **sampling distribution of the mean** start to look like?
- What would happen if we used a larger sample size (n)? How might that affect shape and spread?
- How could this simulation help explain why it's reasonable to use normal probability models, even when the data itself isn't normal?

Remember, the sampling distribution of sample means becomes more normal *even when the population isn't.*

Example: Average Wait Times at a Coffee Shop

Suppose the distribution of wait times at a popular coffee shop is **right-skewed** with an average of 6 minutes and a standard deviation of 3 minutes.

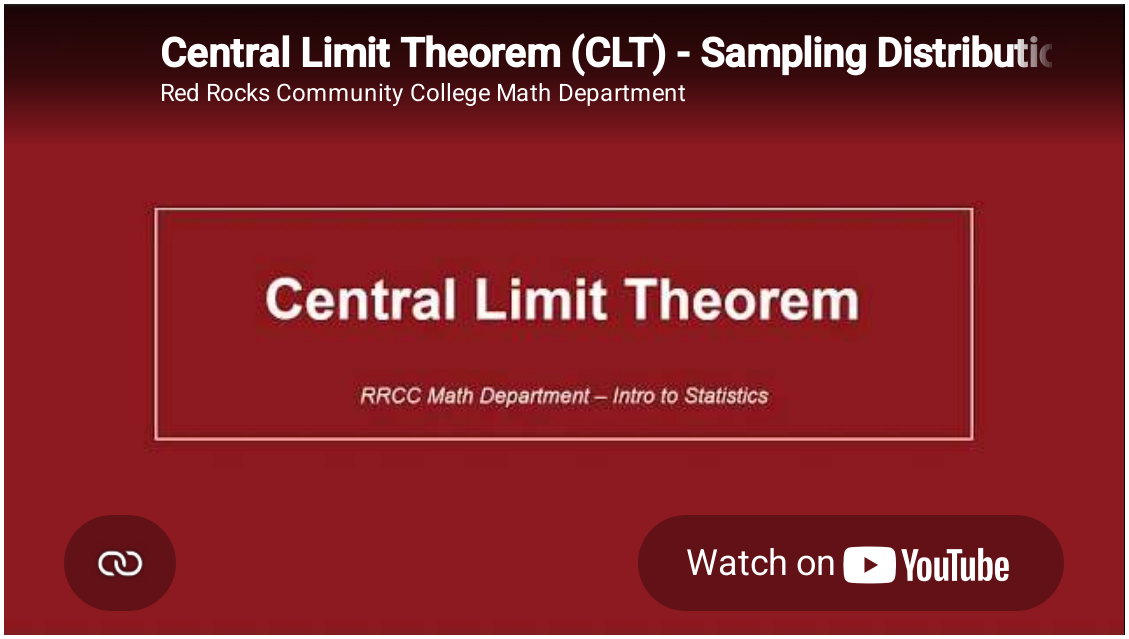
If we take a random sample of 50 customers and calculate the mean wait time, what will the **distribution of the sample mean** look like?

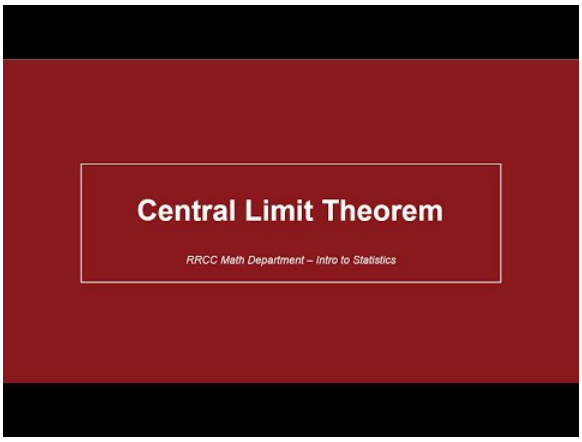
- The individual wait times are skewed
- But the **sampling distribution of the sample mean** will be approximately normal (thanks to the CLT)
- That distribution will be centered at 6 minutes with a standard error:

$$\sigma_{\bar{x}} = \frac{3}{\sqrt{50}} \approx 0.42 \tag{6.3.1}$$

This means that the **average wait time from a sample of 50 people** will usually fall within about 0.42 minutes of 6, and the pattern of those averages will resemble a bell curve, even though the individual data doesn't.

Related Video





In the next section, we'll begin to use the Central Limit Theorem to **solve problems**: calculating probabilities about sample means, building confidence intervals, and making decisions based on data.

This page titled 6.3: Central Limit Theorem- Meaning and Implications is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Mathematics Department.

6.4: CLT for Means and Proportions

In the last two sections, we explored what sampling distributions are and how they behave. The Central Limit Theorem (CLT) states:

Even if the original population isn't normally distributed, the sampling distribution of the sample mean will become approximately normal as sample size increases.

This idea doesn't just apply to means, but also works for **sample proportions**.

Two Types of Statistics

For now, we will discuss the two different statistics:

- **Sample Means**, $\bar{x}$
- **Sample Proportions**, $\hat{p}$

When we take a sample, there is an element of randomness. If we were to take repeated samples, we would see a **sampling distribution** for both mean and proportions start to emerge. This sampling distribution has the following properties:

- Centers around the population value (mean or proportion)
- Gets narrower as sample size increases
- Approximates a normal distribution given the below conditions

---

CLT Summary: Means vs. Proportions

Summary of Central Limit Theorem.

Statistic	Population Parameter	Mean of Sampling Distribution	Standard Error	Normality Conditions
Sample Mean $\bar{x}$	μ	μ	$\frac{\sigma}{\sqrt{n}}$	Approximately normal if $n \geq 30$, or if population normal
Sample Proportion $\hat{p}$	p	p	$\sqrt{\frac{p(1-p)}{n}}$	Approximately normal if $np \geq 10$ and $n(1-p) \geq 10$

What's different:

- The formulas for standard error
- The conditions required for "approximately normal"
- Notation!

What's the same:

- The general shape of the sampling distribution is normal (under the right conditions)
- The center equals the population parameter

Conceptual Comparisons

Example 1: Sample Mean – Commute Times

The population average commute time is 28 minutes with a standard deviation of 6 minutes. You take a sample of 40 commuters.

The sampling distribution of $\bar{x}$ will be approximately normal with:

- Center: 28 minutes
- Standard error: $\frac{6}{\sqrt{40}} \approx 0.95$ minutes

Try to see why this is a silly scenario! If we actually knew the population mean, there would be no need to sample. This simply illustrates that when we sample there is a quantifiable uncertainty.

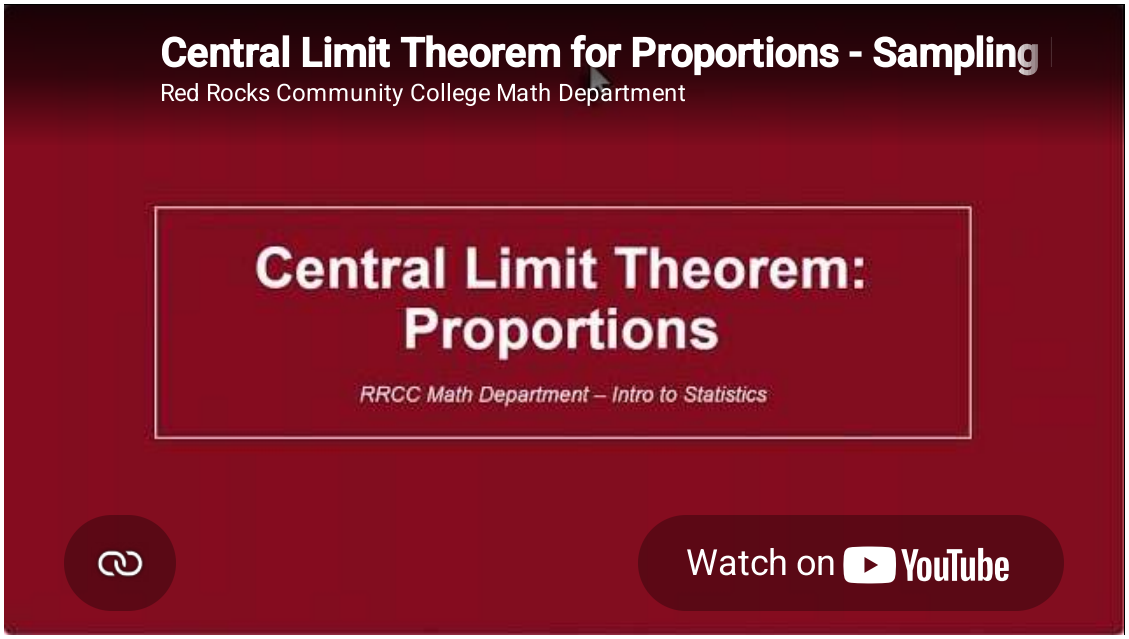
Example 2: Sample Proportion – Product Support

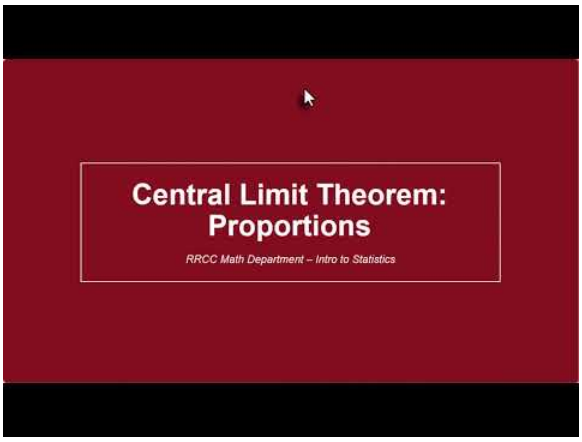
To consider a more realistic scenario, suppose that we already expect a proportion of 60% from past studies. We can utilize the CLT to compare with new data. If we resample with a sample size $n = 100$ we would have:

- Center: 0.60
- Standard error: $\sqrt{\frac{0.6(0.4)}{100}} = 0.049$
- The sampling distribution of $\hat{p}$ will be normal (since $np = 60$ and $n(1-p) = 40$)

Consider if we end up with a value such as 0.63 from our sample. This is within our sample error, and not too surprising. Instead, if we had a sample proportion of 0.7, this is a significantly different value than our original assumption that it may warrent a review!

Related Video





In the next chapter, we use this knowledge to create **confidence intervals** a way to estimate population values using sample statistics and error margins.

This page titled 6.4: CLT for Means and Proportions is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Mathematics Department.

6.5: Approximating Binomial with the Normal Distribution

6.5: Approximating the Binomial with the Normal Distribution

We've now seen that sample means form bell-shaped distributions under the Central Limit Theorem, even if the population isn't normal. The CLT also can help us with **categorical data**, especially when working with the **binomial distribution**.

Sometimes, calculating binomial probabilities directly (especially with large n) can be slow or impossible without technology. This is where the **normal approximation to the binomial** provides a valuable shortcut. Although access to computational power makes binomial calculations much easier today, this approximation was an incredibly valueable historical tool.

Logical Connection

Suppose we run a binomial experiment with:

- n: number of independent trials
- p: probability of success
- X: number of successes

The binomial random variable X is a sum of independent Bernoulli variables:

$$X = X_1 + X_2 + \cdots + X_n, \text{ where } X_i \sim \text{Bernoulli}(p)$$

The mean and standard deviation of a binomial distribution are:

- $\mu = np$
- $\sigma = \sqrt{np(1-p)}$

Since X is a **sum of many independent random variables**, the Central Limit Theorem tells us that that x is approximately normal, with corresponding means and standard deviations.

This approximation gets better as n increases, particularly it is advised to use this approximation when:

- $np \geq 10$
- $n(1-p) \geq 10$

Normal Approximation to the Binomial

If $X \sim \text{Binomial}(n, p)$, and $np \geq 10$, $n(1-p) \geq 10$, then:

$$X \sim N\left(np, \sqrt{np(1-p)}\right)$$

We can approximate $P(X \leq k)$ by computing the corresponding z-score:

$$z = \frac{k + 0.5 - np}{\sqrt{np(1-p)}} \tag{6.5.1}$$

We add the 0.5 as a *continuity corrector*, which guarantees a more accurate result when using a continuous distribution to approximate a discrete.

Example 1: Approximating a Binomial Probability

A fair die is rolled 60 times. What is the probability of getting at most 8 sixes?

- Let X be the number of sixes. Then $X \sim \text{Bin}(n = 60, p = 1/6)$
- $\mu = 60 \cdot \frac{1}{6} = 10$

- $\sigma = \sqrt{60 \cdot \frac{1}{6} \cdot \frac{5}{6}} \approx 2.89$
- Use continuity correction: $P(X \leq 8) \Rightarrow P(X \leq 8.5)$
- Standardize: $z = \frac{8.5-10}{2.89} \approx -0.52$
- From the normal table, $P(Z \leq -0.52) \approx 0.3015$

Answer: $P(X \leq 8) \approx 0.302$ using the normal approximation.

Using a binomial distribution, we find the actual probability to be $P(X \leq 8) = 0.312$.

Coming up in Chapter 7

In the next chapter, we'll use what we've learned to **construct confidence intervals,** helping us estimate unknown population parameters with a known degree of certainty.

This page titled 6.5: Approximating Binomial with the Normal Distribution is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Mathematics Department.

6.6: Vocabulary (Chapter 6)

Vocabulary: Sampling Distributions and Central Limit Theorem

Population

The entire group we are interested in studying or making conclusions about.

Sample

A subset of individuals taken from the population, used to estimate population characteristics.

Statistic

A number calculated from a sample (such as a sample mean or proportion) used to estimate a population parameter.

Sampling Distribution

The distribution of a sample statistic (such as a mean or proportion) based on repeated random samples of the same size. It shows how the statistic would vary across repeated samples.

Sample Mean ($\bar{x}$)

The average value of all data points in a sample. Used to estimate the population mean.

Sample Proportion ($\hat{p}$)

The proportion of observations in a sample that meet a certain condition. Used to estimate the population proportion.

Standard Error (SE)

The standard deviation of a sampling distribution. It measures how much a sample statistic typically varies from sample to sample.

Standard Error of the Mean

The standard deviation of the sampling distribution of the sample mean. Given by $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$.

Standard Error of the Proportion

The standard deviation of the sampling distribution of a sample proportion. Given by $\sqrt{\frac{p(1-p)}{n}}$.

Central Limit Theorem (CLT)

A theorem stating that the sampling distribution of the sample mean (or proportion) becomes approximately normal as the sample size increases, regardless of population distribution shape.

Normal Approximation to the Binomial

A method of using the normal distribution to approximate binomial probabilities when sample size is large and both np and $n(1-p)$ are at least 10.

Continuity Correction

An adjustment used when approximating a discrete binomial distribution with a continuous normal distribution. We add or subtract 0.5 to better align with the binomial probability being approximated.

This page titled 6.6: Vocabulary (Chapter 6) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Mathematics Department.

6.7: Practice (Chapter 6)

6.1: Sampling Distributions

1. Identify each of the following symbols
    a. μ
    b. $\hat{p}$
    c. σ
    d. $\bar{x}$
    e. p
    f. n
2. For each of the following situations, state whether the parameter of interest is a mean or a proportion. It may be helpful to examine whether individual responses are numerical or categorical.
    a. In a survey, one hundred college students are asked how many hours per week they spend on the Internet.
    b. In a survey, one hundred college students are asked: "What percentage of the time you spend on the Internet is part of your course work?"
    c. In a survey, one hundred college students are asked whether or not they cited information from Wikipedia in their papers.
    d. In a survey, one hundred college students are asked what percentage of their total weekly spending is on alcoholic beverages.
    e. In a sample of one hundred recent college graduates, it is found that 85 percent expect to get a job within one year of their graduation date.
3. For each of the following situations, state whether the parameter of interest is a mean or a proportion.
    a. A poll shows that 64% of Americans personally worry a great deal about federal spending and the budget deficit.
    b. A survey reports that local TV news has shown a 17% increase in revenue between 2009 and 2011 while newspaper revenues decreased by 6.4% during this time period.
    c. In a survey, high school and college students are asked whether or not they use geolocation services on their smart phones.
    d. In a survey, internet users are asked whether or not they purchased any Groupon coupons.
    e. In a survey, internet users are asked how many Groupon coupons they purchased over the last year.
4. Explain in your own words what a sampling distribution is.
5. How many samples of 2 can be made from a population of size 10?
6. How many samples of 5 can be made from a population of size 6?
7. What is the difference between the distribution of a population and a sampling distribution of a statistic?
8. Give an example of a situation where you would want to collect many samples and examine the distribution of their means or proportions.
9. Why can't we expect every random sample to exactly match the population?
10. If five different students each collect a sample and calculate a sample mean, will they get the same value? Why or why not?
11. A nonprofit wants to understand the fraction of households that have elevated levels of lead in their drinking water. They expect at least 5% of homes will have elevated levels of lead, but not more than about 30%. They randomly sample 800 homes and work with the owners to retrieve water samples, and they compute the fraction of these homes with elevated lead levels. They repeat this 1,000 times and build a distribution of sample proportions.
    a. What is this distribution called?
    b. Would you expect the shape of this distribution to be symmetric, right skewed, or left skewed? Explain your reasoning.
    c. If the proportions are distributed around 8%, what is the variability of the distribution?
    d. What is the formal name of the value you computed in (c)?
    e. Suppose the researchers budget is reduced, and they are only able to collect 250 observations per sample, but they can still collect 1,000 samples. They build a new distribution of sample proportions. How will the variability of this new distribution compare to the variability of the distribution when each sample contained 800 observations?
12. In a random sample 765 adults in the United States, 322 say they could not cover a $400 unexpected expense without borrowing money or going into debt.

b. What population is under consideration in the data set?
b. What parameter is being estimated?
c. What is the point estimate for the parameter?
d. What is the name of the statistic we use to measure the uncertainty of the point estimate?
e. Compute the value from part (d) for this context.
f. A cable news pundit thinks the value is actually 50%. Should she be surprised by the data?
g. Suppose the true population value was found to be 40%. If we use this proportion to recompute the value in part (e) using p = 04 instead of p, does the resulting value change much?

13. Suppose you are conducting an experiment where you have 10 trials. In each trial, you flip a coin 4 times.

a. Complete the following table by determining the proportion of heads for each sample

Sample	Sample Proportion $\hat{p}$
H H H T	
H H T H	
H T H T	
T T T T	
T H T T	
H T T T	
H T H T	
T T H T	
H H H H	
T H T T	

b. Now complete the following table for the sampling distribution for the sample proportion

Sample Proportion $\hat{p}$	Frequency	$P(\hat{p})$
0		
0.25		
0.5		
0.75		
1		

14. Consider the list of numbers {2,3,4}.

a. Complete the following table by writing the 9 possible combinations of choosing two numbers, with replacement, from this list and the corresponding sample mean.

Sample	Sample Mean $\bar{x}$
2, 2	
2, 3	

b. Now complete the following table for the sampling distribution for the sample proportion

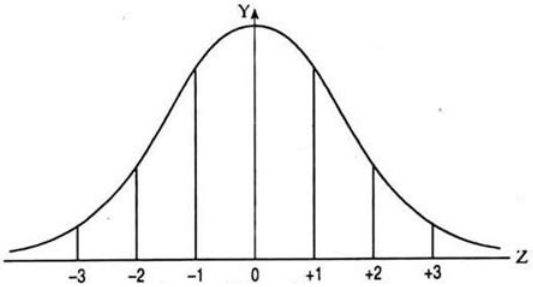
Sample Mean $\bar{x}$	Frequency	$P(\bar{x})$
2		
2.5		
3		
3.5		
4		

15. The closing stock prices of 35 U.S. semiconductor manufacturers are given as follows: 8.625, 30.25, 27.625, 46.75, 32.875, 18.25, 5, 0.125, 2.9375, 6.875, 28.25, 24.25, 21, 1.5, 30.25, 71, 43.5, 49.25, 2.5625, 31, 16.5, 9.5, 18.5, 18, 9, 10.5, 16.625, 1.25, 18, 12.87, 7, 12.875, 2.875, 60.25, 29.25.
    a. Construct a histogram of the distribution of the averages. Start at x = –0.0005. Use bar widths of ten.
    b. In words, describe the distribution of stock prices.
    c. Randomly average five stock prices together. (Use a random number generator.) Continue averaging five pieces together until you have ten averages. List those ten averages.
    d. Construct a histogram of the distribution of the averages. Start at x = -0.0005. Use bar widths of ten.
    e. Does this histogram look like the graph in part c?
    f. In one or two complete sentences, explain why the graphs either look the same or look different?

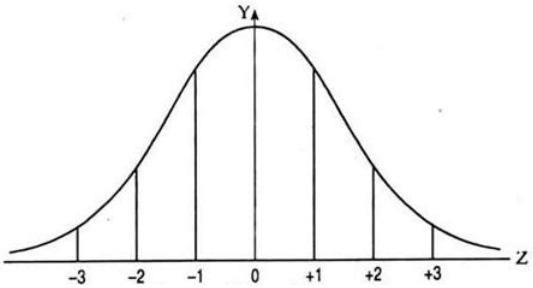
6.2: Mean and Standard Deviation of Sampling Distributions

1. Identify each of the following symbols
    a. $\mu_{\hat{p}}$
    b. $\mu_{\bar{x}}$
    c. $\sigma_{\hat{p}}$
    d. $\sigma_{\bar{x}}$
2. Determine if the following statements are True or False:
    a. The standard deviation for the distribution of sample means is the same as the standard deviation of the parent distribution.
    b. For any distribution with mean μ and standard deviation σ, if samples of size n are taken then the mean of the sampling distribution of the means is also μ.
    c. The standard deviation for the distribution of sample means will get smaller as the sample size gets larger.
    d. As the sample size increases, the mean of the sampling distribution increases.
    e. For sample proportions, the mean of the sampling distribution is equal to the population mean.

3. We say that a point estimator is **unbiased** if (choose any):
    a. its sampling distribution is centered exactly at the parameter it estimates.
    b. the standard deviation of its sampling distribution decreases as the sample size increases.
    c. its sampling distribution is normal.
    d. its value is always equal to the parameter it estimates.
4. Which point estimators are unbiased? Which are biased?
5. How do you determine the mean and standard deviation of the sampling distribution for sample means? Write out the exact formulas and describe the process for using them.
6. A population has a mean of 120 and a standard deviation of 20. What would be the expected mean and standard error of the sample mean based on samples of size 25?
7. Why is the standard deviation of the sampling distribution smaller than the standard deviation of the population from which it came?
8. Explain (with an example) why sample means tend to cluster around the population mean.
9. What does the horizontal axis of a sampling distribution represent?
10. Suppose that the distance of fly balls hit to the outfield (in baseball) is normally distributed with a mean of 250 feet and a standard deviation of 50 feet. We randomly sample 49 fly balls.
    a. What is the mean of the distribution of sample means?
    b. What is the standard deviation of the distribution of sample means?
11. A nonprofit wants determines the fraction of households that have elevated levels of lead in their drinking water is 23%. They randomly sample 800 homes and work with the owners to retrieve water samples, and they compute the fraction of these homes with elevated lead levels. They repeat this 1,000 times and build a distribution of sample proportions.
    a. What is the mean of the distribution of sample proportions?
12. According to the Internal Revenue Service, the average length of time for an individual to complete (keep records for, learn, prepare, copy, assemble, and send) IRS Form 1040 is 10.53 hours (without any attached schedules). The distribution is unknown. Let us assume that the standard deviation is two hours. Suppose we randomly sample 36 taxpayers.
    a. What is the mean of the distribution of sample means?
    b. What is the standard deviation of the distribution of sample means?
13. Let x be normally distributed with mean $\mu = 26$ and standard deviation $\sigma = 6$. Use the given mean and standard deviation to label the empirical rule graphs below.
    a. The first graph represents the x distribution



b. The second graph represents the $\bar{x}$ distribution. Construct the $\bar{x}$ distribution using samples of size $n = 49$.



14. NeverReady batteries has engineered a newer, longer lasting AAA battery. The company claims this battery has an average life span of 17 hours with a standard deviation of 0.8 hours. Your statistics class questions this claim. As a class, you randomly select 30 batteries and find that the sample mean life span is 16.7 hours.
    a. If the process is working properly, what is the probability of getting a random sample of 30 batteries in which the sample mean lifetime is 16.7 hours or less?
    b. Is the company's claim reasonable?
15. Your company has a contract to perform preventive maintenance on thousands of air-conditioners in a large city. Based on service records from previous years, the time that a technician spends servicing a unit averages one hour with a standard deviation of one hour. In the coming week, your company will service a simple random sample of 70 units in the city. You plan to budget an average of 1.1 hours per technician to complete the work. Will this be enough time?
16. John is shopping for wireless routers and is overwhelmed by the number of available options. In order to get a feel for the average price, he takes a random sample of 75 routers and finds that the average price for this sample is $75 and the standard deviation is $25.
    a. Based on this information, how much variability should he expect to see in the mean prices of repeated samples, each containing 75 randomly selected wireless routers?
    b. A consumer website claims that the average price of routers is $80. Is a true average of $80 consistent with John's sample?
17. Students are asked to count the number of chocolate chips in 22 cookies for a class activity. They found that the cookies on average had 14.77 chocolate chips with a standard deviation of 4.37 chocolate chips.
    a. Based on this information, about how much variability should they expect to see in the mean number of chocolate chips in random samples of 22 chocolate chip cookies?
    b. The packaging for these cookies claims that there are at least 20 chocolate chips per cookie. One student thinks this number is unreasonably high since the average they found is much lower. Another student claims the difference might be due to chance. What do you think?

6.3: Central Limit Theorem – Meaning and Implications

1. State the Central Limit Theorem (CLT) in your own words.
2. Why is the Central Limit Theorem important in statistics?
3. What are the requirements on sampling and the population so that the distribution of sample means is approximately normal?
4. When can we say that the sampling distribution of $\bar{x}$ is approximately normal?
5. A population is heavily skewed. What sample size is likely sufficient for the CLT to apply?
6. Give an example of a situation where the CLT helps you use a normal model for inference.
7. Suppose that a category of world-class runners are known to run a marathon (26 miles) in an average of 145 minutes with a standard deviation of 14 minutes. Consider 49 of the races. Let the average of the 49 races.
    a. Does the CLT apply to this situation?
    b. Describe the sampling distribution
    c. Would you be surprised if a runner completed a marathon in 100 minutes?
8. The length of songs in a collector's iTunes album collection is uniformly distributed from two to 3.5 minutes. Suppose we randomly pick five albums from the collection. There are a total of 43 songs on the five albums.
    a. Does the CLT apply to this situation?

b. Describe the sampling distribution
c. Would it be likely to find a song that is less than 1 minute long?

9. The attention span of a two-year-old is exponentially distributed with a mean of about eight minutes. Suppose we randomly survey 60 two-year-olds.
   a. Explain why the sampling distribution is not exponential
   b. Without calculating, which do you think will be higher and why?
      1. The probability that an individual attention span is less than ten minutes.
      2. The probability that the average attention span for the 60 children is less than ten minutes.
10. Lifespans of televisions have a mean of 4.7 years and a standard deviation of 1.2 years. Samples of 36 televisions are randomly selected. Will the distribution of sample means be normally distributed? Why or why not?
11. For students who borrow money to pay for college, the amount of debt for college graduates is normally distributed with mean $37,172 and standard deviation $11,536. Assume that samples of 10 college graduates are randomly selected. Is the distribution of sample means normal? Why or why not?
12. A random variable is not normally distributed, but it is mound shaped. It has a mean of 18 and a standard deviation of 3.
    a. If you take a sample of size 13, can you say what the shape of the sampling distribution for the sample mean is? Why?
    b. For a sample of size 13, state the mean of the sample mean and the standard deviation of the sample mean. Round to two decimal places if necessary.
    c. If you take a sample of size 37, can you say what the shape of the distribution of the sample mean is? Why?
    d. For a sample of size 37, state the mean of the sample mean and the standard deviation of the sample mean. Round to two decimal places if necessary.
13. The distribution of weights of US pennies is approximately normal with a mean of 2.5 grams and a standard deviation of 0.03 grams.
    a. What is the probability that a randomly chosen penny weighs less than 2.4 grams?
    b. Describe the sampling distribution of the mean weight of 10 randomly chosen pennies.
    c. What is the probability that the mean weight of 10 pennies is less than 2.4 grams?
    d. Sketch the two distributions (population and sampling) on the same scale.
14. Could you estimate the probabilities from and if the weights of pennies had a skewed distribution?~A manufacturer of compact fluorescent light bulbs advertises that the distribution of the lifespans of these light bulbs is nearly normal with a mean of 9,000 hours and a standard deviation of 1,000 hours.
    a. What is the probability that a randomly chosen light bulb lasts more than 10,500 hours?
    b. Describe the distribution of the mean lifespan of 15 light bulbs.
    c. What is the probability that the mean lifespan of 15 randomly chosen light bulbs is more than 10,500 hours?
    d. Sketch the two distributions (population and sampling) on the same scale.
    e. Could you estimate the probabilities from parts and if the lifespans of light bulbs had a skewed distribution?

6.4: CLT for Means and Proportions

1. Complete the following table for the difference between sample mean and sample proportion:

Empty table for summarizing the Central Limit Theorem.

	Sample Mean ($\bar{x}$)	Sample Proportion ($\hat{p}$)
Population parameter	?	?
Mean of sampling distribution	?	?
Standard error formula	?	?

2. If a population has proportion $p = 0.6$, and you collect a sample of size 100, compute the expected mean and standard error of the distribution of sample proportions.
3. How do conditions for normal approximation differ between means and proportions?
4. Explain why sampling distributions are usually narrower than the population distribution.

5. Determine if the following sampling distributions are normal.
    a. n=11 p=0.22
    b. n=92 p=0.59
    c. n=48 p=0.19
6. It is known that 64% of Americans personally worry a great deal about federal spending and the budget deficit. Which of the following needs to be satisfied in order to be able to use the Central Limit Theorem to describe the sampling distribution of the proportion of Americans who personally worry a great deal about federal spending and the budget deficit?
    a. The sample needs to be random.
    b. If sampling without replacement, the sample size needs to be less than 10% of the population size.
    c. The sample size should not be less than 50.
    d. The expected numbers of those who do and do not personally worry a great deal about federal spending and the budget deficit need to both be at least 10.
7. Vegetarian college students. Suppose that 8% of college students are vegetarians. Determine if the following statements are true or false, and explain your reasoning.
    a. The distribution of the sample proportions of vegetarians in random samples of size 60 is approximately normal since $n \geq 30$.
    b. The distribution of the sample proportions of vegetarian college students in random samples of size 50 is right skewed.
    c. A random sample of 125 college students where 12% are vegetarians would be considered unusual.
    d. A random sample of 250 college students where 12% are vegetarians would be considered unusual.
    e. The standard error would be reduced by one-half if we increased the sample size from 125 to 250.
8. Suppose that 90% of orange tabby cats are male. Determine if the following statements are true or false, and explain your reasoning.
    a. The distribution of sample proportions of random samples of size 30 is left skewed.
    b. Using a sample size that is 4 times as large will reduce the standard error of the sample proportion by one-half.
    c. The distribution of sample proportions of random samples of size 140 is approximately normal.
    d. The distribution of sample proportions of random samples of size 280 is approximately normal.
9. Yoonie is a personnel manager in a large corporation. Each month she must review 16 of the employees. From past experience, she has found that the reviews take her approximately four hours each to do with a population standard deviation of 1.2 hours. Let X be the random variable representing the time it takes her to complete one review. Assume X is normally distributed. Let $\bar{X}$ be the random variable representing the mean time to complete the 16 reviews. Assume that the 16 reviews represent a random set of reviews.
    a. What is the mean, standard deviation, and sample size?
    b. Find the probability that **one** review will take Yoonie from 3.5 to 4.25 hours. Sketch the graph, labeling and scaling the horizontal axis. Shade the region corresponding to the probability.
    c. Find the probability that the **mean** of a month's reviews will take Yoonie from 3.5 to 4.25 hrs. Sketch the graph, labeling and scaling the horizontal axis. Shade the region corresponding to the probability.
    d. What causes the probabilities in parts B and C to be different?
    e. Find the 95th percentile for the mean time to complete one month's reviews. Sketch the graph.
10. The distribution of income in some Third World countries is considered wedge shaped (many very poor people, very few middle income people, and even fewer wealthy people). Suppose we pick a country with a wedge shaped distribution. Let the average salary be $2,000 per year with a standard deviation of $8,000. We randomly survey 1,000 residents of that country.
    a. Describe the sampling distribution of sample means
    b. How is it possible for the standard deviation to be greater than the average?
    c. Why is it more likely that the average of the 1,000 residents will be from $2,000 to $2,100 than from $2,100 to $2,200?
11. The average length of a maternity stay in a U.S. hospital is said to be 2.4 days with a standard deviation of 0.9 days. We randomly survey 80 women who recently bore children in a U.S. hospital.
    a. Describe the sampling distribution of sample means

b. Is it likely that an individual stayed more than five days in the hospital? Why or why not?
c. Is it likely that the average stay for the 80 women was more than five days? Why or why not?
d. Which is more likely:
    1. An individual stayed more than five days.
    2. The average stay of 80 women was more than five days.
e. If we were to sum up the women's stays, is it likely that, collectively they spent more than a year in the hospital? Why or why not?

12. Suppose that a market research firm is hired to estimate the percent of adults living in a large city who have cell phones. Five hundred randomly selected adult residents in this city are surveyed to determine whether they have cell phones. Of the 500 people surveyed, 421 responded yes - they own cell phones.
    a. Describe the sampling distribution of sample proportions.
    b. Is it likely that 85% of adult residents of this city have cell phones?
13. According to a 2009 Reader's Digest article, people throw away about 9% of what they buy at the grocery store. Assume this is the true proportion and you plan to randomly survey 174 grocery shoppers to investigate their behavior. What is the probability that the sample proportion does not exceed 0.17?
14. Based on historical data, your manager believes that 27% of the company's orders come from first-time customers. A random sample of 168 orders will be used to estimate the proportion of first-time-customers. What is the probability that the sample proportion is greater than than 0.21?
15. Suppose that the efficacy of a certain drug is 0.47. Consider the sampling distribution (sample size n = 93) for the proportion of patients cured by this drug. What is the probability that the sample proportion is between 40% and 50%?
16. A student polls their school to see if students in the school district are for or against the new legislation regarding school uniforms. They survey 600 students and finds that 480 are against the new legislation.
    a. Describe the sampling distribution of sample proportions.
    b. In a sample of 600 students, do we expect that more than 200 are for the new legislation?
17. A housing survey was conducted to determine the price of a typical home in Denver, CO. The mean price of a house was roughly $1.3 million with a standard deviation of $300,000. There were no houses listed below $600,000 but a few houses above $3 million.
    a. Is the distribution of housing prices in Denver symmetric, right skewed, or left skewed? Hint: Sketch the distribution.
    b. Would you expect most houses in Denver to cost more or less than $1.3 million?
    c. Can we estimate the probability that a randomly chosen house in Denver costs more than $1.4 million using the normal distribution?
    d. What is the probability that the mean of 60 randomly chosen houses in Denver is more than $1.4 million?
    e. How would doubling the sample size affect the standard error of the mean?
18. Each year about 1500 students take the introductory statistics course at Red Rocks Community College. This year, scores on the final exam are distributed with a median of 74 points, a mean of 70 points, and a standard deviation of 10 points. There are no students who scored above 100 (the maximum score attainable on the final) but a few students scored below 20 points.
    a. Is the distribution of scores on this final exam symmetric, right skewed, or left skewed?
    b. Would you expect most students to have scored above or below 70 points?
    c. Can we calculate the probability that a randomly chosen student scored above 75 using the normal distribution?
    d. What is the probability that the average score for a random sample of 40 students is above 75?
    e. How would cutting the sample size in half affect the standard error of the mean?

6.5: Approximating the Binomial with the Normal Distribution

1. Which of the following probabilities can be calculated using the normal approximation to binomial?
    a. A Gallup poll suggests that 63% of Americans do not have a great deal of confidence in the mass media to report the news fully, accurately, and fairly. What is the probability that in a random sample of 15 people 9 or more of them have confidence in the mass media?

   b. A clothing store offers store credit cards and only about 37% of the credit card holders are regular customers. If we were to randomly sample 10 store credit card holders to conduct a survey, what is the probability at most 4 of the sampled individuals would be regular customers?
   c. Roughly 600% of adult Americans drink alcohol. What is the probability that in a random sample of 30 people at least 10 are drinkers?
2. You are working with a binomial experiment with n=20 trials, and a fixed probability of success of p=15%. Can the binomial distribution associated with this experiment be approximated by a normal distribution?
3. You are working with a binomial experiment with n=30 trials, and a fixed probability of success of p=40%. Can the binomial distribution associated with this experiment be approximated by a normal distribution?
4. What two conditions must be satisfied before the binomial distribution can be approximated by the normal distribution?
5. A coin is flipped 100 times. Use the normal approximation to estimate the probability of getting fewer than 45 heads.
6. Explain the purpose of the continuity correction.
7. A basketball player hits 85% of their shots. They're going to take 50 free throws. Use the normal approximation to estimate the probability they make at least 40 shots.
8. Why might the normal approximation to the binomial be inaccurate when n is small or p is near 0 or 1?
9. In a city, 46 percent of the population favor the incumbent, Dawn Morgan, for mayor. A simple random sample of 500 is taken. Using the continuity correction factor, find the probability that at least 250 favor Dawn Morgan for mayor.
10. Tony runs a cell phone kiosk in the mall and has 248 potential customers each day. She has a 6% chance to make a sale to each customer. Use continuity correction to normally approximate the following probabilities.
    a. The probability that she makes at least 16 sales
    b. The probability that she makes less than 14 sales
11. Suppose in a local Kindergarten through 12th grade (K - 12) school district, 53 percent of the population favor a charter school for grades K through 5. A simple random sample of 300 is surveyed.
    a. Find the probability that at least 150 favor a charter school.
    b. Find the probability that at most 160 favor a charter school.
    c. Find the probability that more than 155 favor a charter school.
    d. Find the probability that fewer than 147 favor a charter school.
    e. Find the probability that exactly 175 favor a charter school.
12. Four friends, Janice, Barbara, Kathy and Roberta, decided to carpool together to get to school. Each day the driver would be chosen by randomly selecting one of the four names. They carpool to school for 96 days. Use the normal approximation to the binomial to calculate the following probabilities. Round the standard deviation to four decimal places.
    a. Find the probability that Janice is the driver at most 20 days.
    b. Find the probability that Roberta is the driver more than 16 days.
    c. Find the probability that Barbara drives exactly 24 of those 96 days.

Practice problems include parts of:

Introductory Statistics 2e by OpenStax is licensed under CC BY 4.0

Access for free at https://openstax.org/books/introductory-statistics-2e/pages/1-introduction

and

OpenIntro Statistics by David Diez, Christopher Barr, & Mine Çetinkaya-Rundel is licensed under CC BY 3.0

6.7: Practice (Chapter 6) is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

- **Current page** has no license indicated.
- **4.E: Foundations for Inference (Exercises)** by David Diez, Christopher Barr, & Mine Çetinkaya-Rundel has no license indicated. Original source: https://www.openintro.org/book/os.
- **Licensing** by Mathematics Department has no license indicated.

7: Confidence Intervals

Chapter 7: Confidence Intervals

Imagine you're trying to estimate the average rent for a one-bedroom apartment in your city. You can't look at every apartment listing in the entire market, so you take a sample say, 50 listings and calculate the average rent from just those. But here's the big question:

How close is your sample mean to the real population mean?

This is where **confidence intervals** come in. Instead of giving a single number as an estimate, a confidence interval gives a **range** one that likely contains the true population value. It's one of the most important ideas in modern statistics, because it helps us make informed guesses while acknowledging uncertainty.

> **Key idea:** A confidence interval is built around your sample statistic.
>
> It takes the form: **estimate ± margin of error**

Confidence intervals pop up everywhere: opinion polls, clinical studies, economic estimates, even weather forecasts. Instead of saying "we know the exact truth," we say something more honest (and useful): **"We're *this* confident that the truth lies somewhere in here."**

In this chapter, we'll explore:

- How to use a sample to estimate a population **parameter**
- The structure and interpretation of a **confidence interval**
- How to find critical values (like z^* and t^*)
- The difference between intervals for **proportions** and for **means**
- How sample size and confidence level affect the accuracy and width of intervals

---

This page titled 7: Confidence Intervals is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Mathematics Department.

7.1: Point Estimates vs. Interval Estimates

Imagine you're surveying your fellow students to understand how much sleep they are getting each night. You collect a random sample of 35 students and find that the average amount of sleep is 6.3 hours.

Would you say the average for *all* students is exactly 6.3 hours? Not necessarily; your sample of 35 students may have randomly selected more students who sleep less than average, giving you an underestimate of the true average.

There is no way to know this however, so it is better to report a range of values such as "We estimate the average is somewhere between 6.0 and 6.6 hours." There is a methodology to calculating this range of values that guarantees a probability of successfully including the true value in our interval.

Definition: Point Estimate

A **point estimate** is a single value calculated from a sample that is used to estimate a population parameter.

For example: the sample mean, $\bar{x}$, is a point estimate for the population mean, μ.

Definition: Interval Estimate

An **interval estimate** gives a range of values for a population parameter, rather than just one number. The most common example is a **confidence interval**.

Instead of saying "the average GPA is 2.95," we might say, "the average GPA is between 2.85 and 3.05."

Examples: Point and Interval Estimates

Here are a few examples of how both types of estimates show up in real data analysis.

Example 1: Proportion of Voters

You survey 600 voters, and 372 say they support Ballot Measure A. That's a sample proportion of $\hat{p} = \frac{372}{600} = 0.62$.

- **Point estimate:** 0.62 (sample proportion)
- **Interval estimate:** An interval may suggest a proportion of 0.58 to 0.66.

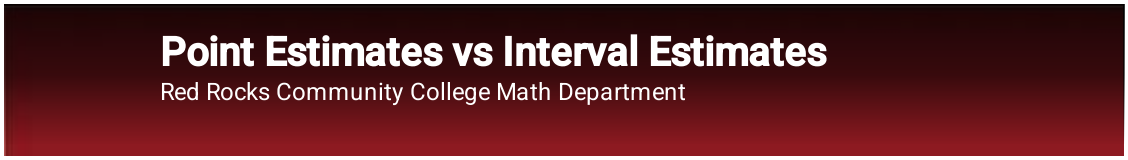
Example 2: Average Daily Water Use

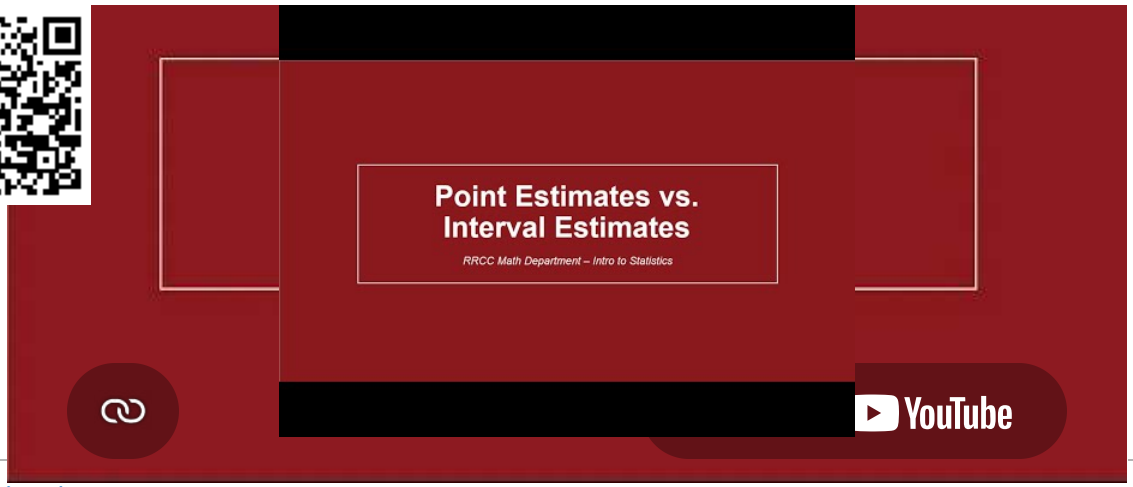
In a study of home water usage, a random sample of 40 households shows an average of 138 gallons per day with a standard deviation of 20 gallons.

- **Point estimate:** 138 gallons (sample mean)
- **Interval estimate:** An interval might be 131.8 to 144.2 gallons/day.

In this example, we could also construct an interval estimate for the standard deviation.

Related Video





Looking Ahead

In the next section, we'll introduce confidence intervals where we combine sample statistics and the standard error to create an interval estimate that has a specific level of confidence (like 95%).

This lets us say: "We're 95% confident that the true value lies in this range."

This page titled 7.1: Point Estimates vs. Interval Estimates is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Mathematics Department.

7.2: Confidence Intervals for a Population Proportion

Confidence Intervals for a Population Proportion

Imagine a psychology researcher is studying anxiety among college students. They survey a **random sample** of 200 students at several two-year colleges. The students complete the Generalized Anxiety Exam (GAD-7) where a score higher than 15 indicates "severe" levels of anxiety. In the sample of 200 students, 28 screen positive for severe anxiety, giving a proportion point estimate of $\hat{p} = 28/200 = 0.14$

The researcher wants to estimate the **true proportion** of all two-year college students that experience severe anxiety, not just the proportion in their sample. The researcher ensured that their sample is unbiased and representative, but they know the sample proportion will probably vary a little, simply due to randomness.

The researcher computes a **95% confidence interval** of $0.09 \leq p \leq 0.19$ which gives an **interval estimate** for the true population proportion. This confidence interval is defined using a technique that guarantees, if this study would be repeated over and over and over with a new random independent sample each time, then 95% of the the constructed confidence intervals would contain the true proportion. In simpler terms, we can say the probability of a random sample being selected that contains the true proportion is 0.95. Using this technique, our confidence interval does not guarantee that the true population value is in the interval, only that we have a certain level of **confidence** in the methodology used to calculate the interval.

We often times simply report that we are 95% confident that the interval $0.09 \leq p \leq 0.19$ contains the true population proportion.

This concept is exceedingly abstract! We will come back to the conceptual ideas after we actually practice the techniques to compute these for proportions.

Definition: Confidence Level

> A **confidence level** is the probability that we pick a random sample, assuming fairly and unbiasedly sampled, that will generate a confidence interval that contains the true population parameter. Common confidence levels are 0.9, 0.95 or 0.99.

This confidence level is chosen, not calculated. We decide upon a confidence level, and then build our confidence interval in a particular way to guarantee this result. It comes from the central limit theorem, where we know the sample distribution fits a normal distribution.

We define $\alpha = 1 - \text{confidence level}$, where α is the lowercase Greek letter **alpha**. For example, if our confidence level is 95%, then $\alpha = 1 - 0.95 = 0.05$

Different research areas have different conventions and expectations for confidence levels. Additionally, we may be limited in the useful confidence level we can have based on the size of our sample.

The Central Limit Theorem Returns

Recall from last chapter, the **Central Limit Theorem** tells us that the sampling distribution approaches a normal distribution as our sample size increases. If we theoretically knew the population proportion, we could center a normal distribution with mean p and standard deviation $\sqrt{p(1-p)/n}$ and determine the sample proportions that would lie at the boundaries of the middle 95%.

However, since we do not know p, we cannot do so. Instead, with the same logic, we could center the same normal distribution at $\hat{p}$. Since there is a 95% chance that our random selection of data yielded a sample proportion in the middle 95% around the unknown population value, then there is a 95% chance a same sized interval centered at $\hat{p}$ lies on top of the population value! This implies that we should try to find the specific values that give us the boundaries of 95% of a normal distribution centered at $\hat{p}$ with standard deviation $\sqrt{p(1-p)/n}$.

You may notice however, a big problem in the standard deviation; it relies upon knowing $\hat{p}$! The final piece of the puzzle is to approximate this standard deviation with our sample values as follows: $\sqrt{\hat{p}(1-\hat{p})/n} \approx \sqrt{p(1-p)/n}$.

Critical Values

A **critical value** is a z or t score that separates regions on a distribution. In our case, it allows us to specify where the boundaries of the confidence interval lie. We often denote the critical value as z^* or more specifically \(z_{\alpha/2\}\) which can be shown in the below diagram.

It is conventional to use the z score on the right of the distribution. If we are interested in the middle 95%, then this means each tail contains 2.5% of the distribution. Therefore we would use a critical value corresponding to the 97.5 percentile value of the normal distribution.

In general, a z-score is defined as the number of standard deviations from the mean:

$$z = \frac{x - mu}{\sigma} \tag{7.2.1}$$

In our case, we are placing the sample distribution around the sample proportion. Therefore we have:

$$z^* = \frac{p - \hat{p}}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \tag{7.2.2}$$

Since we are interested in knowing suggested bounds for p, we can solve for it algebraically:

$$p = \hat{p} + z^* \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \tag{7.2.3}$$

This would give us the upper bound, but due to the symmetry of the normal distribution, we can simply switch the $+$ to a $-$ to recover the lower bound. This is summarized below:

Wald Confidence Interval for a Population Proportion

A **Wald confidence interval for a population proportion** gives a range of values around our sample proportion such that the true population proportion p is captured at a probability of our confidence level. Our interval is given by:

$$\hat{p} - z^* \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < p < \hat{p} + z^* \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \tag{7.2.4}$$

- $\hat{p}$: sample proportion
- n: sample size
- z^*: critical value from the standard normal distribution for the desired confidence level

It is important to note that this confidence interval assumes three crucial things:

- Our sample was unbiased, representative, and random
- Our sample size is sufficiently large
- Our proportions are sufficiently far from extreme (0 or 1)

These last two conditions can be checked by ensuring that:

$$n\hat{p} \geq 10 \quad \text{and} \quad n(1-\hat{p}) \geq 10 \tag{7.2.5}$$

Example 1: Voter Poll

A random sample of 350 voters shows that 195 support Proposition A. Construct a 95% confidence interval for the true population proportion.

First we calculate our statistic: $\hat{p} = \frac{195}{350} \approx 0.557$

To even consider constructing this interval, we need to guarantee our conditions: $n\hat{p} = 195, n(1-\hat{p}) = 155$, so both are greater than 10.

Our standard error can be calculated as: $\sqrt{(0.557)(0.443)/350} \approx 0.0265$

Since we are using a 95% confidence interval, $alpha = 1 - 0.95 = 0.5$ and therefore we need to identify the z-score at the 97.5% percentile (\alpha/2\). We can use technology or a z-table to find our critical value: $z^* \approx 1.96$. It may be worth keeping some common critical values in your notes.

Finally, we calculate the confidence interval: $(0.557 - 0.052, 0.557 + 0.052) = (0.505, 0.609)$

To state this interval in words, we say: We are 95% confident that between 50.5% and 60.9% of all voters support Proposition A.

Example 2: Collegiate AI Usage

In a survey of 480 college students, it was found that 403 students reported using AI tools to help them with their classwork. Let us compute a 99% confidence interval around this statistic.

First we calculate the sample proportion of $\hat{p} = 0.84$ and check the validity of our confidence interval. Since $n\hat{p} = 403$ and $n(1-\hat{p}) = 77$, we are well within capability to ensure an accurate confidence interval. We next compute the standard error:

$$SE = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{0.84 \cdot 0.16}{480}} = 0.017 \tag{7.2.6}$$

This problem has proceeded as the previous, but now we are faced with a 99% confidence interval. This means that $\alpha = 1 - 0.99 = 0.01$. In intuitive terms, we are looking for the middle 99% of the interval, so $alpha$ represents the 1% to each of the tails. In order to use our z-table or spreadsheet formulas, we need the right endpoint, where 0.5% of the distribution will be to the right, and 99.5% to the left. Therefore we find the z-score corresponding to the 99.5 percentile of the standard normal, which is $z^* = 2.576$. Our final confidence interval is calculated as:

$$0.84 \pm (2.576) \cdot (0.017) \rightarrow 0.80 \leq p \leq 0.88 \tag{7.2.7}$$

Example 3: Defective Products

Suppose that you are a quality control engineer and you inspect 160 items and find that 7 are defective.

In this case, $n \cdot \hat{p} = 7$ which is too small to effectively trust the Wald confidence interval method. We would have to use a more complicated technique for calculating this confidence interval that falls outside the scope of the course.

Summary

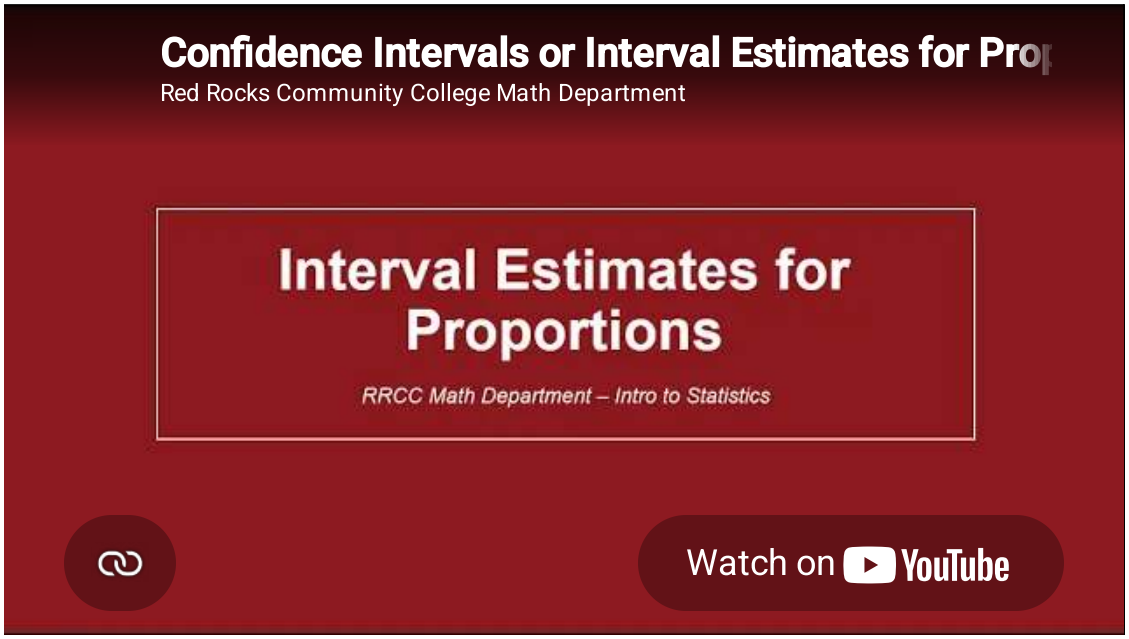
A **point estimate** $\hat{p}$ tells us our best single value guess for a population proportion. A **confidence interval** gives us a range of values with a carefully quantified likelihood of covering the true value. Our formula is:

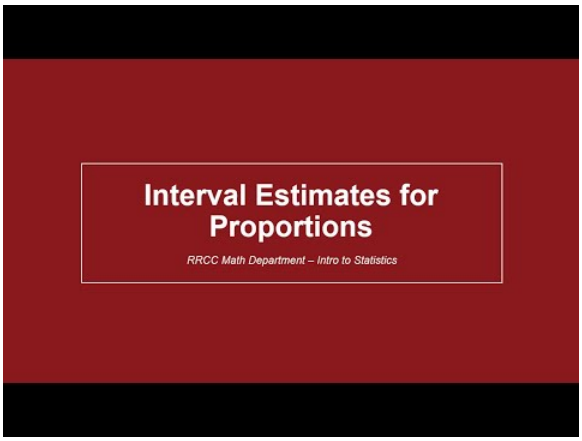
$$\hat{p} \pm z^* \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \tag{7.2.8}$$

We would convey this by saying for example: We are 95% confident that the true population proportion would be found inside our interval.

It's important to remember that the probability statement is about our sampling, where the real randomness occurs. The population value is fixed but unknown, so we talk about the likelihood of capturing it, not the likelihood of where it is.

Related Video





Up Next:

We can do a similar estimation process to estimate a population mean using sample data. You'll see that in the next section.

This page titled 7.2: Confidence Intervals for a Population Proportion is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Mathematics Department.

7.3: Confidence Intervals for a Population Mean

Suppose a health researcher is studying caffeine consumption among college students. She surveys a **random sample** of 20 students and finds they consume an average of 215 milligrams of caffeine per day, with a sample standard deviation of 48 mg/day. She wants to estimate the **true average** caffeine intake for all students on campus at 95% confidence. How can we do a similar estimation for means as we did with proportions in the last section?

The t-Distribution: When the Normal Curve Isn't Enough

Back in Chapter 6, we explored the Central Limit Theorem, which told us that the sampling distribution of sample means forms a normal distribution with certain characteristics. The mean of the sample means ($\mu_{\bar{x}}$) matched the population mean (μ), which is what we're trying to estimate. And the standard error ($\sigma_{\bar{x}}$ or SE) could be calculated based on the population standard deviation ($\frac{\sigma}{\sqrt{n}}$).

But in real-world data, we often don't know the population standard deviation. And sometimes, we're working with smaller sample sizes where the normality of the distribution might not be that precise. That's when the normal model can start to steer us wrong, it doesn't account for these sources of error. We need a better model for this situation.

Enter the **t-distribution**.

What Is the t-Distribution?

The **t-distribution** is a family of curves that look similar to the normal distribution, bell-shaped and symmetric, but with **heavier tails**. That means it gives more room for uncertainty, especially when sample sizes are small.

Just like the normal distribution is used with z-scores, the t-distribution is used with **t-scores**. We'll use it any time we're:

- Estimating a population mean and...
- ...we don't know the population standard deviation

Instead of using σ, we substitute the **sample standard deviation** s, and adjust for the extra variability that comes from estimating.

Degrees of Freedom (df)

There is more than one t-distribution, in fact, there's an infinite set! Each one is described by a number called **degrees of freedom** (usually noted as df).

When estimating a single mean, degrees of freedom equals the sample size minus one:

$$df = n - 1 \tag{7.3.1}$$

As your sample size increases, the t-distribution gets closer and closer to the normal distribution, the extra uncertainty becomes negligible when samples are large (usually $n > 30$).

Visual Insight

Here's how the t-distribution compares to the normal model:

- **Same center:** both are centered at 0
- **Symmetric shape:** both are bell curves
- **Heavier tails:** t-distributions allow for more extreme values, they're "wider" when n is small

[Insert diagram here comparing the standard normal curve and t-distributions with df = 5, 10, 30]
Lighter blue: normal distribution; darker curves: t-distributions with increasing degrees of freedom

Example: Estimating Caffeine Consumption

Consider the health researcher studying caffeine consumption mentioned earlier. Since she doesn't know the population standard deviation, **she will use the t-distribution**, not the normal curve. The degrees of freedom will be $df = 20 - 1 = 19$.

We'll walk through how to calculate confidence intervals with the t-distribution below. For now, just remember:

> *You can use the normal distribution when σ is known, use the t-distribution when σ is unknown and you must estimate with a sample.*

Definition: Confidence Interval for a Population Mean

A **confidence interval** for a population mean is a range of values built around a sample mean, showing the plausible values for the population mean μ.

When the population standard deviation σ is unknown and the sample size is modest, we use the t-distribution:

$$\bar{x} - t^* \left(\frac{s}{\sqrt{n}}\right) < \mu < \bar{x} + t^* \left(\frac{s}{\sqrt{n}}\right) \tag{7.3.2}$$

- $\bar{x}$: sample mean
- s: sample standard deviation
- n: sample size
- t^*: critical value from the t-distribution for the desired confidence level

In this formula, we have $\bar{x}$ which is the point estimate and $t^* \left(\frac{s}{\sqrt{n}}\right)$ which is the margin of error (ME) — and a range we can be reasonably confident contains μ.

What a Confidence Interval Really Means

A confidence interval provides a range of plausible values for a population parameter. It's not only about the calculation, it's about interpretation.

For example, suppose we calculate a 95% confidence interval for mean sleep to be (6.36, 7.44).

What does that actually mean?

We can say: **We are 95% confident that the true average number of hours all students sleep falls between 6.36 and 7.44 hours.**

This does *not* mean there's a 95% chance the true value is inside this specific interval, that's a common misunderstanding.

Instead: If we were to repeat this sampling method many times, 95% of the intervals we construct would contain the true mean. Confidence comes from the **method**, not the individual interval.

Interpretation Tips

When writing or reading a confidence interval conclusion, think carefully about these questions:

- What is the population you're estimating for?
- What parameter are you trying to learn about, a mean or a proportion?
- What does your interval say about that parameter?
- What level of confidence are you working with?

The more clearly you can communicate the variable, population, and level of confidence, the better your conclusion will be.

Choosing a Confidence Level

So far, we've been working with **95%** confidence intervals. That's a common choice, but it's not the only one. Confidence levels are flexible, and we choose them depending on the situation and how much uncertainty we're willing to accept.

Here are some common levels:

- **90% confidence:** narrower interval, less confidence
- **95% confidence:** a standard balance of precision and confidence
- **99% confidence:** wider interval, more confidence

The general rule is:

> Higher confidence → wider interval
> Lower confidence → narrower interval

When the consequences of being wrong are high (safety studies, medication trials, policy decisions), we often want higher confidence. We're willing to have a wider interval if it means we're more certain.

If we're doing a quick exploratory study or working with limited data, we may accept lower confidence in exchange for narrower, more informative ranges.

Thinking Critically About Confidence

- How much certainty do I need before I take action?
- How important is precision vs. accuracy in this case?
- What are the risks if I underestimate or overestimate the true value?

In later chapters, when we use intervals in hypothesis testing or decision-making, we'll return to this idea more formally.

Example 1: Sleep Hours Estimate

A sample of 25 students reports an average of 6.9 hours of sleep per night, with a standard deviation of 1.3 hours. Construct a 95% confidence interval for the average amount of sleep all students get.

- $\bar{x} = 6.9$, $s = 1.3$, $n = 25$, degrees of freedom = 24
- From the t-table, critical value for 95% confidence and df = 24 is $t^* \approx 2.064$
- Standard error: $\frac{1.3}{\sqrt{25}} = 0.26$
- Margin of error: $2.064 \cdot 0.26 = 0.537$
- Interval: $(6.9 - 0.54, 6.9 + 0.54) = (6.36, 7.44)$

Interpretation: We are 95% confident that the average sleep among all students is between 6.36 and 7.44 hours per night.

Example 2: Mean Rent for Off-Campus Housing

A city planner wants to estimate the average rent of apartments near campus. She samples 50 listings and finds a mean rent of $1,165 with a standard deviation of $120.

- $\bar{x} = 1165, s = 120, n = 50$, df = 49 → use $t^* \approx 2.009$
- Standard error: $\frac{120}{\sqrt{50}} \approx 16.97$
- Margin of error: $2.009 \cdot 16.97 \approx 34.08$
- Interval: $(1130.92, 1199.08)$

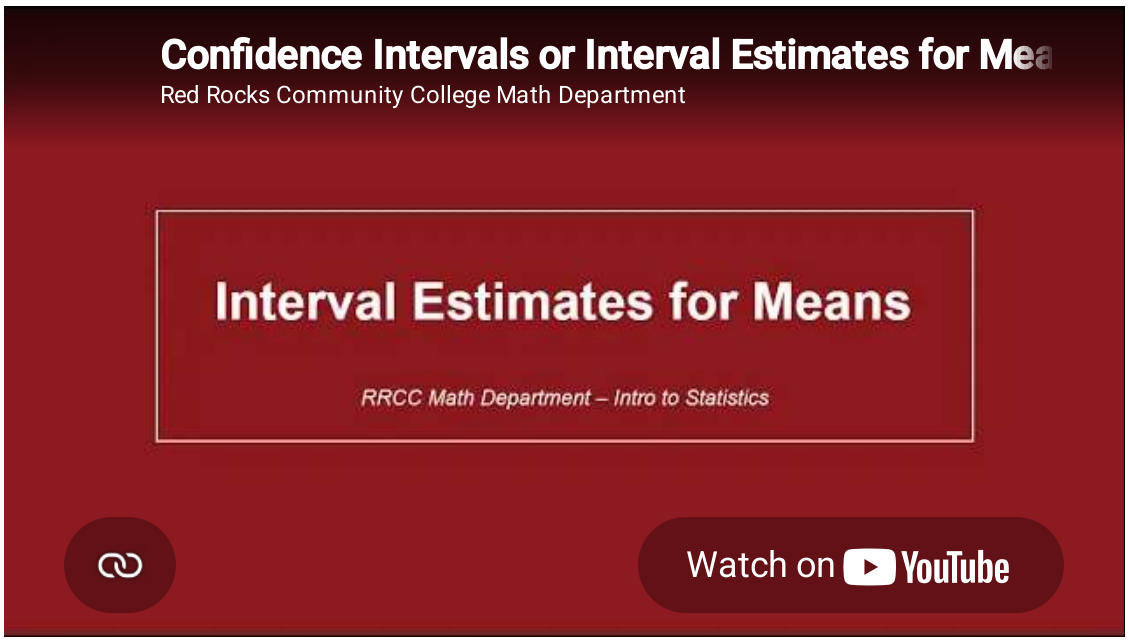
Conclusion: We are 95% confident that the true average rent in this area is between $1,131 and $1,199.

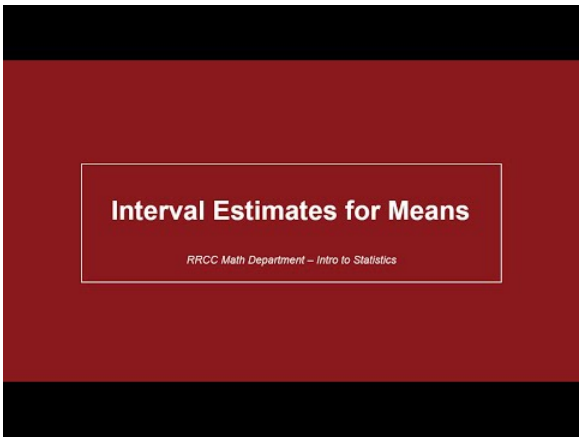
Reflection: Shifting the Statistic

So far, we've built confidence intervals based on the **sample mean** and on the **sample proportion**. In both cases we follow the same procedure (but with different calculations):

- We take a sample
- We compute a statistic (a mean or a proportion)
- We estimate a margin of error
- We build an interval that likely contains the population value

Related Video





Up Next:

In the next section, we'll take a closer look at the connections between the margin of error, the confidence level, and the sample size

---

This page titled 7.3: Confidence Intervals for a Population Mean is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Mathematics Department.

7.4: Effect of Sample Size and Confidence Level

Now that we've built and interpreted confidence intervals, it's time to reflect on something subtle but important: what affects how **wide or narrow** those intervals are?

In this section, we'll focus on two powerful factors:

1. The **confidence level** (like 90%, 95%, 99%)
2. The **sample size** used to build the interval

Higher Confidence Comes with a Cost

When we want to be more confident (say, 99% instead of 95%), we'll need to include a wider range of values. This makes sense to be more sure we've captured the true value, we must cast a wider net.

Here's how it looks with the same sample data:

- Confidence level = 90% → *narrower interval*
- Confidence level = 95% → *standard choice*
- Confidence level = 99% → *wider interval*

The **critical value** (z^* or t^*) increases as confidence increases and so does the *margin of error*.

Larger Samples Give Greater Precision

Sample size also plays a critical role. As the sample size increases:

- The **standard error** decreases
- The **margin of error** gets smaller
- The **confidence interval** becomes narrower

> **Reminder:** Standard error is affected by sample size:
>
> $$\text{SE} = \frac{s}{\sqrt{n}} \text{ for a mean} \quad \text{or} \quad \text{SE} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \text{ for a proportion}$$

Example 1: Different Confidence Levels

You survey 100 students about weekly screen time. The average is 23.4 hours/week with a standard deviation of 5.1 hours.

Compare the width of the following intervals:

- 90% confidence: $t^* \approx 1.660$, ME = 0.85 → Interval = (22.55, 24.25)
- 95% confidence: $t^* \approx 1.984$, ME = 1.02 → Interval = (22.38, 24.42)
- 99% confidence: $t^* \approx 2.626$, ME = 1.35 → Interval = (22.05, 24.75)

Note that the standard error (SE) is the same in all these calculations, only the critical value changes.

Conclusion: As confidence increases, the interval gets wider.

Example 2: Changing the Sample Size

The same data is collected, but from only 25 students. Now compare with:

- Mean = 23.4, SD = 5.1, n = 25 → SE = 1.02

Notice that while the sample standard deviation is the same, the smaller n means there is a *larger* standard error (SE). This makes the margin of error (ME) bigger as well.

- 95% CI: ME = $2.064 \cdot 1.02 \approx 2.11$, Interval = (21.29, 25.51)

If n=100 (the previous example). 95% CI: (22.38, 24.42)

Conclusion: A larger sample size gives a more precise or narrow interval with the same confidence level.

Interactive Demo: Confidence Interval Width

Adjust the sliders below to change the **sample size (n)** and **confidence level**. Watch how the margin of error and confidence interval width respond.

Sample size (n):

100

Confidence level:

95%

```
Sample mean: 50
Sample standard deviation: 15
Standard error:
Margin of error:
Confidence interval:
```

Check Your Understanding: Confidence Intervals

Question 1: Compared to a 90% confidence interval, a 99% confidence interval for the same data is...

- narrower
- wider
- about the same width

Question 2: Compared to a 90% confidence interval, a 99% confidence interval for the same data is...

- narrow
- widen
- stay the same

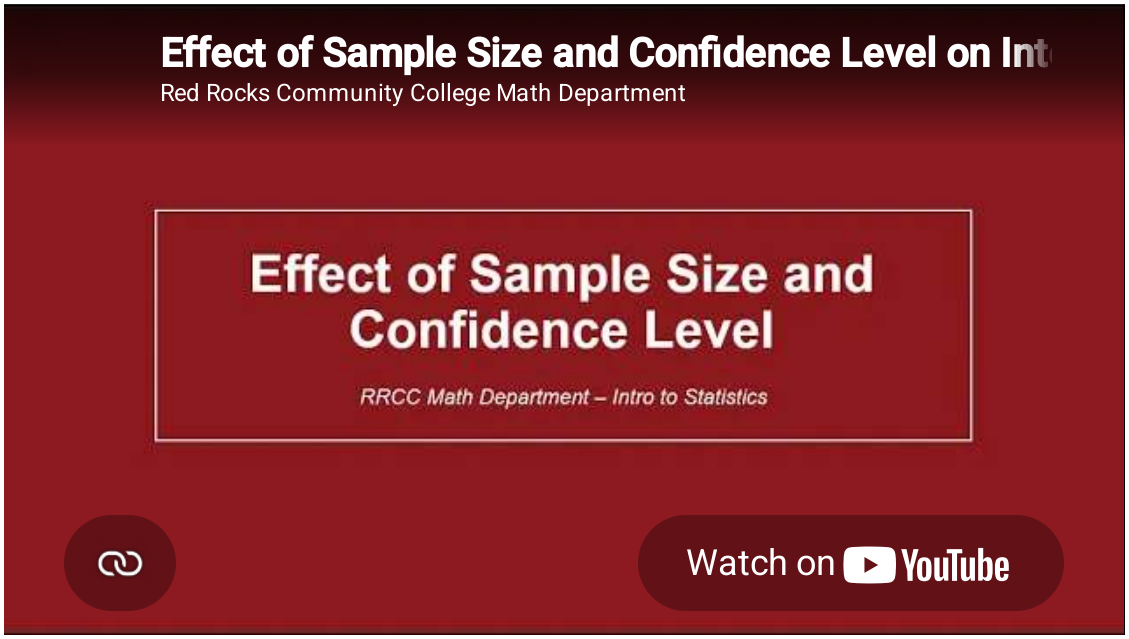
Why This Matters

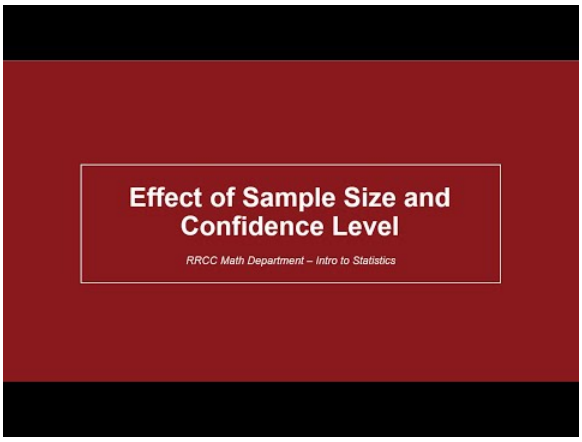
Different studies require different levels of precision and certainty. These tools help researchers design better surveys and polls. You'll use these relationships to:

- Plan studies with a target margin of error
- Decide how trustworthy an estimate may be
- Compare results responsibly

Now that we've built our understanding of confidence intervals, we're ready to apply them in new ways including hypothesis testing. But first, let's reflect on how we interpret and communicate these results effectively.

Related Video





This page titled 7.4: Effect of Sample Size and Confidence Level is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Mathematics Department.

7.5: Minimum Sample Size

Minimum Sample Size for Mean & Proportion Estimates

When planning a study, one critical question is: **How much data do we need?** Too small a sample may lead to inaccurate conclusions due to natural variability. On the other hand, collecting too much data wastes time and resources. Finding the right balance is key, and we can determine this through **sample size calculations** that achieve desired precision and confidence.

Let's explore how to calculate the minimum sample size required to estimate a population mean or proportion with a specified confidence level and margin of error.

Estimating a Population Mean

For estimating a mean, the minimum sample size depends on:

- The desired **margin of error (E)**: the maximum amount of error you're willing to accept in the estimate.
- The **confidence level** (determines the critical value, z*)
- An estimate of the population **standard deviation (σ)**

The formula to calculate the minimum sample size (n) is:

$$n = \left(\frac{z^* \cdot \sigma}{E}\right)^2 \tag{7.5.1}$$

Example: Calculating Sample Size for a Mean

Suppose a researcher wants to determine the average height of sunflowers with a margin of error of 2 cm at a 95% confidence level. If a preliminary study suggests that the standard deviation of sunflower heights is 10 cm, then:

1. Identify: $z^* \approx 1.96$ for 95% confidence
2. Calculate:

$$n = \left(\frac{1.96 \times 10}{2}\right)^2 = 96.04 \tag{7.5.2}$$

Therefore, the researcher needs a sample of at least 97 sunflowers (always round up!) to achieve the desired confidence and precision.

Estimating a Population Proportion

For estimating a proportion, the sample size formula also incorporates:

- The desired **margin of error (E)**
- The **confidence level** (determines the critical value, z*)
- Estimated proportion **($\hat{p}$)**: a guessed proportion based on past research or preliminary data

The minimum sample size formula for a proportion is:

$$n = \left(\frac{z^*}{E}\right)^2 \cdot \hat{p}(1-\hat{p}) \tag{7.5.3}$$

Example: Calculating Sample Size for a Proportion

Let's say a school administrator wants to know the proportion of students who participate in after-school activities with a margin of error of 5% at a 90% confidence level. If a pilot survey showed that about 40% have participated, then:

1. Identify: $z^* \approx 1.645$ for 90% confidence
2. Calculate:

$$n = \left(\frac{1.645}{0.05}\right)^2 \cdot 0.40(1-0.40) = 269.14 \tag{7.5.4}$$

Thus, the administrator needs a sample of at least 270 students to estimate the proportion accurately and precisely.

Note: If you don't have a preliminary estimate, use $\hat{p} = 0.5$ for conservative sample size determination.

Connecting to Real-World Studies

Careful sample size determination is crucial for any research study — whether you're estimating a population mean like average incomes in a region or determining a proportion like voter approval ratings. Appropriate calculations ensure you have enough data to back up your claims confidently.

Plan, calculate, and reassess as you prepare your research to align sample size goals with real-world limitations such as budget, time, or logistical constraints.

7.5: Minimum Sample Size is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

7.6: Interpreting Confidence Intervals

Now that we've learned how to build confidence intervals for means and proportions, it's time to focus on the most important part: **interpreting** what they tell us and what they don't.

Let's work through a real example where a confidence interval helps us answer a bigger question.

Example: Is This Coin Fair?

You're handed a coin that someone claims is **completely fair**. That is, the probability of heads is exactly 0.50.

You want to test this claim by flipping the coin 100 times and calculating the proportion of heads.

You do the test and count 57 heads out of 100 flips. That gives a sample proportion:

$$\hat{p} = \frac{57}{100} = 0.57$$

Step 1: Build a Confidence Interval

The sample proportion is 0.57. We'll use a 95% confidence level to see which values are plausible for the **true proportion of heads**, p.

First, calculate the standard error:

$$SE = \sqrt{\frac{0.57(1-0.57)}{100}} \approx 0.049$$

Find margin of error using $z^* = 1.96$:

$$ME = 1.96 \cdot 0.049 \approx 0.096$$

Confidence interval:

$$(0.57 - 0.096,\ 0.57 + 0.096) = (0.474,\ 0.666)$$

Step 2: Interpret the Interval

We are 95% confident that the **true proportion of heads** is between 0.474 and 0.666.

Now we ask: **Is 0.50 inside this interval?**

Yes! So we *cannot rule out* the possibility that the coin is fair, even though we saw 57 heads. The result is *not strong enough* to say the coin is biased.

The interval leaves open the possibility that $p = 0.5$, since that value is inside the range of likely population values.

Make a Reasoned Conclusion

- The sample proportion was 0.57 (more heads than expected)
- But sampling variability explains small surprises
- Since 0.50 is within the confidence interval → this could happen by chance
- **Conclusion:** No strong evidence the coin is unfair

This is the power of confidence intervals: they don't give us absolute truths, but they help us understand what values are reasonable based on our data.

> **Think About It:**
>
> - Suppose instead we had seen 62 heads out of 100. Would 0.50 still be inside the 95% confidence interval?
> - If you needed strong evidence to prove the coin wasn't fair, would you increase your sample size or your confidence level?

- How would this interval change if we had flipped the coin 1,000 times?

This page titled 7.6: Interpreting Confidence Intervals is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Mathematics Department.

7.6.1: Project Check-In 4

Project Check-In 4: Comparing Means with Confidence Intervals

You've now learned how to build a confidence interval for a sample mean — and in your housing affordability project, that opens up exciting possibilities for comparison!

TODO: Use your collected housing data to construct confidence intervals for the mean price in at least two different cities (or ZIP codes/regions) within your dataset. Then, place them side by side in a graph. You'll reflect both on the technical results and the story they tell.

What to Include:

- Construct **at least two** confidence intervals for average housing prices — each one based on a different group (e.g., city or ZIP code).
- Each confidence interval should include:
  - The sample mean
  - The sample size (n)
  - The sample standard deviation (s)
  - The margin of error and the level of confidence used (use 95% unless you have a reason to choose differently)
  - The resulting confidence interval: (lower bound, upper bound)
- **Graph your confidence intervals** on a number line or bar chart:
  - You can draw this by hand, or use Google Sheets, Excel, Desmos, or another tool
  - Label each interval clearly with its city or region
  - Make sure the graph's axis is scaled to allow visual comparison

Reflection Questions:

- Which group/region had the highest estimated mean housing price? Which one had the lowest?
- Do any of your confidence intervals **overlap**? What might that suggest?
- Is the difference between regions statistically significant? Why or why not?
- What could affect the **precision** of your intervals (i.e., what makes them wider or narrower)?
- How might sampling bias, variation, or small sample size impact your conclusions?

Visualization Example (Optional):

> Use a simple CI graph like this (hand-drawn or digital is fine):
> Midtown (95% CI): ——————————
> West Side (95% CI): ——————
> Suburbs (95% CI): ————————————————
>
> Tip: Align your intervals along a shared price scale so they're easy to interpret together.

Your Submission Should Contain:

- A short paragraph restating your question and what you set out to compare
- Each confidence interval and how it was calculated (with formulas and/or tech screenshots)
- A clearly labeled graph showing multiple confidence intervals
- Answers to the reflection questions above
- Any notes about limitations or interesting discoveries from your dataset

> ***Helpful Tip:*** *You can revisit the worked examples in Chapter 7.4 for step-by-step guidance on calculating confidence intervals using summary statistics. Copy a template, then adjust for each region.*

This page titled 7.6.1: Project Check-In 4 is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Mathematics Department.

7.7: Vocabulary (Chapter 7)

Chapter 7 Vocabulary

Point Estimate

A single value calculated from a sample that serves as the best guess for a population parameter. Example: the sample mean $\bar{x}$ is a point estimate of the population mean μ.

Interval Estimate

A range of values around a sample statistic that likely contains the population parameter. The most common type is a confidence interval.

Confidence Interval

An interval estimate built from a sample statistic plus and minus a margin of error. A 95% confidence interval means the method succeeds in capturing the population parameter about 95% of the time.

Confidence Level

The statistical confidence that the interval contains the population parameter. Common levels include 90%, 95%, and 99%.

Sample Mean ($\bar{x}$)

The average value of a sample; used as a point estimate for the population mean.

Sample Proportion ($\hat{p}$)

The proportion of individuals in a sample with a specific characteristic; used as a point estimate for the population proportion.

Standard Error (SE)

The estimated standard deviation of a sampling distribution. It measures how much the sample statistic varies from sample to sample.

Critical Value

The number of standard errors to extend from the sample statistic when building a confidence interval. Based on the desired confidence level (e.g., $z^* = 1.96$ for 95%).

Margin of Error

The amount added and subtracted to the point estimate to create the confidence interval. It reflects uncertainty due to sampling variability.

Interpretation of a Confidence Interval

A proper interpretation indicates the range within which the population parameter lies and is linked to a confidence level — not the probability that the true value is in that specific interval.

This page titled 7.7: Vocabulary (Chapter 7) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Mathematics Department.

7.8: Practice (Chapter 7)

7.1: Point Estimates vs. Interval Estimates

1. A single-value estimate for a population parameter is called a:
    a. Confidence Interval
    b. Standard Deviation
    c. Point Estimate
    d. Population Mean
2. A 2023 Gallup poll of 1000 Americans found that 45% answered "Three or more" to the question "What is the ideal number of children to have?" The margin of error of the poll was 3%.
    a. What is the sample statistic?
    b. What is the population parameter we wish to estimate?
3. True or False: A point estimate is a single numerical approximation for a population parameter while a confidence interval is a range of values with a level of confidence.
4. Determine if the following are point estimates or interval estimates?
    a. The 2010 General Social Survey asked 1,259 US residents: "Do you think the use of marijuana should be made legal, or not?" 48% of the respondents said it should be made legal.
    b. The true proportion of people whose favorite color is black is between 2% and 6%.
    c. The average weight of a bag of chips is 1.2 oz $\pm$ 0.1oz.
    d. The proportion of those who watch The Daily Show is (0.07, 0.15).
    e. Daily consumer spending for the six-day period after Thanksgiving averaged $84.71.
5. Express the interval (339.8, 548.6) in the form $\bar{x} \pm ME$
6. Express the interval (59%, 67%) in the form $\hat{p} \pm ME$
7. Express the interval estimate 6.379 ± 6.86 in interval notation
8. Express the interval $595.4 < \mu < 851.8$ in the form $\bar{x} \pm ME$
9. What is the interval estimate for a poll where 45% answered "yes" to a question if the margin of error of the poll was 3%?
10. Express the interval estimate 0.35 ± 0.05 in interval notation
11. If a confidence interval for a population mean goes from 21.4 to 25.6, what is the point estimate?
12. For a class project, a political science student at a large university wants to estimate the percentage of students who are registered voters. He surveys 279 students and finds that 54% are registered voters. Answer the following questions.
    a. What is the population proportion?
    b. What is the point estimate for the population proportion?
    c. How many students are registered voters?
    d. What is the sample size?
13. In a random sample of 419 patients in a hospital emergency room, the mean wait time was 42.3 minutes with a standard deviation of 7 minutes. Answer the following questions.
    a. What is the population mean wait time for all emergency room patients at this hospital?
    b. What is the point estimate for the population mean?
    c. What is the population standard deviation?
    d. What is the sample standard deviation?
14. How do you differentiate between a point estimate and an interval estimate?
15. How do you find the margin of error from the lower and upper bounds of a confidence interval?

7.2: Confidence Intervals for a Population Proportion

1. True or False: A confidence interval is used to estimate a parameter using data collected from a sample.
2. State the full formula for a confidence interval for a population proportion.
3. Can we construct a confidence interval if the sample is not random? Why or why not?
4. Name the three conditions that should be met before constructing a confidence interval for a proportion.

5. A polling agency finds that 480 out of 1,200 respondents favor a new policy. Check whether it's appropriate to use a normal model before constructing a confidence interval.
6. If $n\hat{p}$ and $n(1-\hat{p})$ are both less than 10, what should you do?
7. Find the critical value for the common confidence levels of 90%, 95%, and 99%.
8. Describe the critical value and margin of error in your own words.
9. What does the critical value represent in a confidence interval formula?
10. What is the relationship between the standard error and margin of error?
11. Why should you be wary of surveys that do not report a margin of error?
12. In a sample of 200 registered voters, 134 say they plan to vote in the next local election.
    a. What is the point estimate for the population proportion?
    b. Verify that the conditions are met.
    c. What is the critical value for a 95% confidence level?
    d. What is the margin of error?
    e. What is the 95% confidence interval for the proportion of all registered voters who plan to vote?
    f. State the conclusion in a complete sentence.
13. A 2012 survey of 2,254 American adults indicates that 17% of cell phone owners do their browsing on their phone rather than a computer or other device.
    a. What is the point estimate for the population proportion?
    b. Verify that the conditions are met.
    c. What is the critical value for a 90% confidence level?
    d. What is the margin of error?
    e. What is the 90% confidence interval for the population proportion?
    f. State the conclusion in a complete sentence.
14. You survey 120 students and find that 45 of them get at least 8 hours of sleep regularly.
    a. What is the point estimate for the population proportion?
    b. Verify that the conditions are met.
    c. What is the critical value for a 99% confidence level?
    d. What is the margin of error?
    e. What is the 99% confidence interval for the population proportion?
    f. State the conclusion in a complete sentence
15. Suppose that an insurance company conducted a survey. They randomly surveyed 400 drivers and found that 320 claimed they always buckle up. We are interested in the population proportion of drivers who claim they always buckle up.
    a. Which distribution should you use for this problem? Explain your choice.
    b. Construct a 95% confidence interval for the population proportion who claim they always buckle up
        1. Calculate the error bound.
        2. Sketch a graph of the interval estimate
        3. State the confidence interval.
    c. If this survey were done by telephone, list three difficulties we might have in obtaining random results.
16. According to a recent survey of 1,200 people, 16% feel that the president is doing an acceptable job. We are interested in the population proportion of people who feel the president is doing an acceptable job.
    a. Which distribution should you use for this problem? Explain your choice.
    b. Construct a 90% confidence interval for the population proportion who claim they always buckle up
        1. Calculate the error bound.
        2. Sketch a graph of the interval estimate
        3. State the confidence interval.
    c. If this survey were done by telephone, list three difficulties we might have in obtaining random results.

17. A telephone poll of 1,000 adult Americans was reported in an issue of Time Magazine. One of the questions asked was “What is the main problem facing the country?” Twenty percent answered “crime.” We are interested in the population proportion of adult Americans who feel that crime is the main problem.
    a. Construct a 95% confidence interval for the population proportion of adult Americans who feel that crime is the main problem.
    b. Suppose we want to lower the sampling error. What is one way to accomplish that?
    c. The sampling error given by Yankelovich Partners, Inc. (which conducted the poll) is ±3%. In one to three complete sentences, explain what the ±3% represents.
18. In a 2010 Survey USA poll, 70% of the 119 respondents between the ages of 18 and 34 said they would vote in the 2010 general election for Prop 19, which would change California law to legalize marijuana and allow it to be regulated and taxed. At a 95% confidence level, this sample has an 8% margin of error. Based on this information, determine if the following statements are true or false, and explain your reasoning.
    a. We are 95% confident that between 62% and 78% of the California voters in this sample support Prop 19.
    b. We are 95% confident that between 62% and 78% of all California voters between the ages of 18 and 34 support Prop 19.
    c. If we considered many random samples of 119 California voters between the ages of 18 and 34, and we calculated 95% confidence intervals for each, 95% of them will include the true population proportion of Californians who support Prop 19.
    d. In order to decrease the margin of error to 4%, we would need to quadruple (multiply by 4) the sample size.
    e. Based on this confidence interval, there is sufficient evidence to conclude that a majority of California voters between the ages of 18 and 34 support Prop 19.
19. On June 28, 2012 the U.S. Supreme Court upheld the much debated 2010 healthcare law, declaring it constitutional. A Gallup poll released the day after this decision indicates that 46% of 1,012 Americans agree with this decision. At a 95% confidence level, this sample has a 3% margin of error. Based on this information, determine if the following statements are true or false, and explain your reasoning.
    a. We are 95% confident that between 43% and 49% of Americans in this sample support the decision of the U.S. Supreme Court on the 2010 healthcare law.
    b. We are 95% confident that between 43% and 49% of Americans support the decision of the U.S. Supreme Court on the 2010 healthcare law.
    c. If we considered many random samples of 1,012 Americans, and we calculated the sample proportions of those who support the decision of the U.S. Supreme Court, 95% of those sample proportions will be between 43% and 49%.
20. A Washington Post article from 2009 reported that "support for a government-run health-care plan to compete with private insurers has rebounded from its summertime lows and wins clear majority support from the public." More specifically, the article says "seven in 10 Democrats back the plan, while almost nine in 10 Republicans oppose it. Independents divide 52 percent against, 42 percent in favor of the legislation." There were were 819 Democrats, 566 Republicans and 783 Independents surveyed.
    a. Create a 90% confidence interval for the proportion of Independents who oppose the public option plan.
    b. A political pundit on TV claims that a majority of Independents oppose the health care public option plan. Do these data provide strong evidence to support this statement?

7.3: Confidence Intervals for a Population Mean

1. What changes in the construction of a confidence interval when estimating a mean instead of a proportion?
2. A family of distributions resembling the normal distribution but with thicker tails best describes
    a. the bootstrap
    b. degrees of freedom
    c. t-distribution
    d. confidence intervals
3. When should you use the t-distribution instead of a normal (Z) distribution for means?
4. Explain how the shape of the t-distribution affects your margin of error compared to a Z-based interval.

5. Why would you be more likely to use a T-interval in a real-world situation than a Z-interval?
6. What is the formula for the standard error of the sample mean?
7. An e-business manager wants to test the claim that the mean exposure to a certain type of internet ads exceeds 1000 viewers per day. Randomly selecting 20 days she calculated the sample mean and standard deviation.
   a. If the population standard deviation is not available, which distribution should she use?
   b. What is the degrees of freedom in this situation?
8. Given degree of freedom df = 16, find the probability by using the Student's t distribution: $P(t > 0.4) =$
9. You intend to estimate a population mean with a confidence interval. You believe the population to have a normal distribution. Your sample size is 54. Find the critical value that corresponds to a confidence level of 99.5%.
10. You intend to estimate a population mean with a confidence interval. You believe the population to have a normal distribution. Your sample size is 24. While it is an uncommon confidence level, find the critical value that corresponds to a confidence level of 93%.
11. The average time students spend on homework per week is estimated from a sample of 15 students, who report an average of 13.2 hours with a standard deviation of 4.8 hours. Construct a 95% confidence interval.
12. A survey was conducted on 203 undergraduates from Duke University who took an introductory statistics course in Spring 2012. Among many other questions, this survey asked them about the number of exclusive relationships they have been in. The sample average is 3.2 with a standard deviation of 1.97. Estimate the average number of exclusive relationships Duke students have been in using a 90% confidence interval and interpret this interval in context. Check any conditions required for inference, and note any assumptions you must make as you proceed with your calculations and conclusions.
13. The National Survey of Family Growth conducted by the Centers for Disease Control gathers information on family life, marriage and divorce, pregnancy, infertility, use of contraception, and men's and women's health. One of the variables collected on this survey is the age at first marriage. The histogram below shows the distribution of ages at first marriage of 5,534 randomly sampled women between 2006 and 2010. The average age at first marriage among these women is 23.44 with a standard deviation of 4.72. Estimate the average age at first marriage of women using a 95% confidence interval, and interpret this interval in context. Discuss any relevant assumptions.
14. Karen wants to advertise how many chocolate chips are in each Big Chip cookie at her bakery. She randomly selects a sample of 42 cookies and finds that the number of chocolate chips per cookie in the sample has a mean of 7.8 and a standard deviation of 2.3. What is the 98% confidence interval for the number of chocolate chips per cookie for Big Chip cookies?
15. Kyla wants to estimate the average number of hours worked per shift by ER nurses across a large hospital network. The population is known to be approximately normally distributed. Kyla takes a random sample of 19 nurse ER shifts and finds that the sample average is 10.5 hours, with a standard deviation of 0.6 hours.
16. The effectiveness of a blood-pressure drug is being investigated. An experimenter finds that, on average, the reduction in systolic blood pressure is 42.3 for a sample of size 434 and standard deviation 16.6. Estimate how much the drug will lower a typical patient's systolic blood pressure (using a 98% confidence level).
17. You must estimate the mean temperature (in degrees Fahrenheit) with the following sample temperatures. Find the 98% confidence interval.

57.4	86.3	51.9	51.6	99.5	86.5	33.9	84.9	69.2	72.6	67.9	41.9

18. You measure 21 randomly selected textbooks' weights, and find they have a mean weight of 35 ounces. Assume the population standard deviation is 14.5 ounces. Based on this, construct a 95% confidence interval for the true population mean textbook weight.
19. A veterinarian wants to estimate the mean weight of a cat in their city. They take a random sample of 16 cats and found a mean of 9.6 lbs with a sample standard deviation of 0.8 lbs. Find the 95% confidence interval of the mean. Assume that the cat weights are normally distributed.
20. A sample of the length in inches for newborns is given below. Assume that lengths are normally distributed. Find the 90% confidence interval of the mean length.

15.2	22.3	22.5	19.4	18.9	17.7
16.4	23	22.2	21		

21. Private colleges and universities rely on money contributed by individuals and corporations for their operating expenses. Much of this money is invested in a fund called an endowment, and the college spends only the interest earned by the fund. A recent survey of 43 private colleges in the United States revealed the following endowments (in millions of dollars):

Data

276.8	410.9	313.8	243.3	218	87.5	234.5	431	236.8	71.3
151.5	65.6	370.1	88.2	325.8	69.2	97.7	179.5	91.9	180.2
99.8	129.7	257	176.5	174.2	219.2	241	213.4	188.1	64
226.2	34	6.3	253	334.9	301.5	203.1	21	260.1	235.1
194.1	18.6	90.2							

Summary statistics yield the sample mean 188.01 millions of dollars and the sample standard deviation 107.07 millions of dollars. Construct and interpret a 95% confidence interval for the mean endowment of all private colleges in the United States.

22. Suppose you use simple random sampling to select and measure 46 backpacks' weights, and find they have a mean weight of 45 ounces. Assume the population standard deviation is 8 ounces. Based on this, construct a 90% confidence interval for the true population mean backpack weight.

7.4: Effect of Sample Size and Confidence Level

1. How does increasing the confidence level affect a confidence interval?
2. Why do higher confidence levels result in a larger margin of error?
3. What is the effect of decreasing the sample size on confidence interval?
4. Why do larger sample sizes make the margin of error smaller?
5. What is the effect of increasing the standard deviation on the margin of error?
6. Give an example of a situation where you would use each of the following confidence levels: 90%, 95%, and 99%
7. How does a statistician decide on which confidence level to use?
8. Why would we not use a 99.99% confidence intervals?
9. Suppose we calculated the 99% confidence interval for the average age of runners in the 2012 Cherry Blossom Run as (32.7, 37.4) based on a sample of 100 runners. How could we decrease the width of this interval without losing confidence?
10. We are interested in estimating the proportion of graduates at a mid-sized university who found a job within one year of completing their undergraduate degree. Suppose we conduct a survey and nd out that 348 of the 400 randomly sampled graduates found jobs. The graduating class under consideration included over 4500 students.
    a. Describe the population parameter of interest. What is the value of the point estimate of this parameter?
    b. Check if the conditions for constructing a confidence interval based on these data are met.
    c. Calculate a 95% confidence interval for the proportion of graduates who found a job within one year of completing their undergraduate degree at this university, and interpret it in the context of the data.
    d. What does "95% confidence" mean?
    e. Now calculate a 99% confidence interval for the same parameter and interpret it in the context of the data.
    f. Compare the widths of the 95% and 99% confidence intervals. Which one is wider? Explain.
11. Greece has faced a severe economic crisis since the end of 2009. A Gallup poll surveyed 1,000 randomly sampled Greeks in 2011 and found that 25% of them said they would rate their lives poorly enough to be considered "suffering".
    a. Describe the population parameter of interest. What is the value of the point estimate of this parameter?
    b. Check if the conditions required for constructing a confidence interval based on these data are met.
    c. Construct a 95% confidence interval for the proportion of Greeks who are "suffering".
    d. Without doing any calculations, describe what would happen to the confidence interval if we decided to use a higher confidence level.
    e. Without doing any calculations, describe what would happen to the confidence interval if we used a larger sample.

12. In 2010 the General Social Survey asked 1,259 US residents and reported a sample where about 48% of US residents thought marijuana should be made legal. If we wanted to limit the margin of error of a 95% confidence interval to 2%, about how many Americans would we need to survey ?
13. A poll of 783 Independents surveyed evaluating support for the health care public option in 2009, reporting that 52% of Independents in the sample opposed the public option. If we wanted to estimate this number to within 1% with 90% confidence, what would be an appropriate sample size?
14. SAT scores are distributed with a mean of 1,500 and a standard deviation of 300. You are interested in estimating the average SAT score of first year students at your college. If you would like to limit the margin of error of your 95% confidence interval to 25 points, how many students should you sample?
15. A political candidate has asked you to conduct a poll to determine what percentage of people support her. If the candidate only wants a 3% margin of error at a 95% confidence level, what size of sample is needed?
16. You want to obtain a sample to estimate the proportion of a population that possess a particular genetic marker. Based on previous evidence, you believe approximately 44% of the population have the genetic marker. You would like to be 98% confident that your estimate is within 3% of the true population proportion. How large of a sample size is required?
17. Insurance companies are interested in knowing the population percent of drivers who always buckle up before riding in a car.
    a. When designing a study to determine this population proportion, what is the minimum number you would need to survey to be 95% confident that the population proportion is estimated to within 0.03?
    b. If it were later determined that it was important to be more than 95% confident and a new survey was commissioned, how would that affect the minimum number you would need to survey? Why?
18. Why is 0.5 used in place of $\hat{p}$ when determining the minimum sample size necessary for a proportion confidence interval?

7.5: Interpreting Confidence Intervals

1. True or False: If one million 95% confidence intervals for the mean were computed, then approximately 95% of these confidence intervals would contain the population mean.
2. Describe the meaning of a 95% confidence level in the context of estimating a population proportion.
3. Which of the following is the best interpretation of a 95% confidence interval?
    a. There's a 95% chance that the true population parameter is within this one specific interval.
    b. 95% of all possible confidence intervals constructed from samples like this one will contain the true parameter.
    c. This sample is 95% accurate.
    d. We are 95% sure our sample statistic is correct.
4. Which of the following would be the best interpretation of the 99% confidence level?
    a. If we were to take 100 samples, then approximately 95 would contain the true value of the mean and approximately 5 would not
    b. If we were to take 100 samples, then approximately 99 would contain the true value of the mean and approximately 1 would not.
    c. If we were to take 100 samples, then exactly 99 would contain the true value of the mean and exactly 1 would not
    d. There is a 99% probability that the confidence interval contains the true value of the mean
5. Explain why we do not say “there is a 95% probability the population mean falls in this interval.”
6. Why is it misleading to say that a 100% confidence interval would definitely include the population parameter?
7. A hospital administrator hoping to improve wait times decides to estimate the average emergency room waiting time at her hospital. He collects a simple random sample of 64 patients and determines the time (in minutes) between when they checked in to the ER until they were first seen by a doctor. A 95% confidence interval based on this sample is (128 minutes, 147 minutes), which is based on the normal model for the mean. Determine whether the following statements are true or false, and explain your reasoning for those statements you identify as false.
    a. This confidence interval is not valid since we do not know if the population distribution of the ER wait times is nearly normal.
    b. We are 95% confident that the average waiting time of these 64 emergency room patients is between 128 and 147 minutes.

c. We are 95% confident that the average waiting time of all patients at this hospital's emergency room is between 128 and 147 minutes.
d. 95% of such random samples would have a sample mean between 128 and 147 minutes.
e. A 99% confidence interval would be narrower than the 95% confidence interval since we need to be more sure of our estimate.
f. The margin of error is 9.5 and the sample mean is 137.5.
g. In order to decrease the margin of error of a 95% confidence interval to half of what it is now, we would need to double the sample size.

8. The 2009 holiday retail season, which kicked off on November 27, 2009 (the day after Thanksgiving), had been marked by somewhat lower self-reported consumer spending than was seen during the comparable period in 2008. To get an estimate of consumer spending, 436 randomly sampled American adults were surveyed. Daily consumer spending for the six-day period after Thanksgiving, spanning the Black Friday weekend and Cyber Monday, averaged $84.71. A 95% confidence interval based on this sample is ($80.31, $89.11). Determine whether the following statements are true or false, and explain your reasoning.
   a. We are 95% con dent that the average spending of these 436 American adults is between $80.31 and $89.11.
   b. This confidence interval is not valid since the distribution of spending in the sample is right skewed.
   c. 95% of such random samples would have a sample mean between $80.31 and $89.11.
   d. We are 95% confident that the average spending of all American adults is between $80.31 and $89.11.
   e. A 90% confidence interval would be narrower than the 95% confidence interval since we don't need to be as sure about capturing the parameter.
   f. In order to decrease the margin of error of a 95% confidence interval to a third of what it is now, we would need to use a sample 3 times larger.
   g. The margin of error for the reported interval is 4.4.
9. The General Social Survey (GSS) is a sociological survey used to collect data on demographic characteristics and attitudes of residents of the United States. In 2010, the survey collected responses from 1,154 US residents. The survey is conducted face-to-face with an in-person interview of a randomly-selected sample of adults. One of the questions on the survey is "After an average work day, about how many hours do you have to relax or pursue activities that you enjoy?" A 95% confidence interval from the 2010 GSS survey is 3.53 to 3.83 hours.
   a. Interpret this interval in the context of the data.
   b. What does a 95% confidence level mean in this context?
   c. Suppose the researchers think a 90% confidence level would be more appropriate for this interval. Will this new interval be smaller or larger than the 95% confidence interval? Assume the standard deviation has remained constant since 2010.
10. Another question on the General Social Survey is "For how many days during the past 30 days was your mental health, which includes stress, depression, and problems with emotions, not good?" Based on responses from 1,151 US residents, the survey reported a 95% confidence interval of 3.40 to 4.24 days in 2010.
    a. Interpret this interval in context of the data.
    b. What does a 95% confidence level mean in this context?
    c. Suppose the researchers think a 99% confidence level would be more appropriate for this interval. Will this new interval be smaller or larger than the 95% confidence interval?
    d. If a new survey asking the same questions was to be done with 500 Americans, would the standard error of the estimate be larger, smaller, or about the same. Assume the standard deviation has remained constant since 2010.
11. The Marist Poll published a report stating that 66% of adults nationally think licensed drivers should be required to retake their road test once they reach 65 years of age. It was also reported that interviews were conducted on 1,018 American adults, and that the margin of error was 3% using a 95% confidence level.
    a. Verify the margin of error reported by The Marist Poll.
    b. Based on a 95% confidence interval, does the poll provide convincing evidence that more than 70% of the population think that licensed drivers should be required to retake their road test once they turn 65?
12. A survey on 1,509 high school seniors who took the SAT and who completed an optional web survey between April 25 and April 30, 2007 shows that 55% of high school seniors are fairly certain that they will participate in a study abroad program in

college.[43]

a. Is this sample a representative sample from the population of all high school seniors in the US? Explain your reasoning.
b. Let's suppose the conditions for inference are met. Even if your answer to part (a) indicated that this approach would not be reliable, this analysis may still be interesting to carry out (though not report). Construct a 90% confidence interval for the proportion of high school seniors (of those who took the SAT) who are fairly certain they will participate in a study abroad program in college, and interpret this interval in context.
c. What does "90% confidence" mean?
d. Based on this interval, would it be appropriate to claim that the majority of high school seniors are fairly certain that they will participate in a study abroad program in college?

13. In a recent Zogby International Poll, 5 of 48 respondents rated the likelihood of a terrorist attack in their community as “likely” or “very likely.” Create a 97% confidence interval for the proportion of American adults who believe that a terrorist attack in their community is likely or very likely. Explain what this confidence interval means in the context of the problem.
14. In a survey, 600 adults in a certain country were asked how many hours they worked in the previous week. Based on the results, a 95% confidence interval for mean number of hours worked was lower bound: 34 hours and upper bound: 39 hours. Which of the following represents a reasonable interpretation of the result? For those that are not reasonable, explain the flaw.
    a. There is a 95% chance the mean number of hours worked by adults in this country in the previous week was between 34 hours and 39 hours.
    b. We are 95% confident that the mean number of hours worked by adults in a particular area of this country in the previous week was between 34 hours and 39 hours.
    c. 95% of adults in this country worked between 34 hours and 39 hours last week.
    d. We are 95% confident that the mean number of hours worked by adults in this country in the previous week was between 34 hours and 39 hours.
15. A recent survey estimated that 19 percent of all people living in a certain region regularly use sunscreen when going outdoors. The margin of error for the estimate was 1 percentage point. Based on the estimate and the margin of error, which of the following is an appropriate conclusion?
    a. Approximately 1% of all the people living in the region were surveyed.
    b. Between 18% and 20% of all the people living in the region were surveyed.
    c. All possible samples of the same size will result in between 18% and 20% of those surveyed indicating they regularly use sunscreen.
    d. The probability is 0.01 that a person living in the region will use sunscreen when going outdoors.
    e. It is plausible that the percent of all people living in the region who regularly use sunscreen is 18.5%.
16. An advertiser claims that 9 out of 10 dentists (90%) recommend Brand X toothpaste. A consumer group investigates the truth of this claim. They survey a randomly selected sample of 4500 dentists, and find that 3964 of them recommend Brand X toothpaste. Based on this survey, they construct a 95% confidence interval for the proportion of dentists that recommend Brand X toothpaste of (0.871, 0.890). What can be concluded from this survey?

Practice problems include parts of:

Introductory Statistics 2e by OpenStax is licensed under CC BY 4.0

Access for free at https://openstax.org/books/introductory-statistics-2e/pages/1-introduction

and

OpenIntro Statistics by David Diez, Christopher Barr, & Mine Çetinkaya-Rundel is licensed under CC BY 3.0

7.8: Practice (Chapter 7) is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

- **Current page** has no license indicated.
- **4.E: Foundations for Inference (Exercises)** by David Diez, Christopher Barr, & Mine Çetinkaya-Rundel has no license indicated. Original source: https://www.openintro.org/book/os.
- **Licensing** by Mathematics Department has no license indicated.

8: Hypothesis Testing

So far, we've been focused on estimating population values using confidence intervals. But what if instead of estimating, we want to *test whether a claim is true*? That's where hypothesis testing comes in a framework that helps us make decisions using sample data.

Imagine this: A school claims that its students score an average of 80% on a state exam. A new study collects data from a sample of students and finds an average score of 77%. Is that just random variation, or **evidence that the true average is lower than claimed?** We need a way to check.

Hypothesis testing is how we decide whether the differences we see in a sample are statistically significant, or just due to natural variation. It is often used in medicine, science, education, and business any time someone asks, "Is this effect real, or just random?"

> **Key idea:** Hypothesis testing helps us answer questions like:
>
> - Is the new product better than the current one?
> - Is this average different from what we expected?
> - Does this difference matter statistically?

In this chapter, we'll walk step-by-step through the hypothesis testing process for a population mean or proportion. You'll learn how to:

- Write formal **null** and **alternative hypotheses**
- Use sample data to calculate a **test statistic** (like z or t)
- Find and interpret a **p-value**: the probability of seeing your result by random chance
- Decide whether to **reject or fail to reject** the null hypothesis
- Understand what it means to make a Type I or Type II error

> **Why it matters:** Hypothesis testing is the core tool researchers use to draw conclusions. Whether testing a medical treatment, a business strategy, or a public claim, knowing how to interpret evidence correctly is essential and powerful.

By the end of this chapter, you'll be able to look at a sample and say, with statistical confidence: "Yes, this difference looks real" or "No, this could just be random chance."

Let's learn how to test our questions not just ask them.

This page titled 8: Hypothesis Testing is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Mathematics Department.

8.1: Why hypothesis testing and Type I and Type II Errors

Before we dive into numbers or formulas, we need to zoom out and ask an essential question:

> **What are we trying to figure out?**
> In hypothesis testing, we start with a belief or assumption about the world and then use data to test whether that belief is reasonable or needs to be reconsidered.

Example Scenarios: What Are We Testing?

Let's look at four common scenarios where hypothesis testing appears in real life. Each one sets up a basic question we're trying to answer and helps illustrate what a **hypothesis** really is.

1. Is this new drug more effective than the standard treatment?

   This question is *not* about a single patient; it's about a larger population. We're comparing two treatments, looking for **evidence that one is better than the other**, and verifying that any observed difference is unlikely to be due to chance.

2. Does an after-school program improve math scores?

   We hypothesize that students who attend the program will perform better than those who do not. But even if we see higher scores, we have to ask: Is this improvement due to the program, or is it just random variation?

3. Is this machine producing parts that meet design standards?

   The manufacturer assumes the process is working correctly. A hypothesis test might be triggered if defective parts are suspected, to **test whether the process has shifted away** from acceptable levels.

4. Do more than 50% of local voters support the new policy?

   A political scientist sets up a hypothesis involving population proportion: they want to know if support exceeds a threshold. The test helps estimate whether the opinion measured in a sample reflects the entire population.

In all of these examples, we are doing some form of comparison: new vs. old, program vs. no program, actual vs. expected. And in each case, we are trying to use evidence (our sample) to make a judgment about the entire population. A **hypothesis** gives us something to test.

Definition: Null Hypothesis

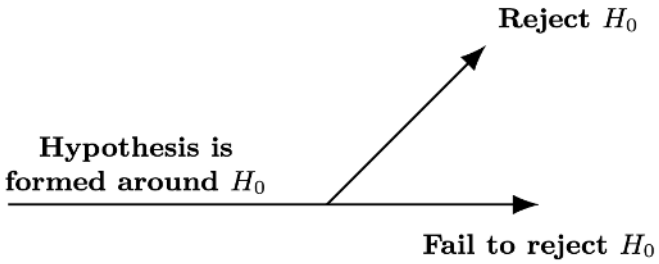
The **null hypothesis**, written as H_0, is a formal statement that there is *no effect*, *no difference*, or *no change* in the population. It represents a starting assumption that we test using data.

In a hypothesis test, we assume the null hypothesis is true and then use sample data to evaluate whether there is strong enough evidence to reject it.

Examples of Null Hypotheses:

- The new drug has the same recovery rate as the current one.
- The average score this year is still 75 (same as last year).
- There is no difference between Group A and Group B.
- The true population proportion is 50%.

Key idea: The null hypothesis is the claim we test against — and either **reject** or **fail to reject** based on the strength of the evidence.



Key idea: Every hypothesis test starts with a question. But that leads to an important follow-up: What if we're wrong?

What If We're Wrong?

When we run a hypothesis test, we examine a claim using sample data. Based on this, we'll make a decision: Do we reject the claim or not? But no matter how carefully we run the test, we have to face this key idea:

> **There's always a chance we're wrong.**
> Our conclusion might not match what's actually true in the population.

We can think of the universe as having a "back of the book" answer. It's hidden from us. We only get to see a sample, and then we try to infer what the correct answer is. Sometimes, we'll get it right. But sometimes we'll make mistakes.

There Are Two Kinds of Mistakes

When we work through a hypothesis test, two outcomes can happen:

- We correctly **reject** a false claim or **fail to reject** a true claim.
- Or, we make one of two classic errors:

Type I Error

This happens when the **null hypothesis is actually true** but our test leads us to wrongly reject it.

"We thought there was an effect, but there wasn't."

Type II Error

This happens when the **null hypothesis is actually false** but the test fails to detect the difference, and we wrongly keep it.

"There really was a difference, but we didn't find it."

Error Decision Table

This table shows all four possible situations:

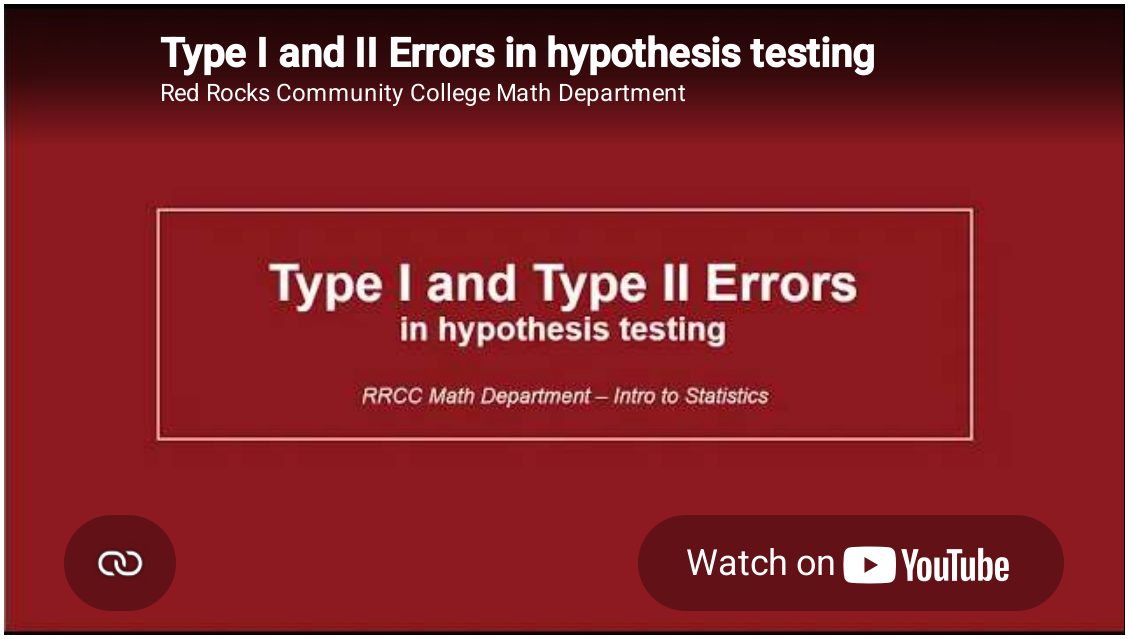
Table for reference for type 1 and type 2 errors.

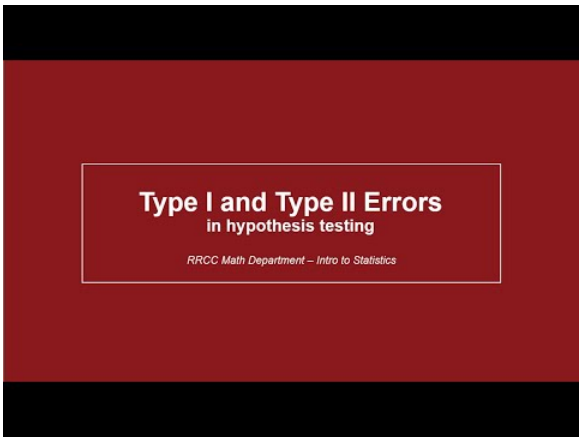
Reality (Truth)	We *Reject* Null Hypothesis	We *Fail to Reject* Null Hypothesis
Null is True	❌ Type I Error	✅ Correct Decision
Null is False	Correct Decision	❌ Type II Error

Understanding error helps us stay humble. Statistics will never “prove” anything with absolute certainty, but it can help us make informed decisions and assess how much risk of error we’re willing to tolerate.

> **Reflection:**
> Think of a time when someone made a wrong assumption based on partial information. What kind of error would that be? How might they have avoided it?

Related Video:





This page titled 8.1: Why hypothesis testing and Type I and Type II Errors is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Mathematics Department.

8.2: Hypothesis Testing Framework

Now that we've built a foundation for understanding what a hypothesis is — and what it means to be wrong — we're ready to outline the **hypothesis testing process** in full.

This process combines the ideas of both Ronald Fisher (who emphasized using data to test a ruling-out hypothesis, H_0) and Jerzy Neyman and Egon Pearson (who structured hypothesis testing as a formal decision-making process based on Type I and Type II errors).

Here's a five-step plan we'll use for every hypothesis test. Whether we're studying proportions, means, or comparing two groups, this structure remains the same — what will change is the type of test statistic we use.

An Outline of the Five Steps of a Hypothesis Test

1. **State the hypotheses**
   Clearly define the **null hypothesis** (H_0) and the **alternative hypothesis** (H_A).
   - H_0: often represents "no effect," "no difference," or "status quo." This is always an "equals" statement.
   - H_A: what you are trying to find evidence for. This will be some kind of inequality (>,<, or $\neq$).
2. **Select the appropriate test statistic and distribution**
   Choose a statistical test that matches the structure of your question, and define the relevant test statistic T, then calculate it using the sample data.
   - Is this a test about **one mean**? Use a z or t test for a single sample mean.
   - Is this comparing **two means**? Use a two-sample t test.
   - Is this a test about **one proportion** or rate? Use a z test for proportions.
   - For other types of tests see - **List of test statistics**
   - Your choice depends on: the type of variable, the number of groups, what statistics we can calculate and the sample size.
3. **Choose a significance level (α)**
   Set the allowable probability of making a **Type I Error** — in other words, the risk you're willing to take of rejecting the null when it's actually true.

   Common values: $\alpha = 0.05, 0.01, 0.10$
4. **Determine where it falls in the sampling distribution (Calculate the p-value)**
   Then determine how extreme that value is under the assumption H_0 is true (either via a p-value or using a critical value from a distribution table).
5. **Make a decision and state your conclusion**
   Use your comparison between t_{obs} and the distribution under H_0 to decide:
   - Reject H_0 (if the evidence against H_0 is strong)
   - Fail to reject H_0 (if the evidence against H_0 isn't strong enough)

   Always frame your decision in the context of the original research question.

Example: "Winters are hotter today than when I was a kid"

A friend says, "I swear winters are warmer now than when we were kids." Can we test this statistically?

You decide to gather historical winter temperatures from 20 years ago and compare them to modern-day winters in your region. You calculate the average January temperature today and compare it to 32 degrees F, the known average from the past.

Let's walk through the structure:

- **Step 1:**

$$H_0 : \mu = 32^\circ\text{F} \quad \text{(no warming)}$$
$$H_A : \mu > 32^\circ\text{F} \quad \text{(today's winters are warmer)} \tag{8.2.1}$$

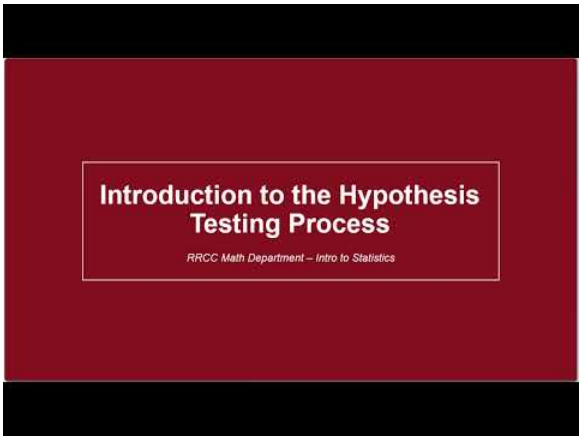
- **Step 2:** Use a one-sample t-test for a population mean (since you're comparing sample data to a known mean).

- **Step 3:** Choose $\alpha = 0.05$. Since there is no excessive worry of the Type I error here (accidentally concluding that winters are warmer when they aren't) we will go with the default $\alpha = 0.05$.
- **Step 4:** Calculate t_{obs} from your sample data (mean, standard deviation, sample size). If we assume that the current mean temperature is still 32 degrees F, how likely or unlikely is our observed sample data?
- **Step 5:** Compare t_{obs} to t-distribution and make your conclusion. If the probability of our results is below $\alpha = 0.05$ we can conclude that winters are indeed warmer. If the probability of our results is higher than that, we don't have strong enough evidence to say winters are warmer.

This structure helps us test the belief (“it’s hotter!”) using real evidence from data.

Related Video





What Comes Next

Now that we’ve outlined the general process, we’re going to build specific examples using data. We’ll start with hypothesis tests involving a **single mean**, and later move to tests comparing **two means** or working with **proportions**.

In each case, the framework we've outlined stays the same — we only change the specific test statistic and distribution used.

This page titled 8.2: Hypothesis Testing Framework is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Mathematics Department.

8.3: Tests for a Single Mean (z and t-tests)

Ready to put the hypothesis testing process to work? In this section, we'll walk through a full hypothesis test using a scenario that compares a sample mean to a known population mean using a **z-test statistic**.

Before we begin, let's clear up the difference between **z-tests** and **t-tests**, since you'll see both in this course.

Z-test vs. T-test

The formula for both is similar, but the test statistic depends on what we know:

- **Z-test:** Use when the population standard deviation σ is known.
- **T-test:** Use when σ is unknown and must be estimated with the sample standard deviation s

Z-test statistic:

$$Z = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

this test statistic will fall somewhere in the normal distribution.

T-test statistic:

$$T = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}$$

this test statistic will fall somewhere in the Student t-distribution.

Both represent standardized distances between your sample mean and the hypothesized mean under H_0 but within different distributions. We'll start with the easier case — when population σ is known — and revisit t-tests soon.

Once we have calculated our test statistic, we want to know: how extreme is this result if the null hypothesis were true? This is where the p-value comes in. The p-value helps us quantify whether our sample provides evidence against the null.

Definition: p-value

The **p-value** is the probability of observing a test statistic as extreme or more extreme than what we found in our sample, **assuming the null hypothesis is true**.

It reflects how surprising our result is under the assumption that H_0 is correct.

Choosing Left, Right, or Two-Tailed

The structure of your alternative hypothesis (H_A) determines which direction you're testing. Before you collect any data, decide what direction your test should examine and clearly state the form of your alternative hypothesis. The direction matters because it tells us where to look on the sampling distribution for unusual results: in the left tail, right tail, or both.

- **Left-tailed test:** Use when you are testing whether the population mean is *less than* the null value. Example: $H_A : \mu < \mu_0$
  For a left tailed test, the p-value will be $P(T < T_{obs})$
- **Right-tailed test:** Use when you are testing whether the mean is *greater than* the null value. Example: $H_A : \mu > \mu_0$
  For a right tailed test, the p-value will be $P(T > T_{obs})$.
- **Two-tailed test:** Use when you are testing whether the mean is simply *different* from the null value — that is, either greater or less. Example: $H_A : \mu \neq \mu_0$
  Two-tailed tests require computing the probability in the "tail" and then doubling it.

The smaller the p-value, the stronger the evidence against the null. Now that we have our test statistic and a way to interpret it, we're ready to step through a full hypothesis test example.

Example 1: "Traffic is so much worse these days"

You've heard it before: someone groans, "Traffic is AWFUL lately. It used to take 30 minutes to drive to work — now it takes forever."

You're curious: Is this backed up by data, or just selective memory?

You record your commute time on the same route over 36 different days. The known historical average commute time is $\mu = 30$ minutes, with population standard deviation $\sigma = 5$ minutes. Your sample has an average time of $\bar{x} = 32$ minutes.

Step 1: State the Hypotheses

We set up hypotheses to test a claim of change:

- $H_0 : \mu = 30$ (the average is still the same)
- $H_A : \mu > 30$ (average commute time is greater today)

Step 2: Select Test and Statistic

We're testing one sample mean against a known μ, and σ is known → use a z-test.

$$Z = \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}} = \frac{32-30}{\frac{5}{\sqrt{36}}} = \frac{2}{\frac{5}{6}} = \frac{2}{0.8333} \approx 2.40$$

Step 3: Set the Significance Level α

Let's use $\alpha = 0.05$ (5%) as the threshold for rejecting H_0.

Step 4: Calculate P-value

Find the area to the right of z = 2.40 in the normal distribution using the z-table or technology. This is a right-tailed test since H_A contains "greater than".

$$P(Z > 2.40) = 0.0082$$

Step 5: Make a Conclusion

The P-value represents the probability of seeing our sample mean or greater if we assume the population mean is still 30min. Since 0.0082 < 0.05, we reject H_0. Our sample results significantly unlikely to have occured randomly if H_0 is true, so we reject the H_0 assumption.

Conclusion: There is strong evidence that commute times are longer now than they used to be.

> **Reflect** Practical Significance vs. Statistical Significance: We have strong statistical evidence that commute times are longer on average now, but what we have not looked at is *how much* longer. Is the increase practically significant?

Example 2: "Are students sleeping less during finals week?"

You suspect that students at your school get less sleep during finals week than during a typical week.

Previous studies say college students average 7 hours of sleep per night (on a non-finals week). You collect a sample of 50 students and find they average 6.6 hours of sleep during finals week. Assume the population standard deviation is $\sigma = 1.2$ hours.

Try walking through the five steps using a z-test!

1. State your null and alternative hypotheses.
2. Compute the z test statistic using $\bar{x} = 6.6, \mu = 7, \sigma = 1.2, n = 50$
3. Choose a significance level (we can just use $\alpha = 0.05$ again)
4. Find the p-value using a z-table or technology.

5. Make a decision using $\alpha = 0.05$ and write your conclusion in the context of the problem.

Hint: This is a left-tailed test (you are testing if sleep is less than usual).

> **Tip:** When doing real-world work, using tools like calculators, spreadsheet functions, or AI assistants can help double-check your work — but always know what your answer means.

Example 3: "Are coffee drinkers at this café drinking more than the national average?"

A nationwide survey reports that the average amount of coffee consumed by adults is 2.7 cups per day. You suspect that customers at your local independent café might drink more than average, so you decide to collect some data.

You randomly sample 15 customers and ask each how many cups of coffee they've had that day. Your sample has an average of $\bar{x} = 3.1$ cups, with a sample standard deviation of $s = 0.75$ cups.

Question: Based on your sample, is there evidence that your local café customers drink more coffee than the national average?

Step 1: State the Hypotheses

- $H_0 : \mu = 2.7$
- $H_A : \mu > 2.7$

This is a one-tailed (right-tail) test. We're looking for evidence that the average is higher than the known population mean.

Step 2: Use a t-test (σ unknown)

Since we do **not** know the population standard deviation and have a small sample size (n = 15), we'll use the **t-distribution**.

Calculate the test statistic:

$$t = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}} = \frac{3.1 - 2.7}{\frac{0.75}{\sqrt{15}}} = \frac{0.4}{0.1938} \approx 2.06 \tag{8.3.1}$$

Step 3: Choose Significance Level

Use $\alpha = 0.05$

Step 4: Two Approaches to Making a Decision

There are two common methods used to reach a decision in hypothesis testing. When using technology, we usually use the p-value method, but if you're using the t-table, you would use the Critical Value method. Both methods use the same test statistic, but frame the conclusion differently. Let's examine both side by side:

Outlines the two approaches to hypothesis testing.

p-value Method	Critical Value Method
• Calculated test statistic: $t_{obs} = 2.06$ • Degrees of freedom: $df = 14$ • Using technology, the p-value for $t = 2.06$ is approximately $P(t > 2.06) = 0.030$ • Compare to $\alpha = 0.05$: ◦ $p = 0.030 < 0.05$ → reject H_0 • **Conclusion:** There is statistically significant evidence that customers drink more.	• Calculated test statistic: $t_{obs} = 2.06$ • Degrees of freedom: $df = 14$ • From the t-table: critical value $t^* = 1.76$ ◦ $t_{obs} = 2.06 > t^* = 1.76$ → reject H_0 • **Conclusion:** The sample mean is far enough to reject the null hypothesis.

> **Note:** Both methods lead to the same decision. Choose the approach best suited to the tools you're using — both are valid and help lead you to a reasoned conclusion.

Step 5: Decision

Since $p = 0.030 < 0.05$, we reject H_0.

Conclusion: There is statistically significant evidence that customers at this café drink more coffee than the national average.

> **Reflect:**
> This result suggests the higher average is more than just random chance, these customers really do seem to consume more coffee than elsewhere in the country.

Looking Ahead: Two Means

What if we want to compare *two groups* — like students in an afterschool program vs. a control group?

To compare two sample means, we'll use the two-sample t-test. If we do not assume equal variances, we use **Welch's t-test**:

$$T = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

We'll explain this formula and when to use it in the next section. But for now, remember: the logic stays the same. Hypotheses, test statistic, p-value, decision, conclusion.

This page titled 8.3: Tests for a Single Mean (z and t-tests) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Mathematics Department.

8.4: Comparing Two Means (Independent Samples)

Sometimes a single sample isn't enough. Instead, we want to compare **two groups** — to test whether one mean is higher (or lower) than the other.

In these situations, we'll use a two-sample t-test. If we do not assume equal variances between the groups, we'll specifically use **Welch's t-test**.

Questions That Involve Comparing Two Means

Here are five real-world questions where comparing group averages is a natural fit:

- Do students who attend tutoring score higher on exams than those who don't?
- Is it colder on average in Leadville, Colorado than in Anchorage, Alaska?
- Do iPhone users spend more on monthly mobile plans than Android users?
- Does a new fertilizer increase the average crop yield compared to the standard?
- Do influencers post more times per day on TikTok than Twitter users do on Twitter?

All of these questions can be framed as tests comparing the **means of two independent samples**.

Welch's t-test Statistic

Used when the sample sizes or variances may be unequal:

$$T = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \tag{8.4.1}$$

We'll use this statistic to test differences between two groups. This test statistic follows the Student t-distribution just like the t-tests we saw in the last section.

Let's try it out in a real example.

Example: "Is it colder in Leadville than Anchorage?"

You're arguing with a friend: They claim it's obviously colder in Alaska. But you're not so sure — Leadville, Colorado is at 10,000 feet after all.

You decide to test whether the average winter temperature in Leadville is colder than in Anchorage.

Step 1: State the Hypotheses

You'd like to test:

- $H_0 : \mu_{Leadville} = \mu_{Anchorage}$
- $H_A : \mu_{Leadville} < \mu_{Anchorage}$

This is a **left-tailed test**. You're looking for evidence that winter temperatures are *lower* in Leadville.

Step 2: Calculate the Appropriate Test Statistic

You collect 10 average January temperatures from each city (in °F):

- Leadville: $\bar{x}_1 = 12.2, s_1 = 4.8, n_1 = 10$
- Anchorage: $\bar{x}_2 = 17.6, s_2 = 6.2, n_2 = 10$

This is a two-sample independent t-test. Sample standard deviations are not equal → We use **Welch's t-test**.

Plug into the equation:

$$T = \frac{12.2 - 17.6}{\sqrt{\frac{(4.8)^2}{10} + \frac{(6.2)^2}{10}}} = \frac{-5.4}{\sqrt{2.304 + 3.844}} = \frac{-5.4}{\sqrt{6.148}} \approx \frac{-5.4}{2.478} \approx -2.18 \quad (8.4.2)$$

Step 3: Choose Significance Level

You select the standard $\alpha = 0.05$ as your cutoff for a Type I error (false positive).

Step 4: Compute P-value Using Technology

The degrees of freedom here are computed using the Welch-Satterthwaite formula(opens in new window). Generally this calculation is included in whatever technology you're using. In some cases we may use a simplified, conservative option where the degrees of freedom is the smaller of $n_1 - 1$ and $n_2 - 1$ but this does typically result in slightly larger P-values.

Use technology to find P(t < -2.18). We get $p \approx 0.023$

Step 5: Make Your Conclusion

Since 0.023 < 0.05, you reject H_0.

Conclusion: There is statistically significant evidence that winters in Leadville are colder than in Anchorage.

> **Interpretation:** While Leadville may not be as famously cold as Anchorage, your data suggest that altitude beats latitude!

Looking Ahead: Comparing Proportions

We've now compared two **means** — but many research questions involve questions about **proportions**, like:

- Will more than 50% of voters support a ballot measure?
- Is one treatment more likely to succeed than another?
- Are more students passing with the new class schedule?

We'll use a new test statistic to for proportions, but the already familiar sampling distribution following the Normal Distribution. Again, the process stays the same — only the test statistic changes.

This page titled 8.4: Comparing Two Means (Independent Samples) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Mathematics Department.

8.4.1: Project Checkpoint

As part of your semester project, you've been working with housing or economic data from two different cities or regions. Now that you know how to run a two-sample hypothesis test, it's time to apply that technique to your own data.

Research Question:

Is one city more expensive than the other?

Think about how this could apply to your dataset:

- Are you comparing **rents**, **home prices**, or another cost-related indicator?
- Is the cost measure available for both cities in your data?
- What exactly are you comparing — the mean monthly rent? Average price per square foot?

Once you have your comparison variable selected, follow the five-step hypothesis testing process with your data.

Instructions

1. **State your hypotheses clearly**
   Use appropriate notation and define what μ_1 and μ_2 represent in the context of your project.
2. **Compute your sample statistics**
   For each city, calculate:
   - Sample mean $\bar{x}_1, \bar{x}_2$
   - Sample standard deviation s_1, s_2
   - Sample size n_1, n_2

   These will be used in the Welch's t-test formula:

$$T = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \tag{8.4.1.1}$$

3. **Choose your significance level (α)**
   Most projects will use $\alpha = 0.05$, but you should briefly justify your choice.
4. **Calculate the test statistic and p-value**
   You can use a statistical tool such as:
   - ****Need link to a tool here****
   - Google Sheets: use `=T.TEST()` function
   - Excel, R, Python, a graphing calculator, or even AI assistance
5. **Make your conclusion**
   Refer back to your hypotheses and use your p-value to decide whether to:
   - Reject the null hypothesis
   - Fail to reject the null hypothesis

   Interpret the result in context: Does your data give strong evidence that one city is more expensive?

Deliverable

Add this analysis to your semester project report. Be sure to include:

- Your clearly written hypotheses (null and alternative)
- A summary of your sample statistics
- A brief explanation of the method and tool you used
- Your calculated test statistic and p-value
- Your written conclusion — in context of your data

> 💡 **Tip:** Be sure to consider your assumptions! Is it reasonable to treat the two samples as independent? Did either group have extreme variation or small sample size?

This page titled 8.4.1: Project Checkpoint is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Mathematics Department.

8.5: Tests for a Single Proportion

Many real-world questions boil down to a single idea: **What proportion of people...** agree with something, own something, choose something, or took some action?

In this section, we'll explore how to test claims about population **proportions** — the fraction or percentage of a population that fits a category. These types of questions are everywhere: in surveys, polls, tracking forms, and policy planning.

Recall: What is a Proportion?

Remember that a **proportion** is a number between 0 and 1 that represents the part of a group that meets a condition. Proportions can also be shown as percentages (multiply by 100).

Here are a few examples:

- You survey 40 students, and 25 say they study off-campus. What's the proportion?
- In a local election, 1,080 voters out of 1,500 said "yes" to a tax measure. What's the proportion who voted yes?
- 20 out of 25 light bulbs passed a quality test. What's the pass rate?

What Questions Lead to a Proportion Test?

Here are five example research questions that naturally lead to testing a single proportion:

- Do more than 60% of local residents support building new bike lanes?
- Is customer satisfaction higher than 90% at a coffee shop chain?
- Are less than half of voters in favor of the proposed tax levy?
- Are more than 1 in 4 high school students getting less than 6 hours of sleep per night?
- Do more than 20% of all shoppers use coupons during checkout?

Each question can be approached by taking a sample, measuring the **sample proportion**, and comparing it to a stated or assumed population proportion using a **z-test for proportions**.

Test Statistic for a Single Proportion

Here's the formula we'll use to test a null hypothesis involving a proportion:

$$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \tag{8.5.1}$$

- $\hat{p}$ = sample proportion
- p_0 = claimed (null hypothesis) population proportion
- n = sample size

This statistic measures how far away your sample result is from the null claim, in terms of standard error. We'll use it just like with z-tests for means → to calculate a p-value and decide whether to reject H_0.

Example: "Don't you think most people want more parks?"

A city planner claims that **most residents support adding more public parks**. "Most," you assume, should mean more than 50%.

You collect a random sample of 120 residents. Out of them, 69 say they support funding for more parks. Can we conclude that more than half of the population feels this way?

Let's use our 5-step testing framework.

Step 1: State the Hypotheses

- $H_0 : p = 0.50$ (half of the population supports it)
- $H_A : p > 0.50$ (more than half support more parks)

This is a one-tailed (right side) test.

Step 2: Select Test and Plug In

Your sample proportion is:

$\hat{p} = \frac{69}{120} = 0.575$

Now compute your test statistic:

$$Z = \frac{0.575 - 0.50}{\sqrt{\frac{0.50 \cdot (1-0.50)}{120}}} = \frac{0.075}{\sqrt{\frac{0.25}{120}}} = \frac{0.075}{0.0456} \approx 1.645 \tag{8.5.2}$$

Step 3: Choose Significance Level

We'll use $\alpha = 0.05$

Step 4: Find the p-value

Using technology or a z-table, $P(z > 1.645) \approx 0.0500$

Step 5: Make Your Decision

Since $p = 0.0500 = \alpha$, we are exactly on the boundary!

- Some instructors will say: reject if $p \leq \alpha$. In that case → reject H_0
- Others require $p < \alpha$. In that case → fail to reject

Conclusion: This result is right on edge. There is borderline evidence that more than half of residents support more parks — but we can't say so with high confidence.

> **Reflect:** With more data, we might be able to know for sure — but in this sample, support seems just above 50%. It's worth noting that at a lower significance level we would conclude that there is not sufficent evidence for majority support of parks, but at a higher significance level there would be sufficient evidence for our conclusion.
>
> This shows how the significance level is not set in stone and can influence our conclusions, which is why we need to set it before performing the study. How strong of evidence to we **want** to have before we make our conclusion? This is why interpreting a p-value is not as straight forward as some study authors might imply.

This page titled 8.5: Tests for a Single Proportion is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Mathematics Department.

8.6: Interpreting p-values and Significance

Once we've completed a hypothesis test and found a p-value, what does that number really tell us?

A p-value is more than a number — it represents a decision point, but also carries the weight of uncertainty.

What is a p-value, really?

Let's review our working definition of a p-value:

> ### Definition: p-value
>
> The **p-value** is the probability of observing a sample statistic as extreme or more extreme than what we saw — **assuming the null hypothesis is true**.
>
> It reflects how surprising our result would be if the null hypothesis were actually correct.

A small p-value means our observed result would be very unlikely under the null. A large p-value means the result is quite consistent with the null.

Important: A small p-value does not prove that the null is false. It just suggests the data are inconsistent with the null — and we may want to reject it.

Drawing Conclusions from Your p-value

We compare the p-value to our significance level α to make a formal decision:

- If $p \leq \alpha$: reject H_0 → there's statistically significant evidence for H_A
- If $p > \alpha$: fail to reject H_0 → there's not enough evidence against it

For example, if $\alpha = 0.05$ and our p-value is 0.03, we reject the null.

> ### Statistical Significance
>
> If the p-value is less than or equal to our α level, we say the result is **statistically significant**.
>
> This means the result is unlikely to be due to chance alone — *according to the assumptions of our model*.

Limits of p-values: p-hacking and Misuse

P-values are useful — but not magical. Let's talk about the risks of misusing them.

What is p-hacking?

P-hacking is when someone tries many comparisons or statistical tweaks, only reporting the p-values that look *significant* — even if those results occurred just by chance.

- Trying dozens of statistical tests until "something sticks"
- Selectively reporting only results with $p < 0.05$
- Stopping a data collection early because $p < 0.05$

This creates false positives — appearing to have evidence when there isn't any. It's why replicable studies and good design matter.

What can we do?

- Plan your test **before** collecting data
- Report all results — not just significant ones
- Understand that one significant p-value doesn't "prove" your claim

> **Remember:** A p-value doesn't tell you the *probability that your hypothesis is true*. It tells you the probability of seeing your data **if the null hypothesis were true**.

Thinking Beyond the P-value

While significance testing is useful, it should never be your only consideration. Always ask:

- Is this result **practically significant**? (Does the effect matter in the real world?)
- What's the **effect size**? (Is the difference meaningful?)
- What are the risks of a Type I or Type II error in context?
- Would this result hold with a larger or different sample?

P-values are just one tool — use with care, context, and critical thinking.

> **Think About It:**
> If you find a statistically significant result ($p = 0.04$), but the difference is only 0.2 points on a 100-point scale, what should you conclude? What other info would you want before acting on that result?

This page titled 8.6: Interpreting p-values and Significance is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Mathematics Department.

8.7: Vocabulary (Chapter 8)

Chapter 8 Vocabulary

Hypothesis

A testable statement about a population parameter. Hypotheses are used to structure statistical tests and are evaluated based on sample data.

Null Hypothesis (H_0)

The default or starting assumption in a hypothesis test — usually a statement of "no effect," "no change," or "no difference." We test this claim using data.

Alternative Hypothesis (H_A)

The rival claim to the null. It represents the outcome we're trying to find evidence for (e.g., an increase, decrease, or difference).

Test Statistic

A calculated value (e.g., z, t) that measures how far the sample statistic is from the null hypothesis value. It's used to determine how extreme the result is.

z-test for a mean

A hypothesis test used when the population standard deviation is known. Assumes normal distribution. The test statistic is:
$Z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$

Welch's t-test

A two-sample test for comparing means when variances are unequal and sample sizes may differ. It uses sample standard deviations and calculates degrees of freedom based on variability.

z-test for a proportion

Used to test a hypothesis about a population proportion. The test statistic is:
$Z = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}}$

p-value

The probability of observing a result as extreme or more extreme than your observed sample statistic, assuming the null hypothesis is true. A small p-value suggests the data are inconsistent with H_0.

Significance Level (α)

The threshold we set for how much risk of a Type I error we are willing to accept. Common values are 0.05, 0.01, and 0.10.

Statistical Significance

If the p-value is less than or equal to α, we say the result is statistically significant — meaning it is unlikely to have occurred by chance under the null.

Practical Significance

A result that is meaningful or impactful in the real world. Even small statistical differences may not matter practically, depending on the context.

Type I Error

Rejecting the null hypothesis when it is actually true. This is a false positive. The probability of this is controlled by α.

Type II Error

Failing to reject the null hypothesis when the alternative is actually true. This is a false negative.

Power of a Test

The probability that a test will correctly reject a false null hypothesis. It is equal to 1 minus the probability of a Type II Error.

One-tailed Test

A hypothesis test in which the alternative hypothesis is directional (e.g., $p > p_0$ or $\mu < \mu_0$) — we are only interested in one direction of difference.

Two-tailed Test

A hypothesis test in which the alternative hypothesis tests for any difference from the null (e.g., $\mu \neq \mu_0$). Both tails of the distribution are considered.

Decision Rule

The rule that guides whether to reject or fail to reject the null hypothesis, usually based on comparing the p-value to α.

p-hacking

A problematic practice in which researchers run many statistical analyses until they obtain a significant ($p < 0.05$) result, often misrepresenting the evidence.

Effect Size

A measure of how big a difference or change is — separate from statistical significance. A small p-value may accompany a small effect size, or vice versa.

Standard Error

The estimated standard deviation of a sampling distribution. Used in many test statistics, including for means and proportions.

This page titled 8.7: Vocabulary (Chapter 8) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Mathematics Department.

8.8: Practice (Chapter 8)

8.1: Why Hypothesis Testing and Type I and Type II Errors

1. What type of statistical question is best suited for a hypothesis test, and why?
2. For each of the following, determine if the statement is a Null Hypothesis. If it is not, rewrite it so that it becomes a Null Hypothesis.
   a. The mean number of years Americans work before retiring is 34.
   b. At most 60% of Americans vote in presidential elections.
   c. The mean starting salary for Red Rocks Community College graduates is not equal to that of Community College of Denver graduates.
   d. Twenty-nine percent of high school seniors get drunk each month.
   e. Fewer than 5% of adults ride the bus to work in Denver.
   f. The mean number of cars a person owns in their lifetime is not more than ten.
   g. About half of Americans prefer to live away from cities, given the choice.
   h. Europeans have a mean paid vacation each year equal to that of Asians.
   i. The chance of developing breast cancer is under 11% for women.
   j. Private universities mean tuition cost is more than $20,000 per year.
   k. The mean price of mid-sized cars in a region is $32,000.
   l. A sleeping bag is tested to withstand temperatures of –15 °F.
3. Write the null and hypothesis in words for the following situations.
   a. New York is known as "the city that never sleeps". A random sample of 25 New Yorkers were asked how much sleep they get per night. Do these data provide convincing evidence that New Yorkers on average sleep less than 8 hours a night?
   b. Employers at a company are worried about the effect of March Madness, a basketball championship held each spring in the US, on employee productivity. They estimate that on a regular business day employees spend on average 15 minutes of company time checking personal email, making personal phone calls, etc. They also collect data on how much company time employees spend on such non-business activities during March Madness. They want to determine if these data provide convincing evidence that employee productivity decreases during March Madness.
   c. The mean entry level salary of an employee at a company is $58,000. You believe it is higher for IT professionals in the company.
4. Write the null and hypothesis in words for the following situation.
   a. Since 2008, chain restaurants in California have been required to display calorie counts of each menu item. Prior to menus displaying calorie counts, the average calorie intake of diners at a restaurant was 1100 calories. After calorie counts started to be displayed on menus, a nutritionist collected data on the number of calories consumed at this restaurant from a random sample of diners. Do these data provide convincing evidence of a difference in the average calorie intake of a diners at this restaurant?
   b. Based on the performance of those who took the GRE exam between July 1, 2004 and June 30, 2007, the average Verbal Reasoning score was calculated to be 462. In 2011 the average verbal score was slightly higher. Do these data provide convincing evidence that the average GRE Verbal Reasoning score has changed since 2004?
5. Fill out the following error table. Label which boxes represent Type I error, Type II error, and correct decisions:

Hypothesis Testing Outcomes

	Reality (Truth)	We Reject H_0	We Fail to Reject H_0
th)	H_0 is true		
th)	H_0 is false		

6. Food inspectors inspect samples of food products to see if they are safe. This can be thought of as a hypothesis test with the null hypothesis "The food is safe". The following is an example of what type of error? *The sample suggests that the food is safe, but it actually is not safe.*

7. You conduct a hypothesis test to determine whether most people believe in UFOs. Given a null hypothesis of: Most people believe in UFOs, the result of your test is to fail to reject. In fact, most people do not believe in UFOs. In this case, you have made what what type of error?
8. Come up with a real-world example where you would:
    a. Reject a true null hypothesis
    b. Fail to reject a false null hypothesis
9. What is a practical situation where rejecting H_0 would lead to a big decision (e.g., policy change, funding reshuffle)?
10. Why does decreasing the probability of making a type one error increase the probability of making a type two error?
11. A machine in a factory is supposed to fill vials with approximately 50 mL of a liquid. The quality control manager wants to test whether the machine is working properly. If the machine is found to be over or under-filling then the assembly line will need to be shut down so that it can be re-calibrated.
    a. Write the Null and Alternative Hypotheses to the scenario in words.
    b. What would it mean to make a Type 1 error in this situation?
    c. What would it mean to make a Type 2 error in this situation?
    d. Which error might be worse in this scenario and why?
12. A patient named Diana was diagnosed with Fibromyalgia, a long-term syndrome of body pain, and was prescribed anti-depressants. Being the skeptic that she is, Diana didn't initially believe that anti-depressants would help her symptoms. However after a couple months of being on the medication she decides that the anti-depressants are working, because she feels like her symptoms are in fact getting better.
    a. Write the hypotheses in words for Diana's skeptical position when she started taking the anti-depressants.
    b. What is a Type 1 error in this context?
    c. What is a Type 2 error in this context?
    d. How would these errors affect the patient?
13. A food safety inspector is called upon to investigate a restaurant with a few customer reports of poor sanitation practices. The food safety inspector uses a hypothesis testing framework to evaluate whether regulations are not being met. If he decides the restaurant is in gross violation, its license to serve food will be revoked.
    a. Write the hypotheses in words.
    b. What is a Type 1 error in this context?
    c. What is a Type 2 error in this context?
    d. Which error is more problematic for the restaurant owner? Why?
    e. Which error is more problematic for the diners? Why?
    f. As a diner, would you prefer that the food safety inspector requires strong evidence or very strong evidence of health concerns before revoking a restaurant's license? Explain your reasoning.
14. Suppose regulators monitored 403 drugs last year, each for a particular adverse response. For each drug they conducted a single hypothesis test with a significance level of 5% to determine if the adverse effect was higher in those taking the drug than those who did not take the drug; the regulators ultimately rejected the null hypothesis for 42 drugs.
    a. Describe the error the regulators might have made for a drug where the null hypothesis was rejected.
    b. Describe the error regulators might have made for a drug where the null hypothesis was not rejected.
    c. Suppose the vast majority of the 403 drugs do not have adverse effects. Then, if you picked one of the 42 suspect drugs at random, about how sure would you be that the drug really has an adverse effect?
    d. Can you also say how sure you are that a particular drug from the 361 where the null hypothesis was not rejected does not have the corresponding adverse response?

8.2: Hypothesis Testing Framework

1. List the five steps in the hypothesis testing process, and write one sentence describing the purpose of each.
2. What factors determine which test statistic (z, t, etc.) you should use?
3. How do you tell whether a test is left, right, or two tailed?
4. How does the type of test e.g., left tailed, relate to the inequality symbols used in the hypotheses?

5. How does a researcher decide the level of significance for a hypothesis test?
6. Why do we make our decision about α before seeing the results of the test?
7. How do you decide whether to reject or fail to reject the null hypothesis?
8. Why can we never accept the null hypothesis?
9. Complete this prompt: “When we reject the null hypothesis, we are concluding that...”
10. What are some cases where “failing to reject” H_0 might still give you useful information?
11. Write the null and alternative hypotheses using equations/inequalities with the correct symbols for the following situations.
    a. New York is known as "the city that never sleeps". A random sample of 25 New Yorkers were asked how much sleep they get per night. Do these data provide convincing evidence that New Yorkers on average sleep less than 8 hours a night?
    b. Employers at a company are worried about the effect of March Madness, a basketball championship held each spring in the US, on employee productivity. They estimate that on a regular business day employees spend on average 15 minutes of company time checking personal email, making personal phone calls, etc. They also collect data on how much company time employees spend on such non-business activities during March Madness. They want to determine if these data provide convincing evidence that employee productivity decreases during March Madness.
    c. The mean entry level salary of an employee at a company is $58,000. You believe it is higher for IT professionals in the company.
12. Write the null and alternative hypotheses using equations/inequalities with the correct symbols for the following situations.
    a. Since 2008, chain restaurants in California have been required to display calorie counts of each menu item. Prior to menus displaying calorie counts, the average calorie intake of diners at a restaurant was 1100 calories. After calorie counts started to be displayed on menus, a nutritionist collected data on the number of calories consumed at this restaurant from a random sample of diners. Do these data provide convincing evidence of a difference in the average calorie intake of a diners at this restaurant?
    b. Based on the performance of those who took the GRE exam between July 1, 2004 and June 30, 2007, the average Verbal Reasoning score was calculated to be 462. In 2011 the average verbal score was slightly higher. Do these data provide convincing evidence that the average GRE Verbal Reasoning score has changed since 2004?
    c. A statistics instructor believes that fewer than 20% of Red Rocks Community College students attended the opening night midnight showing of the latest Harry Potter movie. They survey 84 of their students and finds that 11 of them attended the midnight showing.
13. A study suggests that the average college student spends 2 hours per week communicating with others online. You believe that this is an underestimate and decide to collect your own sample for a hypothesis test. You randomly sample 60 students from your dorm and nd that on average they spent 3.5 hours a week communicating with others online. A friend of yours, who offers to help you with the hypothesis test, comes up with the following set of hypotheses. Indicate any errors you see.
    a. $H_0 : \bar{x} < 2$ hours
    b. $H_A : \bar{x} > 3.5$ hours
14. The proportion of Americans who have frequent migraines is 15.2% according to the CDC. An acupuncturist claims that her treatment can reduce this figure significantly. A random sample of 279 Americans is administered the acupuncturists treatment and 39 report experiencing migraines.
    a. State hypotheses for the scenario using the correct symbols.
    b. Suppose that we fail to reject the null hypothesis. Write a conclusion statement in the context of the problem. Assume no errors were made.
    c. Suppose that we reject the null hypothesis. Write a conclusion statement in the context of the problem. Assume no errors were made
15. A machine in a factory is supposed to fill vials with approximately 50 mL of a liquid. The quality control manager wants to test whether the machine is working properly. If the machine is found to be over or under-filling then the assembly line will need to be shut down so that it can be re-calibrated.
    a. State hypotheses for the scenario using the correct symbols.
    b. Suppose that we fail to reject the null hypothesis. Write a conclusion statement in the context of the problem. Assume no errors were made.

c. Suppose that we reject the null hypothesis. Write a conclusion statement in the context of the problem. Assume no errors were made

16. Previously, an organization reported that teenagers spent 4.5 hours per week, on average, on the phone. The organization thinks that, currently, the mean is higher. Fifteen randomly chosen teenagers were asked how many hours per week they spend on the phone. The sample mean was 4.75 hours with a sample standard deviation of 2.0. Write an appropriate conclusion.
    a. State hypotheses for the scenario using the correct symbols.
    b. Suppose that we fail to reject the null hypothesis. Write a conclusion statement in the context of the problem. Assume no errors were made.
    c. Suppose that we reject the null hypothesis. Write a conclusion statement in the context of the problem. Assume no errors were made

8.3: Tests for a Single Mean (z and t-tests)

1. Which two distributions can you use for a hypothesis test for a single mean?
2. Which distribution do you use when you are testing a population mean and the standard deviation is known? Assume sample size is large.
3. Explain in your own words why knowing the population standard deviation is important for using a z-test.
4. You are testing that the mean speed of your cable Internet connection is more than three Megabits per second. State the null and alternative hypotheses.
5. A population mean is 13. The sample mean is 12.8, and the sample standard deviation is two. The sample size is 20. What distribution should you use to perform a hypothesis test? What assumptions should be made?
6. It is believed that Red Rocks Community College (RRCC) Intermediate Algebra students get less than seven hours of sleep per night, on average. A survey of 22 RRCC Intermediate Algebra students generated a mean of 7.24 hours with a standard deviation of 1.93 hours. At a level of significance of 5%, do RRCC Intermediate Algebra students get less than seven hours of sleep per night, on average? What is the distribution to be used for this test?
7. A researcher claims that the average time people spend on social media each day is more than 2 hours. A sample of 50 people has an average of 2.3 hours and a known population standard deviation of 0.8 hours.
    a. What type of test should the researcher conduct?
    b. What significance level should the researcher choose?
8. A particular brand of tires claims that its deluxe tire averages at least 50,000 miles before it needs to be replaced. From past studies of this tire, the standard deviation is known to be 8,000. A survey of owners of that tire design is conducted. From the 28 tires surveyed, the mean lifespan was 46,500 miles with a standard deviation of 9,800 miles. Using $\alpha = 0.05$, is the data highly inconsistent with the claim?
9. From generation to generation, the mean age when smokers first start to smoke varies. However, the standard deviation of that age remains constant of around 2.1 years. A survey of 40 smokers of this generation was done to see if the mean starting age is at least 19. The sample mean was 18.1 with a sample standard deviation of 1.3. Do the data support the claim at the 1% level?
10. The cost of a daily newspaper varies from city to city. However, the variation among prices remains steady with a standard deviation of 20¢. A study was done to test the claim that the mean cost of a daily newspaper is $1.00. Twelve costs yield a mean cost of 95¢ with a standard deviation of 18¢. Does the data support the claim at the 10% level?
11. The mean number of sick days an employee takes per year is believed to be about ten. Members of a personnel department do not believe this figure. They randomly survey eight employees. The number of sick days they took for the past year are as follows: $12, 4, 15, 3, 11, 8, 6, 8$. Let the $x =$ number of sick days they took for the past year. Should the personnel team believe that the mean number is ten?
12. The mean work week for engineers in a start-up company is believed to be about 60 hours. A newly hired engineer hopes that it's shorter. She asks ten engineering friends in start-ups for the lengths of their mean work weeks. Based on the results that follow, should she count on the mean work week to be shorter than 60 hours? Data (length of mean work week):
    70, 45, 55, 60, 65, 55, 55, 60, 50, 55.
13. True or False: "A large positive z-score always means we reject the null." Explain.
14. A hospital administrator randomly selected 64 patients and measured the time (in minutes) between when they checked in to the ER and the time they were first seen by a doctor. The average time is 137.5 minutes and the standard deviation is 39 minutes.

He is getting grief from his supervisor on the basis that the wait times in the ER increased greatly from last year's average of 127 minutes. However, the administrator claims that the increase is probably just due to chance.

a. Are conditions for inference met? Note any assumptions you must make to proceed.
b. Using a significance level of $\alpha = 0.05$, is the change in wait times statistically significant? Use a two-sided test since it seems the supervisor had to inspect the data before he suggested an increase occurred.
c. Would the conclusion of the hypothesis test change if the significance level was changed to $\alpha = 0.01$?

15. Researchers investigating characteristics of gifted children collected data from schools in a large city on a random sample of thirty-six children who were identified as gifted children soon after they reached the age of four. The distribution of the ages (in months) at which these children first counted to 10 successfully is provided as sample statistics $\bar{x} = 30, \sigma = 1$

a. Are conditions for inference satisfied?
b. Suppose you read on a parenting website that children first count to 10 successfully when they are 32 months old, on average. Perform a hypothesis test to evaluate if these data provide convincing evidence that the average age at which gifted children first count to 10 successfully is different than the general average of 32 months. Use a significance level of 0.10.
c. Interpret the p-value in context of the hypothesis test and the data.
d. Calculate a 90% confidence interval for the average age at which gifted children first count to 10 successfully.
e. Do your results from the hypothesis test and the confidence interval agree? Explain.

16. A manufacturer claims that bearings produced by their machine last 7 hours on average under harsh conditions. A factory worker randomly samples 75 ball bearings, and records their lifespans under harsh conditions. They calculate a sample mean of 6.85 hours, and the standard deviation of the data is 1.25 working hours. Conduct a formal hypothesis test of this claim. Make sure to check that relevant conditions are satisfied.
17. The nutrition label on a bag of potato chips says that a one ounce (28 gram) serving of potato chips has 130 calories and contains ten grams of fat, with three grams of saturated fat. A random sample of 35 bags yielded a sample mean of 134 calories with a standard deviation of 17 calories. Is there evidence that the nutrition label does not provide an accurate measure of calories in the bags of potato chips? Assume we have verified the independence, sample size, and skew conditions are satisfied.

8.4: Comparing Two Means (Independent Samples)

1. What does Welch's t-test account for that a pooled two-sample t-test does not?
2. Do hypothesis tests for two means require equal sample sizes? Why or why not?
3. Give an example of a study that compares two means but would NOT be appropriate for a t-test. Explain why.
4. A fitness company is testing two diets. Group A lost an average of 5.1 lbs ($s = 2.3, n = 15$). Group B lost 4.2 lbs ($s = 2.1, n = 15$). Is there evidence that Diet A works better at a 5% significance level?

a. State the hypotheses
b. Compute the test statistic (use Welch's t-test)
c. Find the approximate p-value using technology.
d. Write the conclusion in complete sentences.

5. A sample of 12 in-state graduate school programs at school A has a mean tuition of $64,000 with a standard deviation of $8,000. At school B, a sample of 16 in-state graduate programs has a mean of $80,000 with a standard deviation of $6,000. On average, is the mean tuition different?
6. A student at a four-year college claims that mean enrollment at four–year colleges is higher than at two–year colleges in the United States. Two surveys are conducted. Of the 35 two–year colleges surveyed, the mean enrollment was 5,068 with a standard deviation of 4,777. Of the 35 four-year colleges surveyed, the mean enrollment was 5,466 with a standard deviation of 8,191. On average, is the mean enrollment of four-year colleges higher than two-year colleges?
7. An economist wants to compare mean hourly rate charged by automobile mechanics in two suburbs. She randomly selects auto repair facilities from both suburbs and records their hourly rates (in dollars). The data are as follows:

Suburb 1 Data

35.3	33.2	37.8	35.6	33.5
36	36.1	31.7	29.6	37.6
37.9				

Suburb 2 Data

47.7	50.7	42.2	41.9	48.7
42.9	42.4	37.3	38.2	45.3
44	43.8	42.4	44.1	42.7
48.6	56.2	46.3	51.6	38.4
51.1	47.5			

Use 10% level of significance to decide whether there is sufficient evidence that the mean hourly rate charged by automobile mechanics in suburb 1 is the same as the mean hourly rate charged by automobile mechanics in suburb 2. You may also assume that the samples are from normal populations.

8.5: Tests for a Single Proportion

1. A sociologist claims the probability that a person picked at random in Times Square in New York City is visiting the area is 0.83. You want to test to see if the claim is correct. State the null and alternative hypotheses.
2. It is thought that more than 50% of respondents in a taste test would prefer Brand *A*. In a particular test of 100 people, 49 preferred Brand *A*. State the null and alternative hypotheses.
3. You are performing a hypothesis test of a single population mean using a Student's t-distribution. What must you assume about the distribution of the data?
4. A policy maker claims that fewer than half of the city's residents support a new tax. A random sample finds that 91 out of 200 residents support the tax.
   a. Find the sample proportion
   b. Compute test statistic
5. Explain each part of the z-test for a proportion: what are $\hat{p}$, p_0, and n?
6. Why do we use p_0 (the null value) in the denominator instead of $\hat{p}$?
7. You are performing a hypothesis test of a single population proportion. What must be true about the quantities of \np\ and \nq\?
8. What are some situations in which you might prefer a two-tailed hypothesis about a proportion instead of a one-tailed one?
9. True or False: "If $p = 0.04$, that means 4% of the population is in favor." Explain.
10. It is believed that nearsightedness affects about 8% of all children. In a random sample of 194 children, 21 are nearsighted.
    a. Construct hypotheses appropriate for the following question: do these data provide evidence that the 8% value is inaccurate?
    b. What proportion of children in this sample are nearsighted?
    c. Given that the standard error of the sample proportion is 0.0195 and the point estimate follows a nearly normal distribution, calculate the test statistic (the Z statistic).
    d. What is the p-value for this hypothesis test?
    e. What is the conclusion of the hypothesis test?
11. The National Institute of Mental Health published an article stating that in any one-year period, approximately 9.5 percent of American adults suffer from depression or a depressive illness. Suppose that in a survey of 100 people in a certain town, seven of them suffered from depression or a depressive illness. Conduct a hypothesis test to determine if the true proportion of people in that town suffering from depression or a depressive illness is lower than the percent in the general adult American population.

12. Your statistics instructor claims that 60 percent of the students who take her Intro to Statistics class go through life feeling more enriched. For some reason that she can't quite figure out, most people don't believe her. You decide to check this out on your own. You randomly survey 64 of her past Intro to Statistics students and find that 34 feel more enriched as a result of her class. Conduct a hypothesis test for the instructor's claim.
13. A poll done for *Newsweek* found that 13% of Americans have seen or sensed the presence of an angel. A contingent doubts that the percent is really that high. It conducts its own survey. Out of 76 Americans surveyed, only two had seen or sensed the presence of an angel. As a result of the contingent's survey, would you agree with the *Newsweek* poll? Conduct a hypothesis test for Newsweek's claim.
14. "Japanese Girls' Names" by Kumi Furuichi. It used to be very typical for Japanese girls' names to end with "ko." (The trend might have started around my grandmothers' generation and its peak might have been around my mother's generation.) "Ko" means "child" in Chinese characters. Parents would name their daughters with "ko" attaching to other Chinese characters which have meanings that they want their daughters to become, such as Sachiko—happy child, Yoshiko—a good child, Yasuko—a healthy child, and so on. However, I noticed recently that only two out of nine of my Japanese girlfriends at this school have names which end with "ko." More and more, parents seem to have become creative, modernized, and, sometimes, westernized in naming their children. I have a feeling that, while 70 percent or more of my mother's generation would have names with "ko" at the end, the proportion has dropped among my peers. I wrote down all my Japanese friends', ex-classmates', co-workers, and acquaintances' names that I could remember. Following are the names. (Some are repeats.) Test to see if the proportion has dropped for this generation.
    a. Ai, Akemi, Akiko, Ayumi, Chiaki, Chie, Eiko, Eri, Eriko, Fumiko, Harumi, Hitomi, Hiroko, Hiroko, Hidemi, Hisako, Hinako, Izumi, Izumi, Junko, Junko, Kana, Kanako, Kanayo, Kayo, Kayoko, Kazumi, Keiko, Keiko, Kei, Kumi, Kumiko, Kyoko, Kyoko, Madoka, Maho, Mai, Maiko, Maki, Miki, Miki, Mikiko, Mina, Minako, Miyako, Momoko, Nana, Naoko, Naoko, Naoko, Noriko, Rieko, Rika, Rika, Rumiko, Rei, Reiko, Reiko, Sachiko, Sachiko, Sachiyo, Saki, Sayaka, Sayoko, Sayuri, Seiko, Shiho, Shizuka, Sumiko, Takako, Takako, Tomoe, Tomoe, Tomoko, Touko, Yasuko, Yasuko, Yasuyo, Yoko, Yoko, Yoko, Yoshiko, Yoshiko, Yoshiko, Yuka, Yuki, Yuki, Yukiko, Yuko, Yuko.
15. According to an article in *Bloomberg Businessweek*, New York City's most recent adult smoking rate is 14%. Suppose that a survey is conducted to determine this year's rate. Nine out of 70 randomly chosen N.Y. City residents reply that they smoke. Conduct a hypothesis test to determine if the rate is still 14% or if it has decreased.
16. Is the proportion of cat lovers in Lakewood more than the proportion in Golden? Out of 498 people surveyed in Lakewood, 284 are cat lovers. Out of 480 people surveyed in Golden, 263 are cat lovers. Conduct a full hypothesis test at the 0.05 level of significance to answer the question.

8.6: Interpreting p-values and Significance

1. Is the p-value for a Z-statistics of 2.5 larger when n=500 or n=1000?
2. Explain in your own words what a p-value means. What does it assume?
3. What does an extremely small p-value (e.g., 0.001) suggest about our result?
4. What is "p-hacking"? Why is it problematic in scientific research?
5. Give an example where a result could be statistically significant but not practically significant.
6. Draw a sketch of the sampling distribution and p-value for the following hypothesis test: $H_0 : \mu = 23, H_1 : \mu > 23, \bar{x} = 25$
7. Draw a sketch of the sampling distribution and p-value for the following hypothesis test: $H_0 : p = 0.3, H_1 : p \neq 0.3, \hat{p} = .25$
8. How does choosing your significance level α relate to the risk of a Type I error?
9. What happens if you lower α (say, from 0.05 to 0.01)? How does that affect:
    a. Type I error risk?
    b. Likelihood of rejecting H_0?
10. Explain how reporting only p-values (and not effect sizes, variability, or context) can be misleading.
11. Interpret: p = 0.075. Which is more correct?
    a. There is fairly strong evidence against H_0
    b. We should reject H_0
    c. We fail to reject H_0 because we lack strong evidence
    d. Explain your choice.

12. The mean age of graduate students at a University is at most 31 years with a standard deviation of two years. A random sample of 15 graduate students is taken. The sample mean is 32 years and the sample standard deviation is three years. Are the data significant at the 1% level?
    a. The p-value is 0.0264. Interpret the *p*-value in complete sentences.
13. A USA Today/Gallup poll conducted between 2010 and 2011 asked a group of unemployed and underemployed Americans if they have had major problems in their relationships with their spouse or another close family member as a result of not having a job (if unemployed) or not having a full-time job (if underemployed). 27% of the 1,145 unemployed respondents and 25% of the 675 underemployed respondents said they had major problems in relationships as a result of their employment status.
    a. The p-value for this hypothesis test is approximately 0.35. Explain what this means in context of the hypothesis test and the data.
14. A running shoe manufacturer believes that the proportion of faulty running shoes from the supplier is greater than 0.1, the proportion stated by the supplier. We perform a hypothesis test at a significance level of 0.05 with null and alternative hypotheses: H_0: p = 0.1 and H_a: p > 0.1. A researcher inspects 100 items and finds 16 faulty ones. She gives a p-value of 0.02 in her analysis. What is the correct interpretation of the p-value?
15. What should you do if \alpha\ equals the p-value?
16. You are given the following hypotheses: $H_0 : \mu = 34, H_A : \mu > 34$. We know that the sample standard deviation is 10 and the sample size is 65. For what sample mean would the p-value be equal to 0.05? Assume that all conditions necessary for inference are satisfied.

Practice problems include parts of:

Introductory Statistics 2e by OpenStax is licensed under CC BY 4.0

Access for free at https://openstax.org/books/introductory-statistics-2e/pages/1-introduction

and

OpenIntro Statistics by David Diez, Christopher Barr, & Mine Çetinkaya-Rundel is licensed under CC BY 3.0

8.8: Practice (Chapter 8) is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

- **Current page** has no license indicated.
- **4.E: Foundations for Inference (Exercises)** by David Diez, Christopher Barr, & Mine Çetinkaya-Rundel has no license indicated. Original source: https://www.openintro.org/book/os.
- **Licensing** by Mathematics Department has no license indicated.

9: Bivariate Data, Correlation, and Regression

What happens to housing prices as the number of bedrooms increases? Do students who study more tend to score higher on exams? As gas prices go up, do fewer people take road trips? These are all examples of questions involving **bivariate data** where we're interested in the relationship between *two* variables.

In this chapter, we'll explore how to visually and numerically investigate associations between two variables. You'll learn how to describe the **direction, strength, and form** of a relationship, and how to interpret the results in context. And just as importantly, we'll also discuss when a relationship *might not* be meaningful or when it might be misleading.

> **Key idea:** Correlation and regression help us explore how one variable may be associated with or even help predict another.
>
> But remember: **correlation ≠ causation**.

We begin with **scatterplots**, the foundational way to visualize relationships between two quantitative variables. From there, you'll learn about the **correlation coefficient (r)**, a numerical measure that tells us how strong and in what direction two variables move together.

Then we'll build a **regression model**: the best-fitting straight line that helps us predict one variable from another. We'll look at equations, interpret slope and intercept, and examine how well our line fits the data using **residuals** (prediction errors).

- Understand what a **linear relationship** looks like and how to detect it
- Calculate and interpret the correlation coefficient (r)
- Construct and interpret a **least-squares regression line**
- Use context to make predictions (within reason!)
- Recognize the limitations of regression models

> **Why it matters:** From economics to healthcare to sports analytics, understanding relationships between variables helps us make better decisions and smarter predictions.

By the end of this chapter, you won't just be spotting patterns you'll be modeling them.

Let's explore what happens when we connect two variables and what those connections can reveal.

This page titled 9: Bivariate Data, Correlation, and Regression is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Mathematics Department.

9.1: Scatterplots and Direction of Association

Much of what we've done so far has involved analyzing one variable at a time whether it was categorical (like favorite restaurant) or numerical (like exam scores). But in the real world, variables almost never exist in isolation. They interact, influence, and move together.

In this chapter, we shift to analyzing relationships between two variables. These are called **bivariate data**. We ask questions like:

- Do older cars have higher mileage?
- Are students who study more scoring higher?
- Do cities with higher rents also report higher income?

To explore the possible relationship between two **quantitative variables**, the first thing we typically do is make a **scatterplot**.

What is a Scatterplot?

A **scatterplot** is a graph that plots paired data points in a coordinate system. Each point represents one individual (or measurement), with one variable on the x-axis and one variable on the y-axis.

For example, suppose we collect data on 10 students' hours of study and exam scores. Each student's study time and score becomes one point on the graph:

[Insert scatterplot figure here: Study Hours vs. Exam Score]
Each point = 1 student

By plotting data this way, we can visually scan for **patterns** are scores tending to rise as study time increases? Or do they stay flat?

Direction of Association

One of the most important patterns to look for is the **direction of association** between the two variables. This tells us how values in one variable tend to change as the other variable increases.

- **Positive association:** As x increases, y tends to increase (upward slope)
- **Negative association:** As x increases, y tends to decrease (downward slope)
- **No association:** No clear pattern in either direction

[Insert Scatterplot: Positive Association]	*[Insert Scatterplot: Negative Association]*	*[Insert Scatterplot: No Association]*
Positive	Negative	No clear direction

Example: Rent and Income

Suppose we collect a dataset with 30 cities, where we record:

- Median monthly rent
- Median household income

We create a scatterplot with rent on the x-axis and income on the y-axis. We may observe that as rent increases, income also increases suggesting a **positive association** between the two variables.

But not all points are exactly on a straight line. The data wobbles as it often does in the real world. This leads us to our next question: Can we *quantify* how strong the relationship is?

We can often recognize a potential connection between two variables, but we need a more formal tool to measure it. In the next section, we'll introduce a number called the **correlation coefficient**, which tells us how strong and how linear the relationship is.

This page titled 9.1: Scatterplots and Direction of Association is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Mathematics Department.

9.1.1: Datasaurus Dozen

The Datasaurus Dozen: Why We Always Plot Our Data

Up to this point, we've talked about plotting two quantitative variables to find direction or pattern. But you might ask: "Do I really have to make a scatterplot if I already have the means, standard deviations, and correlation?"

Yes. And here's why.

The **Datasaurus Dozen** is a group of datasets created by Justin Matejka and George Fitzmaurice in 2017 as an update to Anscombe's classic 1973 datasets. These twelve different datasets all have roughly the same:

- Mean of x
- Mean of y
- Standard deviations
- Correlation between x and y

But when we look at scatterplots of the data, the patterns and stories are completely different.

Shared Descriptive Statistics

Descriptive Statistics for Example Graphs

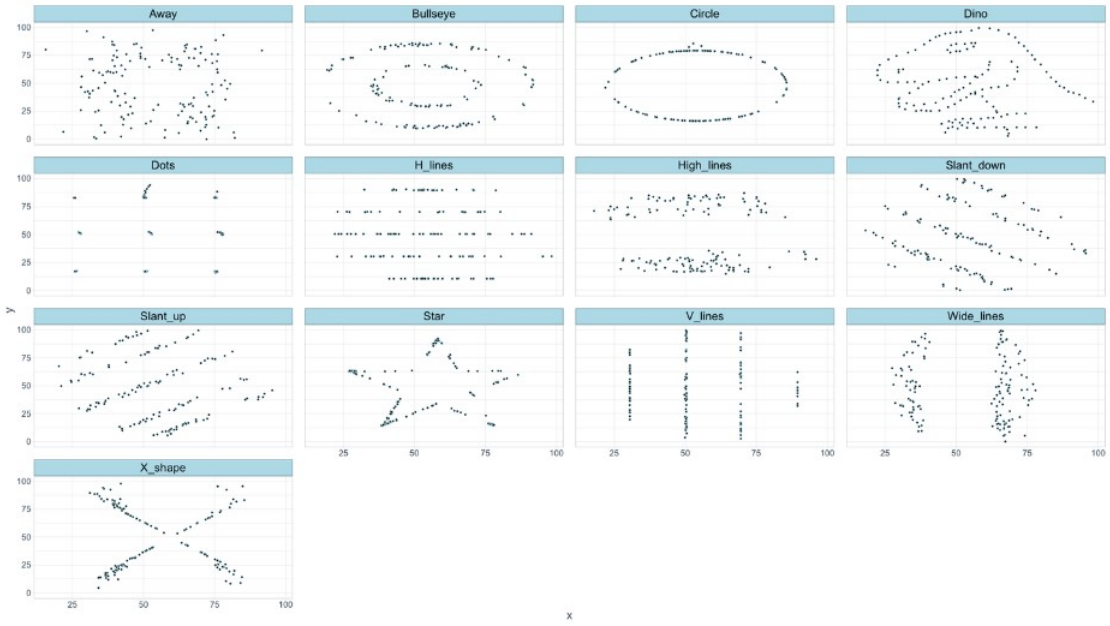
Dataset	Mean of x	Mean of y	Standard Deviation of x	Standard Deviation of y	Correlation (r)
Dino (Original)	54.26	47.83	16.77	26.94	–0.06
Star	54.26	47.83	16.77	26.94	–0.06
X-Shape	54.26	47.83	16.77	26.94	–0.06
H-Shape	54.26	47.83	16.77	26.94	–0.06
Circle	54.26	47.83	16.77	26.94	–0.06
Slant-Up	54.26	47.83	16.77	26.94	–0.06
Slant-Down	54.26	47.83	16.77	26.94	–0.06
Vertical Lines	54.26	47.83	16.77	26.94	–0.06
Horizontal Lines	54.26	47.83	16.77	26.94	–0.06
Tight Cluster	54.26	47.83	16.77	26.94	–0.06
Wide Spread	54.26	47.83	16.77	26.94	–0.06
Target Shape	54.26	47.83	16.77	26.94	–0.06

Plots of all the data

```
dataset     x     y
<chr>   <dbl> <dbl>
dino     55.4  97.2
dino     51.5  96.0
dino     46.2  94.5
dino     42.8  91.4
dino     40.8  88.3
dino     38.7  84.9
dino     35.6  79.9
dino     33.1  77.6
dino     29.0  74.5
dino     26.2  71.4
ℹ 1,836 more rows
```

Common statistical values for each group in the dataset

	Summary statistics					*Regression results*	
Dataset	Mean x	Mean y	Std Dev x	Std Dev y	Corr x y	Intercept	Coefficients
Away	54.27	47.83	16.77	26.94	-0.06	53.43	-0.10
Bullseye	54.27	47.83	16.77	26.94	-0.07	53.81	-0.11
Circle	54.27	47.84	16.76	26.93	-0.07	53.80	-0.11
Dino	54.26	47.83	16.77	26.94	-0.06	53.45	-0.10
Dots	54.26	47.84	16.77	26.93	-0.06	53.10	-0.10
H_lines	54.26	47.83	16.77	26.94	-0.06	53.21	-0.10
High_lines	54.27	47.84	16.77	26.94	-0.07	53.81	-0.11
Slant_down	54.27	47.84	16.77	26.94	-0.07	53.85	-0.11
Slant_up	54.27	47.83	16.77	26.94	-0.07	53.81	-0.11
Star	54.27	47.84	16.77	26.93	-0.06	53.33	-0.10
V_lines	54.27	47.84	16.77	26.94	-0.07	53.89	-0.11
Wide_lines	54.27	47.83	16.77	26.94	-0.07	53.63	-0.11
X_shape	54.26	47.84	16.77	26.93	-0.07	53.55	-0.11



By IngmundForberg - Own work, CC BY-SA 4.0, Link

Why This Matters

If you only look at summary statistics, you might assume that all of these distributions are the same but they're radically different when visualized.

This teaches us something vital: **summary statistics don't tell the full story.** We must **always plot our data** to detect patterns, outliers, structures, or shapes that numbers alone might hide.

This is especially important as we move into correlation where we'll quantify how strong a linear relationship is. But beware: even strong-looking numbers can mislead if we don't first look at our data with our own eyes.

This page titled 9.1.1: Datasaurus Dozen is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Mathematics Department.

9.2: Correlation Coefficient (r)

In the last section, we explored how scatterplots can show whether two variables tend to move together, upward, downward, or not at all.

But now we want to go a step further. Suppose we see a general upward trend in a scatterplot, how strong is that trend? Is it a tight linear pattern, or just a vague upward cloud? We need a way to formally **measure** how strong the linear association is between two variables.

This is where the **correlation coefficient** comes in. It gives us a number between –1 and 1 that describes the direction and strength of the linear relationship between two quantitative variables.

Why Do We Need a Number?

While scatterplots are the best way to detect patterns, human eyes can be subjective. Two people might look at the same cloud of points and come to different conclusions.

That's why we calculate the correlation coefficient, also known as r. This number lets us compare linear associations numerically:

- Is a pattern strong or weak?
- Is it positive or negative?
- Is the linear association meaningful or just noise?

The correlation coefficient was first formally developed by **Karl Pearson** in the early 1900s and is sometimes called "Pearson's r". It's a widely used summary statistic but it must be interpreted with care.

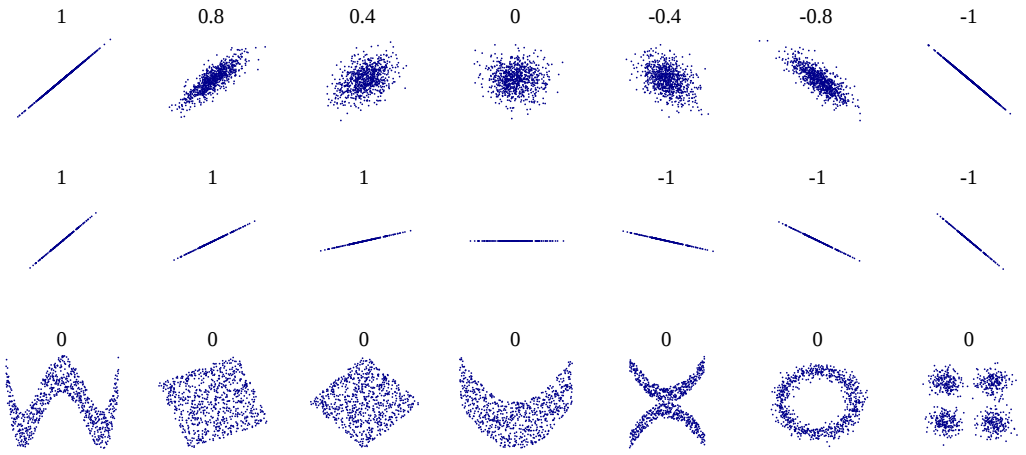
Definition: Correlation Coefficient

The **correlation coefficient**, denoted r, is a number between –1 and 1 that measures the strength and direction of a **linear relationship** between two quantitative variables.

Interpretation of r:

- $r > 0$: Positive linear association
- $r < 0$: Negative linear association
- $|r|$ near 1: Strong linear relationship
- $r = 0$: No linear relationship (though there could still be a curved pattern!)

Important: Correlation only measures the strength of a *linear* relationship. It does not explain cause and effect.



By DenisBoigelot, original uploader was Imagecreator - Own work, original uploader was Imagecreator, CC0, Link

---

Examples: Interpreting Correlation

Here are three examples that help us think about how to interpret r. In each case, imagine a scatterplot of the data alongside the reported value of r.

Example 1: Hours of TV and Grades

A researcher collects data on 100 students, recording how many hours of television they watch per day and their GPA. The correlation is:

$$r = -0.62$$

This is a **moderately strong negative linear relationship**. Students who watch more TV tend to have lower GPAs but the relationship isn't perfect. There's still a lot of variation.

Example 2: Shoe Size and Math Score

In a satire of misused statistics, someone runs a study comparing shoe size and standardized math test scores. They report:

$$r = 0.01$$

This value is close to zero, suggesting **no linear relationship**. There is no real association here and it's a reminder that not all variables will relate, even if we wish they might.

Example 3: Height and Arm Span

A study of 50 adult humans measures height (in cm) and arm span. The analysis reports:

$$r = 0.97$$

This is a **very strong positive linear relationship**. Height and arm span tend to increase together in a nearly perfect line matching biological expectations.

In the next section, we'll explore how to **calculate** the correlation coefficient from raw data or technology and how this leads us directly into linear models and least-squares regression.

This page titled 9.2: Correlation Coefficient (r) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Mathematics Department.

9.3: Least-Squares Regression Line

Once we've established that two variables are related and we've measured that relationship with the correlation coefficient r the next natural step is to **predict**.

We might want to guess someone's height based on their arm span. Or predict a student's test score based on their hours of study. Or estimate housing costs from square footage.

We handle these predictions using a mathematical model: a **line of best fit** that allows us to estimate y-values based on given x-values. That line is called the **least-squares regression line**.

Story Problem: Studying and Test Scores

Suppose you work as a tutor for a group of students studying for an important exam. You record how many hours they studied the night before and the test score they earned (out of 100).

Study Time and Test Scores

Study Hours (x)	Test Score (y)
1	52
2	55
2	59
3	61
4	66
5	73
5	71
6	77
7	80
8	85

Can we find a line that models the relationship between hours studied (x) and test score (y)?

Goal: Finding a Line of Best Fit

The **least-squares regression line** is the line that best fits the data. It minimizes the total amount of squared error between the actual y-values and the predicted y-values from the line.

The line will be in the form:

$$\hat{y} = a + bx$$

- $\hat{y}$: predicted value of y
- a: y-intercept
- b: slope of the line

This line is most useful for:

- **Interpolation:** estimating values *within the range* of the data
- **Extrapolation:** projecting beyond the measured data (with caution!)

Computing the Correlation Coefficient and Regression Line

Let's walk through the process using your study hours and test scores from the dataset above.

Step 1: Calculate the Means

First, compute the mean of x and the mean of y:

- $\bar{x} = 4.3$
- $\bar{y} = 67.9$

Step 2: Compute the Slope (b)

The slope b of the regression line is computed using this formula:

$$b = r \cdot \frac{s_y}{s_x}$$

We already determined (via technology or calculator) that the correlation $r = 0.97$, the standard deviation of y is $s_y = 11.2$, and the standard deviation of x is $s_x = 2.2$.

Now plug in:

$$b = 0.97 \cdot \frac{11.2}{2.2} \approx 4.94$$

Step 3: Compute the Intercept (a)

Use the formula:

$$a = \bar{y} - b \cdot \bar{x}$$

Substitute in the known values:

$$a = 67.9 - 4.94 \cdot 4.3 \approx 67.9 - 21.24 = 46.66$$

Regression Line:

So the full equation of the least-squares regression line is:

$$\hat{y} = 46.66 + 4.94x$$

We can now use this line to:

- Estimate a student's score if they studied for 5 hours
- Predict how much someone might score if they only studied for 1 hour
- Check the reasonableness of someone's test result based on their study time

---

Interpreting the Slope and Intercept

Slope: For each additional hour of study, predicted test score increases by about 4.94 points.

Intercept: The predicted test score for someone who studied 0 hours is about 46.66 a contextual estimate, not always meaningful in real-life scenarios.

In future sections, we'll explore how well this line fits the data, how to check for residuals, and how to use it responsibly for prediction and analysis.

This page titled 9.3: Least-Squares Regression Line is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Mathematics Department.

9.4: Interpreting Slope, Intercept, and Residuals

Once we've calculated the least-squares regression line, we want to understand what the slope and intercept really mean. And beyond that, how good our predictions are which is where residuals come in.

Let's anchor these ideas in a story.

Story Example: Does Practice Help with Typing Speed?

Suppose a typing instructor is studying the relationship between number of practice sessions attended (x) and typing speed at the end of the month (y), measured in words per minute (wpm).

The instructor records data from 12 students, runs a least-squares regression analysis, and finds the equation:

Predicted typing speed: $\hat{y} = 42 + 3.5x$

Interpreting the Equation

- **Slope (3.5):** For every additional practice session, the model predicts an increase of 3.5 words per minute in typing speed.
- **Intercept (42):** If a student attends 0 practice sessions, the model predicts a typing speed of 42 wpm. (This may not be meaningful depending on context — we should always be cautious when interpreting an intercept.)

Definition: Residual

A **residual** is the difference between an observed value and its predicted value from the regression model:

$$\text{Residual} = \text{Actual} - \text{Predicted} = y - \hat{y} \tag{9.4.1}$$

If the residual is positive, the prediction *underestimated* the actual value.
If the residual is negative, the prediction *overestimated* the actual value.

Example: Predictions and Residuals

A student attended 6 practice sessions and achieved a final speed of 62 wpm.

Prediction: $\hat{y} = 42 + 3.5(6) = 63$ wpm

Actual: 62 wpm

Residual: $62 - 63 = -1$

This tells us the model **overestimated** the student's performance by 1 wpm.

Interactive Scatterplot: Explore the Line and Residual

Click in the box below to add a student data point. The regression line is fixed as: $\hat{y} = 42 + 3.5x$

The app will draw the prediction and calculate the residual for your clicked point.

What Should We Take Away?

- Slope tells us the change in predicted y for a one-unit increase in x.
- Intercept tells us the predicted value when $x = 0$. Be cautious when interpreting these — sometimes they're outside the realistic data range.
- Residuals measure prediction error. Large residuals may indicate bad fits or outliers.

Think About It:

- If all the residuals are large, does that mean the model is bad?
- If the residuals are sometimes positive and sometimes negative, what does that suggest?
- Would this model be reliable for predicting the speed of someone who attends 15 practice sessions?

This page titled 9.4: Interpreting Slope, Intercept, and Residuals is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Mathematics Department.

9.5: Cautions with Correlation and Causation

By now, you've seen that two quantitative variables can have a strong positive or negative relationship. That's what correlation helps us measure. But it's also important to know what **correlation does not tell us**.

Definition: Causation

> **Causation** means that a change in one variable directly causes a change in the other.
>
> This means that changing variable x *produces* a predictable change in variable y, through a direct influence or mechanism.

Just because two variables change together doesn't mean one causes the other to change. This is the classic warning:

Correlation does not imply causation.

Examples of Correlated—But Not Causal—Relationships

Example 1: Ice Cream Sales and Drowning

Across many cities, ice cream sales and drowning rates both rise in the summer. The correlation is strong and positive.

- But that doesn't mean buying ice cream causes drowning.
- There's a third factor (weather/season) that increases both.

Example 2: Number of NIC Cage Films and Spelling Bee Winners

From 1999–2009, there was a high correlation between the number of Nicolas Cage films released each year and deaths from falling into pools.

- This is pure coincidence — no one thinks one affects the other.
- But without context, the correlation looks "strong."

Example 3: Shoe Size and Reading Score (in Children)

Shoe size and reading score among children ages 3–10 are strongly correlated.

- Bigger shoe size doesn't improve reading ability.
- They're both *linked to age*. Growing older increases both.

Why This Matters

It's tempting to look at a pattern and say one variable causes the other. But that leap requires more than just a statistic—it needs thoughtful reasoning, study design, and evidence.

Things that help support a causal conclusion include:

- Study design (e.g., randomized controlled trials)
- Plausible mechanism or theory
- Temporal order (cause precedes effect)
- Replication across populations or time

Correlation can be **a clue**, but never the whole story on its own.

> **Think About It:**
>
> - Have you ever read a headline that confused correlation with causation?

- What other variables might explain a strong correlation you've seen?
- How can researchers responsibly use correlation to develop better studies?

This page titled 9.5: Cautions with Correlation and Causation is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Mathematics Department.

9.6: Vocabulary (Chapter 9)

Chapter 9 Vocabulary

Bivariate Data

Data that includes two quantitative variables measured for each individual in a study. Often used to examine relationships between variables.

Scatterplot

A graph that displays pairs of numerical data as points on a coordinate plane. Each point represents an individual's value for two variables (x and y).

Direction of Association

A description of how one variable tends to change as the other changes. Associations can be positive, negative, or have no clear direction.

Positive Association

When values of one variable tend to increase as the values of the other variable also increase. The pattern typically slopes upward from left to right.

Negative Association

When values of one variable increase as the other decreases. The pattern typically slopes downward from left to right.

Correlation (r)

A numerical measure between –1 and 1 that describes the strength and direction of a linear relationship between two variables. A value close to ±1 indicates a strong linear relationship.

Correlation Coefficient (r)

Another name for the correlation measure. It quantifies linear association but does not imply causation.

Least-Squares Regression Line (LSRL)

The line that best fits a scatterplot and minimizes the sum of the squared residuals. It models the linear relationship and can be used for prediction.

Regression Equation

The formula of the form $\hat{y} = a + bx$ where $\hat{y}$ is the predicted value, a is the intercept, and b is the slope.

Slope

In a regression line, the slope tells how much the predicted value of y changes for each one-unit increase in x. It represents the rate of change.

Intercept

The y-value predicted when $x = 0$ in a regression model. It may or may not be meaningful depending on the context and data range.

Residual

The difference between an actual data value and the value predicted by a regression model. Residual = actual – predicted.

Interpolation

Using a regression model to predict a value within the range of the observed data.

Extrapolation

Using a regression model to predict a value outside the range of the observed data. It can be risky because the relationship may not hold beyond the data range.

Causation

The idea that one variable has a direct effect on another. Causation implies a change in one variable produces a change in the other. It requires evidence beyond just statistical association.

Correlation Does Not Imply Causation

A reminder that even strong relationships between variables do not prove one causes the other. Alternate explanations or lurking variables may be involved.

This page titled 9.6: Vocabulary (Chapter 9) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Mathematics Department.

9.7: Practice (Chapter 9)

9.1: Scatter plots and Direction of Association

1. What is a scatterplot and when is it appropriate to use one?
2. Draw a scatterplot of the following data by hand or using technology. What pattern do you observe?

x	y
1	2
2	4
3	6
4	7
5	10

3. The following table shows the poverty rates and cell phone usage in the United States. Construct a scatter plot of the data.

Year	Poverty Rate	Cellular Usage per Capita
2003	12.7	54.67
2005	12.6	74.19
2007	12	84.86
2009	12	90.82

4. Does the higher cost of tuition translate into higher-paying jobs? The table lists the top ten colleges based on mid-career salary and the associated yearly tuition costs. Construct a scatter plot of the data.

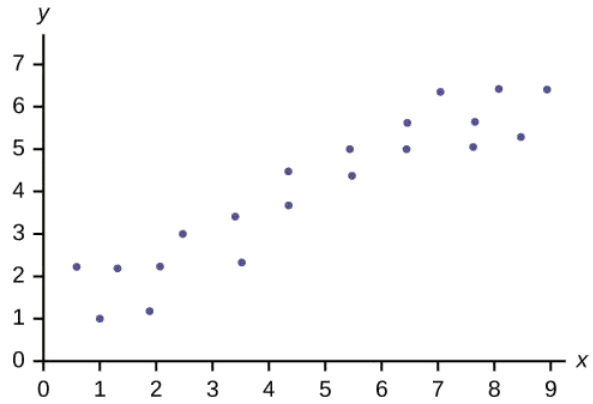
School	Mid-Career Salary (in thousands)	Yearly Tuition
Princeton	137	28,540
Harvey Mudd	135	40,133
CalTech	127	39,900
US Naval Academy	122	0
West Point	120	0
MIT	118	42,050
Lehigh University	118	43,220
NYU-Poly	117	39,565
Babson College	117	40,400
Stanford	114	54,506

5. A random sample of ten celebrities produced the following data where x is the number of endorsements the celebrity has and y is the amount of money made (in millions of dollars). Draw a scatter plot of the data.

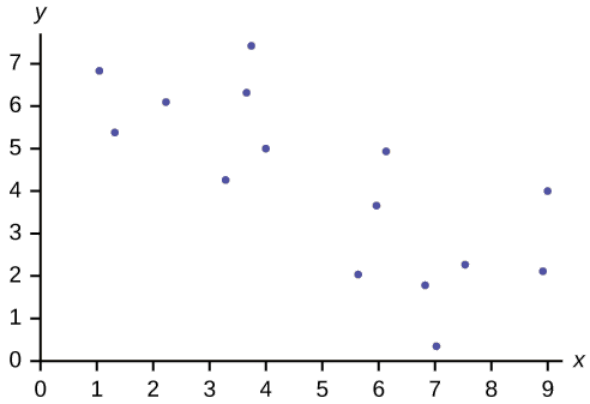
Celebrity endorsements and income

x	y	x	y
0	2	5	12
3	8	4	9
2	7	3	9
1	3	0	3
5	13	4	10

6. Office productivity is relatively low when the employees feel no stress about their work or job security. However, high levels of stress can also lead to reduced employee productivity. Sketch a plot to represent the relationship between stress and productivity.
7. Does the scatter plot below appear to have a positive or negative association?



8. Does the scatter plot below appear to have a positive or negative association?



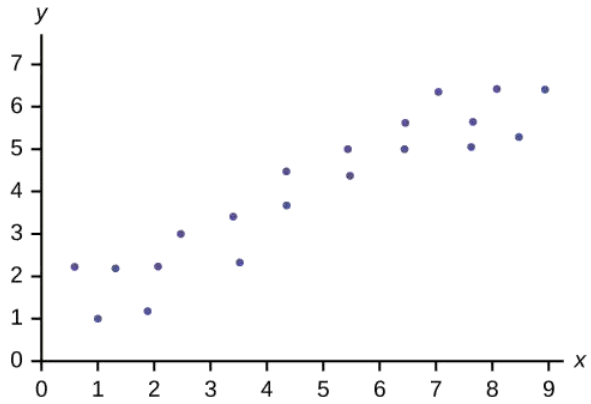
9. Suppose you were to collect data for the following pair of variables, people: age, grip strength. You want to make a scatterplot.
   a. Which variable would you use as the explanatory variable?
   b. Which variable would you use as the response variable?
   c. Would you expect to see a positive or negative association?
10. What does it mean to say two variables are positively associated?
    a. There is a nonlinear relationship between the variables.
    b. When the value of one variable increases, the value of the other variable decreases.
    c. There is a linear relationship between the variables.
    d. When the value of one variable increases, the value of the other variable also increases.
11. Explain the difference between a positive association and a negative association. Sketch a rough example of each.

12. Describe what a scatterplot with no association would look like. What might that tell you about the relationship between two variables?
13. If two variables had a positive association, what would happen if the explanatory and response variables were swapped? What association would result?

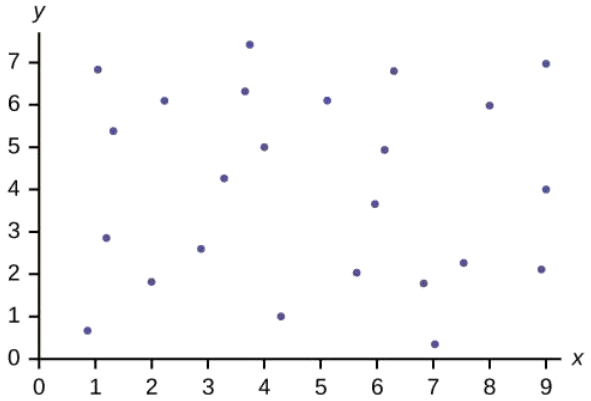
9.2: Correlation Coefficient (r)

1. What are the possible values for the correlation coefficient?
2. What does a correlation coefficient of $r = 0.94$ tell us about the relationship of the two variables in a dataset?
3. How does the value of r change as the points in a scatterplot move further from a perfect straight line?
4. In order to have a correlation coefficient between traits A and B, it is necessary to have:
    a. one group of subjects, some of whom possess characteristics of trait A, the remainder possessing those of trait B
    b. measures of trait A on one group of subjects and of trait B on another group
    c. two groups of subjects, one which could be classified as A or not A, the other as B or not B
    d. two groups of subjects, one which could be classified as A or not A, the other as B or not B
5. Match each r value below with the best description:
    1. Strong negative linear relationship
    2. No linear relationship
    3. Moderate positive correlation
    4. Strong positive linear relationship

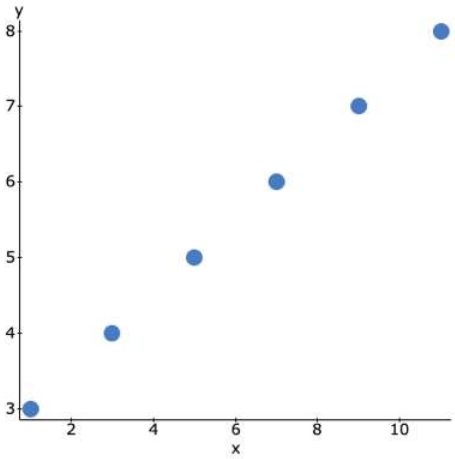
    a. $r = -0.15$
    b. $r = -0.92$
    c. $r = 0$
    d. $r = 0.83$
6. Does the scatter plot below appear to have a strong or weak linear relationship?



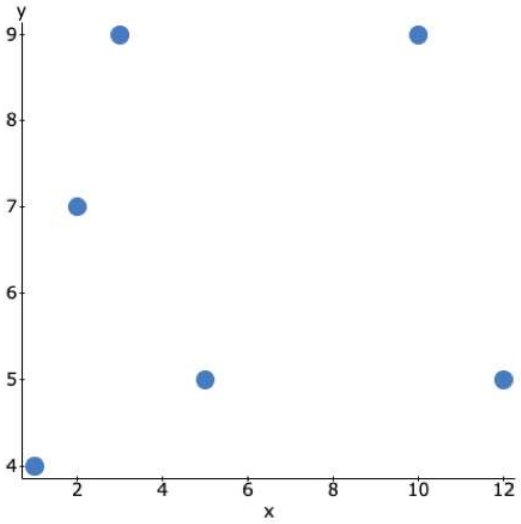
7. Does the scatter plot below appear to have a strong or weak linear relationship?



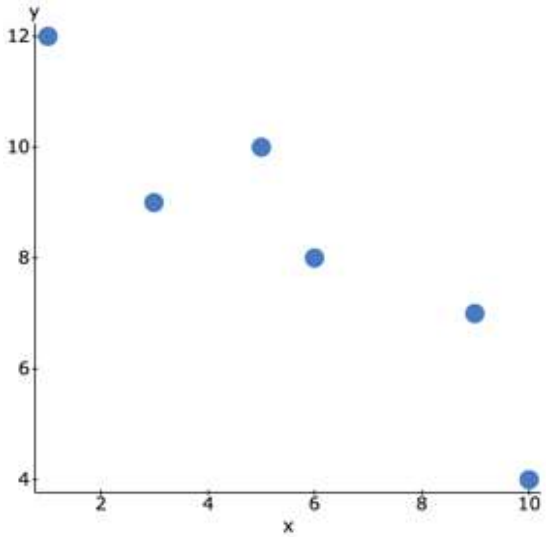
8. Estimate the correlation coefficient for the data represented in the scatterplot.



9. Estimate the correlation coefficient for the data represented in the scatterplot.



10. Estimate the correlation coefficient for the data represented in the scatterplot.



11. If the correlation coefficient is negative, this means
    a. the slope of the regression line is negative
    b. the test statistic can not be calculated

c. there is a strong correlation between the variables
d. there is no correlation between the variables
e. the calculation was done wrong
f. the regression line can be used for prediction

12. Consider a sample least squares regression analysis between a dependent variable (Y) and an independent variable (X). A sample correlation coefficient of −1 (minus one) tells us that
    a. there is no relationship between Y and X in the sample
    b. there is no relationship between Y and X in the population
    c. there is a perfect negative relationship between Y and X in the population
    d. there is a perfect negative relationship between Y and X in the sample.
13. If the correlation between age of an auto and money spent for repairs is +.90
    a. 81% of the variation in the money spent for repairs is explained by the age of the auto
    b. 81% of money spent for repairs is unexplained by the age of the auto
    c. 90% of the money spent for repairs is explained by the age of the auto
    d. none of the above
14. A national consumer magazine reported the following correlations. The correlation between car weight and car reliability is -0.9. The correlation between car weight and annual maintenance cost is 0.3. According to this information, which of the following statements are true?
    a. Heavier cars tend to be less reliable.
    b. Heavier cars tend to cost more to maintain.
    c. Car weight is related more strongly to reliability than to maintenance cost.
15. Calculate the correlation coefficient for the data in the following table. Use the value to describe the relationship between the variables.

Age	Number of Driver Deaths per 100,000
17.5	38
22	36
29.5	24
44.5	20
64.5	18
80	28

16. Calculate the correlation coefficient for the data in the following table. Use the value to describe the relationship between the variables.

Year of Birth	Life Expectancy
1930	59.7
1940	62.9
1950	70.2
1965	69.7
1973	71.4
1982	74.5
1987	75
1992	75.7
2010	78.7

17. Calculate the correlation coefficient for the data in the following table. Use the value to describe the relationship between the variables.

Year	Poverty Rate	Cellular Usage per Capita
2003	12.7	54.67
2005	12.6	74.19
2007	12	84.86
2009	12	90.82

18. A random sample of ten celebrities produced the following data where x is the number of endorsements the celebrity has and y is the amount of money made (in millions of dollars). Calculate the value of the correlation coefficient for this data set. What does this mean in this context?

Celebrity Endorsements and Income

x	y	x	y
0	2	5	12
3	8	4	9
2	7	3	9
1	3	0	3
5	13	4	10

14. Eduardo and Rosie are both collecting data on number of rainy days in a year and the total rainfall for the year. Eduardo records rainfall in inches and Rosie in centimeters. How will their correlation coefficients compare?
15. What are some limitations of using only r to describe a relationship?
16. List three things the correlation coefficient tells us, and two things it does not tell us.

9.3: Least-Squares Regression Line

1. Explain what the "least-squares" in "least-squares regression line" means and why it's useful.
2. Use the following table to calculate the equation of the least squares regression line.

x	y
1	2
2	4
3	6
4	7
5	10

3. The table below shows the life expectancy for an individual born in the United States in certain years. Calculate the equation of the least squares regression line.

Year of Birth	Life Expectancy
1930	59.7
1940	62.9
1950	70.2
1965	69.7
1973	71.4
1982	74.5
1987	75
1992	75.7
2010	78.7

4. The height (sidewalk to roof) of notable tall buildings in America is compared to the number of stories of the building (beginning at street level). Calculate the equation of the least squares regression line.

Height (in feet)	Stories
1,050	57
428	28
362	26
529	40
790	60
401	22
380	38
1,454	110
1,127	100
700	46

5. Does the higher cost of tuition translate into higher-paying jobs? The table lists the top ten colleges based on mid-career salary and the associated yearly tuition costs.
   a. Calculate the equation of the least squares regression line for the full table.
   b. Remove any outliers and then re-calculate the equation of the least squares regression line.

School	Mid-Career Salary (in thousands)	Yearly Tuition
Princeton	137	28,540
Harvey Mudd	135	40,133
CalTech	127	39,900
US Naval Academy	122	0
West Point	120	0
MIT	118	42,050
Lehigh University	118	43,220
NYU-Poly	117	39,565
Babson College	117	40,400
Stanford	114	54,506

6. A random sample of ten celebrities produced the following data where x is the number of endorsements the celebrity has and y is the amount of money made (in millions of dollars). Find the equation of the least-squares regression line for this data.

Celebrity endorsements and income

x	y	x	y
0	2	5	12
3	8	4	9
2	7	3	9
1	3	0	3
5	13	4	10

7. Ornithologists, scientists who study birds, tag sparrow hawks in 13 different colonies to study their population. They gather data for the percent of new sparrow hawks in each colony and the percent of those that have returned from migration. **Percent return:** 74; 66; 81; 52; 73; 62; 52; 45; 62; 46; 60; 46; 38. **Percent new:** 5; 6; 8; 11; 12; 15; 16; 17; 18; 18; 19; 20; 20
   a. Use the data to make a scatter plot.
   b. Find the equation of the least-squares regression line. Add this to your scatter plot from part a.
   c. How well does the regression line fit the data? Explain your response.
8. The following table shows per-capita data for the annual wine consumption in liters and deaths from heart disease rate in a random sample of 10 countries.

Wine Consumption and Heart Disease

Wine consumption in liters	2.5	3.9	2.9	2.4	2.9	0.8	9.1	2.7	0.8	0.7
Deaths from heart disease	221	167	131	191	220	297	71	172	211	300

   a. Make a scatter plot of the data
   b. Find the equation of the least-squares regression line. Add this to your scatter plot from part a.
   c. How well does the regression line fit the data? Explain your response.
   d. Do the data provide convincing evidence that there is a linear relationship between the amount of alcohol consumed and the heart disease death rate?
9. A group of students measure the length and width of a random sample of beans. They are interested in investigating the relationship between the width (x) and length (y). Their summary statistics are displayed in the table below. All units, if applicable, are millimeters.

Mean width:	7.5
Stdev width:	0.909
Mean height:	15.655
Stdev height:	1.512
Correlation coefficient:	0.7769

   a. The students are interested in using the width of the beans to predict the height. Calculate the slope of the regression equation.
   b. Write the equation of the best-fit line that can be used to predict bean heights. Use x to represent width and y to represent height.

c. What fraction of the variability in bean heights can be explained by the linear model of bean height vs width? Express your answer as a decimal.
d. If, instead, the students are interested in using the height of the beans to predict the width, calculate the slope of this new regression equation.
e. Write the equation of the best-fit line that can be used to predict bean widths. Use *x* to represent height and *y* to represent width.

10. Body frame size is determined by a person's wrist circumference in relation to height. A researcher measures the wrist circumference and height (both in inches) of a random sample of individuals. The data is displayed below.

Mean wrist:	6.0509
Stdev width:	0.8332
Mean height:	29.7088
Stdev height:	5.9484
Correlation coefficient:	0.7936

a. calculate the slope b of the regression equation $\hat{y} = a + bx$
b. calculate the intercept a of the regression equation $\hat{y} = a + bx$
c. write out the full regression equation $\hat{y} = a + bx$

11. What is interpolation? What is extrapolation? Which one is more risky — and why?

9.4: Interpreting Slope, Intercept, and Residuals

1. A researcher wishes to examine the relationship between years of schooling completed and the number of pregnancies in young women. Her research discovers a linear relationship, and the least squares line is: $\hat{y} = 1 - 4x$ where x is the number of years of schooling completed and $haty$ is the number of pregnancies. The slope of the regression line can best be interpreted by which of the following way:
   a. When amount of schooling increases by one year, the number of pregnancies increases by 4.
   b. When amount of schooling increases by one year, the number of pregnancies decreases by 1.
   c. When amount of schooling increases by one year, the number of pregnancies increases by 1.
   d. When amount of schooling increases by one year, the number of pregnancies decreases by 4.
2. A manager at an insurance company wishes to examine the relationship between years of college completed and the starting salary at their company. Their research discovers a linear relationship, and the least squares line is: $\hat{y} = 30.956 + 3.3x$ where x is the number of years of college completed and y is the starting salary in thousands of dollars. The slope of the regression line can be interpreted in the following way:
   a. When starting salary goes up by 30.956 thousand dollars, the number of years of college is 3.3.
   b. When amount of college increases by one year, the starting salary increases by 30.956 thousand dollars.
   c. When amount of college increases by one year, the starting salary increases by 3.3 thousand dollars.
   d. When amount of college increases by one year, the starting salary decreases by 30.956 thousand dollars.
   e. When amount of college increases by one year, the starting salary decreases by 3.3 thousand dollars.
   f. When amount of college increases by 3.3 years, the starting salary increases by 30.956 thousand dollars.
3. The equation of the regression line for hours x spent cycling in the summer and hours y spent working in the summer is $\hat{y} = -0.7x + 42$.
   a. Interpret the slope in a complete sentence with correct units.
   b. Interpret the intercepts in a complete sentence with correct units.
4. The equation of the regression line that relates income in dollars of student's parents, x, with cost in dollars of tuition, y, is $\hat{y} = 8700 + 0.03x$.
   a. Interpret the slope in a complete sentence with correct units.

b. Interpret the intercepts in a complete sentence with correct units.

5. The equation of the regression line that relates dollars spent monthly on advertising, x, to monthly revenue, y, is $\hat{y} = 5800 + 10x$.
   a. Interpret the slope in a complete sentence with correct units.
   b. Interpret the intercepts in a complete sentence with correct units.
6. The equation of the regression line that relates one hundredth percent blood alcohol level x to the number of names a person can remember five minutes after hearing them is (\hat{y} = 12 - 0.5x \)
   a. Interpret the slope in a complete sentence with correct units.
   b. Interpret the intercepts in a complete sentence with correct units.
7. A random sample of ten celebrities produced the following data where x is the number of endorsements the celebrity has and y is the amount of money made (in millions of dollars). Interpret the slope and intercept(s) of the line of best fit for this data.

Celebrity endorsements and income

x	y	x	y
0	2	5	12
3	8	4	9
2	7	3	9
1	3	0	3
5	13	4	10

8. Given a regression line $\hat{y} = 30 + 4.2x$, what is the predicted value of y when $x = 6$?
9. An electronics retailer used regression to find a simple model to predict sales growth in the first quarter of the new year (January through March). The model is good for 90 days, where x is the day. The model can be written as follows: $\hat{y} = 101.32 + 2.48x$ where $\hat{y}$ is in thousands of dollars.
   a. What would you predict the sales to be after one month?
   b. What is the meaning of $\hat{y}(100)$?
   c. When do we predict the sales to reach one million dollars?
10. A regression was run to determine if there is a relationship between hours of TV watched per day (x) and number of sit ups a person can do (y). Use the results below to predict the number of sit ups a person who watches 10 hours of TV can do (to one decimal place)

```
y=ax+b
a=-0.948
b=31.264
r²=0.583696
r=-0.764
```

11. Calculate the residual for a point where the actual value is 74 and the predicted value is 68. Interpret the result.
12. The linear regression equation for a data set is $\hat{y} = -4.1 + 1.6x$. The actual value at $x = 15$ is 20.4. What is the residual value at $x = 15$?
13. A regression equation is $\hat{y} = 55 + 0.7x$
    a. What is the predicted value when $x = 10$?
    b. A student with $x = 10$ has an actual value of 66. What is the residual?
14. Suppose one collected the following information where X is diameter of tree trunk and Y is tree height. What is your estimate of the average height of all trees having a trunk diameter of 7 inches?

Tree diameter and height

X	Y
4	8
2	4
8	18
6	22
10	30
6	8

15. The following table shows the life expectancy for an individual born in the United States in certain years.

Life Expectancies

Year of Birth	Life Expectancy
1930	59.7
1940	62.9
1950	70.2
1965	69.7
1973	71.4
1982	74.5
1987	75
1992	75.7
2010	78.7

a. Decide which variable should be the independent variable and which should be the dependent variable.
b. Draw a scatter plot of the ordered pairs.
c. Calculate the least squares line.
d. Find the correlation coefficient. Is it significant?
e. Find the estimated life expectancy for an individual born in 1950 and for one born in 1982.
f. Why aren't the answers to part e the same as the values in the table that correspond to those years?
g. Use the two points in part e to plot the least squares line on your graph from part b.
h. Based on the data, is there a linear relationship between the year of birth and life expectancy?
i. Which point has the largest residual? Explain what the residual means in context. Is this point an outlier? An influential point? Explain.Using the least squares line, find the estimated life expectancy for an individual born in 1850. Does the least squares line give an accurate estimate for that year? Explain why or why not.
j. What is the slope of the least-squares (best-fit) line? Interpret the slope.
k. What does the sign (positive or negative) of a residual tell you about the prediction?

16. The height (sidewalk to roof) of notable tall buildings in America is compared to the number of stories of the building (beginning at street level).

Heights of buildings

Height (in feet)	Stories
1,050	57
428	28
362	26
529	40
790	60
401	22
380	38
1,454	110
1,127	100
700	46

a. Using “stories” as the independent variable and “height” as the dependent variable, make a scatter plot of the data.
b. Does it appear from inspection that there is a relationship between the variables?
c. Calculate the least squares line.
d. Find the correlation coefficient. Is it significant?
e. Find the estimated heights for 32 stories and for 94 stories.
f. Based on the data in the table, is there a linear relationship between the number of stories in tall buildings and the height of the buildings?
g. Are there any outliers in the data? If so, which point(s)?
h. What is the estimated height of a building with six stories? Does the least squares line give an accurate estimate of height? Explain why or why not.
i. Based on the least squares line, adding an extra story is predicted to add about how many feet to a building?
j. What is the slope of the least squares (best-fit) line? Interpret the slope.

17. Explain why residuals are important when assessing the usefulness of a model.
18. What is the problem with using a regression model to predict what will happen in the distant future?

9.5: Cautions with Correlation and Causation

1. What is the difference between correlation and causation?
2. A study of the CO2 concentration and global temperature data found that the correlation between CO2 concentration and global temperature was 0.993. Based on this correlation, can it be concluded that the increased level of CO2 concentration is causing global warming?
3. A study was done on smoking and lung capacity. 200 smokers took part in a study that asked them how many cigarettes a day they smoked and then measured their lung capacity. The correlation was found to be $r = -0.992$. Based solely on this study it can be concluded that smoking causes lung cancer.
4. A study was done asking people how much money they spend per month on their natural gas bill and how much money per month they spend on their electric bill. The correlation r was found to be 0.94 and the p-value for correlation was 0.0003. Does a person with a high natural gas bill also have a high electric bill?
5. Give an example of two variables that might have a strong correlation but are not causally related. Explain what the "lurking variable" might be.
6. Give an example of two variables that might have a strong correlation and are causally related. Explain how the causality might be proven.
7. Select True or False for the following statements.

a. When ice cream sales go up, so do crime rates. A reasonable conclusion is that there is an association between ice cream sales and the crime rate.
b. Association does not imply causation.
c. If there is an association identified between the explanatory and response variable, the relationship is causal.
d. The response variable is always caused by the explanatory variable
e. When ice cream sales go up, so do crime rates. A reasonable conclusion is that the increase in ice cream sales causes an increase in the crime rate.

8. Can correlation help us form hypotheses about cause and effect? Why or why not?
9. Explain why a randomized controlled experiment is better than an observational study when trying to establish causation.
10. Review one recent article or claim you've seen online or in the media that may confuse correlation with causation. What evidence would you need to decide if a causal relationship exists?

Practice problems include parts of:

Introductory Statistics 2e by OpenStax is licensed under CC BY 4.0

Access for free at https://openstax.org/books/introductory-statistics-2e/pages/1-introduction

and

OpenIntro Statistics by David Diez, Christopher Barr, & Mine Çetinkaya-Rundel is licensed under CC BY 3.0

9.7: Practice (Chapter 9) is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

- **Current page** has no license indicated.
- **Licensing** by Mathematics Department has no license indicated.

CHAPTER OVERVIEW

10: Project, Labs, Appendix and Tables

10.1: Project Assignments

10.1.1: Sampling and Data Web Page

10.1.2: Quantitative Variable Analysis

10.1.3: Qualitative Variable Analysis

10.1.4: Interval Estimates

10.1.5: Hypothesis Testing

10.1.6: Correlation

10.2: Tech Labs

10.2.1: TL 1

10.2.2: TL 2

10.2.3: TL 3

10.2.4: TL 4

10.2.5: TL 5

10.2.6: TL 6

10.2.7: TL 7

10.3: Summary Statistic Cards

10.4: Project Reflection

10.5: z-Table

10.6: t-Table

This page titled 10: Project, Labs, Appendix and Tables is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Mathematics Department.

10.1: Project Assignments

10.1: Project Assignments is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

10.1.1: Sampling and Data Web Page

Semester Project: Sampling and Data Collection from Two Cities

In this first step of your semester-long housing project, you will gather data on **1-bedroom apartment rentals** from **two different metro cities** of your choosing. One of these cities must be **Lakewood, CO**. You'll select a second metro area (e.g., Golden, Boulder, Aurora) that will allow for meaningful comparison throughout the project.

Your goal is to apply principles of **random sampling** to gather **40 listings from each city (80 total)**. For each listing, you'll collect a combination of **quantitative** and **qualitative** variables such as rental price, square footage, and a chosen feature like pet policy or parking. This will form the foundation of your dataset for the rest of the course.

Objectives

- Apply statistics terminology to a real-world data collection example (population, sample, parameter, statistic, variables, technique)
- Differentiate between qualitative and quantitative variables
- Utilize appropriate random sampling techniques
- Generate relevant initial questions about your sample data

Assignment Steps

1. **Gather Samples**
   Go to housing rental websites such as apartments.com, Zillow.com, or Craigslist ... . Search for **1-bedroom apartments** currently available in your two selected cities. You may use zip codes or city names.
2. **Apply Random Sampling:**
   Use a clearly defined random sampling technique to select exactly **40 listings from each city**. Record your combined data in a spreadsheet with clearly labeled columns:
   - **City**
   - **Address**
   - **Rental price per month**
   - **Square footage**
   - **Fees**
   - **Where we collected the data from**
   - One or more **qualitative variable** of your choice (e.g., pet policy, parking type, flooring type, in-unit washer/dryer)
3. **Describe Your Process:** In a new word document. (This will become the final paper)
   a. Explain your **sampling method**. Was your approach truly random? What challenges did you encounter?
   b. Clearly identify **which variables** are quantitative and which are qualitative
   c. List at least **three questions** you might explore later based on this data

---

Helpful Tips for Data Collection

- Don't just use the first 40 listings; those may reflect search order biases or sponsored properties
- Standardize your **qualitative variable** categories (e.g., "Street Parking" vs. "Roadside Parking" → choose one version)
- Think forward: make sure your data can be used for comparison between cities in later assignments

---

Submission Instructions

- Submit your Excel file with clearly labeled columns and text box reflection.
- Submit your Word document with the descriptions and any written info.
- Your spreadsheet should contain **80 total listings (40 per city)**
- File name format (Excel): `LastName_SampleData.xlsx`
- File name format (Word): `LastName_FirstDraft.docx`
- Upload to the "Project Assignments" folder in our LMS

10.1.1: Sampling and Data Web Page is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

10.1.2: Quantitative Variable Analysis

Semester Project: Analyzing a Quantitative Variable (Rental Prices)

In this project assignment, you will perform a thorough analysis the **quantitative variables** you collected in your Sampling and Data Project: the **rental prices for 1-bedroom apartments and their square footage**.

Using Excel or another tool, you'll create **frequency tables, visual summaries** (histogram and boxplot), and calculate relevant **descriptive statistics** including mean, median, standard deviation, and five number summary. You'll use these tools to describe the shape, center, and spread of your data, and interpret your findings in plain language.

Objectives

- Construct a frequency table and histogram
- Construct a boxplot and identify potential outliers
- Describe distribution shape (modality, skew), center, and spread
- Calculate and interpret key statistics: mean, median, range, standard deviation, and five number summary
- Use technology to support analysis

Assignment Steps

1. **Create a frequency table** for rental prices using 5 to 10 evenly spaced class intervals.
2. **Create a histogram** from the frequency table in Excel. Your histogram should include:
    - A meaningful chart title
    - Clearly labeled axes
    - Equal-width bins/classes
3. **Create a boxplot** for rental price data to visually assess spread and potential outliers.
4. **Calculate the following descriptive statistics**:
    - Mean
    - Median
    - Standard Deviation
    - Range
    - Five number summary
5. **Repeat #1-4 for the square footage variable**
6. **Summarize the distribution** of each variable. This should be a paragraph or two for each. Include:
    - Shape of the distribution (modal shape, skewness)
    - Noted outliers and what they might represent
    - Comparison of mean and median
    - What the spread tells you about market variability
7. **Real-World Interpretation:** What does this data tell us about 1-bedroom rental prices in Lakewood? Who might use this information and why? Are any listings significantly higher or lower than average, and what might that mean?
8. **Circle back to the question you wrote in the data collection part of the project:** Can you answer any price-related questions you posed earlier using your visuals or stats?

Helpful Tips

- Class intervals in your histogram should be **equal-width, non-overlapping, and intuitive** (e.g., $100 increments)
- **Use labels:** Axis titles and legend titles should match your variable name exactly
- The **boxplot** is great for visualizing skewness and identifying any extreme outliers
- If your mean is much higher than your median, that might signal right skew (a few expensive listings pulling up the average)

10.1.2: Quantitative Variable Analysis is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

10.1.3: Qualitative Variable Analysis

Semester Project – Visualizing a Qualitative Variable

In this portion of your housing data project, you'll analyze a **qualitative (categorical) variable** from your collected rental data.

Your goal is to summarize this variable (e.g.: pet policy, parking type, flooring material) using a **frequency table**, a **pie chart**, a **bar graph**, and written analysis. You'll also begin crafting a new section of your final report: *Visual Summary of a Categorical Variable.*

Objectives

- Analyze graphical and numerical summaries of categorical data
- Use technology tools (e.g., Excel or Google Sheets) to build high-quality visualizations
- Draw conclusions from real-world housing data

Assignment Steps

1. Create a **frequency table** that counts how often each category appears in your selected qualitative variable (e.g., how many listings allowed pets, or what types of parking were offered).
2. Use the table to create **two separate graphs**:
   - A **pie chart**
   - A **bar graph**

   Make sure each chart is clearly labeled with a title, category names, and, if needed, percentage labels or a legend.
3. In a **Text Box** (or a Notes section in your spreadsheet or document), respond to the following:
   a. Which chart better represents your data — bar graph or pie chart? Justify your choice.
   b. What insights or surprises do your visuals reveal? Are there dominant or rare categories?
   c. Who might benefit from this information? (e.g., landlords, renters, city planners).
   d. Look back at any questions you asked about this variable in Assignment 1, can you answer any of them now?

What to Include in Your Final Paper (Draft Version)

This work will become a key section of your final project: **Visual Summary of a Categorical Variable**. In the draft you are building for your paper, make sure to include:

- A brief description of the variable you investigated and why it's relevant
- Your frequency table embedded or described in paragraph form
- Visual charts (labeled and titled)
- Your observations and written interpretation
- Your reflection on why one graph communicates better than the other

---

Helpful Tips

- **Clean your categories:** Check the data for consistency. Example: group "Street Parking" and "Roadside Parking" under one label.
- **Pie charts** are best with fewer categories and when you want to show part–whole relationships.
- **Bar graphs** are better when categories are numerous or differences are large and easy to compare side-by-side.
- **Label everything:** chart titles, axis labels, data labels (if helpful), and consider what colors you're using.

Submission Instructions

- Upload your Excel file or slide with:
  - Frequency table
  - Bar graph and pie chart
  - Written explanation in a Text Box or Notes section

- File format: `LastName_QualitativeGraph.xlsx`
- Submit to the “Project Assignments” folder in our LMS
- Also upload your current draft of the project write-up including any updates to parts 1 and 2 as well as the figures from part 3.

10.1.3: Qualitative Variable Analysis is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

10.1.4: Interval Estimates

Semester Project – Assignment 4: Confidence Interval Analysis (Comparing Two Cities)

In this portion of your semester project, you will apply the concepts of **confidence intervals** to the data you collected for **two cities** (40 listings from each). Confidence intervals help us estimate unknown population parameters, like the **true average rent** or the **proportion of listings with a feature like pet-friendly policies,** based on sample data.

You'll compute **confidence intervals for the mean rental price** of 1-bedroom apartments in each city at the **90%** and **95%** confidence levels. Finally, you'll compare the city results and reflect on how different confidence levels affect interpretation.

Objectives

- Make inferences about two populations using sample data
- Construct confidence intervals for means
- Compare intervals across different confidence levels
- Interpret results in real-world language and context

Assignment Steps

1. **Estimate the mean rental price** for each city at both the **90%** and **95%** confidence levels. Show all work clearly.
2. **Compare the confidence intervals across the two cities:**
    - Do the intervals at 90% confidence overlap? What about at 95%?
    - What do the intervals tell you about the cities rental markets?
    - Discuss the impact of using a 95% vs. 90% confidence level.
3. **Final reflection:** Who might benefit from these insights? How could these results be used in a practical way? Write 3-4 sentences summarizing your results.

What to Include in Your Final Paper (Draft Version)

This section becomes part of your project report under: **Comparing Estimates from Two Cities**. Your draft should include:

- Listed confidence intervals (including level used)
- Include your final reflection from the steps above
- (Optional) Visual aids comparing intervals side by side

This section should only be 1/2-1 page.

10.1.4: Interval Estimates is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

10.1.5: Hypothesis Testing

Semester Project – Assignment 5: Hypothesis Testing – Comparing Lakewood and Golden

In this assignment, you'll conduct a formal **hypothesis test** comparing mean rental prices between **your tow cities** based on the rental data you collected. The focus will be on testing a claim about the difference (or lack of difference) in average rent for 1-bedroom apartments.

Using a **0.05 significance level**, you'll construct your null and alternative hypotheses, perform a two-sample t-test for the means, and interpret your results in plain language. You'll also explore the potential impact of Type I and Type II errors and reflect on how to strengthen your study.

Objectives

- Make inferences about one or more population means using sample data
- Conduct a two-sample hypothesis test comparing population means
- Use technology (e.g., Excel, calculators, or online tools) to perform statistical tests
- Interpret results in context, including discussion of potential errors

---

Assignment Steps

1. **Make a claim** based on your observations:
   - Do you expect Lakewood's average rent to be higher, lower, or about the same as your other city?
   - Write both your **null hypothesis (H_0)** and your **alternative hypothesis (H_1)** using proper statistical notation. Hint: this is a claim about two population means.
2. **Conduct a two-sample t-test** at the **0.05 significance level**:
   - Record your sample means, standard deviations, sample sizes, test statistic, and p-value
   - Use the p-value to make your conclusion (or compare the test statistic to the critical value)
   - Clearly **state the conclusion** in plain language (avoid technical phrases like "fail to reject H_0")
3. **Discuss Type I and Type II errors:**
   - Describe what each type of error represents in your context
   - What are the **real-world consequences** if an error occurs? Consider stakeholders such as community leaders, renters, or housing advocates.
4. **Reflect on your study design:**
   - Would having more data help your conclusion?
   - Could a different sampling method improve your results?
   - What would help you be more confident in your findings?

What to Include in Your Final Paper (Draft Version)

This assignment builds a section of your final project paper titled: **Hypothesis Test: Comparing City Rental Prices**. Your write-up should include:

- The original claim (clearly stated)
- Null and alternative hypotheses
- Test results: test statistic, p-value (or critical value)
- Plain language conclusion
- Real-world meaning of the findings: who it affects and how
- Discussion of potential errors and what could improve confidence

---

Helpful Tips

- **Include unit labels** (e.g., rent in dollars) in your explanation
- Use clear, everyday language to help your reader (and yourself!) explain what happened in the test and how to interpret it

---

10.1.5: Hypothesis Testing is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

10.1.6: Correlation

Semester Project – Assignment 6: Correlation and Scatterplot Analysis

In this portion of your project, you will explore the relationship between two key quantitative variables in your dataset: **rental price** and **square footage** of 1-bedroom apartments. Your goal is to measure the **strength and direction** of any relationship using a scatterplot, calculate the **correlation coefficient (r)**, and interpret your findings in a real-world context.

This analysis introduces the concepts of **correlation** and **regression**, which can inform future predictions and deeper understanding of data patterns. You'll also explore the possibility of other variables influencing rental price and describe how they might positively or negatively relate.

Objectives

- Analyze paired quantitative data using correlation and regression tools
- Create a scatterplot visualizing relationship between square footage and price
- Calculate and interpret the correlation coefficient (r)
- Identify possible outliers or surprising patterns in the plot
- Use technology to deepen understanding of relationships between numerical variables

---

Assignment Steps

1. **Create a scatterplot** of **price vs. square footage** using Excel, Google Sheets, or another tool. Your graph must include:
   - Clearly labeled axes (e.g., X-axis: square footage, Y-axis: price)
   - A descriptive and meaningful title (e.g., "Rental Price vs. Square Footage in Lakewood & Golden")
2. **Calculate the correlation coefficient (r)** for your paired dataset. Use technology to determine:
   - The r-value (strength & direction)
   - Whether it is statistically significant (use a critical value table or p-value)
3. **Answer the following reflection questions:**
   1. **Interpret the correlation:** Is it strong, weak, or moderate? Is it positive or negative? What does that tell you about the relationship between square footage and rental price in your sample?
   2. **Contextualize the scatterplot:** What story does it tell about the local rental market? Are there outliers, holes, or clusters that influence your interpretation?
   3. **Identify a negatively correlated variable:** What variable (not square footage) might be inversely related to rental price? Make a suggestion and support your reasoning.

What to Include in Your Final Paper (Draft Version)

This analysis forms the foundation for your paper section titled: **Exploring Relationships Between Variables**. Be sure to include:

- Your scatterplot image or embedded chart
- The calculated correlation coefficient and significance level (if tested)
- Written interpretation of the strength and direction of the relationship
- Discussion of any visible outliers and their potential impact
- Your thoughts on other possible influencing variables, with reasoning

---

Helpful Tips

- **r-values near +1 or -1** reflect strong correlations. An r of 0.75 means a strong positive trend. An r of -0.6 would suggest a moderately strong negative correlation.
- Don't forget to label ALL axes; "Price" and "Square Footage" should include units (e.g., $USD, sq. ft.)
- **Outliers** can distort your value of r. Check your scatterplot visually for values that "don't fit," which could be apartments much larger or smaller than others in your sample.

- If you’re curious about building a **regression line** (optional), Excel and Google Sheets make it easy to add a trendline and show the equation on the chart.
- Your possible "negatively correlated" influences might include things like being farther from downtown, stricter pet policies, or lack of amenities.

Submission Instructions

- Submit your work as a Google Sheet, Excel file, slide screenshot, or document
- Include your scatterplot, r-value, and all written reflection responses
- File name format: `LastName_CorrelationAnalysis`
- Upload to the “Project Assignments” folder in the LMS

10.1.6: Correlation is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

10.2: Tech Labs

10.2: Tech Labs is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

10.2.1: TL 1

Tech Lab 1 – Introduction to Excel

Purpose of the Technology Labs

Welcome to the Technology Labs! The goal of these activities is to introduce you to Excel and guide you through how we'll use it alongside statistics in this course. You don't need prior experience, we'll start from the ground up.

Why Excel?

We live in a data-driven world where decisions in business, science, healthcare, government, and daily life are based on data. Excel is one of the most widely used tools to understand and process that data. It saves hours of manual work thanks to its ability to quickly perform calculations and organize large datasets.

Getting Excel

As an RRCC student, you can freely download Microsoft Excel using instructions posted in the Tech Labs module on D2L. This is available for both Windows and Mac computers.

- Supports: Desktop Excel on Windows or Mac
- Not compatible: Chromebooks or iPads (unless using Excel Web with limited features)

You can also access the Web version of Excel from most devices, but note that it lacks some charting and advanced tools.

Getting Started with Excel

Basics You Should Know:

- Terminology: Workbook, Worksheet, Formula Bar, Ribbon
- How to identify rows, columns, and cells (e.g., A1, B2)
- How to select multiple cells, rows, or columns
- How to copy and paste data within or across spreadsheets
- How to add new sheets and navigate between them
- How to rename sheets
- How to undo, save your work, and format text/numbers
- Resize columns and rows when needed

Your First Excel Assignment: Personal Budget

This short activity introduces formulas and cell references in Excel and gives you practice with a real-world scenario, a monthly budget. It will help build skills you'll use again throughout the course.

1. In cell A1, type **"Budget"**. Increase the font size and apply bold formatting.
2. Leave a blank row, then create column headers: **Income**, **Amount**, **Expenses**, **Amount**. Bold each heading.
3. List at least **3 sources of income** and input realistic or made-up values next to each. Format as currency.
4. List at least **7 types of expenses** in the same way (e.g., Rent, Groceries, Transportation), with values formatted in currency.
5. Add borders beneath headers and around each table.
6. Underneath the tables, add rows for: **Total Income**, **Total Expenses**, and **Balance**.
7. Use the `=SUM()` function:
   - Next to "Total Income," calculate the sum of all income entries.
   - Next to "Total Expenses," do the same for expenses.
   - Under "Balance," subtract total expenses from total income (e.g., `=B10 - D10`)
8. Try changing some income or expense values — watch how the Balance updates!
9. Bonus: Create a new column to the right of Expenses labeled **"% of Total"**. Use cell references to calculate each expense as a percentage of the total expenses. Example: `=D4/D10` (if D10 is your total expenses).

Submission Instructions

- Save your Excel file as: `LastName_Budget.xlsx`
- Upload it to the Tech Lab folder in D2L (or submit via your instructor's preferred method)

10.2.1: TL 1 is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

10.2.2: TL 2

Tech Lab 2 – Measures of Center, Variation, Boxplots, and Quartiles

Purpose

In this technology lab, you will use Excel to compute descriptive statistics, five-number summaries, and create boxplots. You'll also learn how to interpret measures like the mean, median, range, and standard deviation in the context of real data.

Descriptive Statistics – Measures of Center and Variation

In class, we'll review Excel functions to compute **mean**, **median**, and **mode**. You can also follow along with this video: Measures of Center Use the chapter time stamps in the video to jump to specific parts in Excel.

To compute **range** and **standard deviation**, Excel provides built-in functions. Note that there are different functions depending on whether you're calculating for a *sample* or a *population*

Practice with Descriptive Stats:

Download the "Anonymous Survey Data" from the course module and save a copy on your computer or OneDrive.

1. Find the **mean**, **median**, **range**, and **standard deviation** for two columns:
   - "Number of hours of sleep you got last night"
   - "Number of pets you have"
2. Compare the mean and median for each.
   - Are they close together or different?
   - Is the mean larger than the median or vice versa?
   - How might this help you infer the **skew** of the data?
3. Standard deviation is often used to describe a "typical range" of values or to flag unusual values.
   a. Compute:
      - **mean ± standard deviation**
      - **mean ± 2 × standard deviation**
   b. Interpret:
      - Is your own number of hours of sleep in the "typical" range?
      - Do any class sleep values fall outside the "unusual" range?

Five-Number Summary & Boxplots

Excel can calculate five-number summaries using these formulas:

- Minimum: `=MIN(range)`
- Q1: `=QUARTILE.INC(range, 1)`
- Median: `=MEDIAN(range)`
- Q3: `=QUARTILE.INC(range, 3)`
- Maximum: `=MAX(range)`

Excel can also automatically create a **modified boxplot**. Use this guide to get started: 5 Number Summary and Boxplots in Excel 365 (11:07). For more background knowledge, check out: 5 Number Summary and Boxplots Explained (14:52).

🧪 Practice with Boxplots:

4. Compute the **five-number summary** for the "Number of Pets" column.
5. Create a **boxplot in Excel**. Double-check that the five-number summary matches the visual graph. Add:
   - A clear, descriptive title
   - Labeled axes if possible
6. Interpret your boxplot:

- What values make up the lower 25% of the data? (Between min and Q1)
- What range makes up the middle 50%? (Interquartile Range from Q1 to Q3)
- What range makes up the upper 25%? (Q3 to max)

7. Does the boxplot show skew to the left or right? Justify your answer based on the shape and whiskers.

Interpreting a Published Boxplot

Review the following image from *A Guide to R for Social and Behavioral Science Statistics* by Brian Joseph Gillespie. It shows side-by-side boxplots comparing **number of siblings** for two groups:

- Top boxplot: people who have children
- Bottom boxplot: people who do not have children

The dots (●) represent outliers. While you can't determine exact values or the number of data points from a boxplot, you *can* describe the distribution shape, spread, and central tendencies visually.

Written Reflection:

Write a short paragraph that addresses the following:

- What does each group's boxplot tell us about the number of siblings?
- Which group seems to have more variability?
- How do the median values compare?
- What comparisons and trends can you draw across the two groups?

Submission Instructions

- Include: your descriptive stats, 5-number summary, boxplot(s), and written responses in one Excel file or Word doc
- File name: `LastName_TechLab2.xlsx` or `LastName_TechLab2.docx`
- Submit via the Tech Lab folder on the course LMS

10.2.2: TL 2 is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

10.2.3: TL 3

10.2.3: TL 3 is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

10.2.4: TL 4

10.2.4: TL 4 is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

10.2.5: TL 5

10.2.5: TL 5 is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

10.2.6: TL 6

10.2.6: TL 6 is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

10.2.7: TL 7

10.2.7: TL 7 is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

10.3: Summary Statistic Cards

(under construction)

This page titled 10.3: Summary Statistic Cards is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Mathematics Department.

10.4: Project Reflection

Final Reflection: Looking Back on Your Housing Project

You've completed your semester-long investigation into housing affordability. From posing your original question and collecting real data, to visualizing patterns, estimating with confidence intervals, and testing claims you've now walked through the full process of a statistical study.

Before we wrap up the course, take some time to reflect on what you've learned and how you learned it. You'll have about 15 minutes to write, followed by an informal whole-class conversation.

Individual Reflection Prompts

1. What was your **research question**, and how did you try to answer it using data?
2. What was something **interesting, surprising, or unexpected** that came out of your analysis?
3. What did you find **challenging or frustrating**, and how did you work around it?
4. How did your understanding of the **statistical process** change from Week 1 to now?
5. Which parts of the project do you think were the most important for understanding real-world statistics?
6. What **recommendations or strategies** would you give to future students starting this project for the first time?

Vocabulary & Concept Check

Choose 3–5 of the following terms. In 1–2 sentences each, explain how you used them in your project, and what they mean to you now that you've completed the course.

- Sample
- Population
- Bias
- Distribution
- Mean
- Median
- Standard deviation
- Confidence interval
- Proportion
- Margin of error
- Hypothesis testing
- t-distribution
- p-value
- Scatterplot
- Residual

Looking Ahead

1. How might the skills you learned here show up in future classes, jobs, or personal decisions?
2. What's one thing you wish we'd had more time to explore?
3. What did this course change about how you view statistics, or how statistics get used in the real world?

Whole-Class Conversation: Advice for the Next Group

- What would you tell future students about how to start this project?
- What worked well for you?
- What do you wish you'd done differently?
- Any good resources, websites, or strategies you recommend using?
- What other questions do you think that we could explore as a class?

We'll share some of your advice anonymously with next semester's class! They'll benefit from your insights just like you built on the ideas of students before you.

This page titled 10.4: Project Reflection is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Mathematics Department.

10.5: z-Table

z-Distribution Positive z-Values

z-Distribution Positive z-Values

https://stats.libretexts.org/@go/page/60444

Z \ Decimals	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993

Z \ Decimals	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998

z-Distribution Negative z-Values

z-Distribution Negative z-Values

Z & Decimals	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
-3.5	0.0002	0.0002	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003
-3.4	0.0003	0.0003	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0005	0.0005
-3.3	0.0005	0.0005	0.0005	0.0005	0.0006	0.0006	0.0006	0.0006	0.0006	0.0007
-3.2	0.0007	0.0007	0.0007	0.0008	0.0008	0.0008	0.0008	0.0009	0.0009	0.0009
-3.1	0.0010	0.0010	0.0010	0.0011	0.0011	0.0011	0.0012	0.0012	0.0013	0.0013
-3.0	0.0013	0.0014	0.0014	0.0015	0.0015	0.0016	0.0016	0.0017	0.0018	0.0018
-2.9	0.0019	0.0019	0.0020	0.0021	0.0021	0.0022	0.0023	0.0023	0.0024	0.0025
-2.8	0.0026	0.0026	0.0027	0.0028	0.0029	0.0030	0.0031	0.0032	0.0033	0.0034
-2.7	0.0035	0.0036	0.0037	0.0038	0.0039	0.0040	0.0041	0.0043	0.0044	0.0045
-2.6	0.0047	0.0048	0.0049	0.0051	0.0052	0.0054	0.0055	0.0057	0.0059	0.0060
-2.5	0.0062	0.0064	0.0066	0.0068	0.0069	0.0071	0.0073	0.0075	0.0078	0.0080
-2.4	0.0082	0.0084	0.0087	0.0089	0.0091	0.0094	0.0096	0.0099	0.0102	0.0104
-2.3	0.0107	0.0110	0.0113	0.0116	0.0119	0.0122	0.0125	0.0129	0.0132	0.0136
-2.2	0.0139	0.0143	0.0146	0.0150	0.0154	0.0158	0.0162	0.0166	0.0170	0.0174
-2.1	0.0179	0.0183	0.0188	0.0192	0.0197	0.0202	0.0207	0.0212	0.0217	0.0222
-2.0	0.0228	0.0233	0.0239	0.0244	0.0250	0.0256	0.0262	0.0268	0.0274	0.0281
-1.9	0.0287	0.0294	0.0301	0.0307	0.0314	0.0322	0.0329	0.0336	0.0344	0.0351
-1.8	0.0359	0.0367	0.0375	0.0384	0.0392	0.0401	0.0409	0.0418	0.0427	0.0436
-1.7	0.0446	0.0455	0.0465	0.0475	0.0485	0.0495	0.0505	0.0516	0.0526	0.0537
-1.6	0.0548	0.0559	0.0571	0.0582	0.0594	0.0606	0.0618	0.0630	0.0643	0.0655
-1.5	0.0668	0.0681	0.0694	0.0708	0.0721	0.0735	0.0749	0.0764	0.0778	0.0793
-1.4	0.0808	0.0823	0.0838	0.0853	0.0869	0.0885	0.0901	0.0918	0.0934	0.0951
-1.3	0.0968	0.0985	0.1003	0.1020	0.1038	0.1056	0.1075	0.1093	0.1112	0.1131
-1.2	0.1151	0.1170	0.1190	0.1210	0.1230	0.1251	0.1271	0.1292	0.1314	0.1335
-1.1	0.1357	0.1379	0.1401	0.1423	0.1446	0.1469	0.1492	0.1515	0.1539	0.1562
-1.0	0.1587	0.1611	0.1635	0.1660	0.1685	0.1711	0.1736	0.1762	0.1788	0.1814
-0.9	0.1841	0.1867	0.1894	0.1922	0.1949	0.1977	0.2005	0.2033	0.2061	0.2090
-0.8	0.2119	0.2148	0.2177	0.2206	0.2236	0.2266	0.2296	0.2327	0.2358	0.2389
-0.7	0.2420	0.2451	0.2483	0.2514	0.2546	0.2578	0.2611	0.2643	0.2676	0.2709
-0.6	0.2743	0.2776	0.2810	0.2843	0.2877	0.2912	0.2946	0.2981	0.3015	0.3050
-0.5	0.3085	0.3121	0.3156	0.3192	0.3228	0.3264	0.3300	0.3336	0.3372	0.3409
-0.4	0.3446	0.3483	0.3520	0.3557	0.3594	0.3632	0.3669	0.3707	0.3745	0.3783

Z & Decimals	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
-0.3	0.3821	0.3859	0.3897	0.3936	0.3974	0.4013	0.4052	0.4090	0.4129	0.4168
-0.2	0.4207	0.4247	0.4286	0.4325	0.4364	0.4404	0.4443	0.4483	0.4522	0.4562
-0.1	0.4602	0.4641	0.4681	0.4721	0.4761	0.4801	0.4840	0.4880	0.4920	0.4960

This page titled 10.5: z-Table is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Mathematics Department.

10.6: t-Table

t-Distribution Critical Values

This table shows critical values of the t-distribution for different degrees of freedom (df) at selected significance levels.

Use: One-tailed values for directional tests, Two-tailed for estimating confidence intervals or two-sided tests.

One-tailed t-distribution critical values

df	0.20	0.10	0.05	0.025	0.01	0.005
1	1.376	3.078	6.314	12.706	31.821	63.657
2	1.061	1.886	2.920	4.303	6.965	9.925
3	0.978	1.638	2.353	3.182	4.541	5.841
4	0.941	1.533	2.132	2.776	3.747	4.604
5	0.920	1.476	2.015	2.571	3.365	4.032
6	0.906	1.440	1.943	2.447	3.143	3.707
7	0.896	1.415	1.895	2.365	2.998	3.499
8	0.889	1.397	1.860	2.306	2.896	3.355
9	0.883	1.383	1.833	2.262	2.821	3.250
10	0.879	1.372	1.812	2.228	2.764	3.169
11	0.876	1.363	1.796	2.201	2.718	3.106
12	0.873	1.356	1.782	2.179	2.681	3.055
13	0.870	1.350	1.771	2.160	2.650	3.012
14	0.868	1.345	1.761	2.145	2.624	2.977
15	0.866	1.341	1.753	2.131	2.602	2.947
16	0.865	1.337	1.746	2.120	2.583	2.921
17	0.863	1.333	1.740	2.110	2.567	2.898
18	0.862	1.330	1.734	2.101	2.552	2.878
19	0.861	1.328	1.729	2.093	2.539	2.861
20	0.860	1.325	1.725	2.086	2.528	2.845
21	0.859	1.323	1.721	2.080	2.518	2.831
22	0.858	1.321	1.717	2.074	2.508	2.819
23	0.858	1.319	1.714	2.069	2.500	2.807
24	0.857	1.318	1.711	2.064	2.492	2.797
25	0.856	1.316	1.708	2.060	2.485	2.787
26	0.856	1.315	1.706	2.056	2.479	2.779
27	0.855	1.314	1.703	2.052	2.473	2.771
28	0.855	1.313	1.701	2.048	2.467	2.763
29	0.854	1.311	1.699	2.045	2.462	2.756
30	0.854	1.310	1.697	2.042	2.457	2.750

df	0.40	0.20	0.10	0.05	0.02	0.01
1	0.325	1.376	3.078	6.314	31.821	63.657
2	0.289	1.061	1.886	2.920	6.965	9.925
3	0.277	0.978	1.638	2.353	4.541	5.841
4	0.271	0.941	1.533	2.132	3.747	4.604
5	0.267	0.920	1.476	2.015	3.365	4.032
6	0.265	0.906	1.440	1.943	3.143	3.707
7	0.263	0.896	1.415	1.895	2.998	3.499
8	0.262	0.889	1.397	1.860	2.896	3.355
9	0.261	0.883	1.383	1.833	2.821	3.250
10	0.260	0.879	1.372	1.812	2.764	3.169
11	0.260	0.876	1.363	1.796	2.718	3.106
12	0.260	0.873	1.356	1.782	2.681	3.055
13	0.259	0.870	1.350	1.771	2.650	3.012
14	0.259	0.868	1.345	1.761	2.624	2.977
15	0.259	0.866	1.341	1.753	2.602	2.947
16	0.259	0.865	1.337	1.746	2.583	2.921
17	0.259	0.863	1.333	1.740	2.567	2.898
18	0.259	0.862	1.330	1.734	2.552	2.878
19	0.258	0.861	1.328	1.729	2.539	2.861
20	0.258	0.860	1.325	1.725	2.528	2.845
21	0.258	0.859	1.323	1.721	2.518	2.831
22	0.258	0.858	1.321	1.717	2.508	2.819
23	0.258	0.858	1.319	1.714	2.500	2.807
24	0.258	0.857	1.318	1.711	2.492	2.797
25	0.258	0.856	1.316	1.708	2.485	2.787
26	0.258	0.856	1.315	1.706	2.479	2.779
27	0.258	0.855	1.314	1.703	2.473	2.771
28	0.258	0.855	1.313	1.701	2.467	2.763
29	0.258	0.854	1.311	1.699	2.462	2.756
30	0.258	0.854	1.310	1.697	2.457	2.750

This page titled 10.6: t-Table is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Mathematics Department.

CHAPTER OVERVIEW

11: Applets and Other Stuff



This book is not completely finished, some pages are missing but we are workin' on it. . . (Public Domain ; Public Domain Pictures)

This page titled 11: Applets and Other Stuff is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Mathematics Department.

Index

Glossary

Chapter 1 Vocabulary

Data
: Information collected about individuals, objects, events, or conditions. Each data point represents an observation.

Observation
: A single row or record in a dataset; one instance of data collected from an individual or object.

Dataset
: A structured collection of data, typically organized in rows (observations) and columns (variables).

Variable
: A characteristic that is recorded about each observation and can vary from one observation to another.

Statistics
: The science of collecting, analyzing, interpreting, and communicating data in the presence of variability.

Variability
: The tendency for data values to differ from one another; natural differences observed in real-world data.

Categorical Variable
: A variable that places observations into non-numeric groups or categories.

Numerical Variable
: A variable that represents quantifiable amounts or measurements; values are numeric and meaningful.

Nominal Variable
: A categorical variable with no natural order (e.g., color, gender, blood type).

Ordinal Variable
: A categorical variable with a meaningful order but inconsistent spacing between values (e.g., satisfaction level, fitness difficulty).

Discrete Variable
: A numerical variable based on counts; values are whole numbers (e.g., number of pets, number of bedrooms).

Continuous Variable
: A numerical variable based on measurements; values can include decimals and fall along an interval (e.g., height, income, temperature).

Population
: The entire group of individuals or observations that you're interested in studying or learning about.

Sample
: A subset of the population that is actually studied or measured. Used to make inferences about the population.

Sampling
: The process of selecting individuals or units from a population to create a sample.

Representative Sample
: A sample that accurately reflects the characteristics of the population.

Simple Random Sample (SRS)
: A sample in which every member of the population has an equal chance of being selected.

Stratified Sample
: A sample created by dividing the population into subgroups (strata) and randomly sampling from each subgroup.

Cluster Sample
: A sample created by dividing the population into groups (clusters), then randomly selecting entire clusters.

Systematic Sample
: A sample chosen by selecting every nth individual from a list or process.

Convenience Sample
: A non-random sample chosen based on ease of access, often leading to bias.

Bias
: Systematic error in the way a sample is collected or data is measured, leading to results that do not accurately reflect the population.

Sampling Bias
: A specific type of bias introduced when some members of the population are less likely to be included in the sample than others.

Convenience Bias
: Bias introduced when individuals are selected simply because they are easy to reach.

Nonresponse Bias
: Bias that occurs when individuals selected for a survey do not respond, and their views differ from those who did respond.

Undercoverage Bias
: Bias caused by systematically missing part of the population during sampling.

Voluntary Response Bias
: Bias that occurs when participants choose to respond — often those with strong opinions are overrepresented.

Random Sampling
: A method of sampling where each member of the population has a known and typically equal chance of being selected.

Statistical Question
: A question that can be answered using data and involves variability in the answers.

Observational Study
: A study in which researchers observe and measure variables without affecting or influencing subjects.

Survey
: A method of collecting information from individuals by asking structured questions.

Experiment
: A study in which researchers deliberately apply a treatment and observe the outcomes, often using random assignment.

Treatment Group
: The group that receives the experimental condition or change in an experiment.

Control Group
: The group in an experiment that does not receive the treatment; used as a baseline comparison.

Confounding Variable
: An outside factor that affects both the explanatory and response variable, potentially obscuring results.

Random Assignment
: Assigning individuals to groups in an experiment by chance to ensure fairness and reduce bias.

Blinding
: A method used in experiments to prevent subjects or researchers from knowing which treatment was given, to reduce bias.

Statistical Process
: A step-by-step approach to answering questions with data: ask a question, collect data, organize/summarize, analyze, interpret, communicate.

Data Collection Plan
: A plan that outlines what data will be gathered, how it will be obtained, and what variables will be included.

Variable of Interest
: A specific characteristic or measurement that is central to a particular study or research question.

Semester Project
: A long-term project where students apply the full statistical process to a real-world topic — in this case, housing affordability.

Chapter 2 Vocabulary

Measure of Center
: A single number that represents the “typical” or central value in a dataset. Includes the mean, median, and mode.

Mean (Arithmetic Average)
: The sum of all values divided by the number of values. Sensitive to outliers. Symbol: $\bar{x}$

Median
: The middle value when the data are sorted. If there’s an even number of values, it’s the average of the two middle ones. Resistant to outliers.

Mode

The value (or values) that occur most often in a dataset. There can be no mode, one mode, or multiple modes.

Measure of Spread

A statistic that describes how much the values vary around the center. Includes range, IQR, and standard deviation.

Range

The difference between the maximum and minimum values in a dataset. Sensitive to outliers.

Interquartile Range (IQR)

The range of the middle 50% of values. Calculated as $Q3 - Q1$; resistant to extreme values.

Standard Deviation

The average distance from each data point to the mean. Uses all values in the dataset and is sensitive to outliers.

Variance

The squared average distance from the mean. Used as an intermediate step when calculating standard deviation.

Percentile

A value that separates a certain percent of the data. For example, the 80th percentile is the value that 80% of the dataset lies below.

Quartiles

Special percentiles that divide the data into quarters:

- **Q1:** 25th percentile
- **Q2:** 50th percentile / median
- **Q3:** 75th percentile

Five-Number Summary

A descriptive summary of a dataset using five values: minimum, Q1, median, Q3, and maximum.

Outlier

A data point that is unusually far from the others. Often flagged if it lies beyond 1.5 × IQR below Q1 or above Q3.

Bessel's Correction

The practice of dividing by $n - 1$ instead of n when calculating the sample variance or standard deviation. Helps make estimates less biased when working with sample data.

Box-and-Whisker Plot (Boxplot)

A standardized graph that displays the five-number summary. Visually shows a dataset's spread, skew, and potential outliers.

Whiskers

Lines on a boxplot that extend from Q1 to the minimum and from Q3 to the maximum — unless outliers are present, in which case they stop at the nearest non-outlier.

Skew (in a boxplot)

Asymmetry in the distribution. A longer whisker on one side suggests skew in that direction.

Chapter 3 Vocabulary

Bar Chart

A graph that uses rectangular bars to represent the frequencies or proportions of different categories. The height (or length) of each bar corresponds to its value. Often used with categorical data. Bars do not touch.

Pie Chart

A circular chart divided into slices to represent parts of a whole. Each slice shows the proportion of a category relative to the total dataset. Best used with a small number of categories.

Histogram

A graph that displays the distribution of quantitative data by grouping values into equal-width intervals (bins). The height of each bar represents the frequency within each interval. Bars touch to show continuity.

Bin

An interval used when grouping numerical data for a histogram. Each bin includes a fixed range of values (e.g., 60–70), and all bins should be of equal width unless otherwise noted.

Dot Plot

A simple graph where each data point is shown as a dot placed above a number line. Repeated values are stacked vertically. Ideal for small datasets.

Stem-and-Leaf Plot

A tabular display where quantitative values are split into stems (typically the leading digits) and leaves (the final digit). This chart preserves the original dataset while showing distribution shape.

Frequency

The number of times a specific value or interval appears in a dataset. Frequencies are often shown in tables or bar charts to summarize distributions.

Relative Frequency

The proportion or percentage of total values that fall in a particular category or bin. Calculated as: frequency ÷ total sample size.

Distribution

The way data values are spread or clustered across a range. A distribution can be described by its center, spread, shape, and presence of outliers.

Symmetric Distribution

A distribution where the left and right sides are mirror images of each other. Often has a single peak near the center.

Skewed Distribution

Asymmetry in a dataset's shape. If the right tail (larger values) is longer, it's right-skewed; if the left tail (smaller values) is longer, it's left-skewed.

Uniform Distribution

A distribution where all values (or ranges) are approximately equally frequent. Appears flat with no clear peak.

Bimodal Distribution

A distribution with two clear peaks (modes) in different parts of the data. May suggest the combination of two separate groups.

Time Plot

A graph that shows how a numeric variable changes over time. Time is placed on the x-axis, and values are plotted on the y-axis, often with points connected by a line.

Scatter Plot

A graph that uses points to show the relationship between two numerical variables. Used in time plots (with time on the x-axis) or to display general associations between variables. Lines may or may not be included.

Axis (Plural: Axes)

The reference lines in a graph. The x-axis runs horizontally (typically showing categories or time), and the y-axis runs vertically (often showing numerical values or frequency).

Chapter 4 Vocabulary

Probability

The long-run relative frequency of an event. Describes the proportion of times an event would occur if a random experiment were repeated many times under the same conditions. Values range from 0 to 1.

Relative Frequency

The proportion of times an event occurs out of the total number of trials in a simulation or experiment. Approximates probability in practice.

Experiment

An activity or process with an observable result. In probability, experiments are repeatable under identical conditions but have uncertain outcomes.

Sample Space

The set of all possible outcomes of a random experiment. Usually denoted by the symbol S or Ω.

Event

A set of outcomes (subset of the sample space) that satisfies a particular condition. Events can consist of one or more outcomes.

Complement

The set of all outcomes in the sample space that are *not* part of a given event. Denoted A^c. Represented as "not A".

Mutually Exclusive Events

Two events that cannot occur at the same time. The intersection of mutually exclusive events is the empty set; $P(A \cap B) = 0$.

Addition Rule

A formula to find the probability that at least one of two events occurs:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \quad (1)$$

When A and B are mutually exclusive:

$$P(A \cup B) = P(A) + P(B) \quad (2)$$

Conditional Probability

The probability that an event A occurs, given that another event B has occurred. Notated as $P(A|\ B)$ and calculated as:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} \quad \text{provided } P(B) > 0 \quad (3)$$

Independent Events

Two events A and B are independent if knowing that one occurred does not affect the probability of the other. Mathematically:

$$P(A \cap B) = P(A) \cdot P(B) \quad \text{or equivalently} \quad P(A \mid B) = P(A) \quad (4)$$

Dependent Events

Events for which the outcome or occurrence of one affects the probability of the other. For dependent events, $P(A|\ B) \neq P(A)$.

Counting Rule (Multiplication Rule of Counting)

If one task can be done in m ways and another in n ways, then the total number of ways to complete both tasks is $m \times n$.

Factorial

The product of all positive integers less than or equal to a given number n. Denoted $n!$. For example, $4! = 4 \times 3 \times 2 \times 1 = 24$.

Permutation

An arrangement of items where order matters. The number of ways to arrange r items from n is:

$$P(n, r) = \frac{n!}{(n-r)!} \quad (5)$$

Combination

A selection of items where order does *not* matter. The number of combinations of r items chosen from n is:

$$C(n, r) = \binom{n}{r} = \frac{n!}{r!(n-r)!} \quad (6)$$

Chapter 5 Vocabulary

Random Variable

A variable that takes numerical values determined by the outcome of a random process. Each value is associated with a probability.

Discrete Random Variable

A random variable with a finite or countable set of possible outcomes. Example: number of heads in 3 coin flips.

Continuous Random Variable

A random variable with an infinite number of possible values within a given interval. Example: time to complete a quiz.

Probability Distribution

A table, graph, or formula that assigns probabilities to each possible value of a random variable.

Bernoulli Trial

An experiment with only two possible outcomes: success or failure.

Bernoulli Distribution

A discrete probability distribution for a single Bernoulli trial. Only two values possible: 0 and 1.

Binomial Distribution

A probability distribution that models the number of successes in n independent Bernoulli trials with probability p.

Binomial Coefficient

The number of ways to choose x successes out of n trials: $\binom{n}{x}$.

Probability Mass Function (PMF)

A function that gives the probability that a discrete random variable is exactly equal to some value.

Expected Value

The theoretical mean of a random variable, computed as $E(X) = \sum x \cdot P(x)$.

Variance

The average of the squared differences from the mean: $\sigma^2 = \sum (x - \mu)^2 P(x)$.

Standard Deviation

The square root of the variance. It measures the typical distance between the data values and the mean.

Net Gain

The outcome of a situation after subtracting the initial cost or investment. Often used in lottery and game models.

Normal Distribution

A continuous, symmetric, bell-shaped distribution commonly found in natural and social processes. Defined by its mean and standard deviation.

Density Curve

A curve that represents the probability distribution of a continuous random variable. The area under the curve equals 1.

Symmetric Distribution

A distribution where the left and right halves are mirror images of each other.

Standard Normal Distribution

A normal distribution with mean $\mu = 0$ and standard deviation $\sigma = 1$.

Z-score

A standardized value that indicates how many standard deviations a data point is from the mean: $z = \frac{x - \mu}{\sigma}$.

Empirical Rule (68–95–99.7 Rule)

In a normal distribution: ~68% of data fall within 1 standard deviation, ~95% within 2, and ~99.7% within 3.

Percentile

The percentage of data values below a given value. Example: 90th percentile means 90% scored below that point.

Z-table (Standard Normal Table)

A table showing the cumulative probability up to a given z-score in the standard normal distribution.

Area under the curve

In a normal distribution, the area under the curve corresponds to probability. The total area is 1.

Left Tail

The area under the normal curve to the left of a z-score. Represents $P(Z < z)$.

Right Tail

The area to the right of a z-score: $P(Z > z) = 1 - P(Z < z)$.

Standardization

The process of converting raw data into z-scores so values can be compared across distributions with different scales.

Between Probability

The probability that a random variable falls between two values. Calculated as the difference between cumulative areas.

Chapter 6 Vocabulary

Population

The entire group we are interested in studying or making conclusions about.

Sample

A subset of individuals taken from the population, used to estimate population characteristics.

Statistic

A number calculated from a sample (such as a sample mean or proportion) used to estimate a population parameter.

Sampling Distribution

The distribution of a sample statistic (such as a mean or proportion) based on repeated random samples of the same size. It shows how the statistic would vary across repeated samples.

Sample Mean ($\bar{x}$)

The average value of all data points in a sample. Used to estimate the population mean.

Sample Proportion ($\hat{p}$)

The proportion of observations in a sample that meet a certain condition. Used to estimate the population proportion.

Standard Error (SE)

The standard deviation of a sampling distribution. It measures how much a sample statistic typically varies from sample to sample.

Standard Error of the Mean

The standard deviation of the sampling distribution of the sample mean. Given by $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$.

Standard Error of the Proportion

The standard deviation of the sampling distribution of a sample proportion. Given by $\sqrt{\frac{p(1-p)}{n}}$.

Central Limit Theorem (CLT)

A theorem stating that the sampling distribution of the sample mean (or proportion) becomes approximately normal as the sample size increases, regardless of population distribution shape.

Normal Approximation to the Binomial

A method of using the normal distribution to approximate binomial probabilities when sample size is large and both np and $n(1-p)$ are at least 10.

Continuity Correction

An adjustment used when approximating a discrete binomial distribution with a continuous normal distribution. We add or subtract 0.5 to better align with the binomial probability being approximated.

Chapter 7 Vocabulary

Point Estimate

A single value calculated from a sample that serves as the best guess for a population parameter. Example: the sample mean $\bar{x}$ is a point estimate of the population mean μ.

Interval Estimate

A range of values around a sample statistic that likely contains the population parameter. The most common type is a confidence interval.

Confidence Interval

An interval estimate built from a sample statistic plus and minus a margin of error. A 95% confidence interval means the method succeeds in capturing the population parameter about 95% of the time.

Confidence Level

The statistical confidence that the interval contains the population parameter. Common levels include 90%, 95%, and 99%.

Sample Mean ($\bar{x}$)

The average value of a sample; used as a point estimate for the population mean.

Sample Proportion ($\hat{p}$)

The proportion of individuals in a sample with a specific characteristic; used as a point estimate for the population proportion.

Standard Error (SE)

The estimated standard deviation of a sampling distribution. It measures how much the sample statistic varies from sample to sample.

Critical Value

The number of standard errors to extend from the sample statistic when building a confidence interval. Based on the desired confidence level (e.g., $z^* = 1.96$ for 95%).

Margin of Error

The amount added and subtracted to the point estimate to create the confidence interval. It reflects uncertainty due to sampling variability.

Interpretation of a Confidence Interval

A proper interpretation indicates the range within which the population parameter lies and is linked to a confidence level — not the probability that the true value is in that specific interval.

Chapter 8 Vocabulary

Hypothesis

A testable statement about a population parameter. Hypotheses are used to structure statistical tests and are evaluated based on sample data.

Null Hypothesis (H_0)

The default or starting assumption in a hypothesis test — usually a statement of "no effect," "no change," or "no difference." We test this claim using data.

Alternative Hypothesis (H_A)

The rival claim to the null. It represents the outcome we're trying to find evidence for (e.g., an increase, decrease, or difference).

Test Statistic

A calculated value (e.g., z, t) that measures how far the sample statistic is from the null hypothesis value. It's used to determine how extreme the result is.

z-test for a mean

A hypothesis test used when the population standard deviation is known. Assumes normal distribution. The test statistic is:

$Z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$

Welch's t-test

A two-sample test for comparing means when variances are unequal and sample sizes may differ. It uses sample standard deviations and calculates degrees of freedom based on variability.

z-test for a proportion

Used to test a hypothesis about a population proportion. The test statistic is:

$Z = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}}$

p-value

The probability of observing a result as extreme or more extreme than your observed sample statistic, assuming the null hypothesis is true. A small p-value suggests the data are inconsistent with H_0.

Significance Level (α)

The threshold we set for how much risk of a Type I error we are willing to accept. Common values are 0.05, 0.01, and 0.10.

Statistical Significance

If the p-value is less than or equal to α, we say the result is statistically significant — meaning it is unlikely to have occurred by chance under the null.

Practical Significance

A result that is meaningful or impactful in the real world. Even small statistical differences may not matter practically, depending on the context.

Type I Error

Rejecting the null hypothesis when it is actually true. This is a false positive. The probability of this is controlled by α.

Type II Error

Failing to reject the null hypothesis when the alternative is actually true. This is a false negative.

Power of a Test

The probability that a test will correctly reject a false null hypothesis. It is equal to 1 minus the probability of a Type II Error.

One-tailed Test

A hypothesis test in which the alternative hypothesis is directional (e.g., $p > p_0$ or $\mu < \mu_0$) — we are only interested in one direction of difference.

Two-tailed Test

A hypothesis test in which the alternative hypothesis tests for any difference from the null (e.g., $\mu \neq \mu_0$). Both tails of the distribution are considered.

Decision Rule

The rule that guides whether to reject or fail to reject the null hypothesis, usually based on comparing the p-value to α.

p-hacking

A problematic practice in which researchers run many statistical analyses until they obtain a significant ($p < 0.05$) result, often misrepresenting the evidence.

Effect Size

A measure of how big a difference or change is — separate from statistical significance. A small p-value may accompany a small effect size, or vice versa.

Standard Error

The estimated standard deviation of a sampling distribution. Used in many test statistics, including for means and proportions.

Chapter 9 Vocabulary

Bivariate Data

Data that includes two quantitative variables measured for each individual in a study. Often used to examine relationships between variables.

Scatterplot

A graph that displays pairs of numerical data as points on a coordinate plane. Each point represents an individual's value for two variables (x and y).

Direction of Association

A description of how one variable tends to change as the other changes. Associations can be positive, negative, or have no clear direction.

Positive Association

When values of one variable tend to increase as the values of the other variable also increase. The pattern typically slopes upward from left to right.

Negative Association

When values of one variable increase as the other decreases. The pattern typically slopes downward from left to right.

Correlation (r)

A numerical measure between –1 and 1 that describes the strength and direction of a linear relationship between two variables. A value close to ±1 indicates a strong linear relationship.

Correlation Coefficient (r)

Another name for the correlation measure. It quantifies linear association but does not imply causation.

Least-Squares Regression Line (LSRL)

The line that best fits a scatterplot and minimizes the sum of the squared residuals. It models the linear relationship and can be used for prediction.

Regression Equation

The formula of the form $\hat{y} = a + bx$ where $\hat{y}$ is the predicted value, a is the intercept, and b is the slope.

Slope

In a regression line, the slope tells how much the predicted value of y changes for each one-unit increase in x. It represents the rate of change.

Intercept

The y-value predicted when $x = 0$ in a regression model. It may or may not be meaningful depending on the context and data range.

Residual

The difference between an actual data value and the value predicted by a regression model. Residual = actual – predicted.

Interpolation

Using a regression model to predict a value within the range of the observed data.

Extrapolation

Using a regression model to predict a value outside the range of the observed data. It can be risky because the relationship may not hold beyond the data range.

Causation

The idea that one variable has a direct effect on another. Causation implies a change in one variable produces a change in the other. It requires evidence beyond just statistical association.

Correlation Does Not Imply Causation

A reminder that even strong relationships between variables do not prove one causes the other. Alternate explanations or lurking variables may be involved.

Detailed Licensing

Many of the pages in this book are re-mixes of open content from other books in the LibreTexts Library. Sometimes sections from several sources are used on the same page and there may be no unique open license. Where there is a conflict, the licensing may appear as *undeclared*. On any page, by toggling AutoAttibution on the Tools menu, you can display the detailed attribution on a paragraph level. Material from different books will be color coded and the source displayed by mousing over the paragraph.

Content in the front and backmatter, directories and similar can be uncopywriteable, being simple listings and these are also shown as *undeclared*.

Overview

Title: Introduction to Statistics (RRCC)

Webpages: 114

All licenses found:

- CC BY 4.0: 70.2% (80 pages)
- Undeclared: 29.8% (34 pages)

By Page

- Introduction to Statistics (RRCC) - *CC BY 4.0*
  - Front Matter - *Undeclared*
    - TitlePage - *Undeclared*
    - InfoPage - *Undeclared*
    - Table of Contents - *Undeclared*
    - Licensing - *Undeclared*
  - 1: Introduction to Data - *CC BY 4.0*
    - 1.1: What is Statistics? - *CC BY 4.0*
    - 1.2: Data Types- Categorical vs. Numerical - *CC BY 4.0*
    - 1.3: Populations, Samples, and Sampling Methods - *CC BY 4.0*
    - 1.4: Sources of Bias and Randomness - *CC BY 4.0*
    - 1.5: Exploring Statistical Questions - *CC BY 4.0*
    - 1.6: Designing a Statistical Study - *CC BY 4.0*
      - 1.6.1: Project Introduction - *CC BY 4.0*
    - 1.7: Vocabulary (Chapter 1) - *CC BY 4.0*
    - 1.8: Practice (Chapter 1) - *CC BY 4.0*
  - 2: Exploring and Summarizing Data - *CC BY 4.0*
    - 2.1: Measures of Central Tendency- Mean, Median, and Mode - *CC BY 4.0*
    - 2.2: Measures of Spread- Range, Variance, and Standard Deviation - *CC BY 4.0*
      - 2.2.1: Project Check-In 1 - *CC BY 4.0*
    - 2.3: The Five-Number Summary and Interquartile Range - *CC BY 4.0*
      - 2.3.1: Project Check-In 2 - *CC BY 4.0*
      - 2.3.2: Percentage Review - *CC BY 4.0*
    - 2.4: Identifying Outliers with Boxplots - *CC BY 4.0*
    - 2.5: Vocabulary (Chapter 2) - *CC BY 4.0*
    - 2.6: Practice (Chapter 2) - *Undeclared*
  - 3: Visualizing Data - *CC BY 4.0*
    - 3.1: Graphical Displays for Categorical Data- Bar Charts, Pie Charts - *CC BY 4.0*
    - 3.2: Graphical Displays for Quantitative Data- Dot Plots and Stemplots - *CC BY 4.0*
    - 3.3: Frequency Tables and Relative Frequency - *CC BY 4.0*
    - 3.4: Shapes of Distributions - *CC BY 4.0*
    - 3.5: Time Series Plots and Trends Over Time - *CC BY 4.0*
      - 3.5.1: Project Check-In 3 - *CC BY 4.0*
    - 3.6: Vocabulary (Chapter 3) - *CC BY 4.0*
    - 3.7: Practice (Chapter 3) - *Undeclared*
  - 4: Introduction to Probability - *CC BY 4.0*
    - 4.1: What is Probability? - *CC BY 4.0*
    - 4.2: Introduction to Counting - *CC BY 4.0*
    - 4.3: Foundational Probability Rules - *CC BY 4.0*
    - 4.4: Compound Events, Independence and Conditional Probability - *CC BY 4.0*
    - 4.5: Vocabulary (Chapter 4) - *CC BY 4.0*
    - 4.6: Practice (Chapter 4) - *Undeclared*
  - 5: Probability Distributions - *CC BY 4.0*
    - 5.1: Random Variables- Discrete and Continuous - *CC BY 4.0*
    - 5.2: Expected Value of Discrete Distributions - *CC BY 4.0*

- 5.3: Discrete Distributions- Bernoulli and Binomial - *CC BY 4.0*
- 5.4: The Normal Distribution - *CC BY 4.0*
- 5.5: The Empirical Rule and Standard Normal (Z) Distribution - *CC BY 4.0*
- 5.6: Application of the Normal Distribution - *CC BY 4.0*
- 5.7: Vocabulary (Chapter 5) - *CC BY 4.0*
- 5.8: Practice (Chapter 5) - *Undeclared*

6: Sampling Distributions - *CC BY 4.0*

- 6.1: Sampling Distributions - *CC BY 4.0*
- 6.2: Mean and Standard Deviation of Sampling Distributions - *CC BY 4.0*
- 6.3: Central Limit Theorem- Meaning and Implications - *CC BY 4.0*
- 6.4: CLT for Means and Proportions - *CC BY 4.0*
- 6.5: Approximating Binomial with the Normal Distribution - *CC BY 4.0*
- 6.6: Vocabulary (Chapter 6) - *CC BY 4.0*
- 6.7: Practice (Chapter 6) - *Undeclared*

7: Confidence Intervals - *CC BY 4.0*

- 7.1: Point Estimates vs. Interval Estimates - *CC BY 4.0*
- 7.2: Confidence Intervals for a Population Proportion - *CC BY 4.0*
- 7.3: Confidence Intervals for a Population Mean - *CC BY 4.0*
- 7.4: Effect of Sample Size and Confidence Level - *CC BY 4.0*
- 7.5: Minimum Sample Size - *Undeclared*
- 7.6: Interpreting Confidence Intervals - *CC BY 4.0*
  - 7.6.1: Project Check-In 4 - *CC BY 4.0*
- 7.7: Vocabulary (Chapter 7) - *CC BY 4.0*
- 7.8: Practice (Chapter 7) - *Undeclared*

8: Hypothesis Testing - *CC BY 4.0*

- 8.1: Why hypothesis testing and Type I and Type II Errors - *CC BY 4.0*
- 8.2: Hypothesis Testing Framework - *CC BY 4.0*
- 8.3: Tests for a Single Mean (z and t-tests) - *CC BY 4.0*
- 8.4: Comparing Two Means (Independent Samples) - *CC BY 4.0*
  - 8.4.1: Project Checkpoint - *CC BY 4.0*
- 8.5: Tests for a Single Proportion - *CC BY 4.0*
- 8.6: Interpreting p-values and Significance - *CC BY 4.0*
- 8.7: Vocabulary (Chapter 8) - *CC BY 4.0*
- 8.8: Practice (Chapter 8) - *Undeclared*

9: Bivariate Data, Correlation, and Regression - *CC BY 4.0*

- 9.1: Scatterplots and Direction of Association - *CC BY 4.0*
  - 9.1.1: Datasaurus Dozen - *CC BY 4.0*
- 9.2: Correlation Coefficient (r) - *CC BY 4.0*
- 9.3: Least-Squares Regression Line - *CC BY 4.0*
- 9.4: Interpreting Slope, Intercept, and Residuals - *CC BY 4.0*
- 9.5: Cautions with Correlation and Causation - *CC BY 4.0*
- 9.6: Vocabulary (Chapter 9) - *CC BY 4.0*
- 9.7: Practice (Chapter 9) - *Undeclared*

10: Project, Labs, Appendix and Tables - *CC BY 4.0*

- 10.1: Project Assignments - *Undeclared*
  - 10.1.1: Sampling and Data Web Page - *Undeclared*
  - 10.1.2: Quantitative Variable Analysis - *Undeclared*
  - 10.1.3: Qualitative Variable Analysis - *Undeclared*
  - 10.1.4: Interval Estimates - *Undeclared*
  - 10.1.5: Hypothesis Testing - *Undeclared*
  - 10.1.6: Correlation - *Undeclared*
- 10.2: Tech Labs - *Undeclared*
  - 10.2.1: TL 1 - *Undeclared*
  - 10.2.2: TL 2 - *Undeclared*
  - 10.2.3: TL 3 - *Undeclared*
  - 10.2.4: TL 4 - *Undeclared*
  - 10.2.5: TL 5 - *Undeclared*
  - 10.2.6: TL 6 - *Undeclared*
  - 10.2.7: TL 7 - *Undeclared*
- 10.3: Summary Statistic Cards - *CC BY 4.0*
- 10.4: Project Reflection - *CC BY 4.0*
- 10.5: z-Table - *CC BY 4.0*
- 10.6: t-Table - *CC BY 4.0*

11: Applets and Other Stuff - *CC BY 4.0*

Back Matter - *Undeclared*

- Index - *Undeclared*
- Glossary - *Undeclared*
- Detailed Licensing - *Undeclared*
- Detailed Licensing - *Undeclared*

Detailed Licensing

Overview

Title: Introduction to Statistics (RRCC)

Webpages: 114

All licenses found:

- CC BY 4.0: 70.2% (80 pages)
- Undeclared: 29.8% (34 pages)

By Page

- Introduction to Statistics (RRCC) - *CC BY 4.0*
  - Front Matter - *Undeclared*
    - TitlePage - *Undeclared*
    - InfoPage - *Undeclared*
    - Table of Contents - *Undeclared*
    - Licensing - *Undeclared*
  - 1: Introduction to Data - *CC BY 4.0*
    - 1.1: What is Statistics? - *CC BY 4.0*
    - 1.2: Data Types- Categorical vs. Numerical - *CC BY 4.0*
    - 1.3: Populations, Samples, and Sampling Methods - *CC BY 4.0*
    - 1.4: Sources of Bias and Randomness - *CC BY 4.0*
    - 1.5: Exploring Statistical Questions - *CC BY 4.0*
    - 1.6: Designing a Statistical Study - *CC BY 4.0*
      - 1.6.1: Project Introduction - *CC BY 4.0*
    - 1.7: Vocabulary (Chapter 1) - *CC BY 4.0*
    - 1.8: Practice (Chapter 1) - *CC BY 4.0*
  - 2: Exploring and Summarizing Data - *CC BY 4.0*
    - 2.1: Measures of Central Tendency- Mean, Median, and Mode - *CC BY 4.0*
    - 2.2: Measures of Spread- Range, Variance, and Standard Deviation - *CC BY 4.0*
      - 2.2.1: Project Check-In 1 - *CC BY 4.0*
    - 2.3: The Five-Number Summary and Interquartile Range - *CC BY 4.0*
      - 2.3.1: Project Check-In 2 - *CC BY 4.0*
      - 2.3.2: Percentage Review - *CC BY 4.0*
    - 2.4: Identifying Outliers with Boxplots - *CC BY 4.0*
    - 2.5: Vocabulary (Chapter 2) - *CC BY 4.0*
    - 2.6: Practice (Chapter 2) - *Undeclared*
  - 3: Visualizing Data - *CC BY 4.0*
    - 3.1: Graphical Displays for Categorical Data- Bar Charts, Pie Charts - *CC BY 4.0*
    - 3.2: Graphical Displays for Quantitative Data- Dot Plots and Stemplots - *CC BY 4.0*
    - 3.3: Frequency Tables and Relative Frequency - *CC BY 4.0*
    - 3.4: Shapes of Distributions - *CC BY 4.0*
    - 3.5: Time Series Plots and Trends Over Time - *CC BY 4.0*
      - 3.5.1: Project Check-In 3 - *CC BY 4.0*
    - 3.6: Vocabulary (Chapter 3) - *CC BY 4.0*
    - 3.7: Practice (Chapter 3) - *Undeclared*
  - 4: Introduction to Probability - *CC BY 4.0*
    - 4.1: What is Probability? - *CC BY 4.0*
    - 4.2: Introduction to Counting - *CC BY 4.0*
    - 4.3: Foundational Probability Rules - *CC BY 4.0*
    - 4.4: Compound Events, Independence and Conditional Probability - *CC BY 4.0*
    - 4.5: Vocabulary (Chapter 4) - *CC BY 4.0*
    - 4.6: Practice (Chapter 4) - *Undeclared*
  - 5: Probability Distributions - *CC BY 4.0*
    - 5.1: Random Variables- Discrete and Continuous - *CC BY 4.0*
    - 5.2: Expected Value of Discrete Distributions - *CC BY 4.0*
    - 5.3: Discrete Distributions- Bernoulli and Binomial - *CC BY 4.0*
    - 5.4: The Normal Distribution - *CC BY 4.0*
    - 5.5: The Empirical Rule and Standard Normal (Z) Distribution - *CC BY 4.0*
    - 5.6: Application of the Normal Distribution - *CC BY 4.0*
    - 5.7: Vocabulary (Chapter 5) - *CC BY 4.0*
    - 5.8: Practice (Chapter 5) - *Undeclared*
  - 6: Sampling Distributions - *CC BY 4.0*
    - 6.1: Sampling Distributions - *CC BY 4.0*
    - 6.2: Mean and Standard Deviation of Sampling Distributions - *CC BY 4.0*

- 6.3: Central Limit Theorem- Meaning and Implications - *CC BY 4.0*
- 6.4: CLT for Means and Proportions - *CC BY 4.0*
- 6.5: Approximating Binomial with the Normal Distribution - *CC BY 4.0*
- 6.6: Vocabulary (Chapter 6) - *CC BY 4.0*
- 6.7: Practice (Chapter 6) - *Undeclared*

- 7: Confidence Intervals - *CC BY 4.0*
  - 7.1: Point Estimates vs. Interval Estimates - *CC BY 4.0*
  - 7.2: Confidence Intervals for a Population Proportion - *CC BY 4.0*
  - 7.3: Confidence Intervals for a Population Mean - *CC BY 4.0*
  - 7.4: Effect of Sample Size and Confidence Level - *CC BY 4.0*
  - 7.5: Minimum Sample Size - *Undeclared*
  - 7.6: Interpreting Confidence Intervals - *CC BY 4.0*
    - 7.6.1: Project Check-In 4 - *CC BY 4.0*
  - 7.7: Vocabulary (Chapter 7) - *CC BY 4.0*
  - 7.8: Practice (Chapter 7) - *Undeclared*
- 8: Hypothesis Testing - *CC BY 4.0*
  - 8.1: Why hypothesis testing and Type I and Type II Errors - *CC BY 4.0*
  - 8.2: Hypothesis Testing Framework - *CC BY 4.0*
  - 8.3: Tests for a Single Mean (z and t-tests) - *CC BY 4.0*
  - 8.4: Comparing Two Means (Independent Samples) - *CC BY 4.0*
    - 8.4.1: Project Checkpoint - *CC BY 4.0*
  - 8.5: Tests for a Single Proportion - *CC BY 4.0*
  - 8.6: Interpreting p-values and Significance - *CC BY 4.0*
  - 8.7: Vocabulary (Chapter 8) - *CC BY 4.0*
  - 8.8: Practice (Chapter 8) - *Undeclared*
- 9: Bivariate Data, Correlation, and Regression - *CC BY 4.0*
  - 9.1: Scatterplots and Direction of Association - *CC BY 4.0*
    - 9.1.1: Datasaurus Dozen - *CC BY 4.0*
  - 9.2: Correlation Coefficient (r) - *CC BY 4.0*
  - 9.3: Least-Squares Regression Line - *CC BY 4.0*
  - 9.4: Interpreting Slope, Intercept, and Residuals - *CC BY 4.0*
  - 9.5: Cautions with Correlation and Causation - *CC BY 4.0*
  - 9.6: Vocabulary (Chapter 9) - *CC BY 4.0*
  - 9.7: Practice (Chapter 9) - *Undeclared*
- 10: Project, Labs, Appendix and Tables - *CC BY 4.0*
  - 10.1: Project Assignments - *Undeclared*
    - 10.1.1: Sampling and Data Web Page - *Undeclared*
    - 10.1.2: Quantitative Variable Analysis - *Undeclared*
    - 10.1.3: Qualitative Variable Analysis - *Undeclared*
    - 10.1.4: Interval Estimates - *Undeclared*
    - 10.1.5: Hypothesis Testing - *Undeclared*
    - 10.1.6: Correlation - *Undeclared*
  - 10.2: Tech Labs - *Undeclared*
    - 10.2.1: TL 1 - *Undeclared*
    - 10.2.2: TL 2 - *Undeclared*
    - 10.2.3: TL 3 - *Undeclared*
    - 10.2.4: TL 4 - *Undeclared*
    - 10.2.5: TL 5 - *Undeclared*
    - 10.2.6: TL 6 - *Undeclared*
    - 10.2.7: TL 7 - *Undeclared*
  - 10.3: Summary Statistic Cards - *CC BY 4.0*
  - 10.4: Project Reflection - *CC BY 4.0*
  - 10.5: z-Table - *CC BY 4.0*
  - 10.6: t-Table - *CC BY 4.0*
- 11: Applets and Other Stuff - *CC BY 4.0*
- Back Matter - *Undeclared*
  - Index - *Undeclared*
  - Glossary - *Undeclared*
  - Detailed Licensing - *Undeclared*
  - Detailed Licensing - *Undeclared*