Measures of central tendency: The mean, the median, and the mode

Recently, you learned that **measures of central tendency** are values that represent the center of a dataset. When you're working with a new dataset, identifying the central location of your data helps you quickly understand its basic structure.

In this reading, you'll learn more about three measures of central tendency: the mean, the median, and the mode. We'll go over how to calculate each measure, and discuss which measure is best to use based on your specific data.

Measures of central tendency

The mean, median, and mode all describe the center of a dataset in different ways:

- The **mean** is the average value in a dataset.
- The **median** is the middle value in a dataset.
- The **mode** is the most frequently occurring value in a dataset.

Let's explore how to calculate each measure of central tendency.

Calculate the mean, the median, and the mode

Mean

The **mean** is the average value in a dataset. To calculate the mean, you add up all the values in your dataset and divide by the total number of values.

For example, say you have the following set of values: 10, 5, 3, 50, 12. To find the mean, you add all the values for a total of 80. Then, you divide by 5, the total number of values.

$$(10+5+3+50+12)\div 5=80\div 5=16(10+5+3+50+12)\div 5=80\div 5=16$$

The mean, or average value, is 16.

Median

The **median** is the middle value in a dataset. This means half the values in the dataset are larger than the median, and half the values are smaller than the median.

You can find the median by arranging all the values in a dataset from smallest to largest. If you arrange your five values in this way you get: 3, 5, 10, 12, 50. The median, or middle value, is 10.

If there are an even number of values in your dataset, the median is the average of the two middle values. Let's say you add another value, 8, to your set: 3, 5, 8, 10, 12, 50. Now, the two middle values are 8 and 10. To get the median, take their average.

$$(8+10)\div2=18\div2=9(8+10)\div2=18\div2=9$$

The median is 9.

Mode

The **mode** is the most frequently-occurring value in a dataset. A dataset can have no mode, one mode, or more than one mode.

For example, the set of numbers 1, 12, 33, 54, 75 has no mode because no value repeats. In the set 2, 7, 7, 11, 20 the mode is 7, because 7 is the only value that occurs more than once. The set 3, 12, 12, 40, 40 has two modes: 12 and 40.

When to use the mean, the median, and the mode

Whether you use the mean, median, or mode to describe the center of your dataset depends on the specific data you're working with and what insights you want to gain from your data. Let's discuss some general guidelines for using each measure of central tendency.

Mean versus median

Both the mean and the median describe the central location of a dataset. However, as measures of central tendency, the mean and the median work better for different kinds of data.

The mean has one main disadvantage: it is very sensitive to outliers in your dataset. Recall that an outlier is a value that differs greatly from the rest of the data.

If there are outliers in your dataset, the median is usually a better measure of the center. If there are no outliers, the mean usually works well.

For example, imagine you want to compute the average annual salary for an employee at a small startup company. You have the following salary data:

Employee	**#1**	**#2**	**#3**	**#4**	**#5**	**#6**	**#7**
Salary	$40,000	$45,000	$45,000	$45,000	$45,000	$50,000	$500,000

You can calculate the mean annual salary by adding up all the values in your dataset and dividing by the total number of values. There are seven salaries in total, and their sum is $770,000.

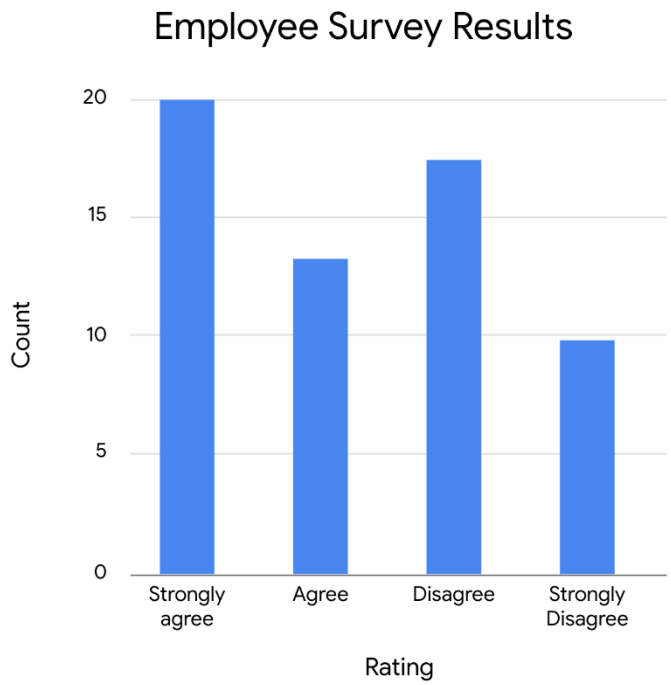
$$\$770,000\div7=\$110,000$$

The mean salary for these seven employees is $110,000. However, the data suggests that this mean value does not accurately reflect the typical salary of an employee at this company. Most employees have a salary between $40,000–$50,000. In fact, only one employee has a salary greater than $50,000. The salary of $500,000 is an outlier that pulls up the average, or skews the mean.

In this case, due to the presence of this outlier value, the median is a better measure of central tendency than the mean. The median, or middle value, in this dataset is $45,000. The median gives you a better idea of the typical salary for an employee at this company.

Mode

The mode is useful when working with categorical data because it clearly shows you which category occurs most frequently. Say a company conducts an employee satisfaction survey. The main item on the survey states, “I am satisfied with the opportunity I have to grow within the company.” Employees choose among four categories for their response: strongly agree, agree, disagree, strongly disagree. A bar chart summarizes the results.



The mode represents the highest bar in the bar chart, which refers to the rating “strongly agree.” This is the most frequently-occurring rating in the dataset. The mode gives the company clear feedback on employee satisfaction; in this case, positive feedback.

Key takeaways

Measures of central tendency such as the mean, median, and mode let you describe the center of your dataset using a single value. As a data professional, knowing the center of your dataset helps you quickly understand its basic structure and determine the next steps in your analysis.

Resources for more information

To learn more about measures of central tendency like the mean, median, and mode, explore the following resource:

- This article from the Australian Bureau of Statistics offers a useful overview of the mean, median, and mode, and discusses how outliers influence measures of central tendency.

Measures of dispersion: Range, variance, and standard deviation

Recently, you learned that **measures of dispersion** let you describe the spread of your dataset, or the amount of variation in your data values. Measures of dispersion like standard deviation can give you an initial understanding of the distribution of your data, and help you determine what statistical methods to apply to your data.

In this reading, you'll learn more about three measures of dispersion: the range, variance, and standard deviation. This reading focuses on the foundational concept of standard deviation. As a data professional, you'll frequently calculate the standard deviation of your data, and use standard deviation as part of more complex data analysis.

Measures of dispersion

Let's examine out the definition of each measure of dispersion: the range, variance, and standard deviation.

Range

The **range** is the difference between the largest and smallest value in a dataset.

For example, imagine you're a biology teacher and you have data on scores for the final exam. The highest score is 99/100, or 99%. The lowest score is 62/100, or 62%. To calculate the range, subtract the lowest score from the highest score.

99 - 62 = 37

The range is 37 percentage points.

The range is a useful metric because it's easy to calculate, and it gives you a very quick understanding of the overall spread of your dataset.

Variance

Another measure of spread is called the **variance,** which is the average of the squared difference of each data point from the mean. Basically, it's the square of the standard deviation. You'll learn more about variance and how to use it in a later course.

Standard deviation

To get a better understanding of the concept of standard deviation, let's explore its definition, visualization, and statistical formula.

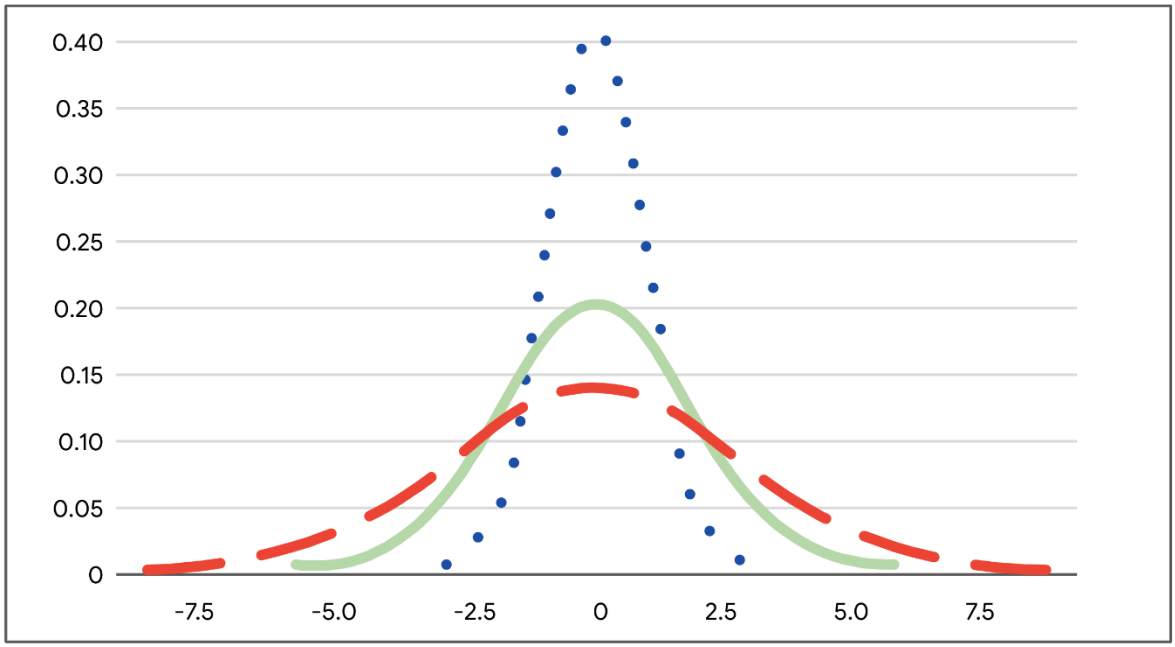
Definition

Standard deviation measures how spread out your values are from the mean of your dataset. It calculates the typical distance of a data point from the mean. The larger the standard deviation, the more spread out your

values are from the mean. The smaller the standard deviation, the less spread out your values are from the mean.

Visualization

Let's examine the plots of three normal probability distributions to get a better idea of spread. Later on, you'll learn about distributions, which map all the values in a dataset. For now, just know that the mean is the highest point on each curve, right in the center.



Each curve has the same mean and a different standard deviation. The standard deviation of the blue dotted curve is 1, the green solid curve is 2, and the red dashed curve is 3. The blue dotted curve has the least spread since most of its data values fall close to the mean. Therefore, the blue dotted curve has the smallest standard deviation. The red dashed curve has the most spread since most of its data values fall farther away from the mean. Therefore, the red dashed curve has the largest standard deviation.

Formula

Now let's discuss how you calculate standard deviation using a formula.
There are different formulas to calculate the standard deviation for a population and a sample. As a reminder, data professionals typically work with sample data, and they make inferences about populations based on the sample. So, let's review the formula for sample standard deviation:

$$s = \sqrt{\frac{\Sigma(x - \overline{x})^2}{n - 1}}$$

In the formula, n is the total number of data values in your sample, x is each individual data value, and $\bar{x}$ (pronounced "x-bar") is the mean of your data values. The Greek letter Sigma is a symbol that means sum.

Note: As a data professional, you'll typically use a computer for calculations. Being able to perform calculations is important for your future career, but being familiar with the concepts behind the calculations will help you apply statistical methods to workplace problems.

To better understand the different parts of the formula, let's calculate the sample standard deviation of a small dataset: 2, 3, 10.

You can do this in five steps:

1. Calculate the mean, or average, of your data values.

(2 + 3 +10) ÷ 3 = 15 ÷ 3 = 5

2. Subtract the mean from each value.

2 - 5 = -3

3 - 5 = -2

10 - 5 = 5

3. Square each result.

-3 * -3 = 9

-2 * -2 = 4

5 * 5 = 25

4. Add up the squared results and divide this sum by one less than the number of data values. This is the variance.

(9 + 4 + 25) ÷ (3 -1) = 38 ÷ 2 = 19

5. Finally, find the square root of the variance.

√19 = 4.36

The sample standard deviation is 4.36.

Now that you know more about the concept of standard deviation, let's check out an example of its practical application.

Example: Real estate prices

Imagine you're a data professional working for a real estate company. The real estate agents on your team would like to inform their clients about the variation in rental prices in different residential areas. Part of your job is calculating the standard deviation of monthly rental prices for apartments in specific neighborhoods and sharing this information with your team. Let's say you have sample data on monthly rental prices for one-

bedroom apartments in two different neighborhoods: Emerald Woods and Rock Park. Assume you calculate the mean and standard deviation for each dataset.

Emerald Woods

Apartment	**#1**	**#2**	**#3**	**#4**	**#5**
Monthly Rent	$900	$950	$1,000	$1,050	$1,100

Mean: $1,000

Standard deviation: $79.05

Rock Park

Apartment	**#1**	**#2**	**#3**	**#4**	**#5**
Monthly Rent	$500	$650	$1,000	$1,350	$1,500

Mean: $1,000

Standard deviation: $431.56

Both neighborhoods have the same mean rental price of $1,000 per month. However, the standard deviation for rental prices in Rock Park ($431.56) is much higher than the standard deviation for rental prices in Emerald Woods ($79.05). This means that there is a lot more variation in rental prices in Rock Park. This is useful information for your agents. For example, they can tell clients that it may be easier for them to find a more affordable apartment in Rock Park that is far below the mean of $1,000. Standard deviation helps you quickly understand the variation in prices in any given neighborhood.

Key takeaways

Data professionals use standard deviation to measure variation in many types of data like ad revenues, stock prices, employee salaries, and more. Measures of dispersion like the standard deviation, variance, and range let you quickly identify the variation in your data values and get a better understanding of the basic structure of your data.

Resources for more information

To learn more about measures of dispersion like the range, variance, and standard deviation, explore the following resources:

- This article from Statistics Canada provides a helpful summary of variance and standard deviation, and discusses the usefulness of standard deviation as a measure of dispersion.

Measures of position: Percentiles and quartiles

Recently, you learned that **measures of position** let you determine the position of a value in relation to other values in a dataset. Along with center and spread, it's helpful to know the relative position of your values. For example, whether one value is higher or lower than another, or whether a value falls in the lower, middle, or upper portion of your dataset.

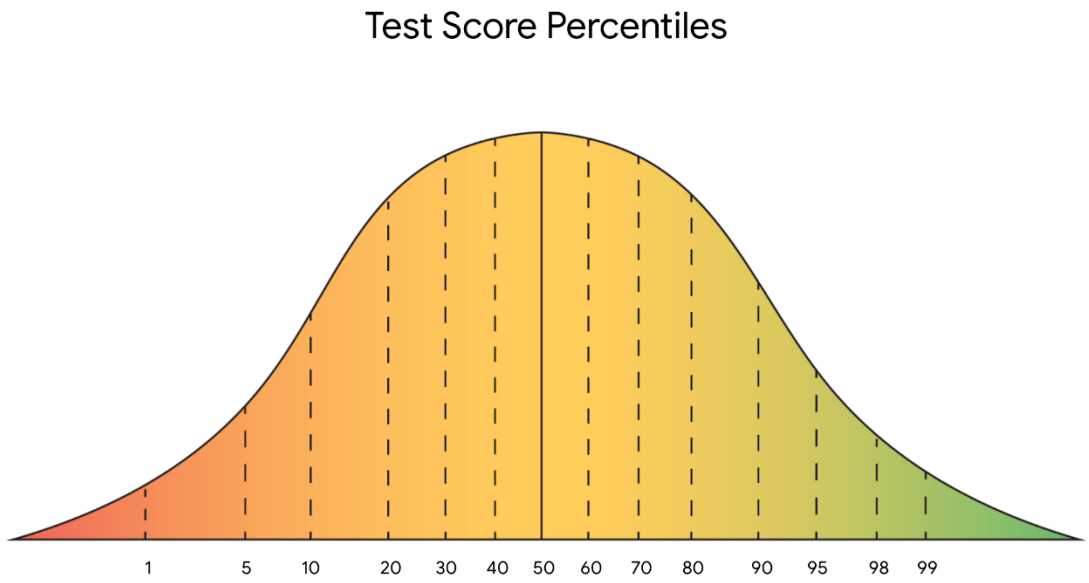
In this reading, you'll learn more about the most common measures of position: percentiles and quartiles. You'll also learn how to calculate the interquartile range, and use the five number summary to summarize your data.

Measures of position

Percentile

A **percentile** is the value below which a percentage of data falls. Percentiles divide your data into 100 equal parts. Percentiles give the relative position or rank of a particular value in a dataset.

For example, percentiles are commonly used to rank test scores on school exams. Let's say a test score falls in the 99th percentile. This means the score is higher than 99% of all test scores. If a score falls in the 75th percentile, the score is higher than 75% of all test scores. If a score falls in the 50th percentile, the score is higher than half, or 50%, of all test scores.



Note: *Percentiles and percentages are distinct concepts. For example, say you score 90/100, or 90%, on a test. This doesn't necessarily mean your score of 90% is in the 90th percentile. Percentile depends on the relative performance of all test takers. If half of all test takers score above 90%, then a score of 90% will be in the 50th percentile.*

Percentiles are useful for comparing values and putting data in context. For example, imagine you want to buy a new car. You'd like a midsize sedan with great fuel economy. In the United States fuel economy is measured

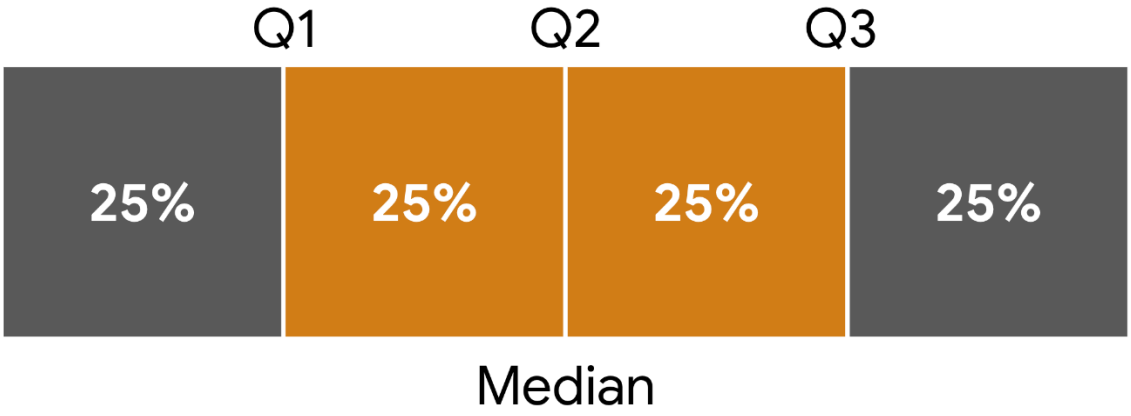
in miles per gallon of fuel, or mpg. The sedan you're considering gets 23 mpg. Is that good or bad? Without a basis for comparison, it's hard to know. However, if you know that 23 mpg is in the 25th percentile of all midsize sedans, you have a much clearer idea of its relative performance. In this case, 75% of all midsize sedans have a higher mpg than the car you're thinking about buying.

Quartile

You can use quartiles to get a general understanding of the relative position of values. A **quartile** divides the values in a dataset into four equal parts.

Three quartiles divide the data into four quarters. Quartiles let you compare values relative to the four quarters of data. Each quarter includes 25% of the values in your dataset.

- The first quartile, Q1, is the middle value in the first half of the dataset. Q1 refers to the 25th percentile. 25% of the values in the entire dataset are below Q1, and 75% are above it.
- The second quartile, Q2, is the median of the dataset. Q2 refers to the 50th percentile. 50% of the values in the entire dataset are below Q2, and 50% are above it.
- The third quartile, Q3, is the middle value in the second half of the dataset. Q3 refers to the 75th percentile. 75% of the values in the entire dataset are below Q3, and 25% are above it.



Example: Car sales

For example, imagine you're a data professional working for an auto dealership. The manager of the sales team wants to compare the performance of each sales representative on the team. The manager asks you to analyze data that provides how many cars each sales representative sold during the past month.

Sales Representative	**#1**	**#2**	**#3**	**#4**	**#5**	**#6**	**#7**	**#8**
Cars Sold	18	13	6	10	15	7	10	9

You can calculate quartiles for your data in four steps:

1. Arrange the values in your dataset from smallest to largest.

[6, 7, 9, 10, 10, 13, 15, 18]

2. Find the median, or middle value, of your entire dataset. This is Q2. There are an even number of values in the dataset, so the median is the average of the two middle values, 10 and 10.

Q2 = (10 + 10) ÷ 2 = 20 ÷ 2 = 10

3. Find the median of the lower half of your dataset [6, 7, 9, 10]. This is Q1. The median is the average of the two middle values, 7 and 9.

Q1 = (7 + 9) ÷ 2 = 16 ÷ 2 = 8

4. Finally, find the median of the upper half of your dataset [10, 13, 15, 18]. This is Q3. The median is the average of the two middle values, 13 and 15.

Q3 = (13 + 15) ÷ 2 = 28 ÷ 2 = 14

Dividing the data into quartiles gives you a clear idea of sales rep performance. You now know that the lower quartile (Q1) of reps sold 8 cars or fewer, and the upper quartile (Q3) sold 14 cars or more. In other words, the lower 25% of reps sold 8 cars or fewer, and the upper 25% sold 14 cars or more. The middle 50% of representatives sold between 8 and 14 cars.

Interquartile range (IQR)

The middle 50% of your data is called the **interquartile range,** or **IQR**. The interquartile range is the distance between the first quartile (Q1) and the third quartile (Q3). This is the same as the distance between the 25th and 75th percentiles. IQR is useful for determining the relative position of your data values. For instance, data values outside the interval Q1 - (1.5 * IQR) and Q3 + (1.5 * IQR) are often considered outliers.

Note: *Technically, IQR is a measure of dispersion because it measures the spread of the middle half or middle 50% of your data (between Q1 and Q3). IQR is less sensitive to outliers than the range because it doesn't include the more extreme values in your dataset.*

IQR = Q3 - Q1. In this case, Q3 = 14 and Q1 = 8.

IQR = 14 - 8 = 6

Five number summary

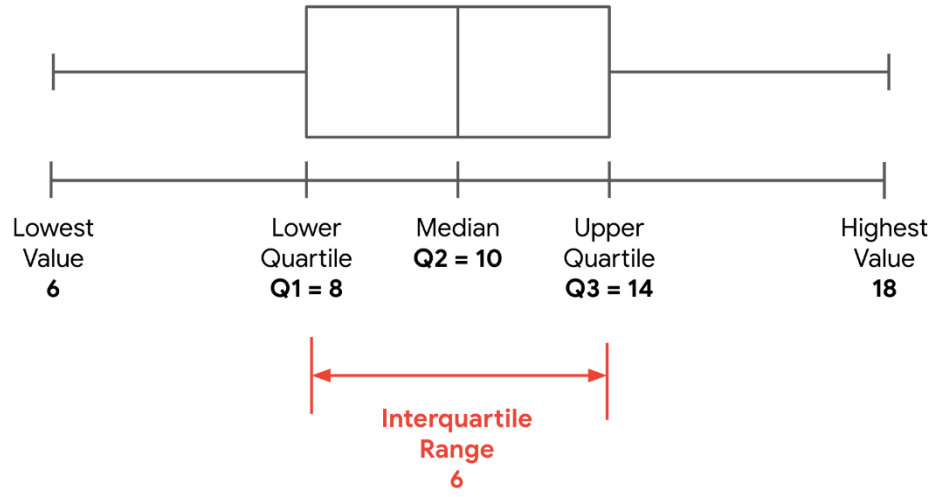
Finally, you can summarize the major divisions in your dataset with the five number summary. The five numbers include:

- The minimum
- The first quartile (Q1)
- The median, or second quartile (Q2)
- The third quartile (Q3)
- The maximum

The five number summary is useful because it gives you an overall idea of the distribution of your data, from the extreme values to the center. You can visualize it with a box plot.

The box part of the box plot goes from Q1 to Q3. The vertical line in the middle of the box is the median (Q2). The horizontal lines on each side of the box, known as whiskers, go from Q1 to the minimum, and from Q3 to the maximum.

The following box plot illustrates the data on car sales. You can find the values on the box plot and determine the interquartile range (IQR). The IQR is the length of the box, or the distance from Q1 to Q3.



Key takeaways

Data professionals use measures of position such as percentiles and quartiles to better understand all types of data, from product sales to household income. Measures of position help you quickly identify the relative location of your data values, and give you a more precise sense of the distribution of your data.

Resources for more information

To learn more about measures of position like percentiles and quartiles, check out the following resource:

- This statistics dictionary from Freie Universität Berlin provides clear definitions and useful examples of measures of position such as percentiles, quartiles, the five number summary, and more.

A/B testing: A way to compare two versions of something to find out which version performs better

Confidence interval: A range of values that describes the uncertainty surrounding an estimate

Descriptive statistics: A type of statistics that summarizes the main features of a dataset

Econometrics: A branch of economics that uses statistics to analyze economic problems

Inferential statistics: A type of statistics that uses sample data to draw conclusions about a larger population

Interquartile range: The distance between the first quartile (Q1) and the third quartile (Q3)

Literacy rate: The percentage of the population in a given age group that can read and write

Mean: The average value in a dataset

Measure of central tendency: A value that represents the center of a dataset

Measure of dispersion: A value that represents the spread of a dataset, or the amount of variation in data points

Measure of position: A method by which the position of a value in relation to other values in a dataset is determined

Median: The middle value in a dataset

Mode: The most frequently occurring value in a dataset

Parameter: A characteristic of a population

Percentile: The value below which a percentage of data falls

Population: Every possible element that a data professional is interested in measuring

Quartile: A value that divides a dataset into four equal parts

Range: The difference between the largest and smallest value in a dataset

Representative sample: A sample that accurately reflects the characteristics of a population

Sample : A subset of a population

Sampling: The process of selecting a subset of data from a population

Standard deviation: A statistic that calculates the typical distance of a data point from the mean of a dataset

Statistic: A characteristic of a sample

Statistical significance: The claim that the results of a test or experiment are not explainable by chance alone

Statistics: The study of the collection, analysis, and interpretation of data

Summary statistics: A method that summarizes data using a single number

Variance: The average of the squared difference of each data point from the mean

Fundamental concepts of probability

Recently, you learned that **probability** uses math to quantify uncertainty, or to describe the likelihood of something happening. For example, there might be an 80% chance of rain tomorrow, or a 20% chance that a certain candidate wins an election.

In this reading, you'll learn more about fundamental concepts of probability. We'll discuss the concept of a random experiment, how to represent and calculate the probability of an event, and basic probability notation.

Probability fundamentals

Foundational concepts: Random experiment, outcome, event

Let's begin with three concepts at the foundation of probability theory:

- Random experiment
- Outcome
- Event

Probability deals with what statisticians call random experiments, also known as statistical experiments. A **random experiment** is a process whose outcome cannot be predicted with certainty.

For example, before tossing a coin or rolling a die, you can't know the result of the toss or the roll. The result of the coin toss might be heads or tails. The result of the die roll might be 3 or 6.

All random experiments have three things in common:

- The experiment can have more than one possible outcome.
- You can represent each possible outcome in advance.
- The outcome of the experiment depends on chance.

In statistics, the result of a random experiment is called an outcome. For example, if you roll a die, there are six possible outcomes: 1, 2, 3, 4, 5, 6.

An event is a set of one or more outcomes. Using the example of rolling a die, an event might be rolling an even number. The event of rolling an even number consists of the outcomes 2, 4, 6. Or, the event of rolling an odd number consists of the outcomes 1, 3, 5.

In a random experiment, an event is assigned a probability. Let's explore how to represent and calculate the probability of a random event.

The probability of an event

The probability that an event will occur is expressed as a number between 0 and 1. Probability can also be expressed as a percent.

- If the probability of an event equals 0, there is a 0% chance that the event will occur.
- If the probability of an event equals 1, there is a 100% chance that the event will occur.

There are different degrees of probability between 0 and 1. If the probability of an event is close to zero, say 0.05 or 5%, there is a small chance that the event will occur. If the probability of an event is close to 1, say 0.95 or 95%, there is a strong chance that the event will occur. If the probability of an event equals 0.5, there is a 50% chance that the event will occur—or not occur.

Knowing the probability of an event can help you make informed decisions in situations of uncertainty. For example, if the chance of rain tomorrow is 0.1 or 10%, you can feel confident about your plans for an outdoor picnic. However, if the chance of rain is 0.9 or 90%, you may want to think about rescheduling your picnic for another day.

Calculate the probability of an event

To calculate the probability of an event in which all possible outcomes are equally likely, you divide the number of desired outcomes by the total number of possible outcomes. You may recall that this is also the formula for classical probability:

# of desired outcomes ÷ total # of possible outcomes

Let's explore the coin toss and die roll examples to get a better idea of how to calculate the probability of a single random event.

Example: Coin toss

Tossing a fair coin is a classic example of a random experiment:

- There is more than one possible outcome.
- You can represent each possible outcome in advance: heads or tails.
- The outcome depends on chance. The toss could turn up heads or tails.

Say you want to calculate the probability of getting heads on a single toss. For any given coin toss, the probability of getting heads is one chance out of two. This is $1 \div 2 = 0.5$, or 50%.

Now imagine that you were to toss a specially designed coin that had heads on both sides. Every time you toss this coin it will turn up heads. In this case, the probability of getting heads is 100%. The probability of getting tails is 0%.

Note that when you say the probability of getting heads is 50%, you aren't claiming that any actual sequence of coin tosses will result in exactly 50% heads. For example, if you toss a fair coin ten times, you may get 4 heads and 6 tails, or 7 heads and 3 tails. However, if you continue to toss the coin, you can expect the long-run frequency of heads to get closer and closer to 50%.

Example: Die roll

Rolling a six-sided die is another classic example of a random experiment:

- There is more than one possible outcome.
- You can represent all possible outcomes in advance: 1, 2, 3, 4, 5, and 6.
- The outcome depends on chance. The roll could turn up any number 1–6.

Say you want to calculate the probability of rolling a 3. For any given die roll, the probability of rolling a 3 is one chance out of six. This is $1 \div 6 = 0.1666$, or about 16.7%.

Probability notation

It helps to be familiar with probability notation as it's often used to symbolize concepts in educational and technical contexts.

In notation, the letter P indicates the probability of an event. The letters A and B represent individual events.

For example, if you're dealing with two events, you can label one event A and the other event B.

- The probability of event A is written as P(A).
- The probability of event B is written as P(B).
- For any event A, $0 \leq P(A) \leq 1$. In other words, the probability of any event A is always between 0 and 1.
- If $P(A) > P(B)$, then event A has a higher chance of occurring than event B.
- If $P(A) = P(B)$, then event A and event B are equally likely to occur.

Key takeaways

Data professionals use probability to help stakeholders make informed decisions about uncertain events. Your knowledge of fundamental concepts of probability will be useful as a building block for more complex calculations of probability.

Resources for more information

To learn more about fundamental concepts of probability, refer to the following resources:

- These lecture notes from Richland Community College provide a useful summary of the fundamental concepts and basic rules of probability.

The probability of multiple events

So far, you've been learning about calculating the probability of single events. Many situations, both in daily life and in data work, involve more than one event. As a future data professional, you'll often deal with probability for multiple events.

In this reading, you'll learn more about multiple events. You'll learn three basic rules of probability: the complement rule, the addition rule, and the multiplication rule. These rules help you better understand the probability of multiple events. First, we'll discuss two different types of events that these rules apply to: mutually exclusive and independent. Then, you'll learn how to calculate probability for both types of events.

Two types of events

The three basic rules of probability apply to different types of events. Both the complement rule and the addition rule apply to events that are mutually exclusive. The multiplication rule applies to independent events.

Mutually exclusive events

Two events are **mutually exclusive** if they cannot occur at the same time.

For example, you can't be on the Earth and on the moon at the same time, or be sitting down and standing up at the same time.

Or, take two classic examples of probability theory. If you toss a coin, you cannot get heads and tails at the same time. If you roll a die, you cannot get a 2 and a 4 at the same time.

Independent events

Two events are **independent** if the occurrence of one event does not change the probability of the other event. This means that one event does not affect the outcome of the other event.

For example, watching a movie in the morning does not affect the weather in the afternoon. Listening to music on the radio does not affect the delivery of your new refrigerator. These events are separate and independent.

Or, take two consecutive coin tosses or two consecutive die rolls. Getting heads on the first toss does not affect the outcome of the second toss. For any given coin toss, the probability of any outcome is always 1 out of 2, or 50%. Getting a 2 on the first roll does not affect the outcome of the second roll. For any given die roll, the probability of any outcome is always 1 out of 6, or 16.7%.

Three basic rules

Now that you know more about the difference between mutually exclusive and independent events, let's review three basic rules of probability:

- Complement rule
- Addition rule
- Multiplication rule

Complement rule

The complement rule deals with mutually exclusive events. In statistics, the complement of an event is the event not occurring. For example, either it snows or it does not snow. Either your soccer team wins the

championship or it does not win the championship. The complement of snow is no snow. The complement of winning is not winning.

The probability of an event occurring and the probability of it not occurring must add up to 1. Recall that a probability of 1 is the same as a 100%.

Another way to think about it is that there is a 100% chance of one event or the other event occurring. There may be a 40% chance of snow tomorrow. However, there is a 100% chance that it will either snow or not snow tomorrow.

The **complement rule** states that the probability that event A does not occur is 1 minus the probability of A. In probability notation, you can write this as:

Complement rule

P(A') = 1 - P(A)

Note: In probability notation, an apostrophe (') symbolizes negation. In other words, if you want to indicate the probability of event A NOT occurring, add an apostrophe after the letter A: P(A'). You can say this as "the probability of not A."

So, if you know there is a 40% chance of snow tomorrow, or a probability of 0.4, you can use the complement rule to calculate the probability that it does not snow tomorrow. The probability of no snow equals one minus the probability of snow.

P(no snow) = 1 - P(snow) = 1 - 0.4 = 0.6.

So, the probability of no snow tomorrow is 0.6, or 60%.

Addition rule (for mutually exclusive events)

The **addition rule** states that if events A and B are mutually exclusive, then the probability of A or B occurring is the sum of the probabilities of A and B. In probability notation, you can write this as:

P(A or B) = P(A) + P(B)

Note that there is also an addition rule for mutually inclusive events. In this course, we focus on the rule for mutually exclusive events.

Let's explore our example of rolling a die.

Die roll (rolling either a 2 or a 4)

Say you want to find the probability of rolling either a 2 or a 4 on a single roll. These two events are mutually exclusive. You can roll a 2 or a 4, but not both at the same time.

The addition rule says that to find the probability of either event occurring, you sum up their probabilities. The odds of rolling any single number on a die are 1 out of 6, or 16.7%.

P(rolling 2 or rolling 4) = P(rolling 2) + P(rolling 4) = ⅙ + ⅙ = ⅓

So, the probability of rolling either a 2 or a 4 is one out of three, or 33%.

Multiplication rule (for independent events)

The **multiplication rule** states that if events A and B are independent, then the probability of both A and B occurring is the probability of A multiplied by the probability of B. In probability notation, you can write this as:

$P(A \text{ and } B) = P(A) \times P(B)$

Note that there is also a multiplication rule for dependent events. In this course, we focus on the rule for independent events.

Let's continue with our example of rolling a die.

Die roll (rolling a 1 and then rolling a 6)

Now imagine two consecutive die rolls. Say you want to know the probability of rolling a 1 and then rolling a 6. These are independent events as the first roll does not affect the outcome of the second roll.

The probability of rolling a 1 and then a 6 is the probability of rolling a 1 multiplied by the probability of rolling a 6. The probability of each event is ⅙, or 16.7%. You can write this as:

P(rolling 1 on the first roll and rolling 6 on the second roll) = P(rolling 1 on the first roll)×P(rolling 6 on the second roll) = ⅙×⅙ = 1/36

So, the probability of rolling a 1 and then a 6 is one out of thirty-six, or about 2.8%.

Key takeaways

The basic rules of probability help you describe events that are mutually exclusive or independent. Understanding basic rules of probability is an essential foundation for more complex analyses you will perform as a future data professional.

Resources for more information

To learn more about probability, refer to the following interactive guide: Seeing Theory.

Calculate conditional probability for dependent events

Recently, you learned that **conditional probability** refers to the probability of an event occurring given that another event has already occurred. Conditional probability allows you to describe the relationship between dependent events, or how the occurrence of the first event affects the likelihood of the second event.

In this reading, you'll learn how to calculate conditional probability for two or more dependent events. Before we discuss calculating conditional probability, we'll go over the concept of dependence.

Conditional probability

Previously, you calculated probability for a single event, and for two or more independent events, such as two consecutive coin flips. Conditional probability applies to two or more dependent events.

Dependent events

Earlier, you learned two events are **independent** if the first event does not affect the outcome of the second event, or change its probability. For example, two consecutive coin tosses are independent events. Getting heads on the first toss doesn't affect the outcome of the second toss.

In contrast, two events are **dependent** if the occurrence of one event changes the probability of the other event. This means that the first event affects the outcome of the second event.

For instance, if you want to get a good grade on an exam, you first need to study the course material. Getting a good grade depends on studying. If you want to eat at a popular restaurant without waiting for a table, you have to arrive early. Avoiding a wait depends on arriving early. In each instance, you can say that the second event is dependent on, or conditional on, the first event.

Now that you have a better understanding of dependent events, let's return to conditional probability and review the formula.

Formula for conditional probability

The formula says that for two dependent events A and B, the probability of event A and event B occurring equals the probability of event A occurring, multiplied by the probability of event B occurring, given event A has already occurred.

Conditional probability

$$P(A \text{ and } B) = P(A) * P(B|A)$$

In probability notation, the vertical bar between the letters B and A indicates dependence, or that the occurrence of event B depends on the occurrence of event A. You can say this as "the probability of B given A."

The formula can also be expressed as the probability of event B given event A equals the probability that both A and B occur divided by the probability of A.

Conditional probability

$$P(B|A) = P(A \text{ and } B) / P(A)$$

These are just two ways of representing the same equation. Depending on the situation, or what information you are given up front, it may be easier to use one or the other.

Note: The conditional probability formula also applies to independent events. When A and B are independent events, P(B|A) = P(B). So, the formula becomes P(A and B) = P(A) * P(B). This formula is also the multiplication rule that you learned about earlier in the course.

Example: playing cards

Let's explore an example of conditional probability that deals with a standard deck of 52 playing cards.

Imagine two events:

- The first event is drawing a heart from the deck of cards.
- The second event is drawing another heart from the same deck.

Say you want to find out the probability of drawing two hearts in a row. These two events are dependent because getting a heart on the first draw changes the probability of getting a heart on the second draw.

A standard deck includes four different suits: hearts, diamonds, spades, and clubs. Each suit has 13 cards. For the first draw, the chance of getting a heart is 13 out of 52, or 25%. For the second draw, the probability of getting a heart changes because you've already picked a heart on the first draw. Now, there are 12 hearts in a deck of 51 cards. For the second draw, the chance of getting a heart is 12 out of 51, or about 23.5%. Getting a heart is now less likely—the probability has gone from 25% to 23.5%.

Now, let's apply the conditional probability formula:

P(A and B) = P(A) * P(B|A)

You want to calculate the probability of both event A and event B occurring. Let's call event A *1st heart,* which refers to getting a heart on the first draw. Let's call event B *2nd heart*, which refers to getting a heart on the second draw, given a heart was drawn the first time. The probability of event A is 13/52, or 25%. The probability of event B is 12/51, or 23.5%.

Let's enter these numbers into the formula:

P(1st heart and 2nd heart) = P(1st heart) * P(2nd heart | 1st heart) = 13/52 * 12/51 = 1/17 = 0.0588, or about 5.9%

So, there is a 5.9% chance of drawing two hearts in a row from a standard deck of playing cards.

Example: online purchases

Let's explore another example. Imagine you are a data professional working for an online retail store. You have data that tells you 20% of the customers who visit the store's website make a purchase of $100 or more. If a customer spends $100, they are eligible to receive a free gift card. The store randomly awards gift cards to 10% of the customers who spend at least $100.

You want to calculate the probability that a customer spends $100 and receives a gift card. Receiving a gift card depends on first spending $100. So, this is a conditional probability because it deals with two dependent events.

Let's apply the conditional probability formula:

P(A and B) = P(A) * P(B|A)

You want to calculate the probability of both event A and event B occurring. Let's call event A *$100* and event B *gift card*. The probability of event A is 0.2, or 20%. The probability of event B is 0.1, or 10%.

P($100 and gift card) = P($100) * P(gift card given $100) = 0.2 * 0.1 = 0.02, or 2%

So, the probability of a customer spending $100 or more and receiving a free gift card is 0.2 * 0.1 = 0.02, or 2%.

Key takeaways

Conditional probability helps you describe the relationship between dependent events. Data professionals often use conditional probability in a business context. For example, they might use conditional probability to predict how an event like a new ad campaign will impact sales revenue. This helps stakeholders make intelligent decisions about the best way to invest their company's resources.

Resources for more information

To learn more about conditional probability, refer to the following resource:

- This article from Investopedia discusses conditional probability in a business context.

Calculate conditional probability with Bayes's theorem

Recently, you learned that **Bayes's theorem** is a math formula for determining conditional probability. The theorem is named after Thomas Bayes, an 18th-century mathematician from London, England. Recall that **conditional probability** refers to the probability of an event occurring given that another event has already occurred. For example, when you draw an ace from a deck of playing cards, this changes the probability of drawing a second ace from the same deck.

In this reading, you'll learn more about the different parts of Bayes's theorem, and how you can use the theorem to calculate conditional probability.

Bayes's theorem

Bayes's theorem provides a way to update the probability of an event based on new information about the event.

Posterior and prior probability

In Bayesian statistics, **prior probability** refers to the probability of an event before new data is collected. **Posterior probability** is the updated probability of an event based on new data.

Bayes's theorem lets you calculate posterior probability by updating the prior probability based on your data.

For example, let's say a medical condition is related to age. You can use Bayes's theorem to more accurately determine the probability that a person has the condition based on age. The prior probability would be the probability of a person having the condition. The posterior, or updated, probability would be the probability of a person having the condition if they are in a certain age group.

The theorem

Let's examine the theorem itself.

Bayes's theorem states that for any two events A and B, the probability of A given B equals the probability of A multiplied by the probability of B given A divided by the probability of B.

Bayes's theorem

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)}$$

In the theorem, prior probability is the probability of event A. Posterior probability, or what you're trying to calculate, is the probability of event A given event B.

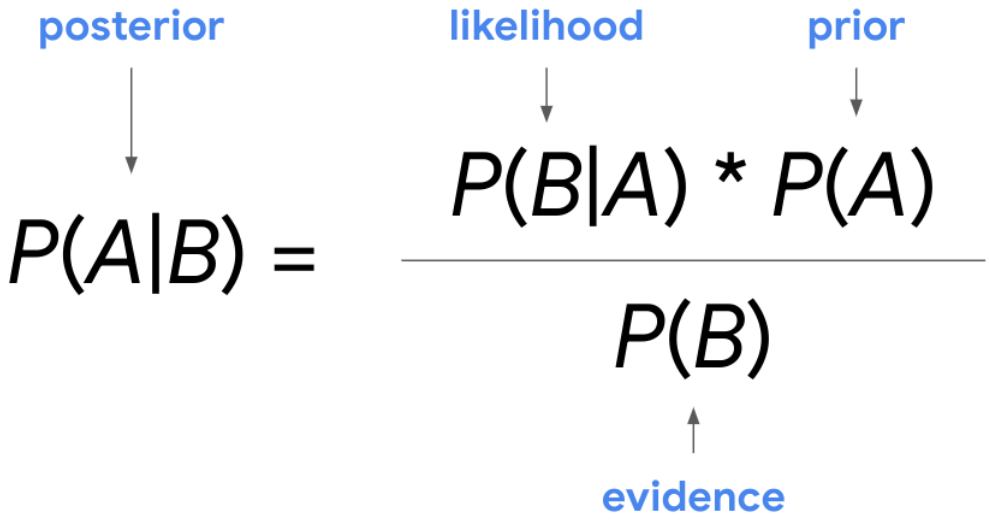
- **P(A)**: Prior probability
- **P(A|B)**: Posterior probability

Sometimes, statisticians and data professionals use the term "likelihood" to refer to the probability of event B given event A, and the term "evidence" to refer to the probability of event B.

- **P(B|A)**: Likelihood
- **P(B)**: Evidence

Using these terms, you can restate Bayes's theorem as:

- Posterior = Likelihood * Prior / Evidence



It can be helpful to think about the calculation from these different perspectives and help to map your problem onto the equation.

One way to think about Bayes's theorem is that it lets you transform a prior belief, P(A), into a posterior probability, P(A|B), using new data. The new data are the likelihood, P(B|A), and the evidence, P(B).

Note: *This reading provides an introduction to the basic concepts and terms associated with Bayes's theorem. A detailed examination of Bayesian statistics is beyond the scope of this course. As you progress in your career as a data professional, you'll have the opportunity to further explore Bayes's theorem and its various applications.*

For now, a key point to remember is that Bayes's theorem includes both the conditional probability of B given A and the conditional probability of A given B. If you know one of these probabilities, Bayes's theorem can help you determine the other.

Let's explore an example to get a better understanding of how the theorem works.

Example: spam filter

A well-known application of Bayes's theorem in the digital world is spam filtering, or predicting whether an email is spam or not. In practice, a sophisticated spam filter deals with many different variables, including the content of the email, its title, whether it has an attachment, the domain type of the sender address (.edu or .org), and more. However, we can use a simplified version of a Bayesian spam filter for our example.

Let's say you want to determine the probability that an email is spam given a specific word appears in the email. For this example, let's use the word "money."

You discover the following information:

- The probability of an email being spam is 20%.

- The probability that the word “money” appears in an email is 15%.
- The probability that the word “money” appears in a spam email is 40%.

In this example, your prior probability is the probability of an email being spam. Your posterior probability, or what you ultimately want to find out, is the probability that an email is spam given that it contains the word “money.” The new data you will use to update your prior probability is the probability that the word “money” appears in an email and the probability that the word “money” appears in a spam email.

When you work with Bayes’s theorem, it’s helpful to first figure out what event A is and what event B is—this makes it easier to understand the relationship between events and use the formula.

Let’s call event A a spam email and event B the appearance of the word “money” in an email. Now, you can re-write Bayes’s theorem using the word “spam” for event A and the word “money” for event B.

$P(A|B) = P(B|A) * P(A) / P(B)$

$P (Spam | Money) = P(Money | Spam) * P(Spam) / P(Money)$

You want to find out the following:

- **P(Spam | Money), or posterior probability**: the probability that an email is spam given that the word “money” appears in the email

Now, enter your data into the formula:

- **P(Spam), or prior probability**: the probability of an email being spam = 0.2, or 20%
- **P(Money), or evidence:** the probability that the word “money” appears in an email = 0.15, or 15%
- **P(Money | Spam), or likelihood:** the probability that the word “money” appears in an email given that the email is spam = 0.4, or 40%

$P (Spam | Money) = P(Money | Spam) * P(Spam) / P(Money) = 0.4 * 0.2 / 0.15 = 0.53333$, or about 53.3%.

So, the probability that an email is spam given that the email contains the word “money” is 53.3%.

Key takeaways

Bayes's theorem is the foundation for the field of Bayesian statistics, also known as Bayesian inference, which is a powerful method for analyzing and interpreting data in modern data analytics. Data professionals use Bayes’s theorem in a wide variety of fields, from artificial intelligence to medical testing.

Having a basic understanding of Bayes’s theorem will enable you to learn more about Bayesian statistics as you advance in your career as a data professional.

Resources for more information

To learn more about Bayes’s Theorem, refer to the following resource:

- Bayes' theorem explained by Pennsylvania State University

For an interesting discussion of the "prosecutor's fallacy," check out this page:

- Explanation of the prosecutor's fallacy by the American Journal of Epidemiology

Discrete probability distributions

Recently, you learned that data professionals use probability distributions to model different kinds of datasets, and to identify significant patterns in their data. Recall that a **probability distribution** describes the likelihood of the possible outcomes of a random event. Discrete probability distributions represent discrete random variables, or discrete events. Often, the outcomes of discrete events are expressed as whole numbers that can be counted. For example, rolling a die can result in a 2 or a 3, but not a decimal value such as 2.575 or 3.184.

In this reading, you'll get an overview of the main attributes of four common discrete probability distributions:

- Uniform
- Binomial
- Bernoulli
- Poisson

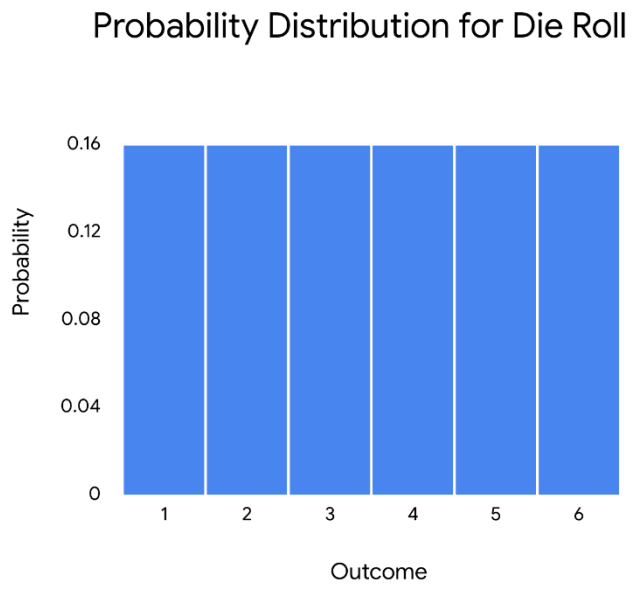
Discrete probability distributions

Uniform distribution

The uniform distribution describes events whose outcomes are all equally likely, or have equal probability.

For example, rolling a die can result in six outcomes: 1, 2, 3, 4, 5, or 6. The probability of each outcome is the same: 1 out of 6, or about 16.7%.

You can visualize a distribution with a graph, such as a histogram. For a discrete distribution, the random variable is plotted along the x-axis, and the corresponding probability is plotted along the y-axis. In this case, the x-axis represents each possible outcome of a single die roll, and the y-axis represents the probability of each outcome.



Note: *Data professionals often use the uniform distribution as part of more complex statistical methods, like Monte Carlo simulations. A detailed discussion of these methods is beyond the scope of this course.*

Note: *The uniform distribution applies to both discrete and continuous random variables.*

Binomial distribution

The **binomial distribution** models the probability of events with only two possible outcomes: success or failure. These outcomes are mutually exclusive and cannot occur at the same time.

This definition assumes the following:

- Each event is independent, or does not affect the probability of the others.
- Each event has the same probability of success.

Remember that success and failure are labels used for convenience. For example, if you toss a coin, there are only two possible outcomes: heads or tails. You could choose to label either heads or tails as a successful outcome based on the needs of your analysis.

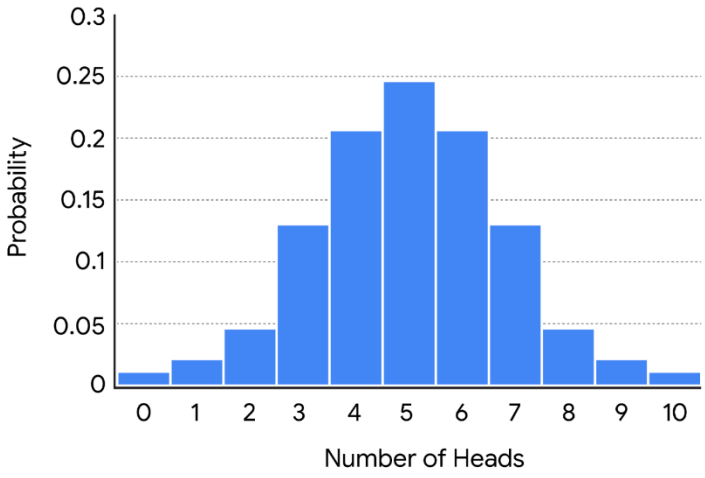
The binomial distribution represents a type of random event called a binomial experiment. A binomial experiment has the following attributes:

- The experiment consists of a number of repeated trials.
- Each trial has only two possible outcomes.
- The probability of success is the same for each trial.
- And, each trial is independent.

An example of a binomial experiment is tossing a coin 10 times in a row. This is a binomial experiment because it has the following features:

- The experiment consists of 10 repeated trials, or coin tosses.
- Each trial has only two possible outcomes: heads or tails.
- Each trial has the same probability of success. If you define success as heads, then the probability of success for each toss is the same: 50%.
- Each trial is independent. The outcome of one coin toss does not affect the outcome of any other coin toss.

On the histogram, the x-axis shows the number of heads, and the y-axis shows the probability of getting each result.



Data professionals might use the binomial distribution to model the probability that:

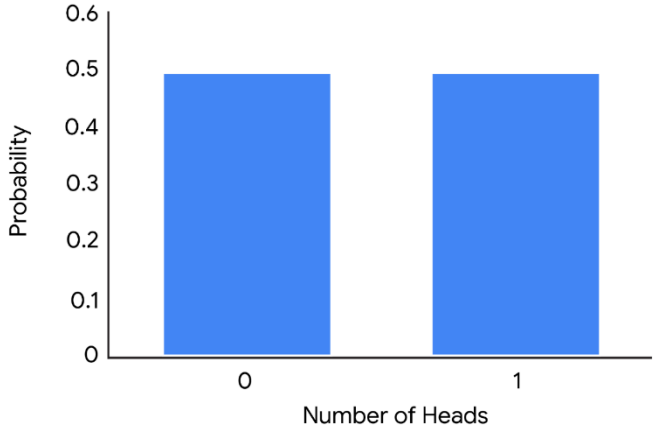
- A new medication generates side effects
- A credit card transaction is fraudulent
- A stock price rises in value

In machine learning, the binomial distribution is often used to classify data. For example, a data professional may train an algorithm to recognize whether a digital image is or is not a specific type of animal, like a cat or a dog.

Bernoulli distribution

The Bernoulli distribution is similar to the binomial distribution as it also models events that have only two possible outcomes (success or failure). The only difference is that the Bernoulli distribution refers to only a single trial of an experiment, while the binomial refers to repeated trials. A classic example of a Bernoulli trial is a single coin toss.

On the histogram, the x-axis represents the possible outcomes of a coin toss, and the y-axis represents the probability of each outcome.



Poisson distribution

The **Poisson distribution** models the probability that a certain number of events will occur during a specific time period.

Note: *The Poisson distribution can also be used to represent the number of events that occur in a specific space, such as a distance, area, or volume. In this course, we focus on time.*

The Poisson distribution represents a type of random experiment called a Poisson experiment. A Poisson experiment has the following attributes:

- The number of events in the experiment can be counted.
- The mean number of events that occur during a specific time period is known.
- Each event is independent.

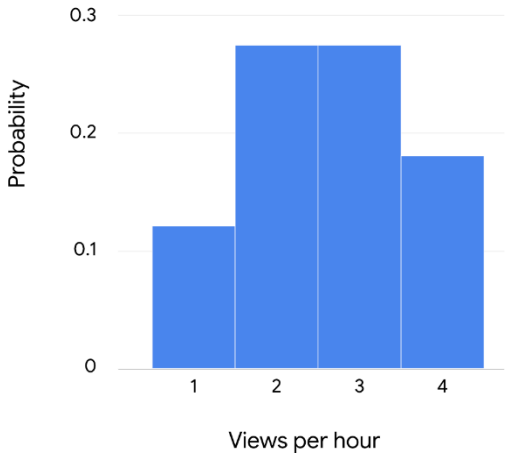
For example, imagine you have an online website where you post content. Your website averages two views per hour. You want to determine the probability that your website will receive a certain number of views in a given hour.

This is a Poisson experiment because:

- The number of events in the experiment can be counted. You can count the number of views.

- The mean number of events that occur during a specific time period is known. There is an average of two views per hour.
- Each outcome is independent. The probability of one person viewing your website does not affect the probability of another person viewing your website.

On the histogram, the x-axis shows the number of views per hour, and the y-axis shows the probability of occurrence.



Data professionals use the Poisson distribution to model data such as the number of:

- Calls per hour for a customer service call center
- Customers per day at a shop
- Thunderstorms per month in a city
- Financial transactions per second at a bank

Key takeaways

Identifying the distribution of your data is a key step in any analysis, and helps you make informed predictions about future outcomes. In your future career as a data professional, you'll use discrete distributions like the binomial and the Poisson to better understand your data. Knowing the probability distribution of your data will also help you choose the statistical method or machine learning model that works best for your analysis.

Resources for more information

To learn more about discrete probability distributions, refer to the following resources:

- This article from Statistics How To provides an overview of the concept of discrete probability distribution and offers links to explore more about specific kinds of distributions such as the binomial and Poisson.

Model data with the normal distribution

Recently, you've been learning about continuous probability distributions, and how they help data professionals model their data. Recall that continuous probability distributions represent continuous random variables, which can take on all the possible values in a range of numbers. Typically, these are decimal values that can be measured, such as height, weight, time, or temperature. For example, you can keep on measuring time with more accuracy: 1.1 seconds, 1.12 seconds, 1.1257 seconds, and so on.

In this course, we focus on a single continuous probability distribution: the normal distribution. In this reading, you'll learn more about the main characteristics of the normal distribution, and how the distribution can help you model your data.

Continuous probability distributions

Before we get to the specific attributes of the normal distribution, let's discuss some general features of all continuous probability distributions.

Probability Density and Probability

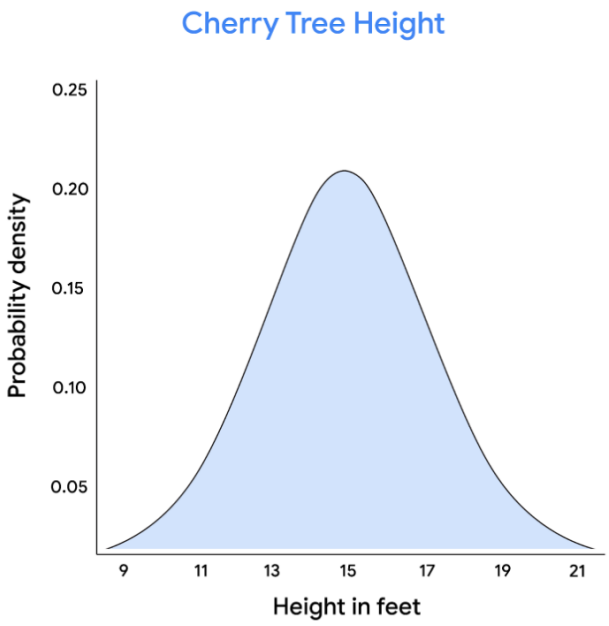
A probability function is a mathematical function that provides probabilities for the possible outcomes of a random variable.

There are two types of probability functions:

- Probability Mass Functions (PMFs) represent discrete random variables
- Probability Density Functions (PDFs) represent continuous random variables

A probability function can be represented as an equation or a graph. The math involved in probability functions is beyond the scope of this course. For now, it's important to know that the graph of a PDF appears as a curve. You've learned about the bell curve, which refers to the graph for a normal distribution.

As an example, imagine you have data on a random sample of cherry trees. Assume that the heights of the cherry trees are approximately normally distributed with a mean of 15 feet and a standard deviation of 2 feet.



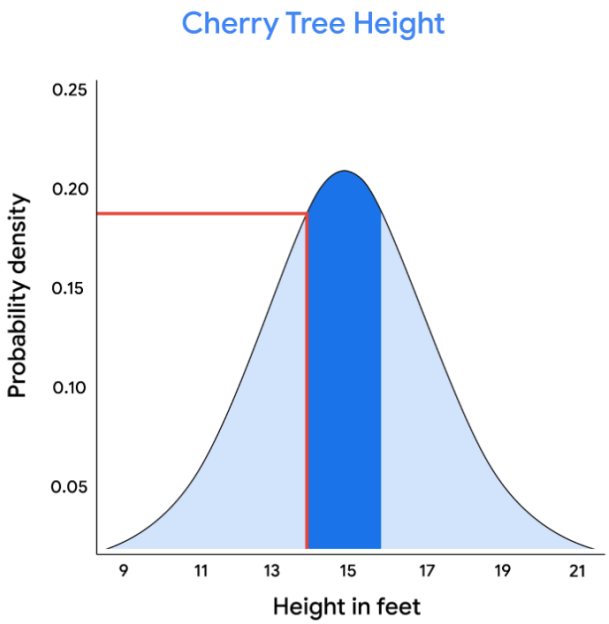
On a continuous distribution, the x-axis refers to the value of the variable you're measuring - in this case, cherry tree height. The y-axis refers to probability density. Note that probability density is not the same thing as probability.

The probability distribution for a continuous random variable can only tell you the probability that the variable takes on a range or interval of values. This is because a continuous random variable may have an infinite number of possible values. For instance, the height of a randomly chosen cherry tree could measure 15 feet, or 15.1 feet, or 15.175 feet, or 15.175245 feet, and so on.

Let's say you want to know the probability that the height of a randomly chosen cherry tree is exactly 15.1 feet. Because the height of the tree could be any decimal value in a given interval, the probability that the tree is exactly any single value is essentially zero.

So, for continuous distributions, it only makes sense to talk about the probability of intervals, such as the interval between 14.5 feet and 15.5 feet.

To find the probability of an interval, you calculate the area under the curve that corresponds to the interval. For example, the probability of a cherry tree having a height between 14.5 feet and 15.5 feet is equal to the area under the curve between the values 14.5 and 15.5 on the x-axis. This area appears as the shaded rectangle in the center of the graph.



In this case, the area of the rectangle is around 0.20. So, there is a 20% chance that the height of a randomly chosen cherry tree is between 14.5 feet and 15. 5 feet.

Note: data professionals typically use statistical software to calculate probabilities on a continuous distribution.

The normal distribution

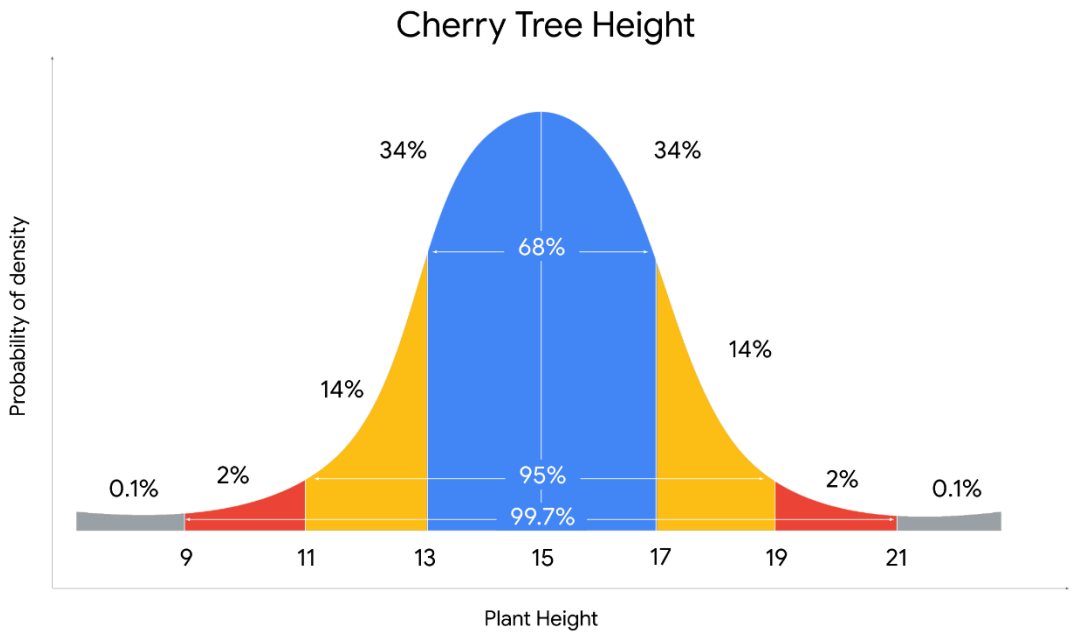
The normal distribution is a continuous probability distribution that is symmetric about the mean and bell-shaped. It is also known as the Gaussian distribution, after the German mathematician Carl Gauss, who first described its formula. The normal distribution is often called the bell curve because its graph has the shape of a bell, with a peak at the center and two downward sloping sides.

The normal distribution is the most common probability distribution in statistics because so many different kinds of datasets display a bell-shaped curve. For example, if you randomly sample 100 people, you will discover a normal distribution curve for continuous variables such as height, weight, blood pressure, shoe size, test scores, and more.

All normal distributions have the following features:

- The shape is a bell curve
- The mean is located at the center of the curve
- The curve is symmetrical on both sides of the mean
- The total area under the curve equals 1

Let's use our cherry tree example to clarify the features of the normal distribution. Recall that the mean height is 15 feet with a standard deviation of 2 feet.



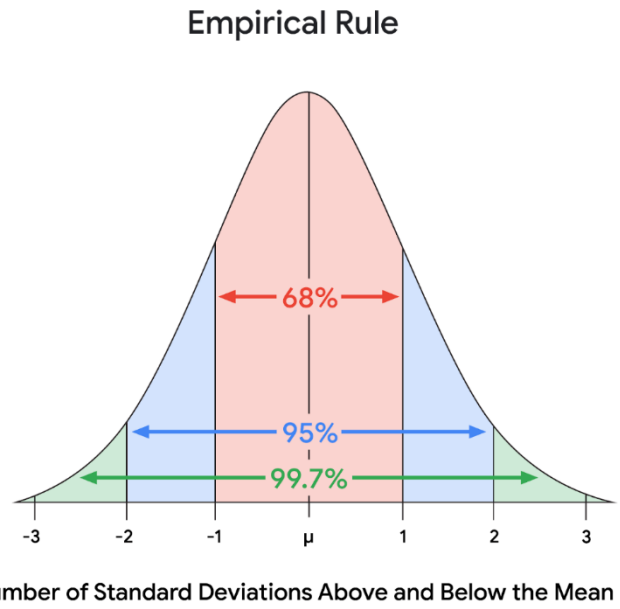
You may notice the following features of the normal curve:

- The mean is located at the center of the curve, and is also the peak of the curve. The mean height of 15 feet represents the most probable outcome in the dataset
- The curve is symmetrical about the mean. 50% of the data is above the mean, and 50% of the data is below the mean
- The farther a point is from the mean, the lower the probability of those outcomes. The points farthest from the mean represent the least probable outcomes in the dataset. These are trees that have more extreme heights, either short or tall
- The area under the curve is equal to 1. This means that the area under the curve accounts for 100% of the possible outcomes in the distribution

The empirical rule

You may also notice that the values on a normal curve are distributed in a regular pattern, based on their distance from the mean. This is known as **the empirical rule.** The rule states that for a given dataset with a normal distribution:

- 68% of values fall within 1 standard deviation of the mean
- 95% of values fall within 2 standard deviations of the mean
- 99.7% of values fall within 3 standard deviations of the mean



If you apply the empirical rule to our cherry tree example, you learn the following:

- Most trees, or 68%, will fall within 1 standard deviation of the mean height of 15 feet. This means that 68% of trees will measure between 13 feet and 17 feet, or 2 feet below the mean and 2 feet above the mean
- 95% of trees will measure between 11 feet and 19 feet, or within 2 standard deviations from the mean
- Almost all trees, or 99.7%, will measure between 9 feet and 21 feet, or within 3 standard deviations of the mean

The empirical rule can give you a quick estimate of how the values in a large dataset are distributed. This saves time and helps you better understand your data.

Knowing the location of your values on a normal distribution is also useful for detecting outliers. Recall that an outlier is a value that differs significantly from the rest of the data. Typically, data professionals consider values that lie more than 3 standard deviations below or above the mean to be outliers. It's important to identify outliers because some extreme values may be due to errors in data collection or data processing, and these false values may skew your results.

Key takeaways

As a data professional, you'll likely use the normal distribution to identify significant patterns in a wide variety of datasets. Understanding the normal distribution is also important for more advanced statistical methods, such as hypothesis testing and regression analysis, which you'll learn about later on.

Resources for more information

To learn more about continuous probability distributions and the normal distribution, check out the following resources:

- This article from Duke University provides a useful summary of the main characteristics of the normal distribution

Addition rule (for mutually exclusive events): The concept that if the events A and B are mutually exclusive, then the probability of A or B happening is the sum of the probabilities of A and B

Bayes' theorem: A math formula for stating that for any two events A and B, the probability of A given B equals the probability of A multiplied by the probability of B given A divided by the probability of B; Also referred to as Bayes' rule

Bayes' rule: (Refer to **Bayes' theorem**)

Bayesian inference: (Refer to **Bayesian statistics)**

Bayesian statistics: A powerful method for analyzing and interpreting data in modern data analytics; Also referred to as Bayesian inference

Binomial distribution: A discrete distribution that models the probability of events with only two possible outcomes: success or failure

Classical probability: A type of probability based on formal reasoning about events with equally likely outcomes

Complement of an event: In statistics, refers to an event not occurring

Complement rule: A concept stating that the probability that event A does not occur is one minus the probability of A

Conditional probability: Refers to the probability of an event occurring given that another event has already occurred

Continuous random variable: A variable that takes all the possible values in some range of numbers

Dependent events: The concept that two events are dependent if one event changes the probability of the other event

Discrete random variable: A variable that has a countable number of possible values

Empirical probability: A type of probability based on experimental or historical data

Empirical rule: A concept stating that the values on a normal curve are distributed in a regular pattern, based on their distance from the mean

False positive: A test result that indicates something is present when it really is not

Independent events: The concept that two events are independent if the occurrence of one event does not change the probability of the other event

Multiplication rule (for independent events): The concept that if the events A and B are independent, then the probability of both A and B happening is the probability of A multiplied by the probability of B

Mutually exclusive: The concept that two events are mutually exclusive if they cannot occur at the same time

Normal distribution: A continuous probability distribution that is symmetrical on both sides of the mean and bell-shaped

Objective probability: A type of probability based on statistics, experiments, and mathematical measurements

Poisson distribution: A probability distribution that models the probability that a certain number of events will occur during a specific time period

Posterior probability: Refers to the updated probability of an event based on new data

Prior probability: Refers to the probability of an event before new data is collected

Probability: The branch of mathematics that deals with measuring and quantifying uncertainty

Probability distribution: A function that describes the likelihood of the possible outcomes of a random event

Random experiment: A process whose outcome cannot be predicted with certainty

Random variable: A variable that represents the values for the possible outcomes of a random event

Sample space: The set of all possible values for a random variable

Standard deviation: A statistic that calculates the typical distance of a data point from the mean of a dataset

Standardization: The process of putting different variables on the same scale

Subjective probability: A type of probability based on personal feelings, experience, or judgment

Z-score: A measure of how many standard deviations below or above the population mean a data point is

The relationship between sample and population

Earlier, you learned that **inferential statistics** use sample data to draw conclusions or make predictions about a larger population. Data professionals use inferential statistics to gain valuable insights about their data.

In this reading, you'll learn about the relationship between sample and population in more detail. We'll also discuss how data professionals use sampling in data work, and the importance of working with a sample that is representative of the population.

Population and Sample

Population vs. sample

In statistics, a **population** includes every possible element that you are interested in measuring, or the entire dataset that you want to draw conclusions about. A statistical population can refer to any type of data, including:

- People
- Organizations
- Objects
- Events
- And more

For instance, a population might be the set of:

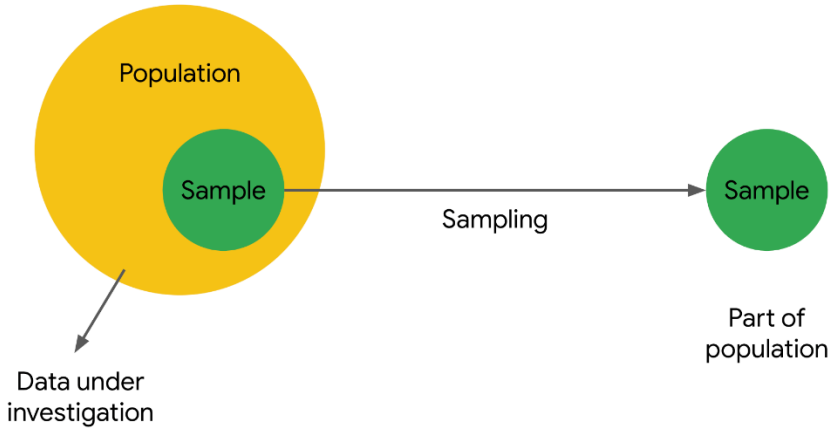
- All students at a university
- All the cell phones ever manufactured by a company
- All the forests on Earth

A **sample** is a subset of a population.

Samples drawn from the above populations might be:

- The math majors at the university
- The cell phones manufactured by the company in the last week
- The forests in Canada

Data professionals use samples to make inferences about populations. In other words, they use the data they collect from a small part of the population to draw conclusions about the population as a whole.



Sampling

Sampling is the process of selecting a subset of data from a population.

In practice, it's often difficult to collect data on every member or element of an entire population. A population may be very large, geographically spread out, or otherwise difficult to access. Instead, you can use sample data to draw conclusions, make estimates, or test hypotheses about the population as a whole.

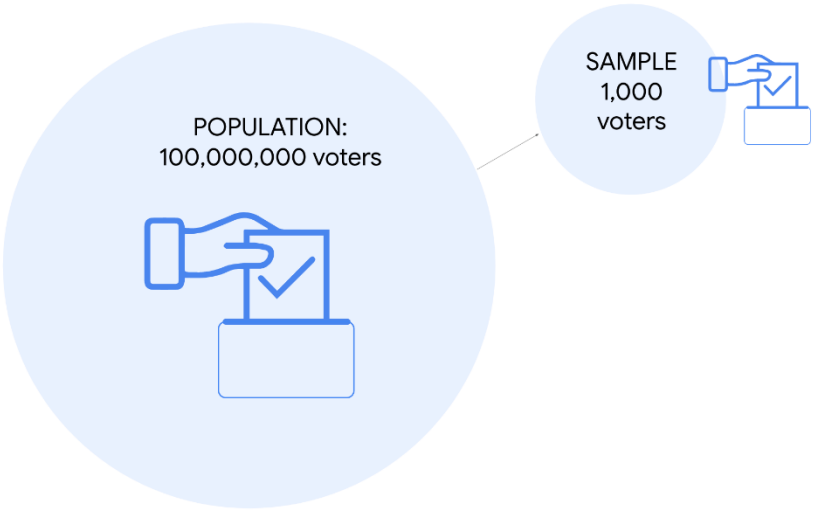
Data professionals use sampling because:

- It's often impossible or impractical to collect data on the whole population due to size, complexity, or lack of accessibility
- It's easier, faster, and more efficient to collect data from a sample
- Using a sample saves money and resources
- Storing, organizing, and analyzing smaller datasets is usually easier, faster, and more reliable than dealing with extremely large datasets

Example: election poll

Imagine you're a data professional working in a country with a large population like India, Indonesia, the United States, or Brazil. There is an upcoming national election for president. You want to conduct an election poll to see which candidate voters prefer. Let's say the population of eligible voters is 100 million people. To survey 100 million people on their voting preferences would take an enormous amount of time, money, and resources – even assuming it would be possible to locate and contact all voters, and that all voters would be willing to participate.

However, it is realistic to survey a sample of 100 or 1000 voters drawn from the larger population of all voters. When you're dealing with a large population, sampling can help you make valid inferences about the population as a whole.



Representative sample

To make valid inferences or accurate predictions about a population, your sample should be representative of the population as a whole. Recall that a **representative sample** accurately reflects the characteristics of a population. The inferences and predictions you make about your population are based on your sample data. If your sample doesn't accurately reflect your population, then your inferences will not be reliable, and your predictions will not be accurate. And this can lead to negative outcomes for stakeholders and organizations.

Statistical methods such as probability sampling help ensure your sample is representative by collecting random samples from the various groups within a population. These methods help reduce sampling bias and increase the validity of your results. You'll learn more about sampling methods later on.

Example: election poll

Ideally, the sample for your election poll will accurately reflect the characteristics of the overall voter population. A voter population in a large country will be diverse in political perspectives, geographic location, age, gender, race, education level, socioeconomic status, etc. Your sample will not be representative if you only collect data from people who belong to certain groups and not others. For example, if you survey people from one political party, or who have advanced degrees, or are older than 70. The results of an election poll based on a non-representative sample will not be accurate. In general, any claims or inferences you make about any population will have more validity if they are based on a representative sample.

Key takeaways

Data professionals work with powerful statistical tools that can model complex datasets and help generate valuable insights. But, if the sample data you're working with doesn't accurately reflect your population—that is, if your sample is not representative—then it doesn't matter how good your model is. If your predictive model is based on a bad sample, then your predictions will not be accurate.

Ultimately, the quality of your sample helps determine the quality of the insights you share with stakeholders. To make reliable inferences about a population, make sure your sample is representative of the population.

The stages of the sampling process

Recently, you've been learning about sampling. As a data professional, you'll work with sample data all the time. Often, this will be sample data previously collected by other researchers; sometimes, your team may collect their own data. Either way, it's important to know how the sampling process works, because it helps determine whether your sample is representative of the population, and whether your sample is unbiased.

In this reading, we'll go over the main stages of the sampling process in more detail. This will give you a better understanding of how the sampling process works and how each step of the process can affect your sample data.

The sampling process

First, let's review the main steps of the sampling process:

1. Identify the target population
2. Select the sampling frame
3. Choose the sampling method
4. Determine the sample size
5. Collect the sample data

Let's explore each step in more detail with an example. Imagine you're a data professional working for a company that manufactures home appliances. The company wants to find out how customers feel about the innovative digital features on their newest refrigerator model. The refrigerator has been on the market for two years and 10,000 people have purchased it. Your

manager asks you to conduct a customer satisfaction survey and share the results with stakeholders.

Step 1: Identify the target population

The first step in the sampling process is defining your target population. The **target population** is the complete set of elements that you're interested in knowing more about. Depending on the context of your research, your population may include individuals, organizations, objects, events, or any other type of data you want to investigate.

A well-defined population reduces the probability of including participants who do not fit the precise scope of your research. For example, you don't want to include all the company's customers, or customers who purchased the company's other refrigerator models.

In this case, your target population will be the 10,000 customers who purchased the company's newest refrigerator model. These are the customers you want to survey to learn about their experience with the newest model.

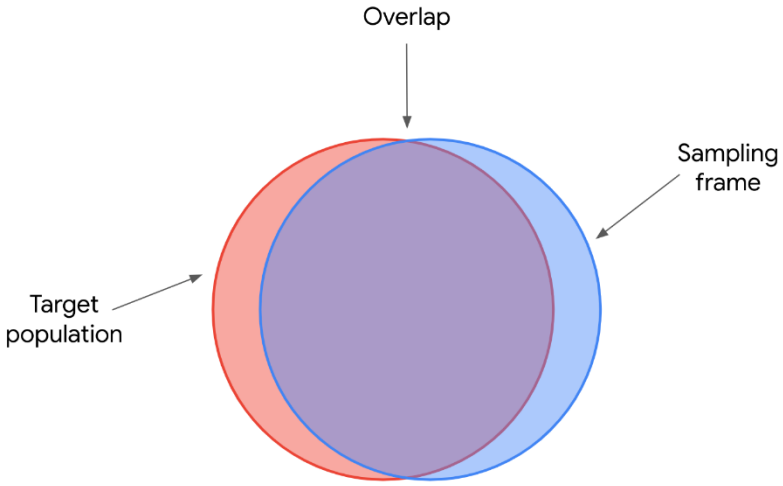
Step 2: Select the sampling frame

The next step in the sampling process is to create a sampling frame. A **sampling frame** is a list of all the individuals or items in your target population.

The difference between a target population and a sampling frame is that the population is general and the frame is specific. So, if your target population is all the customers who purchased the refrigerator, your sampling frame could be an alphabetical list of the names of all these customers. The customers in your sample will be selected from this list.

Ideally, your sampling frame should include the entire target population. However, for practical reasons, your sampling frame may not exactly match your target population, because you may

not have access to every member of the population. For instance, the company's customer database may be incomplete, or contain data processing errors. Or, some customers may have changed their contact information since their purchase, and you may be unable to locate or contact them. Furthermore, sometimes the sampling frame might include elements outside of the target population simply by accident or because it is impossible to know the target population with certainty.



Therefore, generally your sampling frame is the *accessible* part of your target population, but sometimes it will include elements apart from this set.

Step 3: Choose the sampling method

The third step in the sampling process is choosing a sampling method.

There are two main types of sampling methods: **probability sampling** and **non-probability sampling**. Later on, we'll explore the specific methods in more detail. For now, just know that probability sampling uses random selection to generate a sample. Non-probability sampling is often based on convenience, or the personal preferences of the researcher, rather than random selection. Often, probability sampling methods require more time and resources than non-probability sampling methods.

Ideally, your sample will be representative of the population. One way to help ensure that your sample is representative is to choose the right sampling method. Because probability sampling methods are based on random selection, every element in the population has an equal chance of being included in the sample. This gives you the best chance to get a representative sample, as your results are more likely to accurately reflect the overall population.

So, assuming you have the budget, the resources, and the time, you should use a probability sampling method for your survey.

Step 4: Determine the sample size

Step four of the sampling process is to determine the best size for your sample, since you don't have the resources to survey everyone in your sampling frame. In statistics, sample size refers to the number of individuals or items chosen for a study or experiment.

Sample size helps determine the precision of the predictions you make about the population. In general, the larger the sample size, the more precise your predictions. However, using larger samples typically requires more resources.

The sample size you choose depends on various factors, including the sampling method, the size and complexity of the target population, the limits of your resources, your timeline, and the goal of your research.

Based on these factors, you can decide how many customers to include in your sample.

Step 5: Collect the sample data

Now, you're ready to collect your sample data, which is the final step in the sampling process.

You give a customer satisfaction survey to the customers selected for your sample. The survey responses provide useful data on how customers feel about the digital features of the refrigerator. Then, you share your results with stakeholders to help them make more informed decisions about whether to continue to invest in these features for future versions of this refrigerator, and develop similar features for other models.

Key takeaways

Effective sampling ensures that your sample data is representative of your population. Then, when you use sample data to make inferences about the population, you can be reasonably confident that your inferences are reliable.

The decisions you make at each step of the sampling process can affect the quality of your sample data. Understanding the sampling process will make you a better data professional, whether you're analyzing data collected by other researchers or conducting a survey on your own.

Probability sampling methods

Earlier, you learned that there are two main types of sampling methods: probability sampling and non-probability sampling. **Probability sampling** uses random selection to generate a sample. **Non-probability sampling** is often based on convenience, or the personal preferences of the researcher, rather than random selection. The sampling method you use helps determine if your sample is representative of your population, and if your sample is biased. Probability sampling gives you the best chance to create a sample that is representative of the population.

In this reading, you'll learn more about the different methods of probability sampling, and the benefits and drawbacks of each method.

Probability Sampling Methods

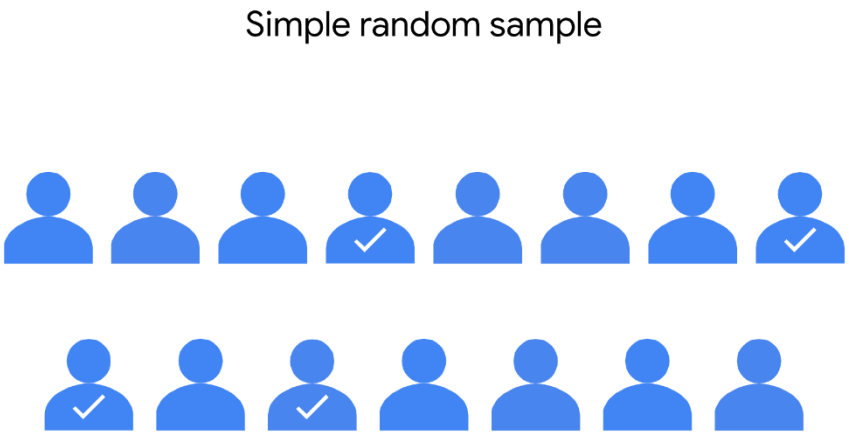
There are four different probability sampling methods:

- Simple random sampling
- Stratified random sampling
- Cluster random sampling
- Systematic random sampling

Let's explore each method in more detail.

Simple random sampling

In a **simple random sample**, every member of a population is selected randomly and has an equal chance of being chosen. You can randomly select members using a random number generator, or by another method of random selection.



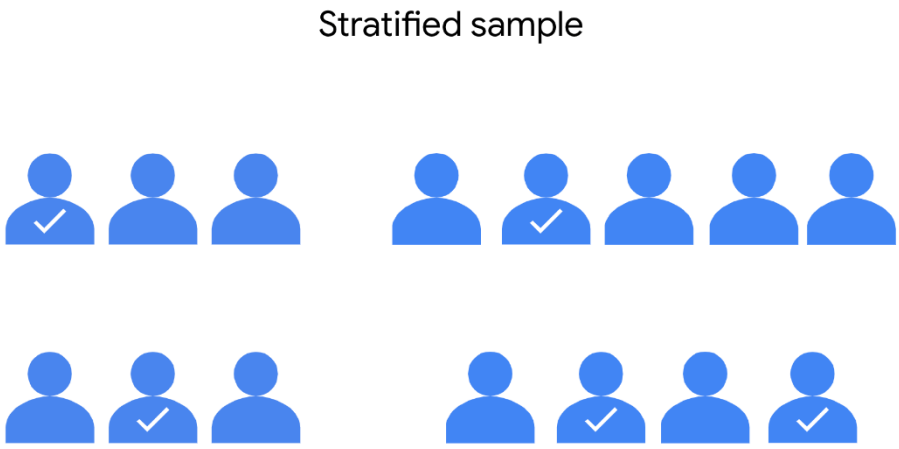
For example, imagine you want to survey the employees of a company about their work experience. The company employs 10,000 people. You can assign each employee in the company database a number from 1 to 10,000, and then use a random number generator to select 100 people for your sample. In this scenario, each of the employees has an equal chance of being chosen for the sample.

The main benefit of simple random samples is that they're usually fairly representative, since every member of the population has an equal chance of being chosen. Random samples tend to avoid bias, and surveys like these give you more reliable results.

However, in practice, it's often expensive and time-consuming to collect large simple random samples. And if your sample size is not large enough, a specific group of people in the population may be underrepresented in your sample. If you use a larger sample size, your sample will more accurately reflect the population.

Stratified random sampling

In a **stratified random sample**,you divide a population into groups, and randomly select some members from each group to be in the sample. These groups are called strata. Strata can be organized by age, gender, income, or whatever category you're interested in studying.



For example, imagine you're doing market research for a new product, and you want to analyze the preferences of consumers in different age groups. You might divide your target population into strata according to age: 20-29, 30-39, 40-49, 50-59, etc. Then, you can survey an equal number of people from each age group, and draw conclusions about the consumer preferences of each age group. Your results will help marketers decide which age groups to focus on to optimize sales for the new product.

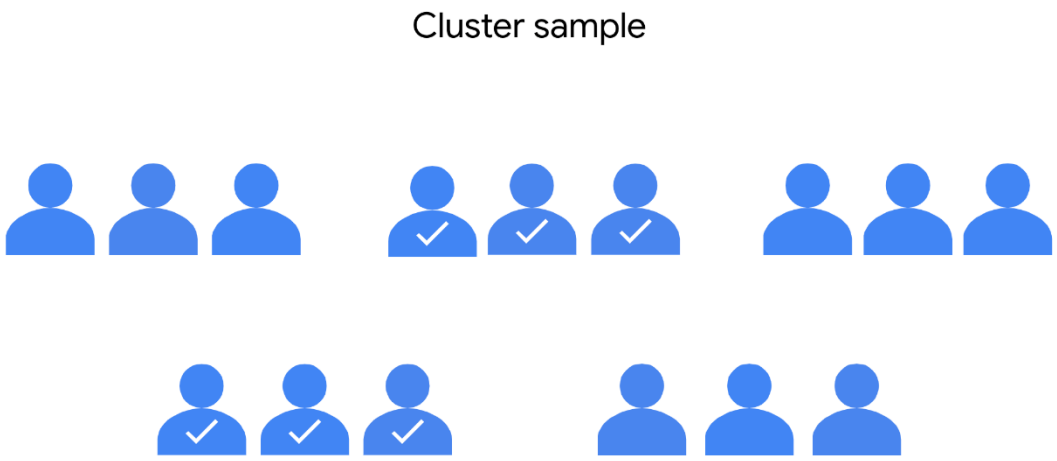
Stratified random samples help ensure that members from each group in the population are included in the survey. This method helps provide equal representation for underrepresented groups, and allows you to draw more precise conclusions about each of the strata. There may be significant differences in the purchasing habits of a 21-year-old and a 51-year-old. Stratified sampling helps ensure that both perspectives are captured in the sample.

One main disadvantage of stratified sampling is that it can be difficult to identify appropriate strata for a study if you lack knowledge of a population. For example, if you want to study median income among a population, you may want to stratify your sample by job type, or industry, or location, or education level. If you don't know how relevant these categories are to median income, it will be difficult to choose the best one for your study.

Cluster random sampling

When you're conducting a **cluster random sample**, you divide a population into clusters, randomly select certain clusters, and include all members from the chosen clusters in the sample.

Cluster sampling is similar to stratified random sampling, but in stratified sampling, you randomly choose *some* members from each group to be in the sample. In cluster sampling, you choose *all* members from a group to be in the sample. Clusters are divided using identifying details, such as age, gender, location, or whatever you want to study.



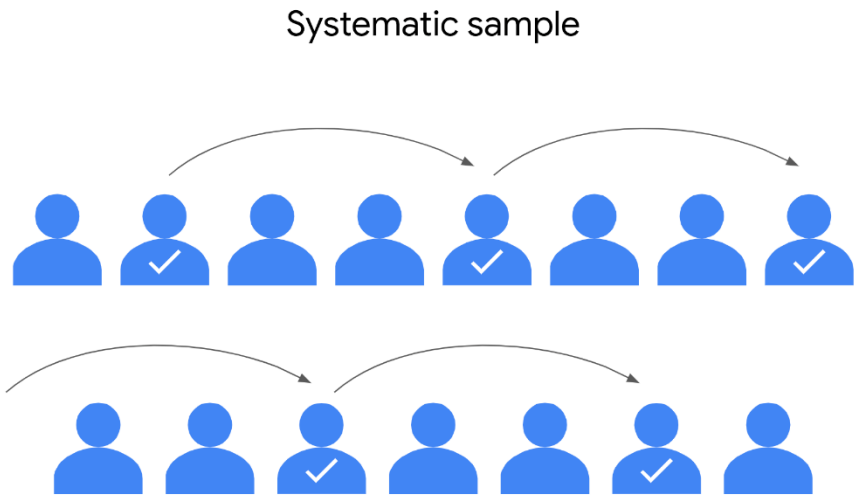
For example, imagine you want to conduct an employee satisfaction survey at a global restaurant franchise using cluster sampling. The franchise has 40 restaurants around the world. Each restaurant has about the same number of employees in similar job roles. You randomly select 4 restaurants as clusters. You include all the employees at the 4 restaurants in your sample.

One advantage of this method is that a cluster sample gets every member from a particular cluster, which is useful when each cluster reflects the population as a whole. This method is helpful when dealing with large and diverse populations that have clearly defined subgroups. If researchers want to learn more about home ownership in the suburbs of Auckland, New Zealand, they can use several well-chosen suburbs as a representative sample of all the suburbs in the city.

A main disadvantage of cluster sampling is that it may be difficult to create clusters that accurately reflect the overall population. For example, for practical reasons, you may only have access to restaurants in England when the franchise has locations all over the world. And employees in England may have different characteristics and values than employees in other countries.

Systematic random sampling

In a **systematic random sample**, you put every member of a population into an ordered sequence. Then, you choose a random starting point in the sequence and select members for your sample at regular intervals.



Imagine you want to survey students at a high school about their study habits. For a systematic random sample, you'd put the students' names in alphabetical order and randomly choose a starting point: say, number 4. Starting with number 4, you select every 10th name on the list (4, 14, 24, 34, …), until you have a sample of 100 students.

One advantage of systematic random samples is that they're often representative of the population, since every member has an equal chance of being included in the sample. Whether the student's last name starts with L or Q isn't going to affect their characteristics. Systematic sampling is also quick and convenient when you have a complete list of the members of your population.

One disadvantage of systematic sampling is that you need to know the size of the population that you want to study before you begin. If you don't have this information, it's difficult to choose consistent intervals. Plus, if there's a hidden pattern in the sequence, you might not get a representative sample. For example, if every 10th name on your list happens to be an honor student, you may only get feedback on the study habits of honor students – and not *all* students.

Key takeaways

The four methods of probability sampling we've covered—simple, stratified, cluster, and systematic—are all based on random selection, which is the preferred method of sampling for most data professionals. Probability sampling methods give you the best chance to create a sample that is representative of the population as a whole. And working with a representative sample allows you to make reliable inferences and accurate predictions about the population you're researching.

Non-probability sampling methods

Recently, you learned that probability sampling methods use random selection, which helps avoid sampling bias. A randomly chosen sample means that all members of the population have an equal chance of being included. In contrast, non-probability sampling methods do not use random selection, so they do not typically generate representative samples. In fact, non-probability methods often result in biased samples. **Sampling bias** occurs when some members of the population are more likely to be selected than other members.

In this reading, you'll learn more about four methods of non-probability sampling, and learn how sampling bias can affect each method. We'll also discuss why non-probability sampling may be useful in certain situations.

Non-probability sampling methods

Non-probability samples use non-random methods of selection, so not all members of a population have an equal chance of being selected. This is why non-probability methods have a high risk of sampling bias. However, non-probability methods are often less expensive and more convenient for researchers to conduct. Sometimes, due to limited time, money, or other resources, it's not possible to use probability sampling. Plus, non-probability methods can be useful for exploratory studies, which seek to develop an initial understanding of a population, rather than make inferences about the population as a whole.

We'll go over four methods of non-probability sampling:

- Convenience sampling
- Voluntary response sampling
- Snowball sampling
- Purposive sampling

Let's explore each method in more detail.

Convenience sampling

For convenience sampling, you choose members of a population that are easy to contact or reach. As the name suggests, conducting a convenience sample involves collecting a sample from somewhere convenient to you, such as your workplace, a local school, or a public park.

For example, to conduct an opinion poll, a researcher might stand at the entrance of a shopping mall during the day and poll people that happen to walk by.

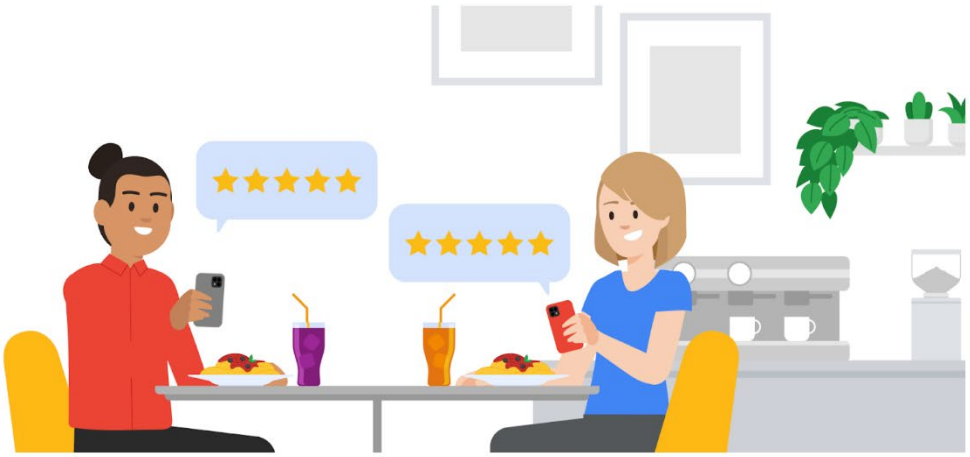
Because these samples are based on convenience to the researcher, and not a broader sample of the population, convenience samples often suffer from undercoverage bias. Undercoverage bias occurs when some members of a population are inadequately represented in the sample. For example, the above sample will underrepresent people who don't like to shop at malls, or prefer to shop at a different mall, or don't visit the mall because they lack transportation.

Convenience sampling is often quick and inexpensive, but it's not a reliable way to get a representative sample.

Voluntary response sampling

A voluntary response sample consists of members of a population who volunteer to participate in a study. Like a convenience sample, a voluntary response sample is often based on convenient access to a population. However, instead of the researcher selecting participants, participants volunteer on their own.

For example, let's say college administrators want to know how students feel about the food served on campus. They email students a link to an online survey about the quality of the food, and ask students to fill out the survey if they have time.

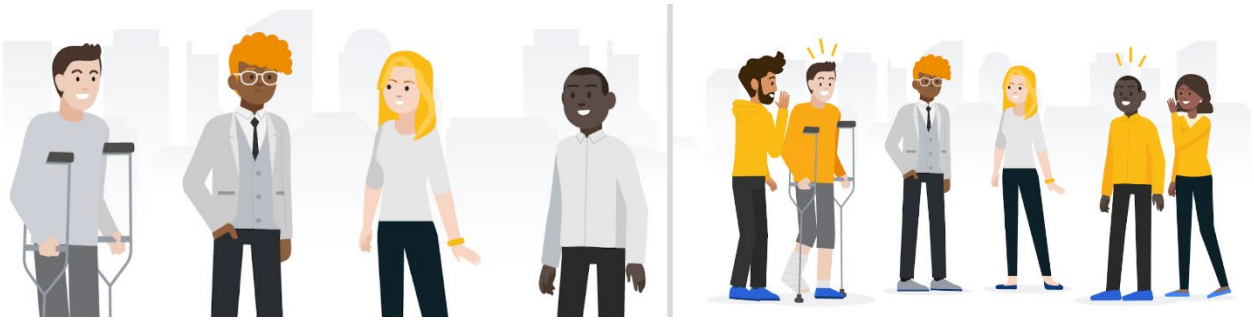


Voluntary response samples tend to suffer from nonresponse bias, which occurs when certain groups of people are less likely to provide responses. People who voluntarily respond will likely have stronger opinions, either positive or negative, than the rest of the population. In this case, only students who really like or really dislike the food may be motivated to fill out the survey. The survey may omit many students who have more mild opinions about the food, or are neutral. This makes the volunteer students unrepresentative of the overall student population.

Snowball sampling

In a snowball sample,researchers recruit initial participants to be in a study and then ask them to recruit other people to participate in the study. Like a snowball, the sample size gets bigger and bigger as more participants join in. Researchers often use snowball sampling when the population they want to study is difficult to access.

For example, imagine a researcher is studying people with a rare medical condition. Due to reasons of confidentiality, it may be difficult for the researcher to obtain contact information for members of this population from hospitals or other official sources. However, if the researcher can find a couple of people willing to participate, these two people may know others with the same condition. The initial participants could then recruit others by sharing the potential benefits of the study.



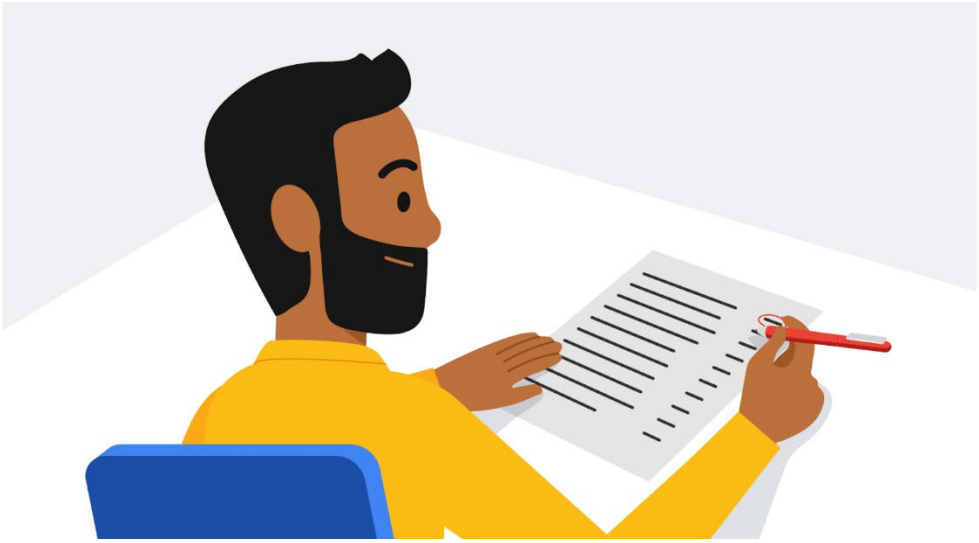
The first illustration shows two researchers sharing information with two people. The second illustration shows those same researchers standing near four people sharing information.

Snowball sampling can take a lot of time, and researchers must rely on participants to successfully continue the recruiting process and build up the "snowball." This type of recruiting can also lead to sampling bias. Because initial participants recruit additional participants on their own, it's likely that most of them will share similar characteristics, and these characteristics might be unrepresentative of the total population under study.

Purposive sampling

In purposive sampling, researchers select participants based on the purpose of their study. Because participants are selected for the sample according to the needs of the study, applicants who do not fit the profile are rejected.

For example, imagine a game development company wants to conduct market research on a new video game before its public release. The research team only wants to include gaming experts in their sample. So, they survey a group of professional gamers to provide feedback on potential improvements.



In purposive sampling, researchers often intentionally exclude certain groups from the sample to focus on a specific group they think is most relevant to their study. In this case, the researcher excludes amateur gamers. Amateur gamers may purchase the new game for different reasons than professional gamers, and enjoy game features that don't appeal to professionals. This could lead to biased results, because the professionals in the sample are not likely to be representative of the overall gamer population.

Purposive sampling is often used when a researcher wants to gain detailed knowledge about a specific part of a population, or where the population is very small and its members all have similar characteristics. Purposive sampling is not effective for making inferences about a large and diverse population.

Key takeaways

Non-probability sampling is useful for collecting data in situations where you have limited time, budget, and other resources. Non-probability sampling is also useful for exploratory research, when you want to get an initial understanding of a population, rather than make inferences about the population as a whole. However, it's important to remember that non-probability sampling methods have a high risk of sampling bias.

As a data professional, you have to think about bias and fairness from the moment you start collecting sample data to the time you present your conclusions. Once you become aware of some common forms of bias, you can remain on the alert for bias in any form.

Infer population parameters with the central limit theorem

Recently, you learned about the central limit theorem and how it can help you work with a wide variety of datasets. Data professionals use the central limit theorem to estimate population parameters for data in economics, science, business, and many other fields.

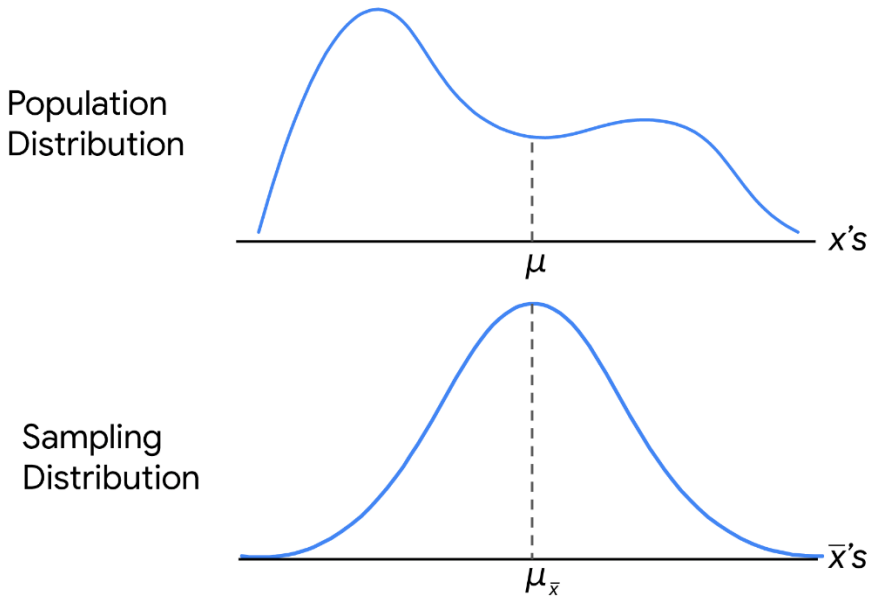
In this reading, you'll learn more about the central limit theorem and how it can help you estimate the population mean for different types of data. We'll go over the definition of the theorem, the conditions that must be met to apply the theorem, and check out an example of the theorem in action.

The central limit theorem

Definition

The **central limit theorem** states that the sampling distribution of the mean approaches a normal distribution as the sample size increases. In other words, as your sample size increases, your sampling distribution assumes the shape of a bell curve. And, as you sample more observations from a population, the sample mean gets closer to the population mean. If you take a large enough sample of the population, the sample mean will be roughly equal to the population mean.

For example, imagine you want to estimate the average weight of a certain class of vehicle, like light-duty pickup trucks. Instead of weighing millions of pickup trucks, you can get data on a representative sample of pickup trucks. If your sample size is large enough, the mean weight of your sample will be roughly equal to the mean weight of the population (adhering to the law of large numbers).



Note: The central limit theorem holds true for any population. You don't need to know the shape of your population distribution in advance to apply the theorem—the distribution could be bell-

shaped, skewed, or have another shape. If you collect enough samples of sufficient size, the shape of the distribution of their means will follow a normal distribution.

Conditions

In order to apply the central limit theorem, the following conditions must be met:

- **Randomization***:* Your sample data must be the result of random selection. Random selection means that every member in the population has an equal chance of being chosen for the sample.
- **Independence***:* Your sample values must be independent of each other. Independence means that the value of one observation does not affect the value of another observation. Typically, if you know that the individuals or items in your dataset were selected randomly, you can also assume independence.
    - **10%***:* To help ensure that the condition of independence is met, your sample size should be no larger than 10% of the total population *when the sample is drawn without replacement* (which is usually the case).
        - **Note**: In general, you can sample with or without replacement. When a population element can be selected only one time, you are sampling without replacement. When a population element can be selected more than one time, you are sampling with replacement. You'll learn more about this topic later on in the course.
- **Sample size**: The sample size needs to be sufficiently large.

Let's discuss the sample size condition in more detail. There is no exact rule for how large a sample size needs to be in order for the central limit theorem to apply. The answer depends on the following factors:

- **Requirements for precision**. The larger the sample size, the more closely your sampling distribution will resemble a normal distribution, and the more precise your estimate of the population mean will be.
- **The shape of the population**. If your population distribution is roughly bell-shaped and already resembles a normal distribution, the sampling distribution of the sample mean will be close to a normal distribution even with a small sample size.

In general, many statisticians and data professionals consider a sample size of 30 to be sufficient when the population distribution is roughly bell-shaped, or approximately normal. However, if the original population is not normal—for example, if it's extremely skewed or has lots of outliers—data professionals often prefer the sample size to be a bit larger. Exploratory data analysis can help you determine how large of a sample is necessary for a given dataset.

Example: Annual salary

Let's explore an example to get a better idea of how the central limit theorem works.

Imagine you're studying annual salary data for working professionals in a large city like Buenos Aires, Cairo, Delhi, or Seoul. Let's say the professional population you're interested in includes 10 million people. You want to know the average annual salary for a professional living in the

city. However, you don’t have the time or money to survey millions of professionals to get complete data on every salary.

Instead of surveying the entire population, you collect survey data from repeated random samples of 100 professionals. Using this data, you calculate the mean annual salary in dollars for your first sample: $40,300. For your second sample, the mean salary is: $41,100. You survey a third sample. The mean salary is $39,700. And so on. Due to sampling variability, the mean of each sample will be slightly different.

In theory, you could take a very large sample and increase the sample size until you’ve surveyed all 10 million people about their salary. The central limit theorem says that as your sample size increases, the shape of your sampling distribution will increasingly resemble a bell curve.

If you take a large enough sample from the population, the mean of your sampling distribution will be roughly equal to the population mean. From this sample of the population, you can precisely estimate the average annual salary for the entire professional population.

Note: In practice, data professionals usually take a single sample. The specific sample size they choose depends on factors like budget, time, resources, and the desired level of confidence for their estimate.

Key takeaways

The central limit theorem can help you infer population parameters like the mean even if you only have available data on a portion of the population. The larger your sample size, the more precise your estimate of the population mean is likely to be. Whether you're measuring vehicle weight or annual salary, the central limit theorem is a useful method for better understanding your data.

The sampling distribution of the mean

Recently, you've learned about how data professionals use sample statistics to estimate population parameters. For example, a data professional might estimate the mean time customers spend on a retail website, or the mean salary of all the people who work in the entertainment industry.

In this reading, you'll learn more about the concept of sampling distribution and how it can help you represent the possible outcomes of a random sample. We'll also discuss how the sampling distribution of the sample mean can help you estimate the population mean.

Sampling distribution of the sample mean

A **sampling distribution** is a probability distribution of a sample statistic. Recall that a **probability distribution** represents the possible outcomes of a random variable, such as a coin toss or a die roll. In the same way, a sampling distribution represents the possible outcomes for a sample statistic. Sample statistics are based on randomly sampled data, and their outcomes cannot be predicted with certainty. You can use a sampling distribution to represent statistics such as the mean, median, standard deviation, range, and more.

Typically, data professionals compute sample statistics like the mean to estimate the corresponding population parameters.

Suppose you want to estimate the mean of a population, like the mean height of a group of humans, animals, or plants. A good way to think about the concept of sampling distribution is to imagine you take repeated samples from the population, each with the same sample size, and compute the mean for each of these samples. Due to sampling variability, the sample mean will vary from sample to sample in a way that cannot be predicted with certainty. The distribution of all your sample means is essentially the sampling distribution. You can display the distribution of sample means on a histogram. Statisticians call this the sampling distribution of the mean.

Note: In practice, due to limited time and resources, data professionals typically collect a single sample and calculate the mean of that sample to estimate the population mean.

Let's explore an example to get a more concrete idea of the sampling distribution of the mean.

Example: Mean length of lake trout

You are a data professional working with a team of environmental scientists. Your team studies the effects of water pollution on fish species. Currently, your team is researching the effects of pollution on the trout population in Lake Superior, one of the Great Lakes in North America. As part of this research, they ask you to estimate the mean length of a trout. Let's say there are 10 million trout in the lake. Instead of collecting and measuring millions of trout, you take sample data from the population.

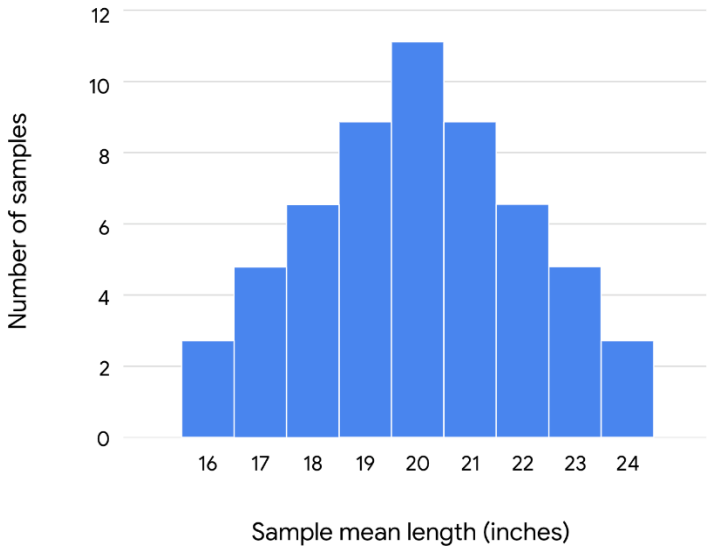
Let's say you take repeated simple random samples of 100 trout each from the population. In other words, you randomly choose 100 trout from the lake, measure them, and then repeat this

process with a different set of 100 trout. For your first sample of 100 trout, you find that the mean length is 20.2 inches. For your second sample, the mean length is 20.5 inches. For your third sample, the mean length is 19.7 inches. And so on. Due to sampling variability, the mean length will vary randomly from sample to sample.

For the purpose of this example, let's assume that the true mean length of a trout in this population is 20 inches. Although, in practice, you wouldn't know this unless you measured every single trout in the lake.

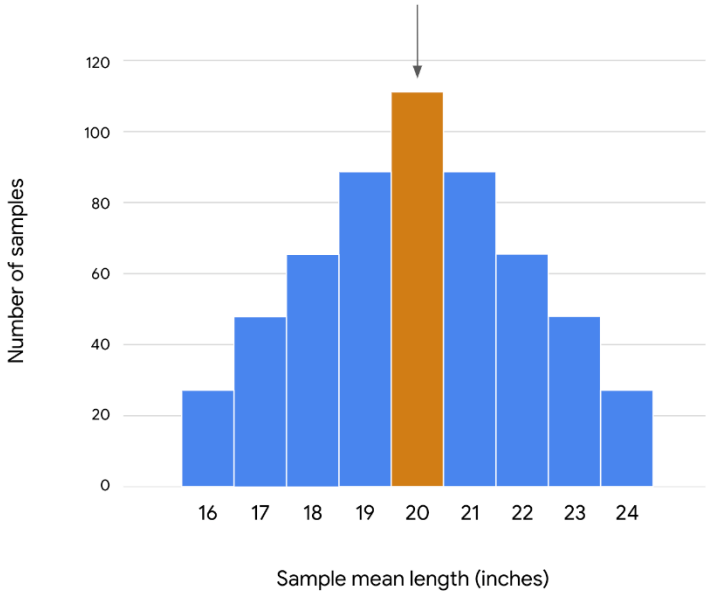
Each time you take a sample of 100 trout, it's likely that the mean length of the trout in your sample will be close to the population mean of 20 inches, but not exactly 20 inches. Every once in a while, you may get a sample full of shorter than average trout, with a mean length of 16 inches or less. Or, you might get a sample full of longer than average trout, with a mean length of 24 inches or more.

You can use a sampling distribution to represent the frequency of all your different sample means. For example, if you take 10 simple random samples of 10 trout each from the population, you can show the sampling distribution of the mean as a histogram. The most frequently occurring value in your sample data will be around 20 inches. The values that occur least frequently will be the more extreme lengths, such as 16 inches or 24 inches.



As you increase the size of a sample, the mean length of your sample data will get closer to the mean length of the population. If you sampled the entire population—in other words, if you actually measured every single trout in the lake—your sample mean would be the same as the population mean.

However, you don't need to measure millions of fish to get an accurate estimate of the population mean. If you take a large enough sample size from the population— – say, 1000 trout —– your sample mean will be a precise estimate of the population mean (20 inches).



Standard error

You can also use your sample data to estimate how precisely the mean length of any given sample represents the population mean.

This is useful to know because the sample mean varies from sample to sample, and any given sample mean is likely to differ from the true population mean. For example, the mean length of the trout population might be 20 inches. The mean length for any given sample of trout might be 20.2 inches, 20.5 inches, 19.7 inches, and so on.

Data professionals use the standard deviation of the sample means to measure this variability. In statistics, the standard deviation of a sample statistic is called the **standard error**. The standard error provides a numerical measure of sampling variability. The standard error of the mean measures variability among all your sample means. A larger standard error indicates that the sample means are more spread out, or that there's more variability. A smaller standard error indicates that the sample means are closer together, or that there's less variability.

In practice, using a single sample of observations, you can apply the following formula to calculate the estimated standard error of the sample mean: $s / \sqrt{n}$. In the formula, s refers to the sample standard deviation, and n refers to the sample size.

For example, in your study of trout lengths, imagine that a sample of 100 trout has a mean length of 20 inches and a standard deviation of 2 inches. You can calculate the estimated standard error by dividing the sample standard deviation, 2, by the square root of the sample size, 100:

$2 \div \sqrt{100} = 2 \div 10 = 0.2$

This means you should expect that the mean length from one sample to the next will vary with a standard deviation of about 0.2 inches.

The standard error helps you understand the precision of your estimate. In general, you can have more confidence in your estimates as the sample size gets larger and the standard error gets smaller. This is because, as your sample size gets larger, the sample mean gets closer to the population mean.

Key takeaways

Estimating population parameters through sampling is a powerful form of statistical inference. Sampling distributions describe the uncertainty associated with a sample statistic, and help you make proper statistical inferences. This is important because stakeholder decisions are often based on the estimates you provide.

Central Limit Theorem: The idea that the sampling distribution of the mean approaches a normal distribution as the sample size increases

Cluster random sample: A probability sampling method that divides a population into clusters, randomly selects certain clusters, and includes all members from the chosen clusters in the sample

Convenience sample: A non-probability sampling method that involves choosing members of a population that are easy to contact or reach

Descriptive statistics: A type of statistics that summarizes the main features of a dataset

Inferential statistics: A type of statistics that uses sample data to draw conclusions about a larger population

Non-probability sampling: A sampling method that is based on convenience or the personal preferences of the researcher, rather than random selection

Nonresponse bias: Refers to when certain groups of people are less likely to provide responses

Point estimate: A calculation that uses a single value to estimate a population parameter

Population: Every possible element that someone is interested in measuring

Population proportion: The percentage of individuals or elements in a population that share a certain characteristic

Probability sampling: A sampling method that uses random selection to generate a sample

Purposive sample: A non-probability sampling method that involves researchers selecting participants based on the purpose of their study

Random seed: A starting point for generating random numbers

Representative sample: A sample that accurately reflects the characteristics of a population

Sample: A subset of a population

Sample size: The number of individuals or items chosen for a study or experiment

Sampling: The process of selecting a subset of data from a population

Sampling bias: Refers to when a sample is not representative of the population as a whole

Sampling distribution: A probability distribution of a sample statistic

Sampling frame: A list of all the items in a target population

Sampling variability: Refers to how much an estimate varies between samples

Sampling with replacement: Refers to when a population element can be selected more than one time

Sampling without replacement: Refers to when a population element can be selected only one time

Simple random sample: A probability sampling method in which every member of a population is selected randomly and has an equal chance of being chosen

Snowball sample: A method of non-probability sampling that involves researchers recruiting initial participants to be in a study and then asking them to recruit other people to participate in the study

Standard error: The standard deviation of a sample statistic

Standard error of the mean: The sample standard deviation divided by the square root of the sample size

Stratified random sample: A probability sampling method that divides a population into groups and randomly selects some members from each group to be in the sample

Systematic random sample: A probability sampling method that puts every member of a population into an ordered sequence, chooses a random starting point in the sequence, and selects members for the sample at regular intervals

Target population: The complete set of elements that someone is interested in knowing more about

Undercoverage bias: Refers to when some members of a population are inadequately represented in a sample

Voluntary response sample: A method of non-probability sampling that consists of members of a population who volunteer to participate in a study

Confidence intervals: Correct and incorrect interpretations

Recently, you learned that data professionals use confidence intervals to help describe the uncertainty surrounding an estimate. To better understand your data, and effectively communicate your results to stakeholders, it's important to know how to correctly interpret a confidence interval.

In this reading, we'll review the correct way to interpret a confidence interval. We'll also discuss some common forms of misinterpretation and how to avoid them.

Correct interpretation

Example: mean weight

Let's explore an example to get a better understanding of how to interpret a confidence interval. Imagine you want to estimate the mean weight of a population of 10,000 penguins. Instead of weighing every single penguin, you select a sample of 100 penguins. The mean weight of your sample is 30 pounds. Based on your sample data, you construct a 95% confidence interval between 28 pounds and 32 pounds.

95 CI [28, 32]

Interpret the confidence interval

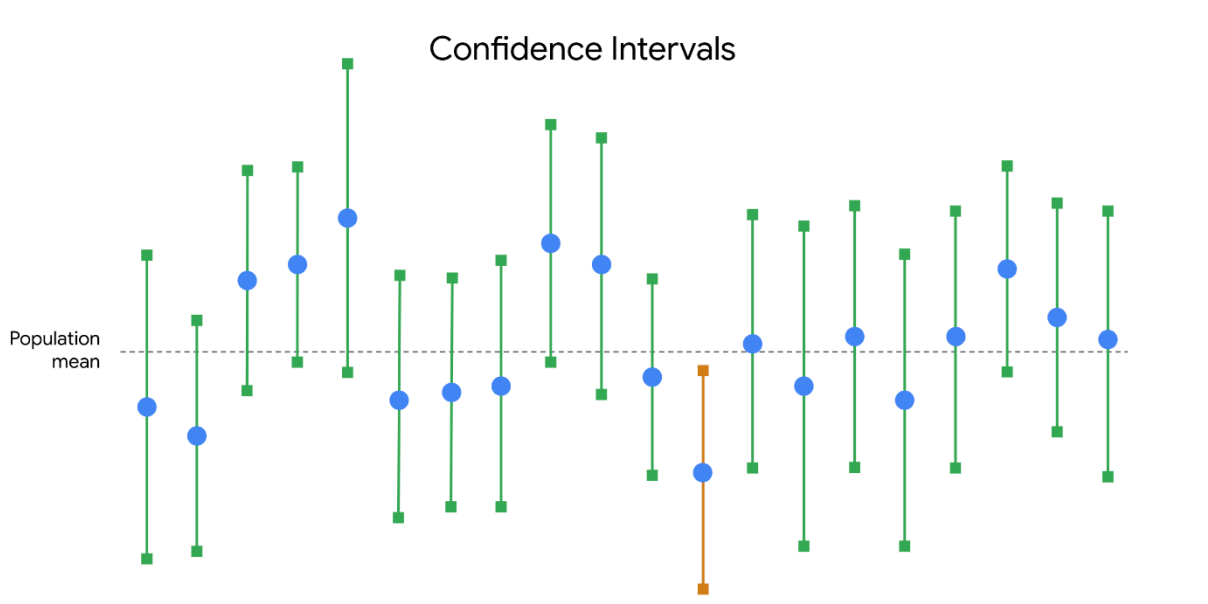
Earlier, you learned that the confidence level expresses the uncertainty of the estimation process. Let's discuss what 95% confidence means from a more technical perspective.

Technically, 95% confidence means that if you take repeated random samples from a population, and construct a confidence interval for each sample using the same method, you can expect that 95% of these intervals will capture the population mean. You can also expect that 5% of the total will *not* capture the population mean.

The confidence level refers to the long-term success rate of the **method**, or the estimation process based on random sampling.

For the purpose of our example, let's imagine that the mean weight of all 10,000 penguins is 31 pounds, although you wouldn't know this unless you actually weighed every penguin. So, you take a sample of the population.

Imagine you take 20 random samples of 100 penguins each from the penguin population, and calculate a 95% confidence interval for each sample. You can expect that approximately 19 of the 20 intervals, or 95% of the total, will contain the actual population mean weight of 31 pounds. One such interval will be the range of values between 28 pounds and 32 pounds.



In practice, data professionals usually select one random sample and generate one confidence interval, which may or may not contain the actual population mean. This is because repeated random sampling is often difficult, expensive, and time-consuming. Confidence intervals give data professionals a way to quantify the uncertainty due to random sampling.

Incorrect interpretations

Now that you have a better understanding of how to properly interpret a confidence interval, let's review some common misinterpretations and how to avoid them.

Misinterpretation 1: 95% refers to the probability that the population mean falls within the constructed interval

One incorrect statement that is often made about a confidence interval at a 95% level of confidence is that there is a 95% probability that the population mean falls within the constructed interval.

In our example, this would mean that there's a 95% chance that the mean weight of the penguin population falls in the interval between 28 pounds and 32 pounds.

This is incorrect. The population mean is a constant.

Like any population parameter, the population mean is a constant, not a random variable. While the value of the sample mean varies from sample to sample, the value of the population mean does not change. The probability that a constant falls within any given range of values is always 0% or 100%. It either falls within the range of values, or it doesn't.

For example, any given random sample of 100 penguins may have a different mean weight: 32.8 pounds, 27.3 pounds, 29.6 pounds, and so on. You can use a sampling distribution to assign a specific probability to each of your sample means because these are random variables. However,

the population mean weight is considered a constant. In our example, if you weigh all 10,000 penguins, you'll find that the population mean is 31 pounds. This value is fixed, and does not vary from sample to sample.

Sample Mean (100 penguins)	Population Mean (10,000 penguins)
32.8 lbs	31 lbs
27.3 lbs	31 lbs
29.6 lbs	31 lbs

So, it's not strictly correct to say there is a 95% chance that your confidence interval captures the population mean because this implies that the population mean is variable. Intervals change from sample to sample, but the value of the population mean you're trying to capture does not.

What you can say is that if you take repeated random samples from the population, and construct a confidence interval for each sample using the same method, you can expect 95% of your intervals to capture the population mean.

Pro tip: Remember that a 95% confidence level refers to the success rate of the estimation process.

Misinterpretation 2: 95% refers to the percentage of data values that fall within the interval

Another common mistake is to interpret a 95% confidence interval as saying that 95% of all of the data values in the population fall within the interval. This is not necessarily true. A 95% confidence interval shows a range of values that likely includes the actual population mean. This is *not* the same as a range that contains 95% of the data values in the population.

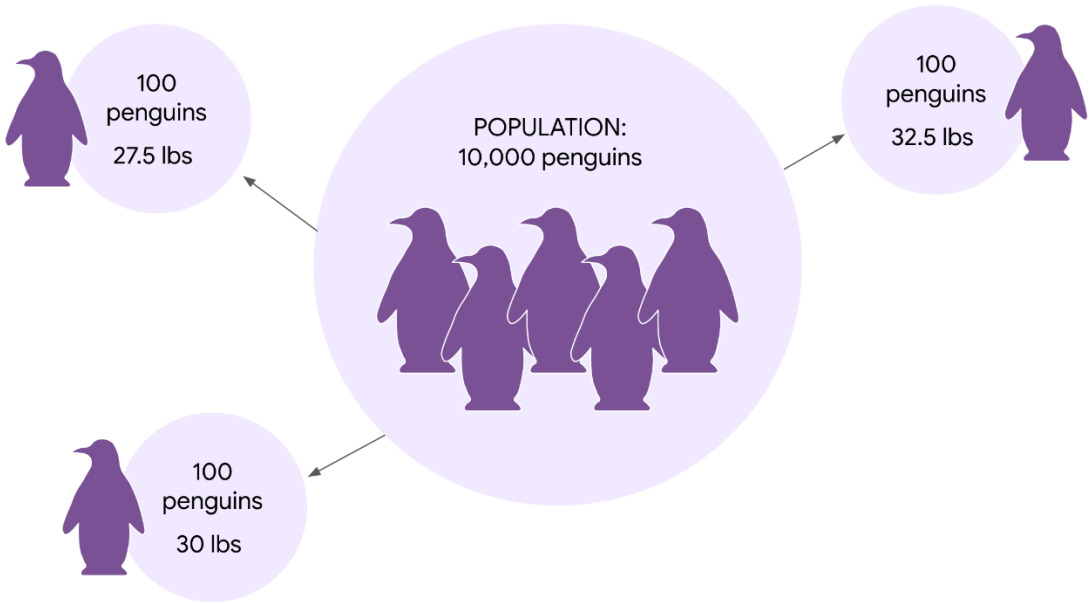
For example, your 95% confidence interval for the mean penguin weight is between 28 pounds and 32 pounds. It may not be accurate to say that 95% of all weight values fall within this interval. It's possible that over 5% of the penguin weights in the population are outside this interval—either less than 28 pounds or greater than 32 pounds.

95% CI [28, 32]

Penguin weight
26.9 lbs
27.7 lbs
28.5 lbs
29.9 lbs
30.6 lbs
31.1 lbs
32.3 lbs
33.4 lbs

Misinterpretation 3: 95% refers to the percentage of sample means that fall within the interval

A third common misinterpretation is that a 95% confidence interval implies that 95% of all possible sample *means* fall within the range of the interval. This is not necessarily true. For example, your 95% confidence interval for mean penguin weight is between 28 pounds and 32 pounds. Imagine you take repeated samples of 100 penguins and calculate the mean weight for each sample. It's possible that *over* 5% of your sample means will be less than 28 pounds or greater than 32 pounds.



Key takeaways

Knowing how to correctly interpret confidence intervals will give you a better understanding of your estimate, and help you share useful and accurate information with stakeholders. You may need to explain the common misinterpretations too, and why they're incorrect. You don't want your stakeholders to base their decisions on a misinterpretation. Understanding how to effectively communicate your results to stakeholders is a key part of your job as a data professional.

Resources for more information

To learn more about interpreting confidence intervals, check out the following resources:

- This article from ThoughtCo. contains a useful list of common mistakes that are made when interpreting confidence intervals.

Construct a confidence interval for a small sample size

So far, you've constructed confidence intervals for large sample sizes, which are usually defined as sample sizes of 30 or more items. For example, when you estimated the mean battery life of a new cell phone, you used a random sample of 100 phones. On the other hand, small sample sizes are usually defined as having fewer than 30 items. Typically, data professionals try to work with large sample sizes because they give more precise estimates. But, it's not always possible to work with a large sample. In practice, collecting data is often expensive and time-consuming. If you don't have the time, money, or resources to take a large sample, you may end up working with a small sample.

In this reading, you'll learn how to construct a confidence interval for a small sample size. We'll go step-by-step through an example involving mean emission levels for a new car engine.

Large versus small sample sizes

First, let's briefly discuss the different methods you use to construct confidence intervals for large and small sample sizes.

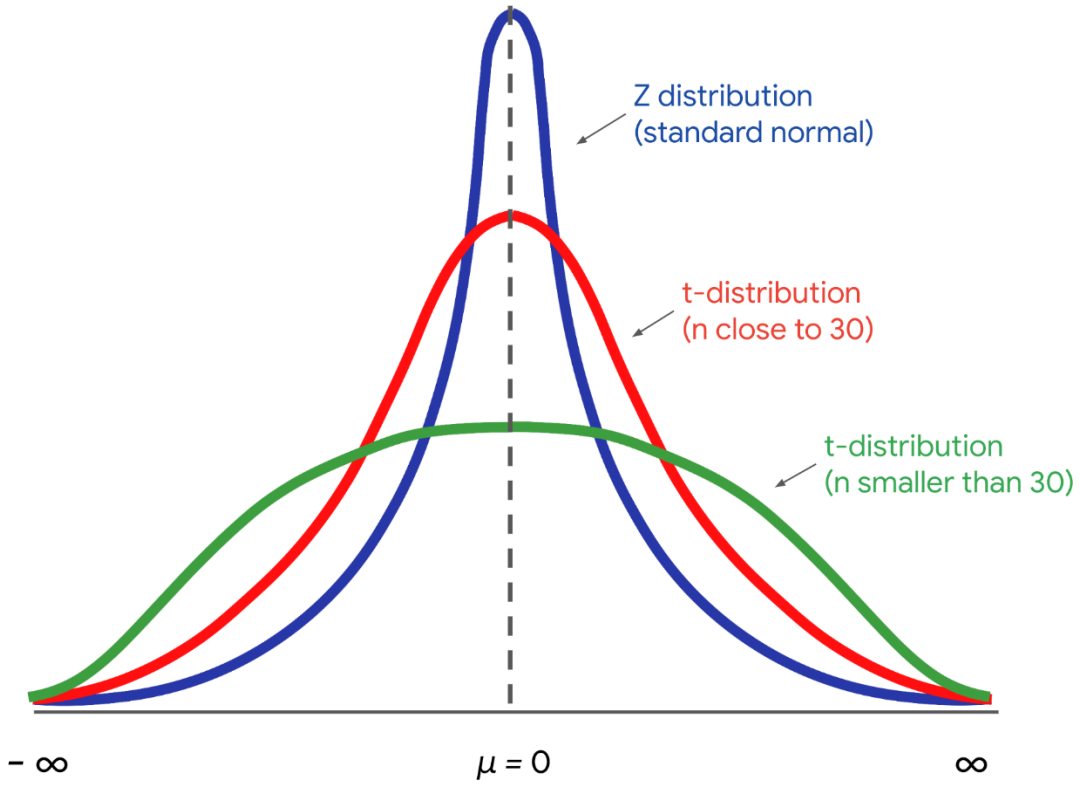
Large sample: Z-scores

For large sample sizes, you use **z-scores** to calculate the margin of error, just like you did earlier to estimate mean battery life for cell phones. This is because of the central limit theorem: for large sample sizes, the sample mean is approximately normally distributed. For a standard normal distribution, also called a **z-distribution**, you use z-scores to make calculations about your data.

Small sample: T-scores

For small sample sizes, you need to use a different distribution, called the **t-distribution**. Statistically speaking, this is because there is more uncertainty involved in estimating the standard error for small sample sizes. You don't need to worry about the technical details, which are beyond the scope of this course. For now, just know that if you're working with a small sample size, and your data is approximately normally distributed, you should use the t-distribution rather than the standard normal distribution. For a t-distribution, you use t-scores to make calculations about your data.

The graph of the t-distribution has a bell shape that is similar to the standard normal distribution. But, the t-distribution has bigger tails than the standard normal distribution does. The bigger tails indicate the higher frequency of outliers that come with a small dataset. As the sample size increases, the t-distribution approaches the normal distribution. When the sample size reaches 30, the distributions are practically the same, and you can use the normal distribution for your calculations.



Example: Mean emission levels

Now that you know a little bit about the t-distribution and t-scores, let's construct a confidence interval for a small sample size.

Context

Imagine you're a data professional working for an auto manufacturer. The company produces high performance cars that are sold around the world. Typically, the engines in these cars have high emission rates of carbon dioxide, or CO_2, which is a greenhouse gas that contributes to global warming. The engineering team has designed a new engine to reduce emissions for the company's best-selling car.

Goal

The goal is to keep emissions below 460 grams of CO_2 per mile. This will ensure the car meets emissions standards in every country it's sold in. Plus, the lower emissions rate is good for the environment, which will appeal to new customers.

Ask

The engineering team asks you to provide a reliable estimate of the emissions rate for the new engine. Due to production issues, there are only a limited number of engines available for testing. So, you'll be working with a small sample size.

Sample

The engineering team tests a random sample of 15 engines and collects data on their emissions. The mean emission rate is 430 grams of CO_2 per mile, and the standard deviation is 35 grams of CO_2 per mile.

Your single sample may not provide the actual mean emissions rate for *every* engine. The population mean for emissions could be above or below 430 grams of CO_2 per mile. Even though you only have a small sample of engines, you can construct a confidence interval that likely includes the actual emission rate for a large population of engines. This will give your manager a better idea of the uncertainty in your estimate. It will also help the engineering team decide if they need to do more work on the engine to lower the emissions rate.

Construct the confidence interval

Let's review the steps for constructing a confidence interval:

1. Identify a sample statistic.
2. Choose a confidence level.
3. Find the margin of error.
4. Calculate the interval.

Step 1: Identify a sample statistic

First, identify your sample statistic. Your sample represents the average emissions rate for 15 engines. You're working with a sample *mean*.

Step 2: Choose a confidence level

Next, choose a confidence level. The engineering team requests that you choose a 95% confidence level.

Step 3: Find the margin of error

Your third step is to find the margin of error. For a small sample size, you calculate the margin of error by multiplying the t-score by the standard error.

The t-distribution is defined by a parameter called the degree of freedom. In our context, the degree of freedom is the sample size - 1, or 15-1 = 14. Given your degree of freedom and your confidence level, you can use a programming language like Python or other statistical software to calculate your t-score.

Based on a degree of freedom of 14, and a confidence level of 95%, your t-score is 2.145.

Now you can calculate the standard error, which measures the variability of your sample statistic.

Here's the formula for the standard error of the mean that you've used before:

Standard Error (Means)

$$SE(x) = s/\sqrt{(n)}$$

In the formula, the letter s refers to sample standard deviation, and the letter n refers to sample size.

Your sample standard deviation is 35, and your sample size is 15. The calculation gives you a standard error of about 9.04.

The margin of error is your t-score multiplied by your standard error. This is 2.145 * 9.04 = 19.39.

Step 4: Calculate the interval

Finally, calculate your confidence interval. The upper limit of your interval is the sample mean plus the margin of error. This is 430 + 19.39 = 449.39 grams of CO_2 per mile.

The lower limit is the sample mean minus the margin of error. This is 430 − 19.39 = 410.61 grams of CO_2 per mile.

You have a 95% confidence interval that stretches from 410.61 grams of CO_2 per mile to 449.39 grams of CO_2 per mile.

95 CI [410.61, 449.39]

The confidence interval gives the engineering team important information. The upper limit of your interval is below the target of 460 grams of CO_2 per mile. This result provides solid statistical evidence that the emissions rate for the new engine will meet emissions standards.

Note: Confidence intervals for small sample sizes only deal with population means, and not population proportions. The statistical reason for this distinction is rather technical, so you don't need to worry about it for now.

Key takeaways

As a data professional, you'll work with both large and small sample sizes. Although large samples give more precise estimates than small samples, collecting a small sample is usually less expensive and time-consuming than collecting a large sample. Knowing how to construct confidence intervals for different sample sizes will help you manage any dataset that you may encounter in your future career.

Resources for more information

To learn more about constructing confidence intervals for different sample sizes, refer to the following resources:

- This article from Scribbr contains a useful overview of the t-distribution and how to use it when constructing a confidence interval.

Terms and definitions from Course 4, Module 4

Confidence interval: A range of values that describes the uncertainty surrounding an estimate

Confidence level: A measure that expresses the uncertainty of the estimation process **Interval**: A sample statistic plus or minus the margin of error

Interval estimate: A calculation that uses a range of values to estimate a population parameter

Lower limit: When constructing an interval, the calculation of the sample means minus the margin of error

Margin of error: The maximum expected difference between a population parameter and a sample estimate

Method: A function that defines and performs behaviors like computation

Point estimate: A calculation that uses a single value to estimate a population parameter

Standard error of the mean: The sample standard deviation divided by the square root of the sample size

Standard error of the proportion: The square root of the sample proportion times one minus the sample proportion divided by the sample size

Upper limit: When constructing an interval, the calculation of the sample means plus the margin of error

Differences between the null and alternative hypotheses

Recently, you learned that **hypothesis testing** uses sample data to evaluate an assumption about a population parameter. Data professionals conduct a hypothesis test to decide whether the evidence from their sample data supports either the null hypothesis or the alternative hypothesis.

In this reading, we'll go over the main differences between the null hypothesis and the alternative hypothesis, and how to formulate each hypothesis in different scenarios.

Statistical hypotheses

Let's review the steps for conducting a hypothesis test:

1. State the null hypothesis and the alternative hypothesis.
2. Choose a significance level.
3. Find the p-value.
4. Reject or fail to reject the null hypothesis.

The first step for any hypothesis test is to state the null and alternative hypotheses. The null and alternative hypotheses are mutually exclusive, meaning they cannot both be true at the same time.

The **null hypothesis** is a statement that is assumed to be true unless there is convincing evidence to the contrary. The null hypothesis typically assumes that there is no effect in the population, and that your observed data occurs by chance.

The **alternative hypothesis** is a statement that contradicts the null hypothesis, and is accepted as true only if there is convincing evidence for it. The alternative hypothesis typically assumes that there is an effect in the population, and that your observed data does *not* occur by chance.

Note: The null and alternative hypotheses are always claims about the population. That's because the aim of hypothesis testing is to make inferences about a population based on a sample.

For example, imagine you're a data professional working for a car dealership. The company implements a new sales training program for their employees. They ask you to evaluate the effectiveness of the program.

- Your **null hypothesis (H_0):** the program had no effect on sales revenue.
- Your **alternative hypothesis (H_a)**: the program increased sales revenue.

Let's explore each hypothesis in more detail.

Null hypothesis

The null hypothesis has the following characteristics:

- In statistics, the null hypothesis is often abbreviated as H sub zero (H_0).
- When written in mathematical terms, the null hypothesis always includes an equality symbol (usually =, but sometimes ≤ or ≥).
- Null hypotheses often include phrases such as "no effect," "no difference," "no relationship," or "no change."

Alternative hypothesis

The alternative hypothesis has the following characteristics:

- In statistics, the alternative hypothesis is often abbreviated as H sub a (H_a).
- When written in mathematical terms, the alternative hypothesis always includes an inequality symbol (usually $\neq$, but sometimes $<$ or $>$).
- Alternative hypotheses often include phrases such as "an effect," "a difference," "a relationship," or "a change."

Example scenarios

Typically, the null hypothesis represents the *status quo*, or the current state of things. The null hypothesis assumes that the status quo hasn't changed. The alternative hypothesis suggests a new possibility or different explanation. Let's check out some examples to get a better idea of how to write the null and alternative hypotheses for different scenarios:

Example#1: Mean weight

An organic food company is famous for their granola. The company claims each bag they produce contains 300 grams of granola—no more and no less. To test this claim, a quality control expert measures the weight of a random sample of 40 bags.

- **H_0**: $\mu = 300$ (the mean weight of all produced granola bags is equal to 300 grams)
- **H_a:** $\mu \neq 300$ (the mean weight of all produced granola bags is not equal to 300 grams)

Example#2: Mean height

Suppose it's assumed that the mean height of a certain species of tree is 30 feet tall. However, one ecologist claims the actual mean height is greater than 30 feet. To test this claim, the ecologist measures the height of a random sample of 50 trees.

- **H_0**: $\mu \leq 30$ (the mean height of this species of tree is equal to or less than 30 feet)
- **H_a:** $\mu > 30$ (the mean height of this species of tree is greater than 30 feet)

Example#3: Proportion of employees

A corporation claims that at least 80% of all employees are satisfied with their job. However, an independent researcher believes that less than 80% of all employees are satisfied with their job. To test this claim, the researcher surveys a random sample of 100 employees.

- **H_0**: $p \geq 0.80$ (the proportion of all employees who are satisfied with their job is equal to or greater than 80%)
- **H_a:** $p < 0.80$ (the proportion of all employees who are satisfied with their job is less than 80%)

Summary: Null versus alternative

The following table summarizes some important differences between the null and alternative hypotheses:

	Null hypothesis (H_0)	**Alternative hypothesis (H_a)**
Claims	There is no effect in the population.	There is an effect in the population.
Language	• No effect • No difference	• An effect • A difference

	Null hypothesis (H_0)	Alternative hypothesis (H_a)
	• No relationship • No change	• A relationship • A change
Symbols	Equality (=, ≤, ≥)	Inequality (≠, <, >)

Key takeaways

The null hypothesis and the alternative hypothesis are foundational concepts in hypothesis testing. To conduct an effective hypothesis test, it's important to understand the differences between the null and alternative hypotheses, and how to properly state each hypothesis.

Resources for more information

To learn more about the null hypothesis and the alternative hypothesis, refer to the following resources:

- This article from Statistics How To features a detailed discussion of the null hypothesis.

Type I and type II errors

Earlier, you learned that you can use a hypothesis test to help determine if your results are statistically significant, or if they occurred by chance. However, because hypothesis testing is based on probability, there's always a chance of drawing the wrong conclusion about the null hypothesis. In hypothesis testing, there are two types of errors you can make when drawing a conclusion: a Type I error and a Type II error.

In this reading, we'll discuss the difference between Type I and Type II errors, and the risks involved in making each error.

Errors in statistical decision-making

Let's review the steps for conducting a hypothesis test:

1. State the null hypothesis and the alternative hypothesis.
2. Choose a significance level.
3. Find the p-value.
4. Reject or fail to reject the null hypothesis.

When you decide to reject or fail to reject the null hypothesis, there are four possible outcomes–two represent correct choices, and two represent errors. You can:

- Reject the null hypothesis when it's actually true (**Type I error)**
- Reject the null hypothesis when it's actually false (Correct)
- Fail to reject the null hypothesis when it's actually true (Correct)
- Fail to reject the null hypothesis when it's actually false (**Type II error**)

	Null Hypothesis is TRUE	Null Hypothesis is FALSE
Reject null hypothesis	Type I Error (False positive)	Correct Outcome! (True positive)
Fail to reject null hypothesis	Correct Outcome! (True negative)	Type II Error (False negative)

Example: Clinical trial

Let's explore an example to get a better understanding of Type I and Type II errors. Hypothesis tests are often used in clinical trials to determine whether a new medicine leads to better outcomes in patients. Imagine you're a data professional who works for a pharmaceutical company. The company invents a new medicine to treat the common cold. The company tests a random sample of 200 people with cold symptoms. Without medicine, the typical person experiences cold symptoms for 7.5 days. The average recovery time for people who take the medicine is 6.2 days.

You conduct a hypothesis test to determine if the effect of the medicine on recovery time is statistically significant, or due to chance.

In this case:

- Your **null hypothesis** (H_0) is that the medicine has no effect.
- Your **alternative hypothesis** (H_a) is that the medicine is effective.

Type I error

A **Type 1 error**, also known as a false positive, occurs when you reject a null hypothesis that is actually true. In other words, you conclude that your result is statistically significant when in fact it occurred by chance.

For example, in your clinical trial, if the null hypothesis is true, that means the medicine has no effect. If you make a Type I error and reject the null hypothesis, you incorrectly conclude that the medicine relieves cold symptoms when it's actually ineffective.

The probability of making a Type I error is called alpha (α). Your significance level, or alpha (α), represents the probability of making a Type I error. Typically, the significance level is set at 0.05, or 5%. A significance level of 5% means you are willing to accept a 5% chance you are wrong when you reject the null hypothesis.

Reduce your risk

To reduce your chance of making a Type I error, choose a lower significance level.

For instance, if you want to minimize the risk of a Type I error, you can choose a significance level of 1% instead of the standard 5%. This change reduces the chance of making a Type I error from 5% to 1%.

Significance level (α)	**Chance of making Type I error**
0.05	5%
0.01	1%

Type II error

However, reducing your risk of making a Type I error means you are more likely to make a Type II error, or false negative. A **Type II error** occurs when you fail to reject a null hypothesis which is actually false. In other words, you conclude your result occurred by chance, when in fact it didn't.

For example, in your clinical study, if the null hypothesis is false, this means that the medicine is effective. If you make a Type II error and fail to reject the null hypothesis, you incorrectly conclude that the medicine is ineffective when it actually relieves cold symptoms.

The probability of making a Type II error is called beta (β), and beta is related to the power of a hypothesis test (power = 1- β). Power refers to the likelihood that a test can correctly detect a real effect when there is one.

Reduce your risk

You can reduce your risk of making a Type II error by ensuring your test has enough power. In data work, power is usually set at 0.80 or 80%. The higher the statistical power, the lower the probability of making a Type II error. To increase power, you can increase your sample size or your significance level.

Note: A detailed discussion of the concept of statistical power is beyond the scope of this course. Power is something you'll learn more about as you advance in your career as a data professional and grow your knowledge of statistics.

Potential risks of Type I and Type II errors

As a data professional, it's important to be aware of the potential risks involved in making the two types or errors.

A Type I error means rejecting a null hypothesis which is actually true. In general, making a Type I error often leads to implementing changes that are unnecessary and ineffective, and which waste valuable time and resources.

For example, if you make a Type I error in your clinical trial, the new medicine will be considered effective even though it's actually ineffective. Based on this incorrect conclusion, an ineffective medication may be prescribed to a large number of people. Plus, other treatment options may be rejected in favor of the new medicine.

A Type II error means failing to reject a null hypothesis which is actually false. In general, making a Type II error may result in missed opportunities for positive change and innovation. A lack of innovation can be costly for people and organizations.

For example, if you make a Type II error in your clinical trial, the new medicine will be considered ineffective even though it's actually effective. This means that a useful medication may not reach a large number of people who could benefit from it.

Key takeaways

As a data professional, it helps to be aware of the potential errors built into hypothesis testing and how they can affect your results. Depending on the specific situation, you may choose to minimize the risk of either a Type I or Type II error. Ultimately, it's your responsibility as a data professional to determine which type of error is riskier based on the goals of your analysis.

Resources for more information

To learn more about Type I and Type II errors, refer to the following resources:

- This article from Simply Psychology contains a helpful summary of the differences between Type I and Type II errors.

Determine if data has statistical significance

Recently, you learned that **statistical significance** is the claim that the results of a test or experiment are not explainable by chance alone. A hypothesis test can help you determine whether your observed data is statistically significant, or likely due to chance. For example, in a clinical trial of a new medication, a hypothesis test can help determine if the medication's positive effect on a sample group is statistically significant, or due to chance.

In this reading, you'll learn more about the concept of statistical significance and its role in hypothesis testing.

Statistical significance in hypothesis testing

Data professionals use hypothesis testing to determine whether a relationship between variables or a difference between groups is statistically significant.

Let's explore an example to get a better understanding of the role of statistical significance in hypothesis testing.

Example: Mean battery life

Let's review the steps for conducting a hypothesis test:

1. State the null hypothesis and the alternative hypothesis.
2. Choose a significance level.
3. Find the p-value.
4. Reject or fail to reject the null hypothesis.

Imagine you're a data professional working for a computer company. The company claims the mean battery life for their best selling laptop is 8.5 hours with a standard deviation of 0.5 hours. Recently, the engineering team redesigned the laptop to increase the battery life. The team takes a random sample of 40 redesigned laptops. The sample mean is 8.7 hours.

The team asks you to determine if the increase in mean battery life is statistically significant, or if it's due to random chance. You decide to conduct a z-test to find out.

Step 1: State the null hypothesis and alternative hypothesis

The null hypothesis typically assumes that your observed data occurs by chance, and it is not statistically significant. In this case, your null hypothesis says that there is no actual effect on mean battery life in the population of laptops.

The alternative hypothesis typically assumes that your observed data does *not* occur by chance, and is statistically significant. In this case, your alternative hypothesis says that there is an effect on mean battery life in the population of laptops.

In this example, you formulate the following hypotheses:

- $\mathbf{H_0}$: $\mu = 8.5$ (the mean battery life of all redesigned laptops is equal to 8.5 hours)
- $\mathbf{H_a}$**:** $\mu > 8.5$ (the mean battery life of all redesigned laptops is greater than 8.5 hours)

Step 2: Choose a significance level

The **significance level**, or alpha (α), is the threshold at which you will consider a result statistically significant. The significance level is also the probability of rejecting the null hypothesis when it is true.

Typically, data professionals set the significance level at 0.05, or 5%. That means results at least as extreme as yours only have a 5% chance (or less) of occurring when the null hypothesis is true.

Note: 5% is a conventional choice, and not a magical number. It's based on tradition in statistical research and education. Other common choices are 1% and 10%. You can adjust the significance level to meet the specific requirements of your analysis. A lower significance level means an effect has to be larger to be considered statistically significant.

Pro tip: As a best practice, you should set a significance level before you begin your test. Otherwise, you might end up in a situation where you are manipulating the results to suit your convenience.

In this example, you choose a significance level of 5%, which is the company's standard for research.

Step 3: Find the p-value

P-value refers to the probability of observing results as or more extreme than those observed when the null hypothesis is true.

Your p-value helps you determine whether a result is statistically significant. A low p-value indicates high statistical significance, while a high p-value indicates low or no statistical significance.

Every hypothesis test features:

- A test statistic that indicates how closely your data match the null hypothesis. For a z-test, your test statistic is a z-score; for a t-test, it's a t-score.
- A corresponding p-value that tells you the probability of obtaining a result at least as extreme as the observed result if the null hypothesis is true.

As a data professional, you'll almost always calculate p-value on your computer, using a programming language like Python or other statistical software. In this example, you're conducting a z-test, so your test statistic is a z-score of 2.53. Based on this test statistic, you calculate a p-value of 0.0057, or 0.57%.

Step 4: Reject or fail to reject the null hypothesis

In a hypothesis test, you compare your p-value to your significance level to decide whether your results are statistically significant.

There are two main rules for drawing a conclusion about a hypothesis test:

- If your p-value is less than your significance level, you reject the null hypothesis.
- If your p-value is greater than your significance level, you fail to reject the null hypothesis.

Note: Data professionals and statisticians always say "fail to reject" rather than "accept." This is because hypothesis tests are based on probability, not certainty—acceptance implies certainty. In general, data professionals avoid claiming certainty about results based on statistical methods.

In this example, your p-value of 0.57% is less than your significance level of 5%. Your test provides sufficient evidence to conclude that the mean battery life of all redesigned laptops has increased from 8.5 hours. You reject the null hypothesis. You determine that your results are statistically significant.

Key takeaways

As a data professional, it's important to understand the concept of statistical significance to effectively conduct a hypothesis test and interpret the results. Insights based on statistically significant results can help stakeholders make more informed business decisions.

Resources for more information

To learn more about statistical significance, refer to the following resources:

- This article from Scribbr provides a helpful overview of statistical significance and discusses some critiques of the concept in contemporary research.

One-tailed and two-tailed tests

Earlier, you learned that a hypothesis test can be either one-tailed or two-tailed. A tail in hypothesis testing refers to the tail at either end of a distribution curve.

In this reading, we'll go over the main differences between one-tailed and two-tailed tests, and discuss the procedure for conducting each test.

One-tailed and two-tailed tests

First, let's discuss the differences between one-tailed and two-tailed tests.

A **one-tailed test** results when the alternative hypothesis states that the actual value of a population parameter is either less than or greater than the value in the null hypothesis.

A one-tailed test may be either left-tailed or right-tailed. A left-tailed test results when the alternative hypothesis states that the actual value of the parameter is less than the value in the null hypothesis. A right-tailed test results when the alternative hypothesis states that the actual value of the parameter is greater than the value in the null hypothesis.

A **two-tailed test** results when the alternative hypothesis states that the actual value of the parameter does not equal the value in the null hypothesis.

For example, imagine a test in which the null hypothesis states that the mean weight of a penguin population equals 30 lbs.

- In a left-tailed test, the alternative hypothesis might state that the mean weight of the penguin population is less than ("<") 30 lbs.
- In a right-tailed test, the alternative hypothesis might state that the mean weight of the penguin population is greater than (">") 30 lbs.
- In a two-tailed test, the alternative hypothesis might state that the mean weight of the penguin population is not equal ("≠") to 30 lbs.

Let's explore a more detailed example to get a better understanding of the difference between one-tailed and two-tailed tests.

Example: One-tailed tests

Imagine you're a data professional working for an online retail company. The company claims that *at least* 80% of its customers are satisfied with their shopping experience. You survey a random sample of 100 customers. According to the survey, 73% of customers say they are satisfied. Based on the survey data, you conduct a z-test to evaluate the claim that *at least* 80% of customers are satisfied.

Let's review the steps for conducting a hypothesis test:

1. State the null hypothesis and the alternative hypothesis.
2. Choose a significance level.
3. Find the p-value.
4. Reject or fail to reject the null hypothesis.

First, you state the null and alternative hypotheses:

- $\mathbf{H_0}$: P >= 0.80 (the proportion of satisfied customers is greater than or equal to 80%)
- $\mathbf{H_a}$: P < 0.80 (the proportion of satisfied customers is less than 80%)

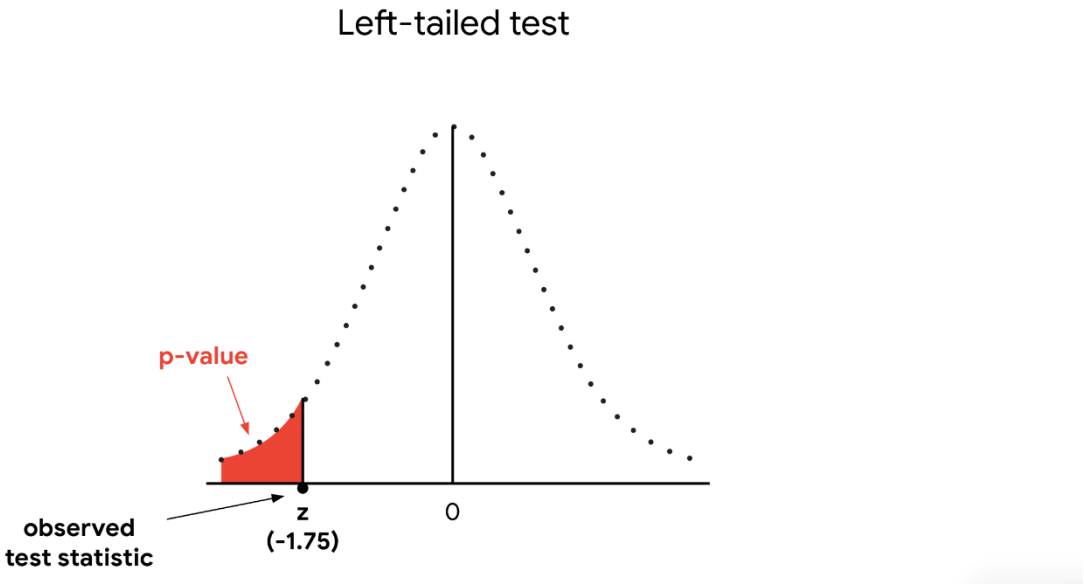
Note: This is a one-tailed test as the alternative hypothesis contains the less than sign ("<").

Next, you choose a significance level of 0.05, or 5%.

Then, you calculate your p-value based on your test statistic. Recall that p-value is the probability of observing results as or more extreme than those observed when the null hypothesis is true**.** In the context of hypothesis testing, "extreme" means extreme in the direction(s) of the alternative hypothesis.

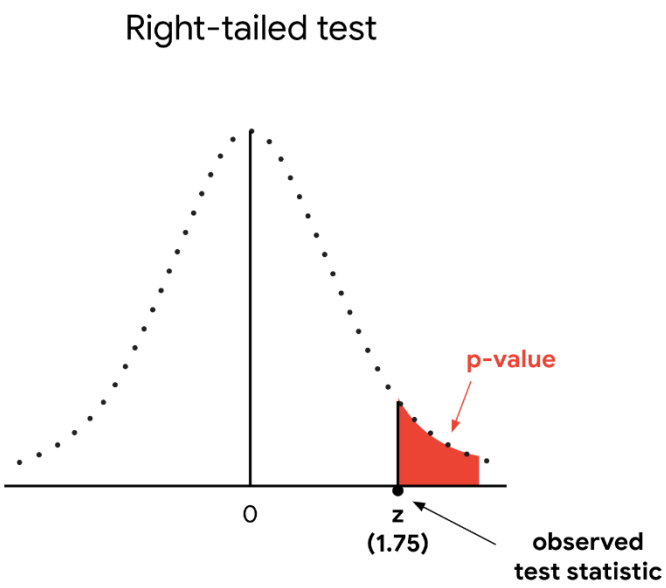
Your test statistic is a z-score of 1.75 and your p-value is 0.04.

Since this is a left-tailed test, the p-value is the probability that the z-score is less than 1.75 standard units away from the mean to the left. In other words, it's the probability that the z-score is less than -1.75. The probability of getting a value less than your z-score of -1.75 is calculated by taking the area under the distribution curve to the left of the z-score. This is called a left-tailed test, because your p-value is located on the left tail of the distribution. The area under this part of the curve is the same as your p-value: 0.04.



Finally, you draw a conclusion. Since your p-value of 0.04 is less than your significance level of 0.05, you *reject* the null hypothesis.

Note: In a different testing scenario, your test statistic might be positive 1.75, and you might be interested in values as great or greater than the z-score 1.75. In that case, your p-value would be located on the right tail of the distribution, and you'd be conducting a right-tailed test.



Example: Two-tailed tests

Now, imagine our previous example has a slightly different set up. Suppose the company claims that 80% of its customers are satisfied with their shopping experience. To test this claim, you survey a random sample of 100 customers. According to the survey, 73% of customers say they are satisfied. Based on the survey data, you conduct a z-test to evaluate the claim that 80% of customers are satisfied.

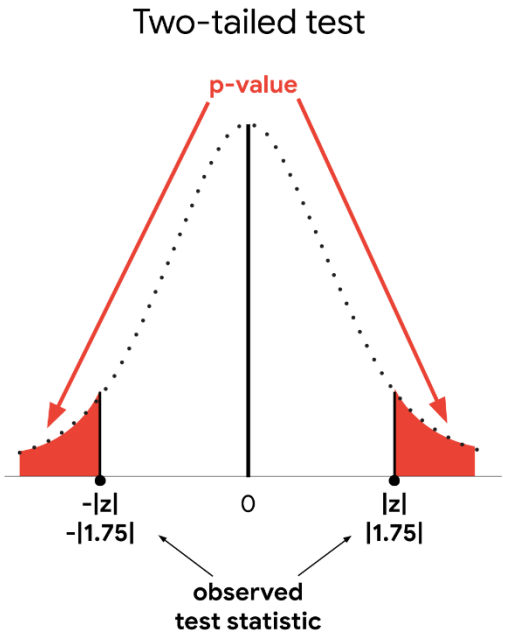
First, you state the null and alternative hypotheses:

- $\mathbf{H_0}$: P = 0.80 (the proportion of satisfied customers equals 80%)
- $\mathbf{H_a}$: P ≠ 0.80 (the proportion of satisfied customers does not equal 80%)

Note: This is a two-tailed test as the alternative hypothesis contains the not equal sign ("≠").

Next, you choose a significance level of 0.05, or 5%.

Then, you calculate your p-value based on your test statistic. Your test statistic is a z-score of 1.75. Since this is a two-tailed test, the p-value is the probability that the z-score is less than -1.75 *or* greater than 1.75. Note that the *p*-value for a two-tailed test is always two times the *p*-value for a one-tailed test. So, in this case, your p-value = 0.04 + 0.04 = 0.08. In a two-tailed test, your p-value corresponds to the area under the curve on *both* the left tail and right tail of the distribution.



Finally, you draw a conclusion. Since your p-value of 0.08 is greater than your significance level of 0.05, you **fail to reject** the null hypothesis.

One-tailed versus two-tailed

You can use one-tailed and two-tailed tests to examine different effects.

In general, a one-tailed test may provide more power to detect an effect in a single direction. However, before conducting a one-tailed test, you should consider the consequences of missing an effect in the other direction. For example, imagine a pharmaceutical company develops a new medication they believe is more effective than an existing medication. As a data professional analyzing the results of the clinical trial, you may wish to choose a one-tailed test to maximize your ability to detect the improvement. In doing so, you fail to test for the possibility that the new medication is less effective than the existing medication. And, of course, the company doesn't want to release a less effective medication to the public.

A one-tailed test may be appropriate if the negative consequences of missing an effect in the untested direction are minimal. For example, imagine that the company develops a new, less expensive medication that they believe is at least as effective as the existing medication. The lower price gives the new medication an advantage in the market. So, they just want to make sure the new medication is not *less* effective than the existing medication. Testing whether it's *more* effective is not a priority. In this case, a one-tailed test may be appropriate.

Key takeaways

Understanding the differences between a one-tailed and a two-tailed test is an important part of conducting a hypothesis test. Depending on the context of your analysis, you may want to use a one-tailed test to examine effects in a single direction, or a two-tailed test to examine effects in both directions.

A/B testing

Earlier, you learned that A/B testing is a way to compare two versions of something to find out which version performs better. For example, a data professional might use A/B testing to compare two versions of a web page or two versions of an online ad. You also learned that A/B testing utilizes statistical methods such as sampling and hypothesis testing.

In this reading, you'll learn more about the general purpose and design of an A/B test and how A/B testing uses statistical methods to analyze data.

Business context

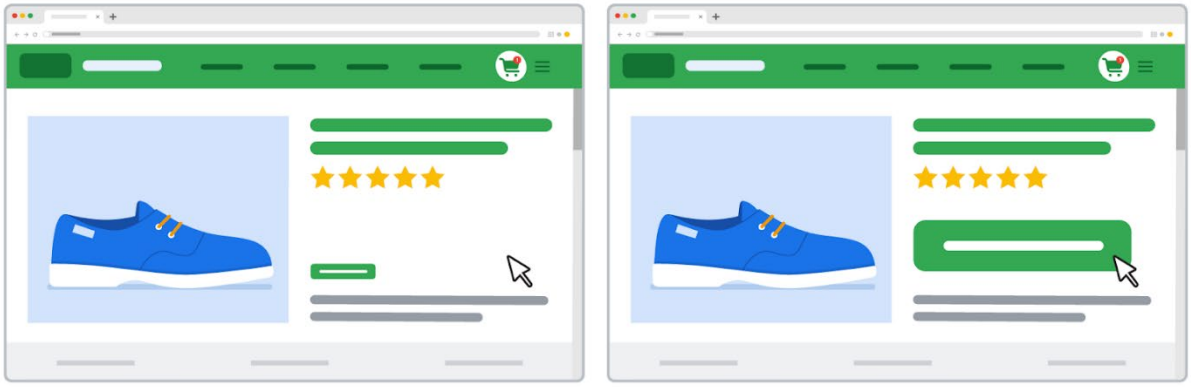
Data professionals often use A/B testing to help stakeholders choose the best design for a website or app to optimize marketing, increase revenue, or enhance customer experience. In practice, A/B testing involves randomly selecting a sample of users and dividing them into two groups (A and B). The two groups visit different versions of a company's website. The two versions are identical except for a single design feature. For instance, the "Purchase" button on Group A's version might have a different size, shape, or color than the "Purchase" button on Group B's version. An A/B test uses statistical analysis to determine whether the change in the feature (e.g., a larger button) affects user behavior for a specific metric. A data professional might use an A/B test to analyze one of the following metrics:

- *Average revenue per user*: How much revenue does a user generate for a website?
- *Average session duration*: How long does a user remain on a website?
- *Click rate:* If a user is shown an ad, does the user click on it?
- *Conversion rate*: If a user is shown an ad, will that user convert into a customer?

Let's explore an example to get a better understanding of how A/B testing works.

Example: Average revenue per user

Imagine you're a data professional who works for an online footwear retailer. The company is trying to grow its business and is researching the average revenue per user on its website. Your team leader asks you to conduct an A/B test to determine whether increasing the size of the "Purchase" button has any effect on average revenue. You randomly select a sample of users and divide them into two groups, A and B. Group A visits the standard version of the company website. Group B visits a version of the website that is identical to the standard version except for the larger "Purchase" button. You run the test online and collect your sample data. The results indicate that average revenue per user is higher for Group B. Finally, you conduct a two-sample hypothesis test to determine whether the observed difference in average revenue is statistically significant or due to chance.



A typical A/B test has at least three main features:

1. Test design
2. Sampling
3. Hypothesis testing

Let's examine each feature in more detail using our example.

Test design

First, let's discuss the fundamental design of an A/B test.

Randomized controlled experiment

An A/B test is a basic version of what's known as a randomized controlled experiment. In a **randomized controlled experiment**, test subjects are randomly assigned to a control group and a treatment group. The **treatment** is the new change being tested in the experiment. The **control group** is not exposed to the treatment. The **treatment group** is exposed to the treatment. The difference in metric values between the two groups measures the treatment's effect on the test subjects.

Note: Ideally, exposure to the treatment is the only significant difference between the two groups. This test design allows researchers to control for other factors that might influence the test results and draw causal conclusions about the effect of the treatment.

In our example, group A is the control group, group B is the treatment group, and the treatment is displaying a larger "Purchase" button. Users in the control group (A) visit the standard version of the company's website. Users in the treatment group (B) visit an alternative version with a larger "Purchase" button (i.e., are exposed to the treatment). By making the website versions for A and B identical except for the size of the "Purchase" button, you minimize the chance that any observed difference in average revenue is due to other features such as page layout or background. This allows you to measure the effect of the larger button by comparing the difference in average revenue per user for group A and group B.

Randomization, or randomly assigning test subjects to the control group or treatment group, also helps control the potential effect of other factors on the outcome of the experiment. In practice,

many different factors might influence whether a user clicks the “Purchase” button or not. For example, perhaps super wealthy users are much more likely to make purchases in general, regardless of button size. If your treatment group consists *only* of super wealthy users, you won’t get valid test results. Any observed increase in average revenue might be due to wealth, not to the larger size of the “Purchase” button (the factor you’re interested in testing). Randomization helps minimize the chance that other factors, such as wealth, will significantly influence your results on average.

Sampling

Random selection helps you create a representative sample that reflects the characteristics of the overall user population. In our example, this is the population of online customers of the company you work for. Using a representative sample for your A/B test will give you valid results that are generalizable, or applicable to the overall population.

You’ll also need to choose a sample size that is appropriate for your A/B test. The larger the sample size, the more precise the results, and the more likely you’ll get results that are statistically significant when there is a difference between group A and group B. However, working with large samples can be expensive and time-consuming. Data professionals determine sample size based on both the goal of the analysis and their available budget.

Hypothesis testing

For the purpose of our example, let’s say you run the online test, collect your data, and discover that group B has a higher average revenue per user than group A. Recall that group B is the treatment group (larger “Purchase” button), and group A is the control group. The next step is to determine whether the observed difference in your data is statistically significant or due to chance. A/B tests use two-sample hypothesis tests to draw conclusions about statistical significance. To determine whether the observed difference in average revenue per user is statistically significant, you conduct a two-sample t-test. You formulate your hypotheses as follows:

- $\mathbf{H}_0$: There is no difference in average revenue per user between A and B
- $\mathbf{H}_a$: There is a difference in average revenue per user between A and B

Results

Based on the results of your t-test, you reject the null hypothesis and conclude that the observed increase in average revenue per user is statistically significant.

The results of your A/B test help you decide whether or not to recommend a design change for your company’s website. In this case, when you present your results to company stakeholders, you suggest implementing the larger “Purchase” button to increase average revenue per user going forward.

Key takeaways

A/B testing is one of the most popular applications of statistics for business purposes. Data professionals use A/B testing to help business leaders optimize product performance, improve customer experience, and grow their online business. Understanding the general purpose and design of an A/B test will be useful in your future career as a data professional.

Experimental Design

Throughout this course, we've discussed how data professionals use hypothesis testing to determine whether the results of an experiment are statistically significant. In previous scenarios, we analyzed the results of experiments such as clinical trials and A/B tests. For instance, we imagined a clinical trial that tests the effectiveness of a new medicine and an A/B test that examines how changing the design of a web page affects the average time customers spend on the page.

Data professionals often work with experimental data previously collected by other researchers. However, the right data for a specific project might not always be available or accessible. In this case, data professionals can design their own experiments and collect their own data.

In this reading, we'll discuss how data professionals design experiments to collect data, test hypotheses, and discover relationships between variables. You'll learn more about the basic concepts and procedures of experimental design.

Context: Experimental design

Experimental design refers to planning an experiment in order to collect data to answer your research question.

Researchers conduct experiments in many fields: medicine, physics, psychology, manufacturing, marketing, and more. The typical purpose of an experiment is to discover a cause-and-effect relationship between variables. For example, a data professional might design an experiment to discover whether:

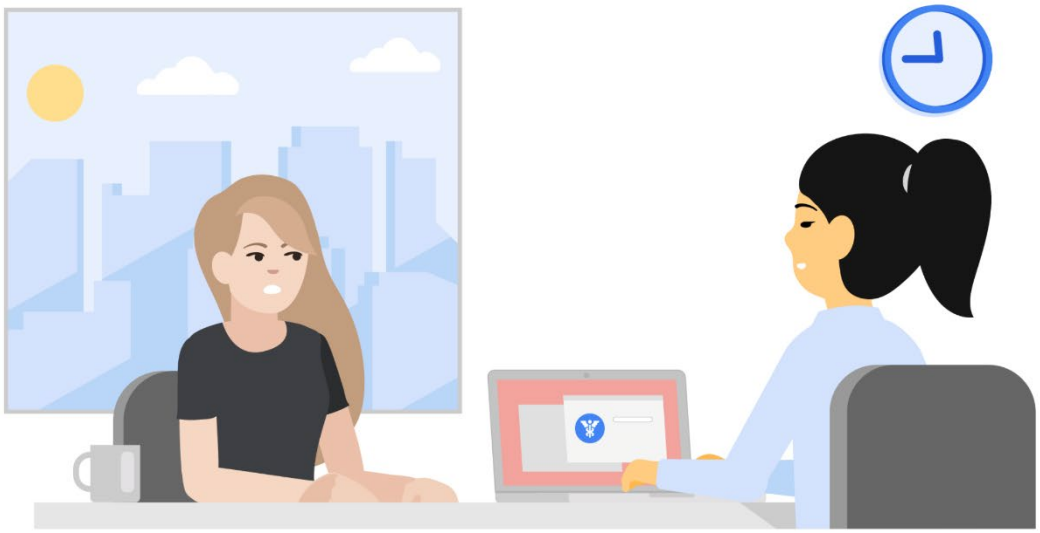
- A new medicine leads to faster recovery time
- A new website design increases product sales
- A new fertilizer increases crop growth
- A new training program improves athletic performance

It's important to understand experimental design because it affects the quality of your data, and the validity of any conclusions you draw based on your results. A poor design might lead to invalid results, which can be costly for companies and consumers. Based on the results of a flawed experiment, a company can spend years developing a medicine that is ineffective or invest heavily in a manufacturing process that is inefficient. A well-designed experiment will give you reliable data that helps answer your research question.

You can explore an example to get a better understanding of experimental design.

Example: Clinical trial

Imagine you're a data professional who works for a pharmaceutical company. The company invents a new medicine to treat the common cold. Your team leader asks you to design an experiment to test the effectiveness of the medicine. You want to find out whether taking the medicine leads to faster recovery time.



There are at least three key steps in designing an experiment:

1. Define your variables
2. Formulate your hypothesis
3. Assign test subjects to treatment and control groups

Note: These are basic steps that apply to controlled experiments (more below). Experimental design is a complex topic, and a more detailed discussion is beyond the scope of this course.

Next, examine each step in more detail using our example.

Step 1: Define your variables

Data professionals often begin by defining the independent and dependent variables in their experiment. This helps clarify the relationship between the variables.

- The **independent variable** refers to the cause you're interested in investigating. A researcher changes or controls the independent variable to determine how it affects the dependent variable. "Independent" means it's not influenced by other variables in the experiment.
- The **dependent variable** refers to the effect you're interested in measuring. "Dependent" means its value is influenced by the independent variable.

In your clinical trial, you want to find out how the medicine affects recovery time. Therefore:

- Your independent variable is the medicine—the cause you want to investigate.
- Your dependent variable is recovery time—the effect you want to measure.

In a more complex experiment, you might test the effect of different medicines on recovery time, or different doses of the same medicine. In each case, you manipulate your independent variable (medicine) to measure its effect on your dependent variable (recovery time).

Note: Later in this certificate program, when we discuss regression analysis, you'll have the chance to learn about independent and dependent variables in more detail.

Step 2: Formulate your hypothesis

The next step is to formulate a hypothesis. Your hypothesis states the relationship between your independent and dependent variables and predicts the outcome of your experiment. Earlier, you learned that data professionals formulate both null and alternative hypotheses when they conduct research that involves statistical testing. Recall that the null hypothesis typically assumes that there is no effect on the population, and the alternative hypothesis assumes the opposite. For your clinical trial:

- Your null hypothesis (H_0) is that the medicine has no effect.
- Your alternative hypothesis (H_a) is that the medicine is effective.

Step 3: Assign test subjects to treatment and control groups

Treatment and control groups

Experiments such as clinical trials and A/B tests are controlled experiments. In a **controlled experiment**, test subjects are assigned to a treatment group and a control group. The **treatment** is the new change being tested in the experiment. The **treatment group** is exposed to the treatment. The **control group** is not exposed to the treatment. The difference in metric values between the two groups measures the treatment's effect on the test subjects.

In your clinical trial, the treatment is the medicine that the subjects in the treatment group are given. The subjects in the control group are not given the medicine. Imagine your results show that mean recovery time is lower in the treatment group (6.2 days) than in the control group (7.5 days). The difference between the two groups, 7.5 - 6.2 = 1.3 days, measures the treatment's impact. In other words, the medicine decreases mean recovery time by 1.3 days.

Note: After a data professional designs and runs their experiment, they use statistical testing to analyze the results. As a next step, you might conduct a two-sample t-test to determine whether the observed difference in recovery time is statistically significant or due to chance.

Ideally, exposure to the treatment is the only significant difference between the two groups. This design allows researchers to control for other factors that might influence the test results and draw causal conclusions about the effect of the treatment.

For example, imagine the subjects in your treatment group have a much healthier diet than the subjects in your control group. Any observed decrease in recovery time for the treatment group might be due to their healthier diet—and not to the medicine. In this case, you cannot say with confidence that the medicine alone is the *cause* of the faster recovery time.

Randomization

Typically, data professionals randomly assign test subjects to treatment and control groups. Randomization helps control the effect of other factors on the outcome of an experiment. Two common methods for assigning subjects to treatment and control groups are completely randomized design and randomized block design.

In a **completely randomized design**, test subjects are assigned to treatment and control groups using a random process. For example, in a clinical trial, you might use a computer program to label each subject with a number and then randomly select numbers for each group.

Sometimes, however, a completely randomized design might not be the most effective approach. When designing an experiment, data professionals must account for **nuisance factors**. These are factors that can affect the result of an experiment, but are not of primary interest to the researcher.

Researchers can use a **randomized block design** to minimize the impact of known nuisance factors. **Blocking** is the arranging of test subjects in groups, or blocks, that are similar to one another. In a block design, you first divide subjects into blocks, and then you randomly assign the subjects within each block to treatment and control groups.

For example, suppose you know that age is a significant factor in recovery time from the common cold. In particular, you know that people under the age of 35 tend to recover faster than older people. In this scenario, age is a nuisance factor because it might affect the results of your experiment. For example, in a clinical trial with a completely randomized design and a smaller sample size, you might randomly get a large proportion of young people in the treatment group. This will make it more difficult to determine whether any observed decrease in recovery time is due to the treatment (medicine) or to the nuisance factor (age).

In this case, blocking for the age factor is a more effective way to design your experiment. First, you divide the test subjects into blocks based on age, such as 21-35, 36-50, and 51-65. Next, you randomly assign the subjects within each block to treatment and control groups. This way, if there is a significant difference in recovery time within a specific block, you can be more confident that this result is due to the treatment (medicine) and not to the nuisance factor (age).

Key takeaways

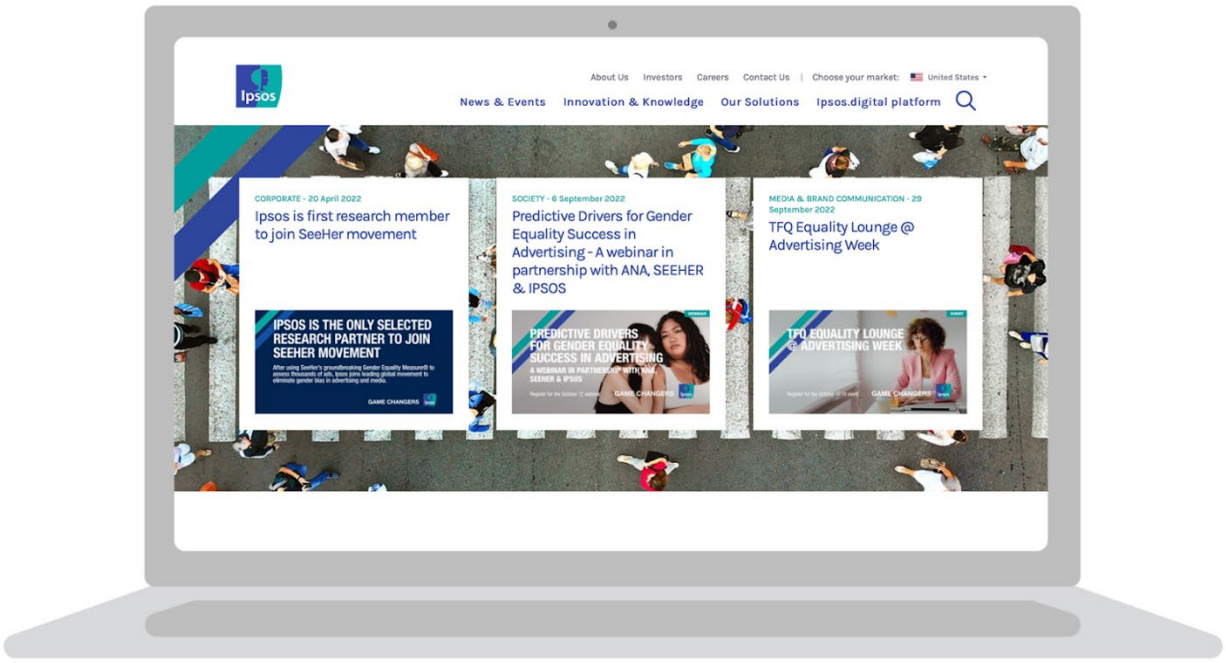
Data professionals use experimental design to plan experiments and collect data that helps answer their research questions. The design of an experiment affects the quality of your data and the validity of your conclusions. Whether you're designing your own experiment, or using data collected by others, it's important to understand the basic principles of experimental design. This knowledge will help you analyze data from experiments such as clinical trials, A/B tests, and more.

Case study: Ipsos: How a market research company used A/B testing to help advertisers create more effective ads

Previously, you learned that A/B testing is a way to compare two versions of something to find out which version performs better. For example, a data professional might use A/B testing to compare two versions of a web page or two versions of an online ad. You also learned that A/B testing utilizes statistical methods such as sampling and hypothesis testing. This case study describes how Ipsos used A/B testing to compare two different online ad formats: ads presented in a sequenced narrative vs. a traditional 30-second ad presented multiple times. You'll learn how data-driven market research reveals important insights about the impact of different ad formats on the effectiveness of a digital ad campaign.

Company background

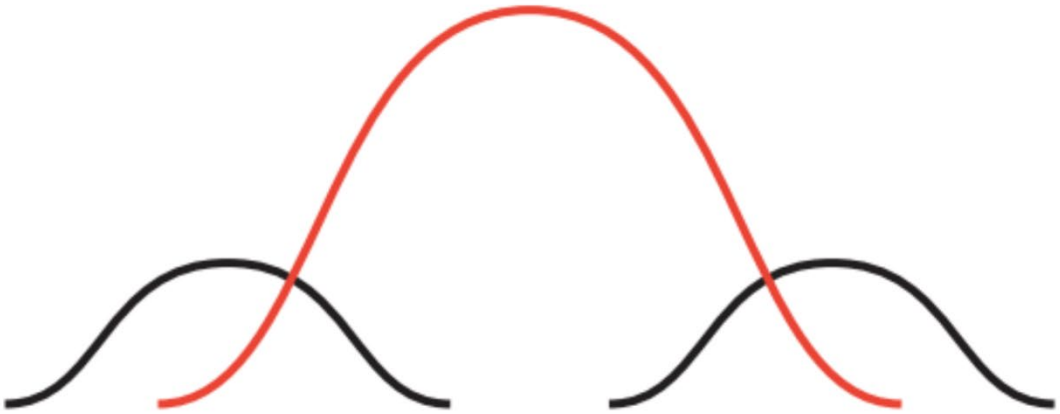
Ipsos is a full-service market research company. Founded in France in 1975, Ipsos is now a global company with 18,000 staff operating in 90 countries. Ipsos provides research services across numerous private and public sector domains. These services include brand building; advertising effectiveness; product development; reputation; customer and user experience; and public opinion, election, and crisis management. Ipsos uses a combination of data sources for their research, from primary data collection to social listening, mobility, and satellite imagery.



Project background

Ipsos' client for this project was an online media company that allows users to post their own video content. The media company wanted to help its own clients - the advertisers on their platform - create the most effective ads. In particular, they wanted to find out whether investing in sequenced ads would increase the likelihood of viewers recalling an ad and purchasing a product. Video ad sequencing lets advertisers show ads in an order based on the most engaging and memorable story structures. The media company commissioned Ipsos to conduct research to measure the impact of five different sequence structures on brand lift.

Note: For the purpose of this case study, we will focus on only one sequence structure: Tease, Amplify, Echo. This sequence starts with a short ad to spark viewer curiosity (Tease); then, it moves on to a longer ad with more information to secure viewer engagement (Amplify); finally, it ends with a shorter ad that recaps the story and spurs viewers to action (Echo).



Tease, Amplify, Echo

Graph with three overlapping curves represents the structure of the Tease, Amplify, Echo ad sequence: first, short video; second, long video; third, short video.

Project framework

Ipsos developed the following research question to guide their project: Does an ad sequence with the Tease, Amplify, Echo structure increase ad recall and purchase intent compared to repeated viewings of a traditional 30-second ad?

Ipsos' initial hypothesis was that ads presented in a sequenced narrative would be more effective than repeated viewings of a traditional ad. To test this hypothesis for these two advertising approaches, Ipsos conducted an A/B test. The A/B test set up an experiment for two groups of users: one group was shown the Tease, Amplify, Echo ads and the other group was shown a traditional ad multiple times. In each case, the different ad formats were based on the same brand content. You'll learn more about the details of the testing process in what follows.

The challenges

At the outset of the project, Ipsos identified two main challenges. The first challenge involved properly designing the A/B test. The second challenge involved creating the test ads in the appropriate test environment.

Test design

Ipsos's primary concern was for the results of the A/B test to be generalizable, or applicable to the overall population of the media company's users. In other words, Ipsos wanted to make valid inferences about the larger user population based on the smaller sample of test participants. To obtain valid test results, Ipsos needed to do the following;

1. Create a representative sample of test participants that mirrored the overall population of the media company's users.
2. Create an online test environment that mirrored the media company's online environment. This also implied creating test ads from multiple brands to reproduce the diversity of ads featured on the media company's platform.

The approach

Despite these challenges, Ipsos conducted the A/B test and achieved their research goals. Ipsos' successful approach to their project included the following elements:

- Team
- Sampling
- Testing process
- Hypothesis testing

Team

To build an effective team, Ipsos created a cross-functional operation, including video production to create test ads based on the Tease, Amplify, Echo structure, and technology to build a realistic online test environment.

To facilitate collaboration among project participants, Ipsos outlined a clear set of editing rules and organized a shared site to house links to videos and edit notes. This allowed rapid feedback and adjustment throughout the development process. Finally, Ipsos had senior client service project managers own and monitor the project design and workflow from beginning to end.

Sampling

Ipsos used random selection from consumer panels to generate a representative sample that accurately reflected the characteristics of the overall user population. Ipsos also made sure that each test group included the same proportion of respondents for key categories such as age and gender. Further, Ipsos used a relatively large sample size of 7,500 respondents to obtain more precise results.

Testing process

To get valid test results, Ipsos conducted A/B testing in an online environment where respondents used the media company's platform as they typically would in their everyday lives. To mirror the diversity of ads on the platform, Ipsos developed test ads across 30 brand categories, from airline tickets to fast food to laundry detergent.

The testing process was organized in the following way:

Surveys were administered online via respondents' smartphones in November and December of 2018. After initial screening, respondents were taken to a browser-based version of the platform where they were free to search for and watch videos as they normally would. Ipsos dynamically inserted test ads at the beginning of the videos selected by respondents in the live test environment. After the browsing session, respondents completed a survey to measure brand lift for ad recall and product intent.

Hypothesis testing

The survey data indicated that the Tease, Amplify, Echo ad sequence led to higher levels of ad recall and purchase intent among respondents than repeated viewings of a traditional ad. To determine whether their

observed results were statistically significant, Ipsos conducted a two-sample t-test for each category: one for ad recall and one for purchase intent. They formulated their hypotheses as follows:

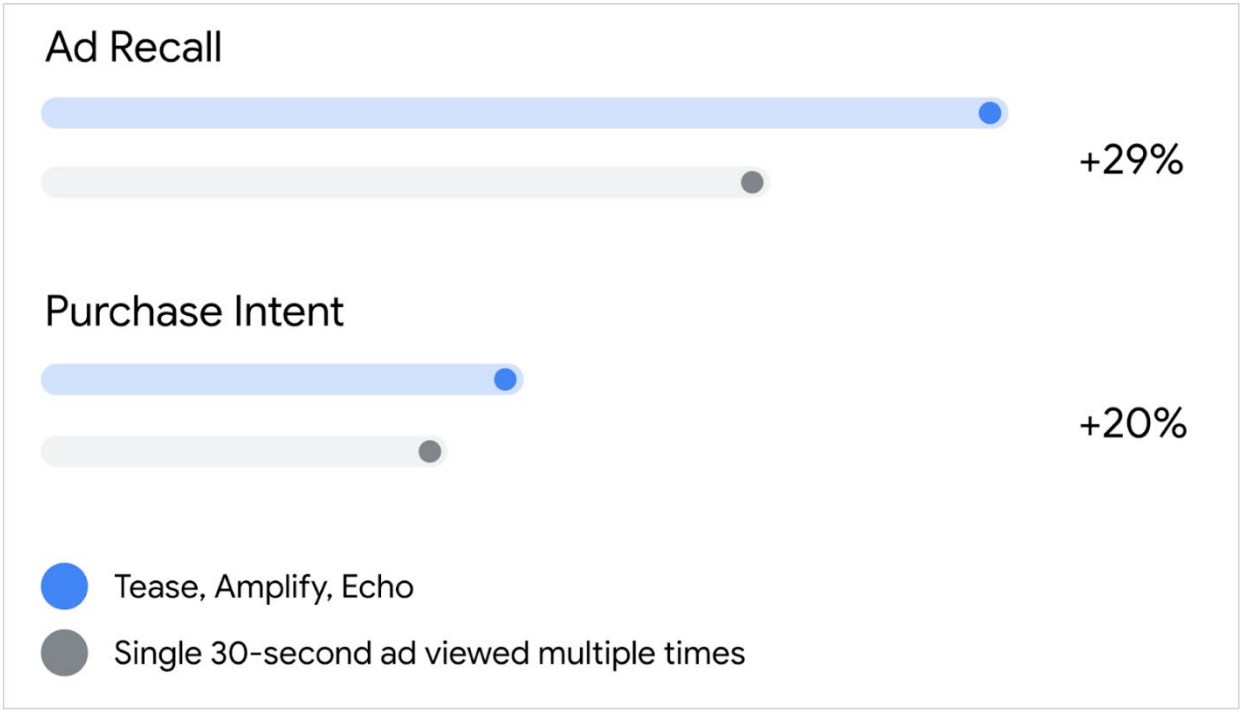
- $\mathbf{H}_0$: There is no difference in ad recall between sequenced ads and a repeated traditional ad.
- $\mathbf{H}_a$: There is a difference in ad recall intent between sequenced ads and a repeated traditional ad.
- $\mathbf{H}_0$: There is no difference in purchase intent between sequenced ads and a repeated traditional ad.
- $\mathbf{H}_a$: There is a difference in purchase intent between sequenced ads and a repeated traditional ad.

For both tests, Ipsos rejected the null hypothesis. They concluded that there are statistically significant and substantively meaningful differences in ad recall and purchase intent between sequenced ads and a repeated traditional ad.

The results

The results of the A/B test indicated that ad sequencing works!

The Tease, Amplify, Echo ad sequence had a significantly greater impact on ad recall and purchase intent than repeated viewings of a traditional ad. For example, across all product categories, 54% of viewers recalled the ad after being exposed to the “Tease, Amplify, Echo” sequence as compared to 42% with a repeated traditional ad. Further, 30% of viewers expressed intent to purchase after being exposed to the “Tease, Amplify, Echo” ad sequence as compared to 25% with a repeated traditional ad.



Bar chart showing the increase in ad recall and purchase intent for the Tease, Amplify, Echo ad sequence as compared with a repeated 30-second ad.

Overall, the results suggest that advertisers should invest in ad sequencing for their digital campaigns to increase brand lift.

Conclusion

This case study on Ipsos' A/B testing demonstrates the power of data-driven research to generate key business insights. The results of the A/B test clearly show how ad sequencing increases ad recall and purchase intent compared to repeated viewings of a traditional ad. Ipsos's research on the benefits of ad sequencing helped the media company improve the experience and performance of the advertisers on their platform and added value to the media company's brand.

Glossary terms from module 5
Terms and definitions from Course 4, Module 5

Alternative hypothesis: A statement that contradicts the null hypothesis and is accepted as true only if there is convincing evidence for it

Hypothesis testing: A statistical procedure that uses sample data to evaluate an assumption about a population parameter

Null hypothesis: A statement that is assumed to be true unless there is convincing evidence to the contrary

One-sample test: A hypothesis test that determines whether or not a population parameter like a mean or proportion is equal to a specific value

One-tailed test: In a hypothesis test, results when the alternative hypothesis states that the actual value of a population parameter is either less than or greater than the value in the null hypothesis

P-value: The probability of observing results as or more extreme than those observed when the null hypothesis is true

Significance level: The threshold at which a result is considered statistically significant

Statistical significance: The claim that the results of a test or experiment are not explainable by chance alone

Test statistic: A value that shows how closely the observed data matches the distribution expected under the null hypothesis

Two-sample test: A hypothesis test that determines whether or not two population parameters such as two means or two proportions are equal to each other

Two-tailed test: In a hypothesis test, results when the alternative hypothesis states that the actual value of the parameter does not equal the value in the null hypothesis

Type I error (false positive): The rejection of a null hypothesis that is actually true

Type II error (false negative): The failure to reject a null hypothesis which is actually false

Z-score: A measure of how many standard deviations below or above the population mean a data point is

Terms and definitions from Course 4

A

A/B testing: A way to compare two versions of something to find out which version performs better

Addition rule (for mutually exclusive events): The concept that if the events A and B are mutually exclusive, then the probability of A or B happening is the sum of the probabilities of A and B

B

Bayes' rule: (Refer to **Bayes' theorem**)

Bayes' theorem: A math formula for stating that for any two events A and B, the probability of A given B equals the probability of A multiplied by the probability of B given A divided by the probability of B; also referred to as Bayes' rule

Bayesian inference: (Refer to **Bayesian statistics)**

Bayesian statistics: A powerful method for analyzing and interpreting data in modern data analytics; also referred to as Bayesian inference

Binomial distribution: A discrete distribution that models the probability of events with only two possible outcomes: success or failure

C

Central Limit Theorem: The idea that the sampling distribution of the mean approaches a normal distribution as the sample size increases

Classical probability: A type of probability based on formal reasoning about events with equally likely outcomes

Cluster random sample: A probability sampling method that divides a population into clusters, randomly selects certain clusters, and includes all members from the chosen clusters in the sample

Complement of an event: In statistics, refers to an event not occuring

Complement rule: A concept stating that the probability that event A does not occur is one minus the probability of A

Conditional probability: The probability of an event occurring given that another event has already occurred

Confidence interval: A range of values that describes the uncertainty surrounding an estimate

Confidence level: A measure that expresses the uncertainty of the estimation process

Continuous random variable: A variable that takes all the possible values in some range of numbers

Convenience sample: A non-probability sampling method that involves choosing members of a population that are easy to contact or reach

D

Dependent events: The concept that two events are dependent if one event changes the probability of the other event

Descriptive statistics: A type of statistics that summarizes the main features of a dataset

Discrete random variable: A variable that has a countable number of possible values

E

Econometrics: A branch of economics that uses statistics to analyze economic problems

Empirical probability: A type of probability based on experimental or historical data

Empirical rule: A concept stating that the values on a normal curve are distributed in a regular pattern, based on their distance from the mean

F

False positive: A test result that indicates something is present when it really is not

I

Independent events: The concept that two events are independent if the occurrence of one event does not change the probability of the other event

Inferential statistics: A type of statistics that uses sample data to draw conclusions about a larger population

Interquartile range: The distance between the first quartile (Q1) and the third quartile (Q3)

Interval: A sample statistic plus or minus the margin of error

Interval estimate: A calculation that uses a range of values to estimate a population parameter

L

Literacy rate: The percentage of the population in a given age group that can read and write

Lower limit: When constructing an interval, the calculation of the sample means minus the margin of error

M

Margin of error: The maximum expected difference between a population parameter and a sample estimate

Mean: The average value in a dataset

Measure of central tendency: A value that represents the center of a dataset

Measure of dispersion: A value that represents the spread of a dataset, or the amount of variation in data points

Measure of position: A method by which the position of a value in relation to other values in a dataset is determined

Median: The middle value in a dataset

Method: A function that defines and performs behaviors like computation

Mode: The most frequently occurring value in a dataset

Multiplication rule (for independent events): The concept that if the events A and B are independent, then the probability of both A and B happening is the probability of A multiplied by the probability of B

Mutually exclusive: The concept that two events are mutually exclusive if they cannot occur at the same time

N

Non-probability sampling: A sampling method that is based on convenience or the personal preferences of the researcher, rather than random selection

Nonresponse bias: When certain groups of people are less likely to provide responses

Normal distribution: A continuous probability distribution that is symmetrical on both sides of the mean and bell-shaped

O

Objective probability: A type of probability based on statistics, experiments, and mathematical measurements

P

Parameter: A characteristic of a population

Percentile: The value below which a percentage of data falls

Point estimate: A calculation that uses a single value to estimate a population parameter

Poisson distribution: A probability distribution that models the probability that a certain number of events will occur during a specific time period

Population: Every possible element that a data professional is interested in measuring

Population proportion: The percentage of individuals or elements in a population that share a certain characteristic

Posterior probability: The updated probability of an event based on new data

Prior probability: The probability of an event before new data is collected

Probability: The branch of mathematics that deals with measuring and quantifying uncertainty

Probability distribution: A function that describes the likelihood of the possible outcomes of a random event

Probability sampling: A sampling method that uses random selection to generate a sample

Purposive sample: A method of non-probability sampling that involves researchers selecting participants based on the purpose of their study

Q

Quartile: A value that divides a dataset into four equal parts

R

Random experiment: A process whose outcome cannot be predicted with certainty

Random seed: A starting point for generating random numbers

Random variable: A variable that represents the values for the possible outcomes of a random event

Range: The difference between the largest and smallest value in a dataset

Representative sample: A sample that accurately reflects the characteristics of a population

S

Sample: A subset of a population

Sample size: The number of individuals or items chosen for a study or experiment

Sample space: The set of all possible values for a random variable

Sampling: The process of selecting a subset of data from a population

Sampling bias: When a sample is not representative of the population as a whole

Sampling distribution: A probability distribution of a sample statistic

Sampling frame: A list of all the items in a target population

Sampling variability: How much an estimate varies between samples

Sampling with replacement: When a population element can be selected more than one time

Sampling without replacement: When a population element can be selected only one time

Simple random sample: A probability sampling method in which every member of a population is selected randomly and has an equal chance of being chosen

Snowball sample: A method of non-probability sampling that involves researchers recruiting initial participants to be in a study and then asking them to recruit other people to participate in the study

Standard deviation: A statistic that calculates the typical distance of a data point from the mean of a dataset

Standard error: The standard deviation of a sample statistic

Standard error of the mean: The sample standard deviation divided by the square root of the sample size

Standard error of the proportion: The square root of the sample proportion times one minus the sample proportion divided by the sample size

Standardization: The process of putting different variables on the same scale

Statistic: A characteristic of a sample

Statistical significance: The claim that the results of a test or experiment are not explainable by chance alone

Statistics: The study of the collection, analysis, and interpretation of data

Stratified random sample: A probability sampling method that divides a population into groups and randomly selects some members from each group to be in the sample

Subjective probability: A type of probability based on personal feelings, experience, or judgment

Summary statistics: A method that summarizes data using a single number

Systematic random sample: A probability sampling method that puts every member of a population into an ordered sequence, chooses a random starting point in the sequence, and selects members for the sample at regular intervals

T

Target population: The complete set of elements that someone is interested in knowing more about

U

Undercoverage bias: When some members of a population are inadequately represented in a sample

Upper limit: When constructing an interval, the calculation of the sample means plus the margin of error

V

Variance: The average of the squared difference of each data point from the mean

Voluntary response sample: A method of non-probability sampling that consists of members of a population who volunteer to participate in a study

Z

Z-score: A measure of how many standard deviations below or above the population mean a data point is