The Power of Statistics

Contents

Week 1

You'll explore the role of statistics in data science and identify the difference between descriptive and inferential statistics. You'll learn how descriptive statistics can help you quickly summarize a dataset and measure the center, spread, and relative position of data.

- Use Python to compute descriptive statistics
- Determine measures of relative position such as percentile, quartile, and interquartile range
- Determine measures of dispersion such as range, variance, and standard deviation
- Determine measures of central tendency such as mean, median, and mode
- Explain the relationship between parameter and statistic in inferential statistics
- Explain the relationship between population and sample in inferential statistics
- Explain the difference between descriptive statistics and inferential statistics

1 Measures of Central Tendency: Mean, Median and 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.
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.

1.1 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.

$$Mean = \frac{10 + 5 + 3 + 50 + 12}{5} = 16$$

1.2 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.

1.3 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.

1.4 When to Use Mean, Median and 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.

1. 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.

2. The mode is useful when working with categorical data because it clearly shows you which category occurs most frequently.

2 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.

2.1 Range

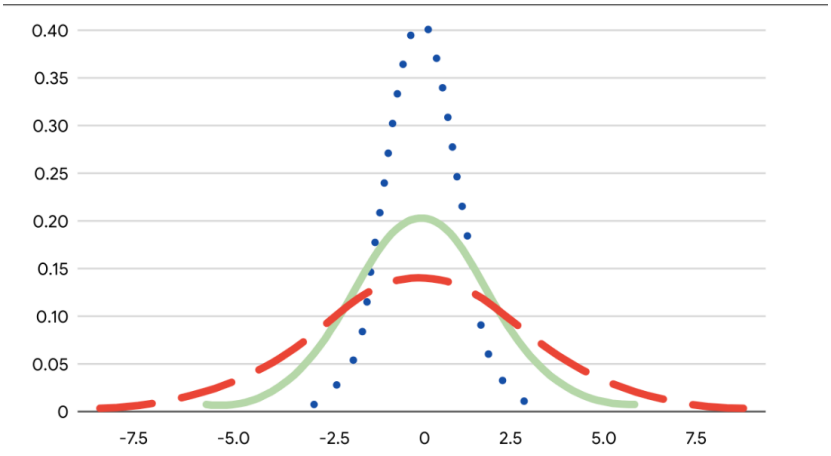
The range is the difference between the largest and smallest value in a dataset. For example, imagine you're a biology teacher and you have data on scores for the final exam. The highest score is 99/100, or 99%. The lowest score is 62/100, or 62%. To calculate the range, subtract the lowest score from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

2.2 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.

2.3 Standard Deviation

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.



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.

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:

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

3 Measures of Position - Percentiles and Quantiles

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.

3.1 Percentiles

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.

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.

3.2 Quartiles

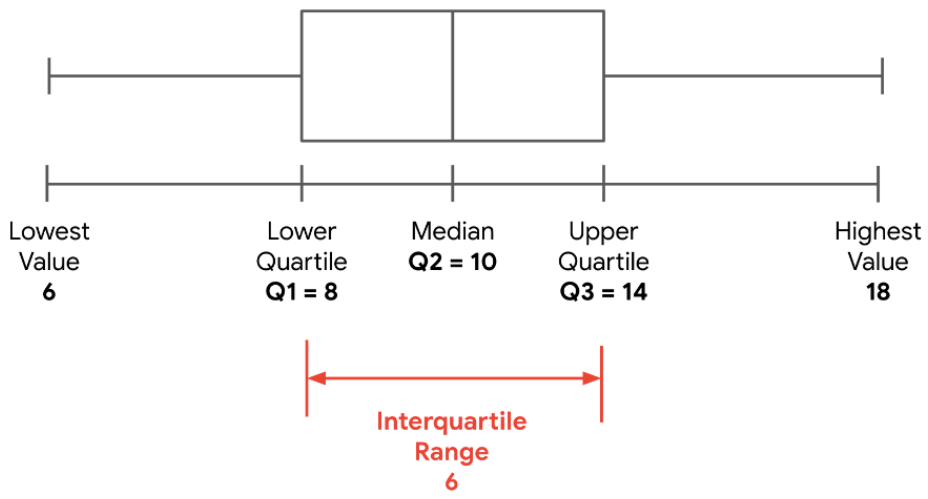
You can use quartiles to get a general understanding of the relative position of values. A quartile divides the values in a dataset into four equal parts. Three quartiles divide the data into four quarters. Quartiles let you compare values relative to the four quarters of data. Each quarter includes 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.

3.3 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.

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.



Week 2

You'll learn the basic rules for calculating probability for single events. Next, you'll discover how data professionals use methods such as Bayes' theorem to describe more complex events. Finally, you'll learn how probability distributions such as the binomial, Poisson, and normal distribution can help you better understand the structure of data.

- Use Python to model data with a probability distribution
- Describe the significance and use of z-scores
- Define the Empirical Rule
- Describe the features and uses of continuous probability distributions such as the normal distribution
- Describe the features and uses of discrete probability distributions such as the binomial and Poisson distributions
- Explain the difference between discrete and continuous random variables
- Describe Bayes' theorem and its applications
- Define dependent events
- Describe conditional probability and its applications
- Define different types of events such as mutually exclusive and independent events
- Apply basic rules of probability such as the complement, addition, and multiplication rules
- Describe basic probability in mathematical terms
- Explain the difference between objective and subjective probability

4 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.

4.1 Foundational concepts: 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.

4.2 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.

4.3 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.

Tossing a Fair coin
Tossing a fair coin is a classic example of a random experiment because -

- 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%.

4.4 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.

5 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.

5.1 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.

5.1.1 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.

5.1.2 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%

5.2 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

5.2.1 The 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:

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

5.2.2 The Addition Rule

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:

$$P(AUB) = P(A) + P(B)$$

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.
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) = $\frac{1}{6} + \frac{1}{6} = \frac{1}{3}$

5.2.3 The Multiplication Rule

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:

$$P(AB) = P(A) * P(B)$$

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.
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 $\frac{1}{6}$, 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) = $\frac{1}{6} * \frac{1}{6} = \frac{1}{36}$

6 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.

6.1 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.
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.

6.2 Conditional Probability

$$P(A \cap B) = P(A) * P(B|A)$$

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.

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

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.

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:

$$P(A \cap B) = P(A) * P(B|A)$$

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

7 Calculate conditional probability with Bayes's theorem

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

7.1 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.

7.2 Bayes Theorem

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.

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

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. 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(A)** - Prior Probability
- $P(A|B)$ - Posterior Probability
- $P(B|A)$ - Likelihood
- **P(B)** - Evidence

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, 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.

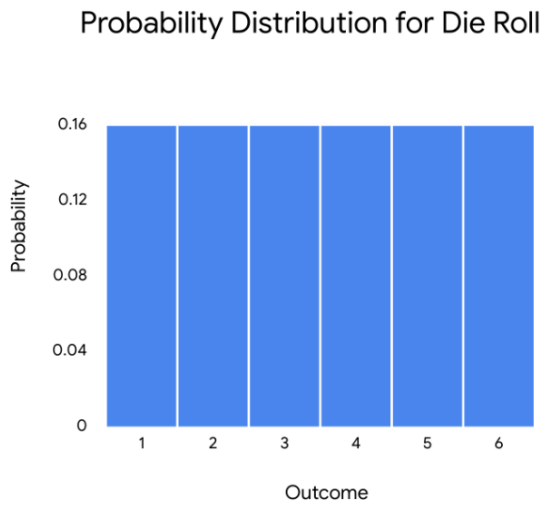
$$P(Spam|Money) = \frac{P(Money|Spam) * P(Spam)}{P(Money)}$$

$$P(Spam|Money) = \frac{0.4 * 0.2}{0.15} = 0.533$$

8 Discrete Probability Distributions

8.1 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.



8.2 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:

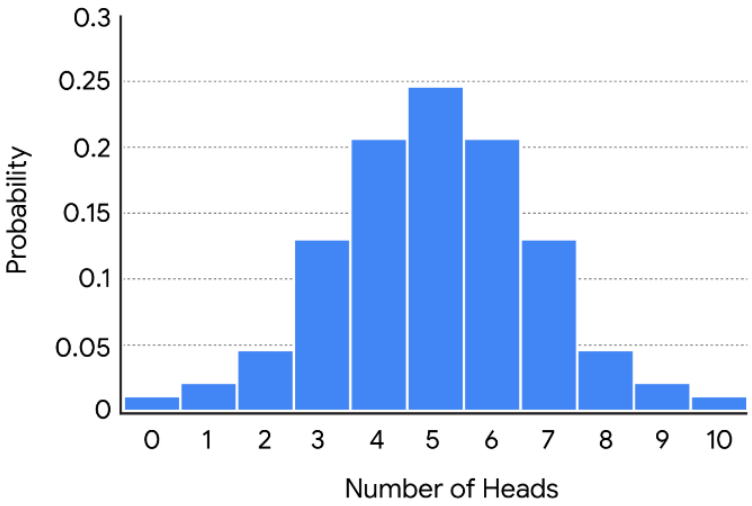
- Each event is independent, or does not affect the probability of the others.
- Each event has the same probability of success.

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.

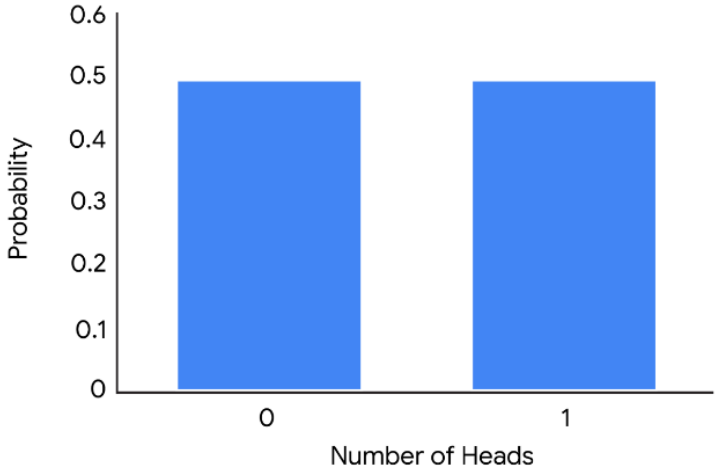


Data professionals might use the binomial distribution to model the probability that:

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

8.3 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.



8.4 Poisson Distribution

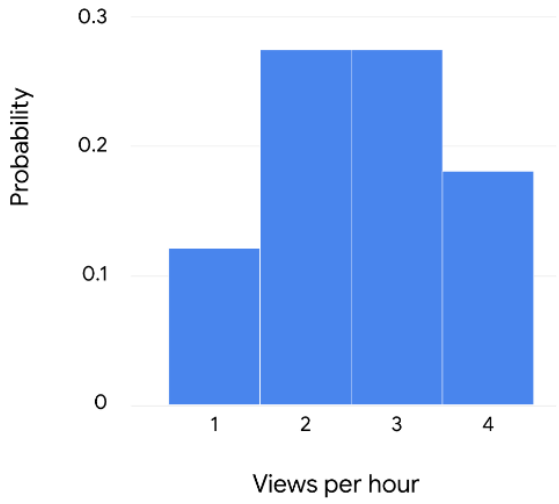
The Poisson distribution models the probability that a certain number of events will occur during a specific time period. The Poisson distribution represents a type of random experiment called a Poisson experiment. A Poisson experiment has the following attributes:

- 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

9 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.

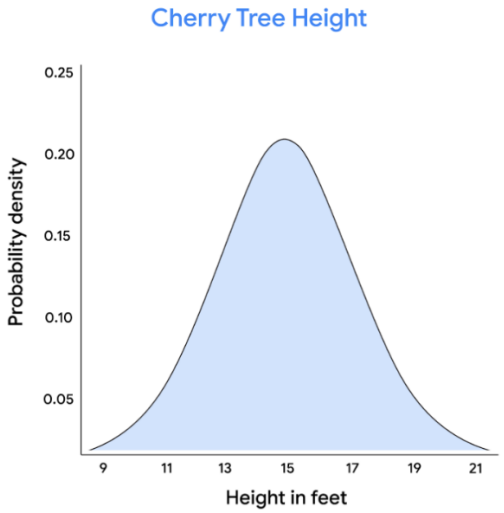
In this course, we focus on a single continuous probability distribution: the normal distribution. In this reading, you'll learn more about the main characteristics of the normal distribution, and how the distribution can help you model your data.

9.1 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:

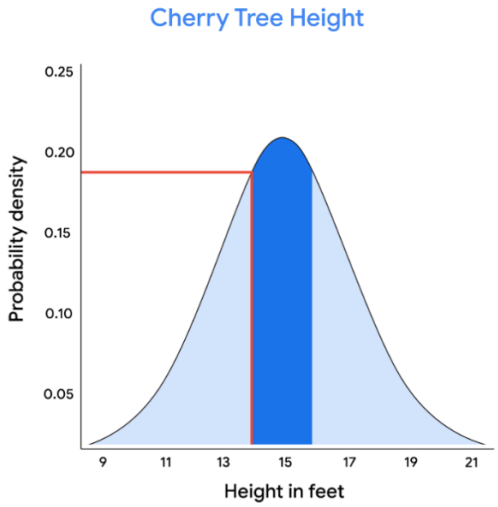
- **Probability Mass Functions (PMFs)** represent discrete random variables
- **Probability Density Functions (PDFs)** represent continuous random variables

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 on 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.



9.2 The Normal Distribution

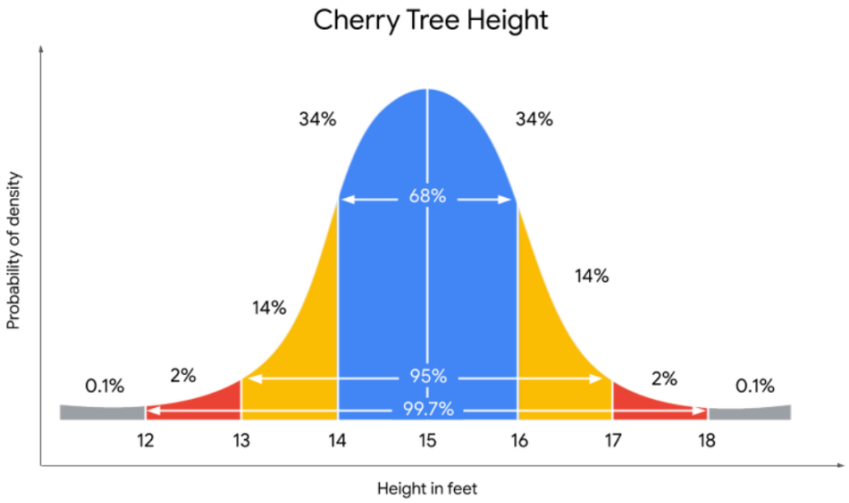
The normal distribution is a continuous probability distribution that is symmetric about the mean and bell-shaped. It is also known as the Gaussian distribution, after the German mathematician Carl Gauss, who first described its formula. The normal distribution is often called the bell curve because 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.



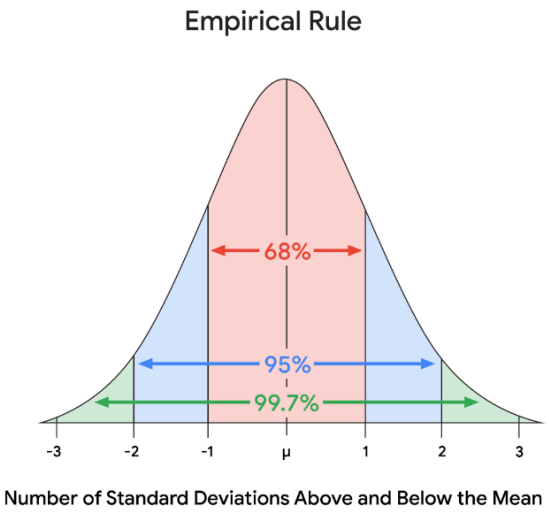
You may notice the following features of the normal curve:

- 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

9.3 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



Week 3

Data professionals use smaller samples of data to draw conclusions about large datasets. You'll learn about the different methods they use to collect and analyze sample data and how they avoid sampling bias. You'll also learn how sampling distributions can help you make accurate estimates.

- Use Python for sampling
- Explain the concept of standard error
- Define the central limit theorem
- Explain the concept of sampling distribution
- Explain the concept of sampling bias
- Describe the benefits and drawbacks of non-probability sampling methods such as convenience, voluntary response, snowball, and purposive
- Describe the benefits and drawbacks of probability sampling methods such as simple random, stratified, cluster, and systematic
- Explain the difference between probability sampling and non-probability sampling
- Describe the main stages of the sampling process
- Explain the concept of a representative sample

10 The Relationship between Sample and Population

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.

10.1 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.

10.2 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.

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

- Identify the target population
- Select the sampling frame
- Choose the sampling method
- Determine the sample size
- 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.

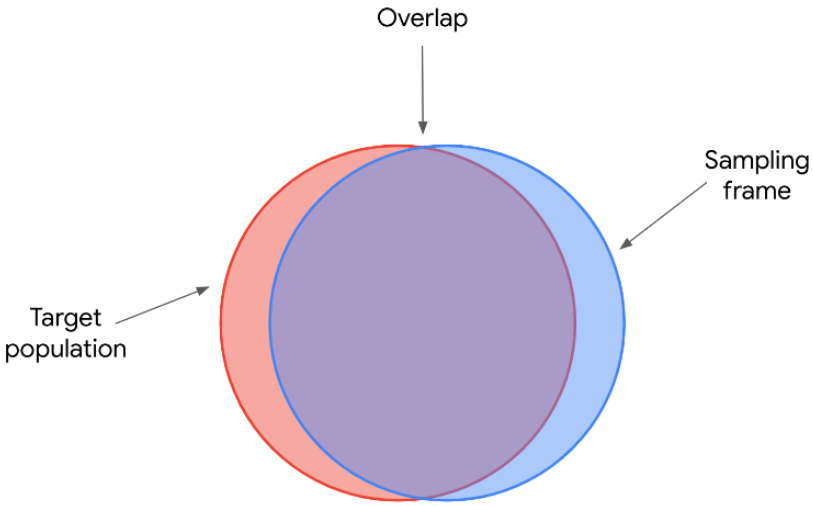
11.1 Step 1 - Identify the Target Population

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.

11.2 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.



11.3 Step 3 - Choose the Sampling Frame

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.

11.4 Step 4 - Determine the Sampling 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.

11.5 Step 5 - Collect your Sample Data

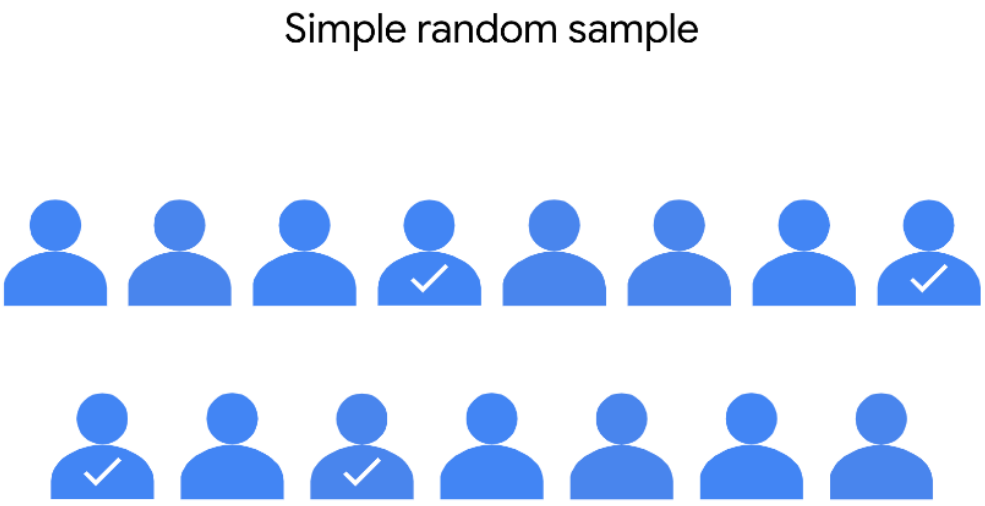
Now, you're ready to collect your sample data, which is the final step in the sampling process. You give a customer satisfaction survey to the customers selected for your sample. The survey responses provide useful data on how customers feel about the digital features of the refrigerator. Then, you share 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.

12 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.

12.1 Simple Random Sampling

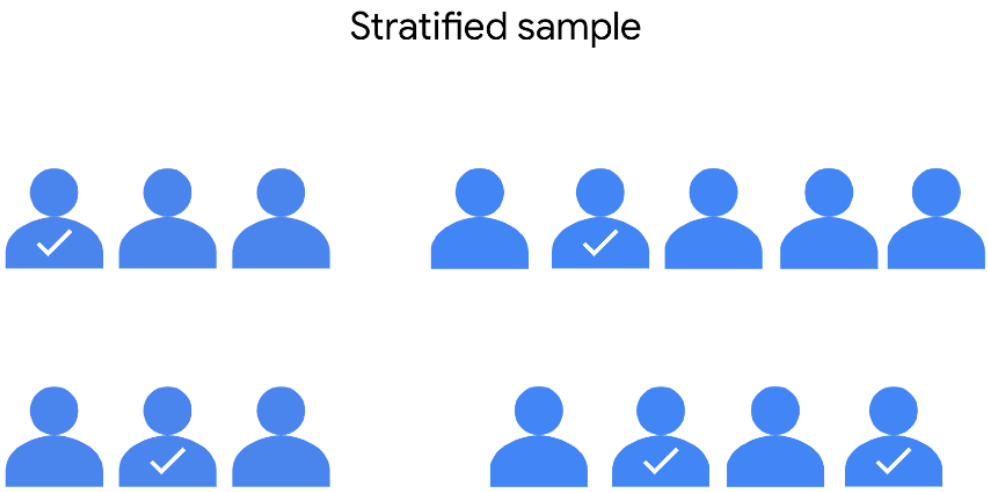
In a simple random sample, every member of a population is selected randomly and has an equal chance of being chosen. You can randomly select members using a random number generator, or by another method of random selection. For example, imagine you want to survey the employees of a company about their 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.

12.2 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.

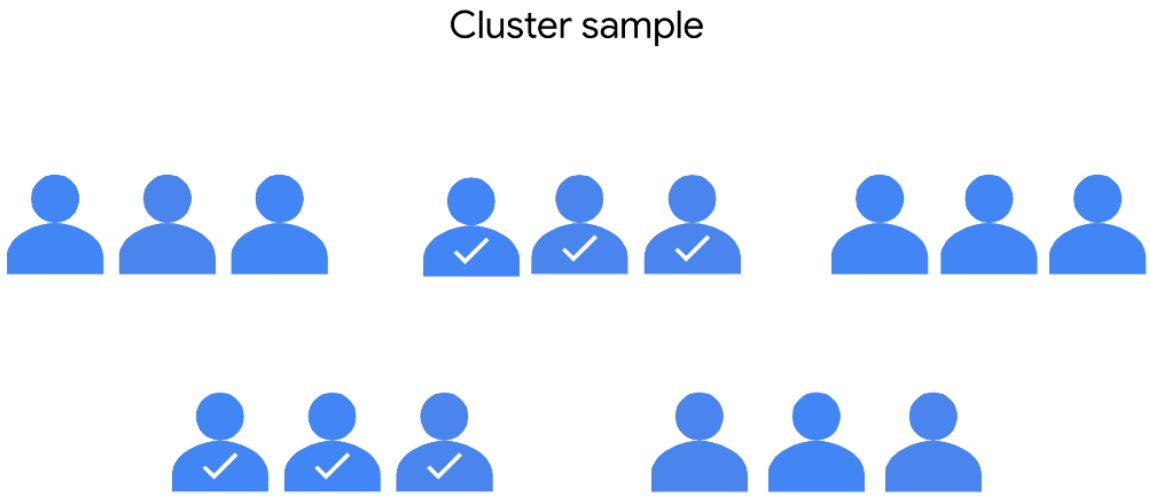


For example, imagine you're doing market research for a new product, and you want to analyze the preferences of consumers in different age groups. You might divide your target population into strata according to age: 20-29, 30-39, 40-49, 50-59, etc. Then, you can survey an equal number of people from 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.

12.3 Cluster Random Sampling

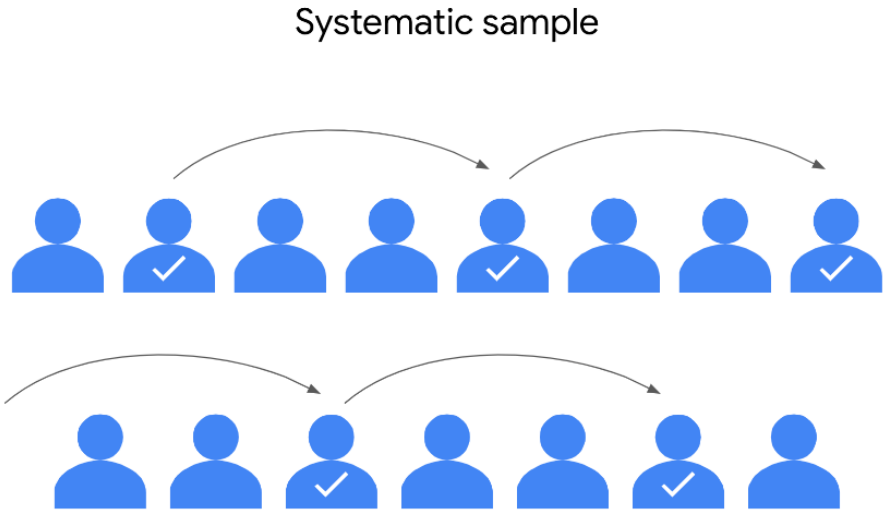
When you're conducting a cluster random sample, you divide a population into clusters, randomly select certain clusters, and include all members from the chosen clusters in the sample.
Cluster sampling is similar to stratified random sampling, but in stratified sampling, you randomly choose some members 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.

12.4 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.

13 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.

13.1 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.

13.2 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.

13.3 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.
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.

13.4 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.

14 Infer Population Parameters with the Central Limit Theorem

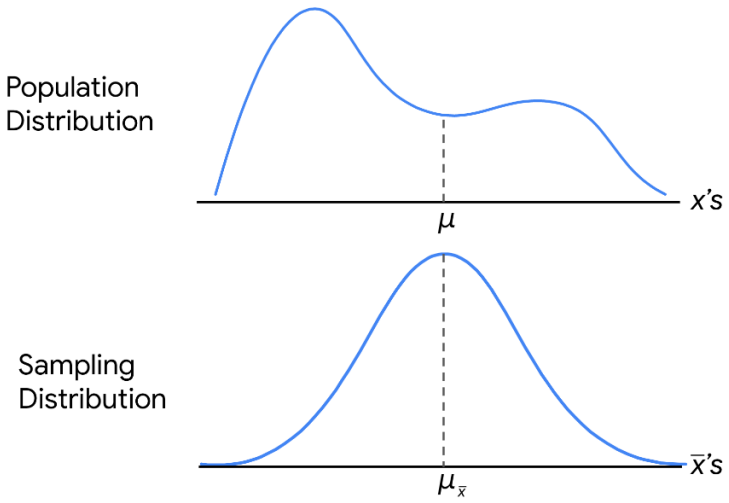
Recently, you learned about the central limit theorem and how it can help you work with a wide variety of datasets. Data professionals use the central limit theorem to estimate population param-

eters 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.

14.1 Central Limit Theorem

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.



14.1.1 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%**: Your sample size should be no larger than 10% of the total population. This applies when the sample is drawn without replacement, which is usually the case.
- **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.

15 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.

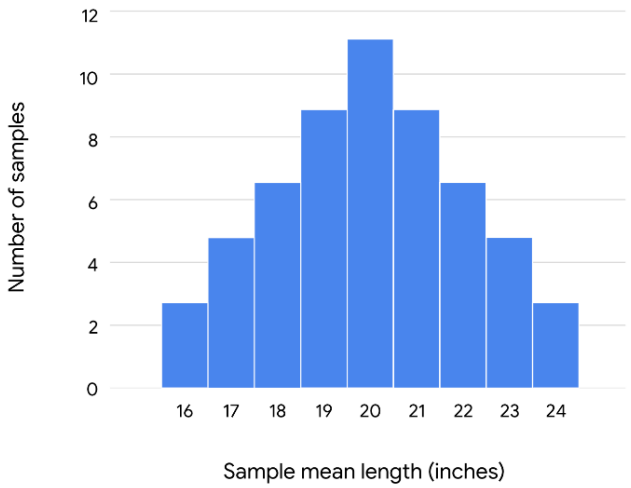
15.1 Sampling Distribution of the Sample Mean

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.

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.

15.2 The 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.

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.

Week 4

You'll explore how data professionals use confidence intervals to describe the uncertainty of their estimates. You'll learn how to construct and interpret confidence intervals — and how to avoid some common misinterpretations.

- Use Python to construct a confidence interval
- Describe how to construct a confidence interval for means and proportions
- Identify common forms of misinterpretation associated with confidence intervals
- Describe how to properly interpret a confidence interval
- Define concepts related to confidence intervals such as confidence level and margin of error
- Explain the difference between a point estimate and an interval estimate

16 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.

16.1 Correct Interpretation - Penguins example

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.

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.

16.2 Incorrect Misinterpretations

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.

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.

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

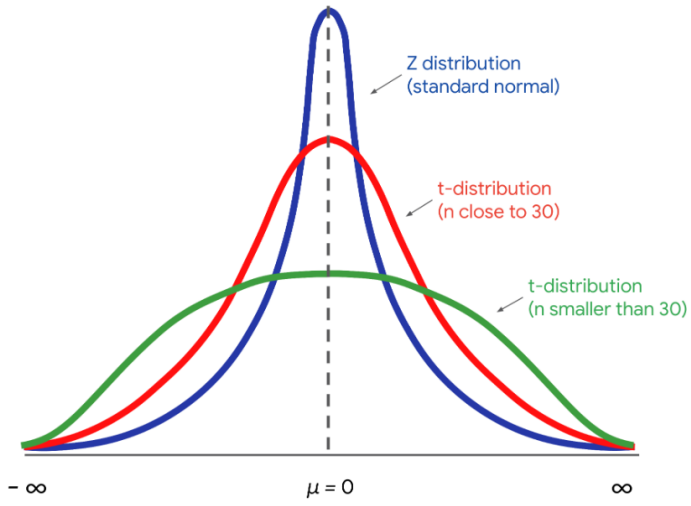
the interval A third common misinterpretation is that a 95% confidence interval implies that 95% of all possible sample means fall within the range of the interval. This is not necessarily true. For example, your 95% confidence interval for mean penguin weight is between 28 pounds and 32 pounds. Imagine 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.

17 Construct a Confidence Interval for 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.

17.1 Large samples: 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.



17.2 Small samples: 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. 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.

17.3 Creating a Confidence Interval

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 CO2, which is an air pollutant. The engineering team has designed a new engine to reduce emissions for the company's best-selling car. The goal is to keep emissions below 460 grams of CO2 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. 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.

The engineering team tests a random sample of 15 engines and collects data on their emissions. The mean emission rate is 430 grams of CO2 per mile, and the standard deviation is 35 grams of CO2 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 CO2 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.

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.

2. **Choose a confidence level** - Next, choose a confidence level. The engineering team requests that you choose a 95% confidence level.

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.

   $$S.E(x) = \frac{\sigma}{\sqrt{n}}$$

   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.

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 CO2 per mile. The lower limit is the sample mean minus the margin of error. This is 43019.39 = 410.61 grams of CO2 per mile. You have a 95% confidence interval that stretches from 410.61 grams of CO2 per mile to 449.39 grams of CO2 per mile.

Week 5

Hypothesis testing helps data professionals determine if the results of a test or experiment are statistically significant or due to chance.

- Use Python to conduct a hypothesis test
- Describe how to conduct a two-sample hypothesis test
- Describe how to conduct a one-sample hypothesis test
- Explain the difference between a type I error and a type II error
- Define concepts related to hypothesis testing such as significance level and p-value
- Understand the role of statistical significance in hypothesis testing
- Explain the difference between the null hypothesis and the alternative hypothesis

18 Differences between Null and Alternative Hypothesis

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.

18.1 Statistical Hypothesis

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 **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.

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 (H0)**: the program had no effect on sales revenue.
Your **alternative hypothesis (Ha)**: the program increased sales revenue.

18.2 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 $\leq$ or $\geq$).
- Null hypotheses often include phrases such as "no effect," "no difference," "no relationship," or "no change."

18.3 Alternate 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."

18.4 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.

H0: $\mu = 300$ (the mean weight of all produced granola bags is equal to 300 grams)
Ha: $\mu \neq 300$ (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.

H0: $\mu \leq 30$ (the mean height of this species of tree is equal to or less than 30 feet)
Ha: $\mu > 30$ (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.

H0: $p \geq 0.80$ (the proportion of all employees who are satisfied with their job is equal to or greater than 80%)
Ha: $P < 0.80$ (the proportion of all employees who are satisfied with their job is less than 80%)

19 Type 1 and Type 2 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.

19.1 Errors in Statistical Tests

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

- State the null hypothesis and the alternative hypothesis.
- Choose a significance level.
- Find the p-value.
- 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)**

19.2 Type 1 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. 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. 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%.

19.3 Type 2 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 β 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.

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.

19.4 Potential Risks of Type 1 and Type 2 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.

20 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.

20.1 Statistical Singificance in Hypothesis Testing - Example

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
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.

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.

   H0: $\mu = 8.5$ (the mean battery life of all redesigned laptops is equal to 8.5 hours)
   Ha: $\mu > 8.5$ (the mean battery life of all redesigned laptops is greater than 8.5 hours)

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.

3. **Choose a 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.

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.

21 One-Tailed and Two-Tailed Test

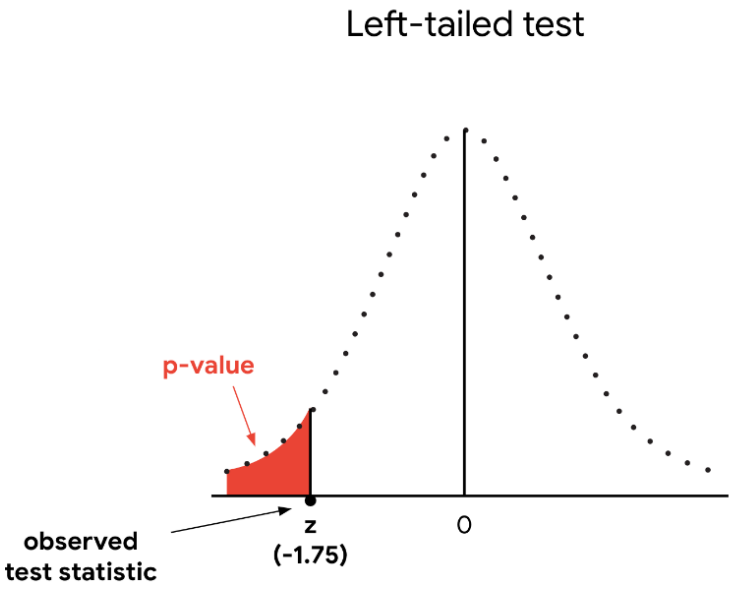
A **one-tailed** test results when the alternative hypothesis states that the actual value of a population parameter is either less than or greater than the value in the null hypothesis. A one-tailed test may be either left-tailed or right-tailed. A left-tailed test results when the alternative hypothesis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

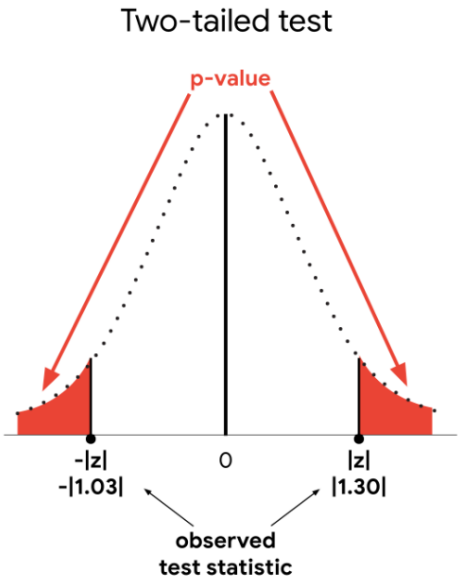
A **two-tailed** test results when the alternative hypothesis states that the actual value of the parameter does not equal the value in the null hypothesis. For example, imagine a test in which the null hypothesis states that the mean weight of a penguin population equals 30 lbs.

- 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 ($\neq$) to 30 lbs.

Example - One-Tailed test 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.
First, you state the null and alternative hypotheses:
H0: $P \geq 0.80$ (the proportion of satisfied customers is greater than or equal to 80%)
Ha: $P < 0.80$ (the proportion of satisfied customers is less than 80Next, 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 one-tailed test, the p-value is 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.



Example - Two-Tailed test
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:
H0: P = 0.80 (the proportion of satisfied customers equals 80%)
Ha: P $\neq$ 0.80 (the proportion of satisfied customers does not equal 80%)
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.



22 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.

22.1 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.