**Google Advanced Data Analytics Certificate**

Course 4: The Power of Statistics

A close-up of a logo

Description automatically generated

**Note**: Check out my Certifications repository on Github to find the specific labs for each course and my end of course portfolio project.

# Module 1: Introduction to statistics

In this section of the course, you’ll learn about the foundational role of statistics in data science. This section focuses on fundamental concepts of descriptive statistics, such as measures of central tendency, dispersion, and position.

**Statistics** is the study of the collection, analysis, and interpretation of data.

## Video 1.1: What You Will Learn

* Fundamental Stats Concepts
* A/B test
* Descriptive statistics
* Inferential statistics
* Types of descriptive statistics(mean, standard deviation, percentiles)
* Python

## Video 1.2: The Role of Statistics in Data Science

**Data professionals use statistical methods to**

* Identify meaningful patterns and relationships in data
* Analyze and quantify uncertainty
* Generate insights from data
* Make informed predictions about the future
* Solve complex problems

**A data professional might use probability to**

* Predict the future rate of return of an investment
* Estimate the annual average sales revenue for a company
* Calculate the margin of error to quantify the uncertainty of an employee satisfaction survey
* Used percentiles to rank median home prices in different cities

**Methods**

* Hypothesis testing
* Classification
* Regression
* Time series analysis

## Video 1.3: Statistics in Action: A/B testing

**A/B testing** is a way to compare two versions of something to find out which version performs better.

**For example**, businesses often use A/B testing to create two versions of a webpage to find out which one gets more clicks, purchases, or subscriptions. Even small changes to a webpage, like changing the color, size, or location of a button can increase financial gains. A/B test help business leaders optimize product performance and improve customer experience.

**A/B testing**

* Marketing email (which email version works best)
* Online ads (which adds generate more user engagements and revenue)

Once you have conducted A/B testing, you can use the gathered data to make permanent changes to your product

**Step By Step Example of A/B testing**

Scenario: Imagine you run an online store and 10 percent of visitors to your website make a purchase. You want to run an A/B test to find out if changing the size of the add to cart button will increase the conversion rate, or the percentage of customers who purchase a product.

The test presents two versions of your webpage, known as version A and version B, to a group of randomly selected users. Version A is the original webpage. Version B is the webpage with the larger add to cart button.

A screenshot of a browser window

Description automatically generated

The test directs half the users to version A and half to version B. The test runs for two weeks. When the test is over, a statistical analysis of the results indicates that the larger button in version B resulted in an increase in purchases. The conversion rate for version B is 30 percent.

A screenshot of a computer

Description automatically generated

This is three times greater than the conversion rate for version A, which is 10 percent. That's a notable increase. **Because of your A/B test, your company has a data-driven reason for replacing the current webpage with version B and increasing the size of an add to cart button.**

**A/B test steps:**

* Analyzes a small group of users (sample)
* Decide on the sample size (choosing the right sample size helps you get valid test results and avoid statistical errors)
* Determine the statistical significance

**Inferential statistics** is when statisticians make inferences about a dataset based on a sample of the data.

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

**Confidence Interval**: A range of values that describes the uncertainty surrounding an estimate.

## Video 1.4: Descriptive Statistics Versus Inferential Statistics

**Descriptive statistics**: Describe or summarize the main features of a dataset.

* Useful to quickly understand large amounts of data

**Forms of descriptive stats**

* Visuals, like graphs and tables
* Summary stats

**Measures of central tendency**

* Describe the center of your dataset (example, the mean)

**Measures of dispersion**

* Describe the spread of your dataset, or the amount of variation in your datapoints (example: standard deviation.

**Inferential statistics:** Make inferences about a dataset based on a sample of data.

**Population**: Every possible element that you are interested in measuring.

**Statistical Population**

* People
* Objects
* Events

Keep in mind that your sample should be **representative of your population**. Otherwise, the conclusions you draw from your sample will be unreliable and possibly biased.

A **representative sample** is a sample that accurately reflects the population.

For example, if you only survey math majors or only student athletes, your sample will not be representative of all college students.

Parameter: A characteristic of a population Statistic: A characteristic of a sample

A diagram of a baby's population

Description automatically generated A diagram of a blue animal

Description automatically generated with medium confidence

## Video 1.5: Measures of Central Tendency

* Measures of central tendency are values that represent the center of the dataset.
* Measures of dispersion are values that represent the spread of a dataset.

**Measures of central tendency**

* Mean
* Median
* Mode

**Mean**

* The average value of all values

**A blue text on a white background

Description automatically generatedA number with blue numbers

Description automatically generated with medium confidence**

**Median**

* The middle value in a dataset.

**A number and mathematical equation

Description automatically generated with medium confidence**

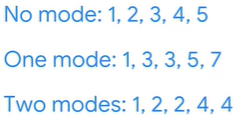
In the previous example we see that the median and the mean are really far apart from each other, this is due to the 70 which is an outlier, an observation that greatly differs from the rest of the data.

**When to use the median or mean**

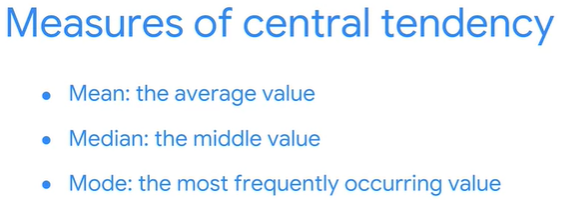
* If outliers, use median
* If no outliers, use the mean

**Mode**

* The most frequently occurring value in a dataset

****

* The mode is more useful when we are dealing with categorical data as we can see what category occurs the most.

****

## Video 1.6: Measures of Dispersion

* Datasets with the same central value can have different levels of variability.

A screenshot of a computer

Description automatically generated

**Measures of dispersion**

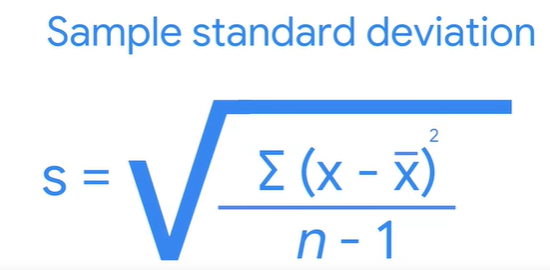
* Range
* Standard deviation
* Variance

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

**Standard deviation**: measures how spread out your values are from the mean of your dataset.

* The larger the SD, the more spread out your values are from the mean.

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

A graph with lines and dots

Description automatically generated

* In the graph below, the blue curve has the smallest standard deviation while the red curve has the largest standard deviation.

**Data Professionals use standard deviation to measure variation in**

* Ad revenues
* Stock prices
* Employee salaries

## Video 1.7: Measures of Position

**Measures of position**: Determine the position of a value in relation to other values in a dataset.

**Measures of position**

* Percentiles
* Quartiles
* Interquartile range
* Five number summary

**Percentile**: The value below which a percentage of data falls

* Percentiles show the relative position or rank of a particular value in a dataset.

**Quartile**: Divides the values in a dataset into four equal parts.

**A diagram of a number of squares

Description automatically generated with medium confidence**

**Five number summary**

* The minimum
* The first quartile (Q1)
* The median, or second quartile (Q2)
* The third quartile (Q4)
* The maximum

A diagram of a box plot

Description automatically generated

## Glossary Terms from Module 1

Terms and definitions from Course 4, Module 1

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

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

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

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

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

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

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

**Mean**: The average value in a dataset

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

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

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

**Median**: The middle value in a dataset

**Mode**: The most frequently occurring value in a dataset

**Parameter**: A characteristic of a population

**Percentile**: The value below which a percentage of data falls

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

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

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

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

**Sample** : A subset of a population

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

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

**Statistic**: A characteristic of a sample

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

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

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

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

# Module 2: Probability

You will learn about fundamental concepts in probability. The first half of this section covers basic rules of probability (complement, addition, multiplication), conditional probability, and Bayes's theorem. The second part concentrates on three probability distributions: the binomial, Poisson, and normal distributions.

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

## Video 2.1: What You Will Learn

* Types of probability
* Basic rules of probability
* Conditional probability
* Bayes’ Theorem
* Probability distributions
* Discrete probability distributions
* Continuous probability
* Z-scores

## Video 2.2: Objective Versus Subjective Probability

**Probability use cases**

* A company will sell a certain amount of product in a given time period.
* A financial investment will have a positive return
* A political candidate will win an election
* A medical test will be accurate

**Types of Probability**

* Objective (Classical, Empirical)
* Subjective

**Objective probability**: Based on statistics, experiments, and mathematical measurements.

**Classical probability**: Based on formal reasoning about events with equally likely outcomes.



**Example**: Rolling some dice, pulling an ace from a stack of cards, flipping heads/tails.

These events have an equally outcome probability.

**Empirical probability**: Based on experimental or historic data.



Example: An ice-cream taste test.

A group of people eating ice cream

Description automatically generated

**Subjective probability**: Based on personal feelings, experience, or judgement.

**Example**: Believing that your team will win the championship or that a certain horse will win the race. There are no statistical analyses behind this assumptions and are based on personal feelings or judgement.

**Example**: The CEO of an auto company might feel confident that using a new technology to manufacture their pickup truck will cut costs and increase profits. But if their prediction is only based on personal feeling or subjective probability, it may not be reliable.

Data science based on statistical analysis or objective probability can help accurately predict the potential impact of the new technology and help the CEO make an informed, data-driven decision about adopting the technology.

## Video 2.3: The Principles of Probability

**Events in probability**

* 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% probability that the event will occur.
* If the probability of an event is close to zero, there's a small chance that the event will occur.
* If the probability of an event is close to one, there's a strong chance that the event will occur.
* For any event A, 0 ≤ P(A) ≤ 1. In other words, the probability of any event A is always between 0 and 1.

Probability measures the likelihood of random events.

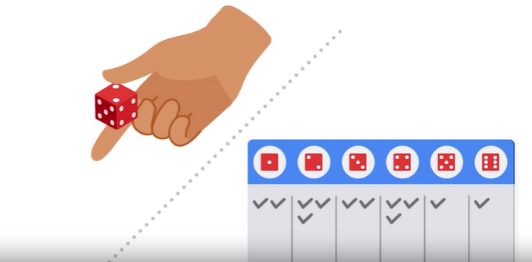
**Random experiment**: A process whose outcome cannot be predicted with certainty.

**Random experiments**

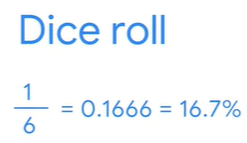
* The experiment can have **more than one possible outcome**
* You can represent each possible outcome **in advance**
* The outcome of the experiment **depends on chance**

**Examples:**

**A hand holding gold coins

Description automatically generated**

In both examples we can have more than one possible outcome (H/T or 1-6) and we can represent each possible outcome in advance as their outcome depends on chance. **These are examples of classical probability**.



Another example is a jar of marbles where we want to know the chances of pulling a green marble.



In this case we have a 3/10 probability of pulling a green marble.

## Video 2.4: The Basic Rules of Probability and Events

**Basic rules of probability**

* Complement rule
* Addition rule
* Multiplication rule

**Types of events**

* Mutually exclusive events
* Independent events

**Probability notation**

* P(A) = probability of event A
* P(B) = probability of event B
* P(A’) = probability of **not** event A

**Complement of an event**: The event not occurring.

**Example:** The complement of winning an event is not winning, the complement of rain is no rain.

* The two probabilities if an event happening and not happening must add to 1.

A math equation with blue text

Description automatically generated

**Example:** Probability of no rain tomorrow.



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

**Example:** You cannot visit Argentina and China at the same time or turn left and right at the same time.

A close up of a sign

Description automatically generated

**Example:** Rolling a 2 or a 4 on a 6-sided die.

A math equation with blue text

Description automatically generated

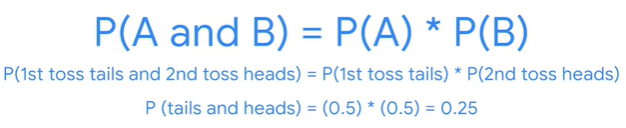
**Independent events:** Two events are independent if the occurrence of one event does not change the probability of the other event.

**Examples:** For example, checking out a book from your local library does not affect tomorrow's weather. Drinking coffee in the morning does not affect the delivery of your mail in the afternoon. These events are separate and independent.

**A blue text on a white background

Description automatically generated**

**Example:** Get tails on your first toss and tails on your second toss. These are independent events since the first toss does not affect the outcome of the second toss.



**In summary**

* **Addition rule** applies to mutually exclusive events.
* **Multiplication rule** applies to independent events.

## Video 2.5: Conditional Probability

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

**Conditional probability is used in**

* Finance
* Insurance
* Science
* Machine Learning

**Dependent events:** Two events are dependent if the occurrence of one event changes the probability of another event.

**Example**: To travel to another country, you first must get a passport, to access a webpage you first must have internet access. The second event depends on the outcome of the first event.

**Example:** Chance of getting an ace from a deck of cards (4/52). Second, chance of getting a second ace (3/51), the probability chances because we have already pulled an ace, thus it is less likely.

A white background with blue text

Description automatically generatedA white background with blue text

Description automatically generated

A white background with blue text

Description automatically generated

## Video 2.6: Baye’s Theorem

A math equation with blue text

Description automatically generated

**Prior probability**: The probability of an event before new data is collected.

**Example**: For example, let's say a medical condition is related to age. You can use Bayes' 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're in a certain age group. Bayes' theorem is the foundation for the field of **Bayesian statistics**, also known as Bayesian inference, which is a powerful method for analyzing and interpreting data in modern data analytics.

A white background with blue text

Description automatically generated**Bayes’ Theorem applications**

* Artificial intelligence
* Medical testing
* Financial institutions
* Online retailers
* Marketers

**Example:** Say you're planning a big outdoor event like a graduation party. The success of the event depends on good weather. On the day of the event, you notice that the morning is cloudy. You want to find out the chance of rain, given that this day starts off cloudy. If there's a high probability of rain, you may decide to move the event indoors or even cancel it. You know the following information: at this time of year, the **overall chance of rain is 10 percent**. However, cloudy mornings are common. About **40 percent of all days start off cloudy** and **50 percent of all rainy days start off cloudy**. In this example, your prior probability is the overall probability of a rainy day. New data will update this probability, in this case, the knowledge that the morning is cloudy, and that rain may be coming. What you ultimately want to find out is the probability that it will rain given that it's cloudy. This is your posterior probability.

A diagram of a weather forecast

Description automatically generated with medium confidenceA math equations on a white background

Description automatically generated

A math equations on a white background

Description automatically generated

## Video 2.7: The expanded version of Bayes’ theorem

A close-up of a text

Description automatically generated

**Bayes’ Theorem is used to evaluate**

* Medical diagnostic test
* Quality control tests
* Software tests

## Video 2.8: Introduction to Probability Distributions

**Probability distribution**: Describes the likelihood of the possible outcomes of a random event.

**Random variable:** Represents the values for the possible outcomes of a random event.

**Random variable:**

* Discrete
* Continuous

**Discrete random variable**: Has a countable number of possible values.

* A discrete random variable has a countable number of possible values. Often discrete variables are whole numbers that can be counted.

**Example:** If you roll a die five times you can count the number of times the die lands on two. If you toss a coin five times you can count the number of times it lands on heads.

**Continuous random variable**: Takes all the possible values in some range of numbers.

* When it comes to continuous variables, you're dealing with decimal values rather than whole numbers.

**Example**: All the decimals’ values between one and two, such as 1.1, 1.12, 1.125 and so on.

* These values are not countable since there is no limit to the possible number of decimal values between one and two.

**Discrete or continuous variables**

* COUNT the number of outcomes = discrete
* MEASURE the outcome = continuous

**Discrete or continuous distributions**

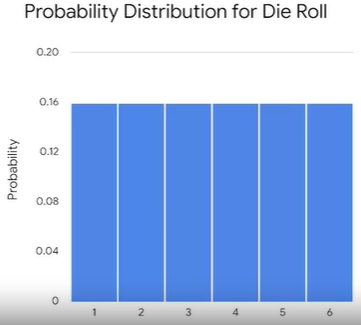
* Discrete distributions represent random variables
* Continuous distributions represent continuous random variables

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

**Example**: A sample space for a single coin toss = {Heads, Tails}

**Example**: A sample space for single die roll = {1, 2, 3, 4, 5, 6}

A blue squares with black numbers

Description automatically generated

**Example**: A continuous random variable may have an infinite number of possible values. Imagine you want to measure the height of an Oak tree you picked at random from a nearby forest. In this example, the height of the tree is a continuous random variable. The tree's height could be say 15 ft or 15.2 ft or 15.2187 ft and so on. You can keep on adding another decimal place to the measurement without limit. Now say you want to know the probability that the height of the oak tree is exactly 15.2 ft. Because the height of the tree could be any decimal value between the range of 15 ft and 16 ft. The probability that the tree is exactly any single value is essentially zero. In this example you'll need to use a continuous probability distribution to tell you the probability that the height of the oak tree is in a certain range or interval.

A cartoon trees with numbers and a ruler

Description automatically generatedA graph of a function of tree heifer

Description automatically generated

## Video 2.9: The Binomial Distribution

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

**This definition assumes**

* Each event is independent
* The probability of success is the same for each event
* The outcomes must be mutually exclusive

**Example:** The binomial distribution applies to an event like tossing the same coin 10 times in a row. Keep in mind that success and failure are labels used for convenience. For example, each toss has 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.

A screenshot of a computer

Description automatically generated**Binomial experiment**

* 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
* Each trial is independent

A number of equations on a white background

Description automatically generated

## Video 2.10: The Poisson Distribution

**Poisson Distribution:** Models the probability that a certain number of events will occur during a specific time period, area, or volume.

**Data professionals use the Poisson distribution to model data such as**

* Calls per hour for a customer service call center
* Visitors per hour for a website
* Customers per day at a restaurant
* Severe storms per month in a city

**Poisson experiment**

* 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

**Example:** Imagine you're a data professional working for a large restaurant chain that serves fast food. You know that the drive-through service at a restaurant receives an average of two orders per minute. You want to determine the probability that a restaurant will receive a certain number of orders in a given minute.

* This is a Poisson experiment because the number of events in the experiment can be counted. You can count the number of orders.
* The mean number of events that occur during a specific time period is known. There is an average of two orders per minute, each outcome is independent.
* The probability of one person placing an order does not affect the probability of another person placing an order.

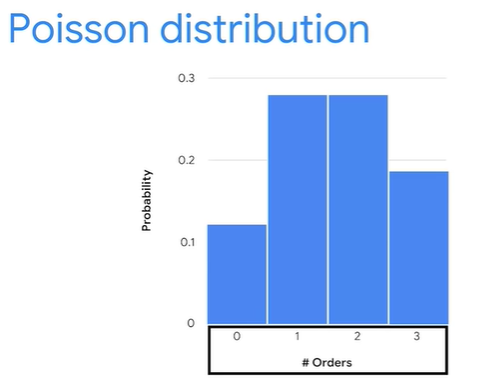
Once you know that you're working with the Poisson distribution, you can apply the Poisson distribution formula to calculate the probability.

A white text with blue letters

Description automatically generated

Let’s apply the Poisson distribution to our restaurant orders

**A math equations and numbers

Description automatically generated**

**Comparison**

**A diagram of a number of events

Description automatically generated**

## Video 2.11: The Normal Distribution

**Normal Distribution**: A continuous probability distribution that is symmetrical on both sides of the mean and bell-shaped (Gaussian)

**Normal distribution**

* Height
* Weight
* Blood pressure
* IQ scores
* Salaries

**A yellow and orange curve with Devils Tower in the background

Description automatically generated with medium confidence**

**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 center
* The total area under the curve equals 1

**Example:** The weight of apples

A diagram of a weight distribution

Description automatically generated

**The empirical rule**

* 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

The empirical rule is useful for estimating data, especially for large data sets like height and weight data for an entire population. You can use the empirical rule to get an initial estimate of the distribution of values in your data set such as what percentage of values will fall within one, two, or three standard deviations of the mean. This saves time and helps you better understand your data. Plus, knowing the location of your values on a normal distribution is useful for detecting outliers

## Video 2.12: Standardize Data Using Z-scores

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

* The z-score is 0 if the value is equal to the mean
* The z-score is positive if the value is greater than the mean
* The z-score is negative if the value is less than the mean

Z-scores help you standardize your data.

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

A diagram of a normal distribution

Description automatically generated

Standardization is useful because it lets you compare scores from different data sets that may have different units, mean values and standard deviations. Data professionals use z-scores to better understand the relationship between data values within a single data set and between different data sets.

For example, data professionals often use z-scores for anomaly detection, which finds outliers in datasets. Applications of anomaly detection include finding fraud in financial transactions, flaws in manufacturing products, intrusions in computer networks and more.

**Example**: Different customer satisfaction surveys may have different rating scales. One survey could rate a product or service from 1 to 20, another from 500 to 1,500, and a third from 130 to 180. Let's say the same product got a score of 9 on the first survey, 850 on the second and 142 on the third. These numbers don't mean much by themselves, but if you know, they all have a z-score of 1, or one standard deviation above the mean, you can meaningfully compare ratings across surveys.

A diagram of a mathematical equation

Description automatically generated

## Glossary Terms from Module 2

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

# Module 3: Sampling

In this section of the course, you’ll learn about the concept of sampling and its applications in data work. The section begins with an overview of inferential statistics and the relationship between sample and population. This is followed by a summary of the sampling process, and the benefits and drawbacks of specific sampling methods. You’ll also learn about sampling distributions for both means and proportions.

## Video 3.1: Introduction to Sampling

**Sampling**: The process of drawing a subset of data from a population.

**Questions answered by sampling**

* How many products in an app store do we need to test to feel confident that all the products are secure from malware?
* How do we select a sample of users to run an effective A/B test for an online retail store?
* How do we select a sample of customers of a video streaming service to get reliable feedback on the shows they watch?

**Sampling is useful because**

* Requires less time
* Saves money and resources
* More practical than analyzing an entire population

**A diagram of a sample

Description automatically generated**

Let’s keep in mind that your sample should be representative of your population.

**Representative sample**: Accurately reflects the characteristics of a population.

Since your inferences and predictions are based on your sample data, a non-representative dataset will not accurately reflect your population and will be subject to negative outcomes.

**Example:** Let’s say that you want to know how many residents use a computer in X city. If your sample group which only includes data scientists, your sample will not be reflective of the entire population as data professionals are very likely to have a computer to do their work. A representative sample would include people with different levels of computer knowledge and access.

**Example**: Let's consider another scenario. Imagine you want to find out the average height of every adult in the United States, that's a lot of people. It would take an incredible amount of time, energy, and money to even attempt to measure every person in the country. Instead, you can take a sample of 100 people and use that sample data to draw conclusions about the entire population. Now, let's say you have sample data only from professional basketball players. Pro basketball players are really tall, some are over seven feet tall. On average, they're much taller than almost everybody else in the population. Their average height does not accurately reflect the average height of the overall population. A sample that includes only pro basketball players is not representative of every adult in the US.

A pie chart with a basketball in the center

Description automatically generatedA person and a child standing next to a measuring tape

Description automatically generated

**Having a representative sample is super important**

A good model can’t overcome a bad sample.

Data professionals work with powerful statistical tools that can model complex datasets and help generate valuable insights. But if the sample data you're working with does not accurately reflect your population, that is if your sample is not representative, then it ultimately doesn't matter how good your model is.

## Video 3.2: The Sampling Process

The sampling process helps determine whether your sample is representative of the population and whether your sample is unbiased.

**A diagram of steps

Description automatically generated**

**Example**: Let's consider a public opinion poll. Imagine the city government of Vancouver, Canada wants to build a new subway system. The public will vote on whether or not to move forward with the project. The city government wants to find out if there's public support for the project. They ask you to take a poll and estimate the percentage of adult residents that support the project. Legal adults are 18 years or older. The first stage in the sampling process is to identify your target population.

The **target population** is the complete set of elements that you're interested in knowing more about.

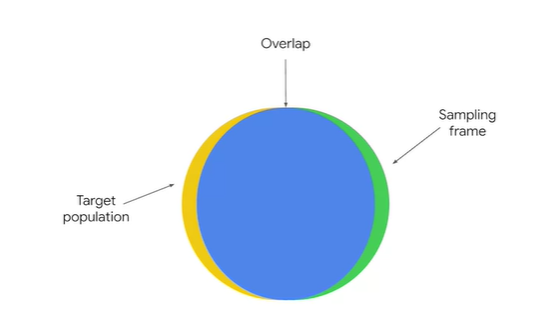
A blue circle with white text

Description automatically generated**Step 1 - Target population**: The target population includes every resident in the city who is 18 years or older and eligible to vote.

**Step 2 - Sampling frame**: A list of all the items in your target population.

The difference between target population is **general** and the sampling frame is **specific**.

The sampling frame will not entirely match your target population, this is because the city might not have complete records for the entire population that is able to vote.



**In other words, your sampling frame is the accessible part of your target population.**

Next is to choose the right sampling method.

**Two main sampling methods**

* Probability sampling
* Non-probability sampling

**Probability sampling**: Uses random selection to generate a sample.

**Non-probability sampling**: Based on convenience or personal preference.

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

Because probability sampling methods are based on random selection, every person 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.

**Step 3 – Sampling Method**: Assuming you have the budget and the time, you can use a probability sampling method for your poll about the subway project.

Sample size helps determine the accuracy of the predictions you make about the population. In general, the larger the sample size, the more accurate your predictions.

**Step 4 – Sample Size**: Based on the desired level of accuracy for your survey, you can decide how many eligible voters to include in your sample.

**Step 5 – Collect the Sample Data**: To pull the residents selected for your sample, you decide to conduct a survey. Based on the survey responses, you determine the percentage of eligible voters 18 and over who favor the proposed subject project. Then, you share this information with city leaders to help them make a more informed decision.

Effective sampling ensures that your sample data is representative of your target population. Then, when you use sample data to make inferences about the population, you can be reasonably confident that your inferences are reliable. Your poll will give city leaders a better idea of public support for the new subway and help inform future decisions about the project.

## Video 3.3: Compare sampling methods

**Probability sampling methods**

* Simple random sampling
* Stratified random sampling
* Cluster random sampling
* Systematic random sampling

A group of blue people

Description automatically generated**Simple random sampling**: Every member of a population is selected randomly and has an equal chance of being chosen.

**Example**: Job satisfaction poll among employees in a company.

**Simple random sample**

* Fairly representative (since every member of the population has an equal chance of being chosen)
* A group of blue people

  Description automatically generatedAvoid bias

**Stratified random sample**: Divide a population into groups and randomly select some members from each group to be in a sample.

Strata, groups organized by age, gender, sex, or any category you are interested in studying.

**Example**: Let’s say you want to survey high school students about how much time they spend studying on weekends. You might divide the student population according to age: 14, 15, 16, and 17-year-olds. Then you can survey an equal number of students from each age group.

**Stratified random sample**

* Members from each group are included.

A group of blue people

Description automatically generatedOne con about stratified random sampling is that you might strata difficult if you lack knowledge over a population.

**Cluster random sample**: Divide a population into clusters, randomly select certain clusters, and include ALL members from the chosen cluster in the sample.

Cluster sampling is similar to stratified sampling, but the main difference is that some sample groups will be left out while stratified sampling makes sure to consider every single group.

**Example**: For example, imagine you want to conduct a survey of employees at a global company using cluster sampling. The company has 10 offices in different cities around the world. Each office has about the same number of employees and similar job roles. You randomly select three offices in three different cities as clusters. You include all the employees at the three offices 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.

**Cluster sample**

* Helpful when dealing with large and diverse populations that have clearly defined subgroups.

One con is that it might be difficult to clearly define groups in a large diverse population.

A group of blue people

Description automatically generated**Systematic random sample**: 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.

**Example**: Let's assume you want to survey students at a community college. For a systematic random sample, you'd put the students' names in alphabetical order, randomly choose a starting point, and pick every fifth name to be in the sample. Systematic random samples are 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 B or R isn't going to affect their characteristics.

**Systematic random sampling**

* Representative
* Quick and convenient when you know the information of your population.

One con is that you need to know the size of your population you want to study before you begin, if you don’t know this exactly, it might be hard to get consistent intervals.

**ALL OF THE PREVIOUS SAMPLING METHODS ARE BASED ON RANDOM SELECTION**

## Video 3.4: The Impact of Bias Sampling

**Sampling Bias**: When a sample is not representative of a population as a whole.

Nonprobability sampling is often less expensive and more convenient for researchers to conduct.

**Non-probability sampling methods**

* Convenience sampling
* Voluntary response sampling
* Snowball sampling
* Purposive sampling

**Convenience sampling**: Choose members of a population that are easy to contact or reach.

Example: For example, to conduct an opinion poll, a researcher might stand in front of a local high school during the day and poll people that happened to walk by. Because these symbols are based on convenience to the researcher and not a broader sample of the population, convenience samples often show undercoverage bias. Undercoverage bias occurs when some members of a population are inadequately represented in the sample. For instance, people who don't work at or attend the school will not be represented as much in this sample.

**Undercoverage bias**: When some members of a population are inadequately represented in the sample.

**Voluntary response sample**: Consists of members of a population who volunteer to participate in a study.

**Example**: The owners of a restaurant want to know how people feel about their dinner options. They ask their regular customers to take an online survey about the quality of the restaurant's food. 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. This makes the volunteer customers at the restaurant an unrepresentative sample.

**Nonresponse bias:** When certain groups of people are less likely to provide answers.

**Snowball sampling**: Researchers recruit initial participants to be in a study and then ask them to recruit other people to participate in the study.

**Example**: For example, if a study was investigating cheating among college students, potential participants might not want to come forward. But if a researcher can find a couple of students willing to participate, these two students may know others who have also cheated on exams. The initial participants could then recruit others by sharing the benefits of the study and reassuring them of confidentiality.

**Purposive sampling**: Researchers select participants based on the purpose of their study.

**Example**: For example, a researcher wants to survey students on the effectiveness of certain teaching methods at the university. The researcher only wants to include students who regularly attend class and have an established record of academic achievement. So, they select the students with the highest-grade point averages to participate in the study. Purposive sampling, the researcher often intentionally excludes 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 students who don't have high grade point averages.

## Video 3.5: How Sampling Affects Your Data

**Statistic vs. parameter**

* The mean weight of a random sample of 100 penguins is a **statistic**.
* The mean weight of the total population is a **parameter**.

**Point Estimate**: Uses a single value to estimate a population parameter.

**Sampling distribution**: A probability distribution of a sample statistic.

A group of penguins in a circle

Description automatically generated

For your first sample, you find the mean weight of the 10 penguins is 3.1 pounds. For your second sample, the mean weight of the 10 penguins is 2.9 pounds. For your third sample, the mean weight is 2.8 pounds, and so on. Imagine that the true mean weight of a penguin in this population is three pounds. Although in practice, you wouldn't know this unless you weighed every single penguin. Each time you take a sample of 10 penguins, it's likely that the mean weight of the penguins in your sample will be close to the population mean of three pounds, but not exactly 3 pounds. Every once in a while, you may get a sample full of smaller than average penguins with a mean weight of 2.5 pounds or less. Or you might get a simple full of larger than average penguins with a mean weight of 3.5 pounds or more.

**Sampling variability**: How much an estimate varies between samples.

We can use a sampling distribution to represent the frequency of all sample means.

A graph of a red and blue bar chart

Description automatically generated with medium confidence

As you increase your sample size, the mean of the sample data will get closer to the mean weight of the population.

**If your sample is large enough, your sample mean will be closer to the mean of the population.**

**Standard error**

* Large standard error = Sample means are more spread out
* Smaller standard error = Sample means are closer together

A graph with green and blue squares

Description automatically generated

A close-up of a white background

Description automatically generatedA math equation with blue text

Description automatically generated

## Video 3.6: The Central Limit Theorem

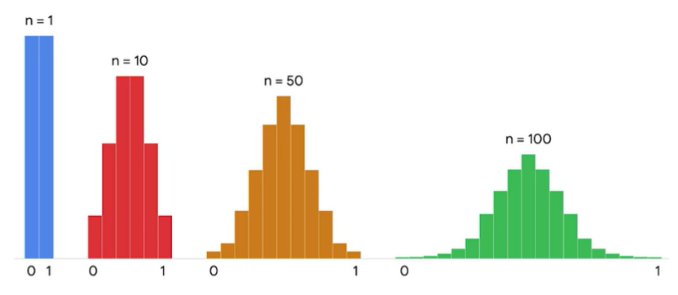
**The Central Limit Theorem can be used to estimate**

* The mean annual household income for an entire city or country
* The mean height and weight for an entire animal or plant population
* The mean commute time for all the employees of a large corporation

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

If you collect sufficient information, your data should follow a normal distribution.

A diagram of a coffee drink

Description automatically generated

**Example**: Imagine you're studying the population of coffee drinkers in the United States. You want to know the average amount of coffee each person drinks per day, but you don't have the time or money to survey every single coffee drinker in the US, which, by the way, is around 150 million people. Instead of surveying the entire population, you collect repeated random samples of up to 100 coffee drinkers. Using this data, you calculate the mean amount of coffee consumed per day for your first sample, 22.5 ounces. For your second sample, the mean amount is 28.2 ounces. You take a third sample; the mean amount is 25.4 ounces and so on. In theory, you could take 10, 50 or 100 samples and keep increasing the sample size until you've surveyed all 150 million people about their coffee consumption.

The central limit theorem says that as your sample size increases, the shape of your sampling distribution will increasingly resemble a bell curve. In practice, this specific sample size you choose will depend on factors like budget, time, resources, and the desire level of confidence for your estimate. If you take a large enough sample from the population, the mean of your sampling distribution will equal the population mean.

## Glossary Terms from Module 3

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

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

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

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

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

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

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

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

**Population**: Every possible element that someone is interested in measuring.

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

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

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

**Random seed**: A starting point for generating random numbers.

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

**Sample**: A subset of a population.

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

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

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

**Sampling distribution**: A probability distribution of a sample statistic.

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

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

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

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

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

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

**Standard error**: The standard deviation of a sample statistic.

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

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

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

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

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

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

# Module 4: Confidence intervals

This section explores how data professionals use confidence intervals to describe the uncertainty in their estimates. First, you’ll get an overview of the general procedure for constructing a confidence interval. This is followed by an explanation of how to properly interpret a confidence interval. Finally, you’ll be presented with detailed examples of how to construct a confidence interval for both means and proportions.

## Video 4.1: Confidence Intervals

**Point estimate**: Uses a single value to estimate a population parameter.

**Interval estimate**: Uses a range of values to estimate a population parameter.

**Confidence intervals**

* Penguin weight: 95% CI[28,32]
* Voters: 99% CI [51,57]

A single estimate can be useful, but it does not reflect uncertainty. This uncertainty is due to the random sampling.

**Example:** You may happen to weigh a sample of penguins that have recently struggled to find food. They only weigh 28 pounds, or you may weigh a sample of penguins that recently fed on a fish buffet and are above average, at 32 pounds. Either way, your sample estimate will not equal the population mean of 31 pounds. If you only provide a sample statistic or point estimate, it won't be as accurate.

**Confidence intervals give data professionals a way to express the uncertainty caused by randomness and provide a more reliable estimate. Along with the sample statistic, a confidence interval includes a margin of error and a confidence level.**

**Main components of a confidence interval**

* Sample statistic
* Margin of error
* Confidence level

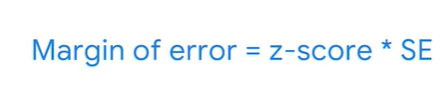
Let’s explore the idea with our penguin’s example

* The sample mean weight of penguins is 30 pounds.

**Interval**: sample statistic +/- margin of error

**Margin of error**: The maximum expected difference between the population parameter and a sample estimate. (This is the amount that a data professional expects their estimate might vary from the actual amount)

A white background with blue text

Description automatically generated

Sample statistic = 30, Margin of error = 2, Lower limit = 28 Upper limit = 32

**Confidence level** = Describes the likelihood that a particular sampling method will produce a confidence interval that includes the population parameter.

**Common confidence intervals**

* 90%
* 95% (Most popular choice)
* 99%

**Example**: Imagine you're a data professional working for a fashion company. Your manager asks you to estimate sales revenue for the new line of spring clothing. When you meet with stakeholders, you might say, "I think we'll do $1 million in sales," or you might say, "**Based on a 95% confidence level, I estimate that our sales revenue will be between $950,000 and $1,050,000.”**

First Statement

* $1,000,000

Second Statement

* 95% CI [$950,000, $,050,000]

The first statement offers a point estimate. The second statement provides a confidence level and an interval estimate and communicates the uncertainty in the estimate. It gives your stakeholders more information and helps them make more informed decisions about issues related to future sales revenue. As a data professional, you also have to make sure your stakeholders understand your results, so it's your job to clearly communicate how to interpret a confidence interval.

## Video 4.2: Interpret Confidence Intervals

Recall that data professionals use confidence intervals to express the uncertainty in their results.

A screenshot of a data analysis

Description automatically generated**Example**: Imagine you are re data professional who works for an urban planning company in a large city. The city government asked your team to design new parks and walkways that feature red maple trees. For planning purposes, your manager asks you to estimate the mean height of all the red maple trees in the city, that's approximately 10,000 trees. Instead of measuring every single tree, you collect a sample of 50 trees. The mean height of the sample is 50ft with a standard deviation of 7.5ft. Based on a 95% confidence level, you calculate a confidence interval for mean height that stretches between 48ft and 52ft. This interval estimate will help your team design new parks and walkways that meet city ordinances for landscaping.

**What does it mean to have 95% confidence interval?**

* 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.

Confidence intervals give data professionals a way to quantify the uncertainty due to random sampling.

**In our example Red maple tree data**

* You have a 95% confidence interval that the mean height is between 48 and 52 ft.
* For the purpose of this example, let's say the actual mean height of all, 10,000 red maples is 51 ft.

A diagram of a graph

Description automatically generatedIn practice, you would have no way of knowing this unless you measured every single tree in the city. This means that if you were to take 20 random samples of 50 trees and calculate a confidence interval for each sample. You can expect 19 out of your 20 intervals or 95% of the total will capture the population mean of 51 ft. One such interval will be the range of values between 48 and 52 ft.

A diagram of trees with text and images

Description automatically generated**Example**: Imagine you take another 20 random samples of 50 trees using the same sampling method. Because each sample is randomly selected from a large population, the meaning will vary from one sample to the next. Remember this is called **sampling variability**. For your first sample of 50 trees, the mean height is 50 ft. For your second sample of 50 trees, the mean height turns out to be 49.5 ft. For your third sample you get a mean height of 51.5 ft and so on. **Because of sampling variability, the mean height for any given sample will not necessarily be equal to the actual population mean. Confidence intervals help express this uncertainty**.

**Three common CI misinterpretations**

* The first common misinterpretation of confidence intervals is that a 95% confidence interval means that 95% of all the data values in your data set fall within the interval. This is not necessarily true. For example, your 95% confidence interval for tree heights is between 48 ft and 52. ft It may not be accurate to say that 95% of all the values in your data set fall in this interval. It's possible that over 5% of the tree heights in your dataset are outside this interval either shorter than 48 ft or taller than 52 ft.
* The second 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 tree height is between 48 ft and 52 ft. Imagine you take repeated samples using the same sampling method. It's possible that over 5% of your sample means will be less than 48 ft or greater than 52 ft.
* The third common misinterpretation is to assume that a confidence interval refers to the only possible source of error in your results. While every confidence interval includes a margin of error, many other kinds of errors can be entered into statistical analysis.

**So, when you're interpreting a confidence interval, remember that the uncertainty lies in an estimation process based on random sampling. A 95% confidence level refers to the success rate of that process. In other words, you can expect 95% of the random intervals you generate to capture the population Parameter.**

## Video 4.3: Construct a Confidence Interval for a Proportion

## Video 4.4: Construct a Confidence Interval for a Mean

# Module 5: Introduction to hypothesis testing

This section of the course describes how data professionals use hypothesis testing to help determine whether their results are statistically significant. First, you’ll get an overview of the general procedure for conducting a hypothesis test, and guidelines for interpreting the results. Then, you’ll examine detailed examples of how to conduct both one-sample and two-sample tests.

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

## Video 5.1: Introduction to Hypothesis Testing

**Example**: For a clinical trial of a new medicine, a hypothesis test can help you determine if the effect of the medicine on an average recovery time of your sample group is statistically significant or due to chance.

**Steps for performing a hypothesis test**

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

**A person's hands holding a pen and a piece of paper

Description automatically generatedExample:** You are given a coin to use in a game. You're not sure if the coin is fair or rigged. That is, you don't know if it's a standard coin or if it's been specifically waited to affect the outcome of a toss, for example, to always land on tails. Before using it in the game, you want to find out whether the coin is fair or not. You decide to test the coin by tossing it six times in a row and recording the outcomes. As you may recall from our earlier discussion of probability:

* if the coin is fair, the chance of landing on heads or tails is 0.5 or 50 percent for any given toss.
* If the coin is rigged for tails, the chance of landing on tails for any given toss will be much higher, perhaps 90 or even 100 percent.

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

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

**1- Hypothesis**

* **Null:** The coin is fair
* **Alternative**: The coin is not fair

**Significance level**: The probability of rejecting the null hypothesis when true

**2-Significance level**

* 5% is commonly used, it can be adjusted depending on the type of study

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

**3-P-Value**

* Let’s say that we calculate the probability of a fair coin landing tails 6 times (0.5^6 = 1.56%)
* If we assume that the null hypothesis is true and the coin is fair, the P-value in our example is 1.56%. Anything lower than that would mean that there is stronger evidence for the alternative hypothesis. Remember, your alternative hypothesis is that the coin is not fair.

**A lower p-value means there is stronger evidence for the alternative hypothesis**

**4-Decide**

* **Reject** or **fail to reject** the null hypothesis.

Statisticians always say, fail to reject rather than except. This is because hypothesis tests are based on probability, not certainty. Acceptance implies certainty. In general, as data professionals, we try not to claim certainty about results based on statistical methods.

A close-up of a white background

Description automatically generated

In our example

P-value: 1.56%

Significance level: 5%

1.56% < 5%

Conclusion: We reject the null hypothesis

**The significance level plays a major role in hypothesis testing.**

A statistically significant result cannot prove with 100 percent certainty that our hypothesis is correct. 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.

**Types of errors in hypothesis testing**

* Type I error (false positive)
* Type II error (false negative)

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

In our example this can be if we conclude that the coin is fair when it’s actually rigged.

**To minimize the risk of a type I error, choose a significance level of 1%.**

## Video 5.2: One-sample test for means

**One-sample test**: Determines whether or not a population parameter like a mean or proportion is equal to a specific value.

**Two-sample test:** Determines whether or not two population parameters such as two means or two proportions are equal to each other.

**One-sample hypothesis test can be used to determine**

* A company's average sales revenue is equal to its target value
* A medical treatment's average rate of success is equal to a set goal
* A stock portfolio's average rate of return is equal to a market benchmark.

A close-up of a test

Description automatically generated

**Example**: Imagine you are a data professional working for an online delivery company.

* Population
  + **Mean =** 40 min
  + **Standard deviation =** 5 min

Recently, the company management launched a new training program to make the delivery process more efficient. After delivery drivers completed the training program, management tracked a random sample of 50 deliveries to understand how long a delivery takes.

* Sample
  + **Mean =** 38 min
  + **Standard deviation =** 5 min

There is an observed difference of two minutes between the population mean of 40 minutes and the sample mean of 38 minutes. The management team asks you to **determine** **if the decrease in average delivery time is statistically significant or if it's due to chance**. If the decrease is statistically significant, the company wants to invest in developing and implementing the training program in other regions. You decide to conduct a one-sample z-test to analyze the data.

**Recall the steps for performing a hypothesis test**

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

In a one-sample z-test, the null hypothesis states that the population mean is equal to an observed value.

**Null hypothesis**: The average delivery time equals 40 minutes.

In a one-sample test, there are three main options for the alternative hypothesis: the population mean is not equal to, less than, or greater than an observed value. In this case, you want to test whether the training has decreased the average delivery time.

**Alternative hypothesis:** The average delivery time is less than 40 minutes

**Significance level**: 5%

**P-value**:

If the p-value < 5%, then **reject** the null hypothesis

A diagram of a normal distribution

Description automatically generatedThe p-value is from a **test statistic**,a value that shows how closely your observed data matches the distribution expected under the null hypothesis.

If you assume the null hypothesis is true and the mean delivery time is 40 minutes, the data for delivery times follows a normal distribution. The test statistic shows where your observed data, a sample mean delivery time of 38 minutes, will fall on that distribution.

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

Z-scores tell you where your values lie on a normal distribution. The following formula gives you a test statistic z based on your sample data:

* A mathematical equation with numbers and symbols

  Description automatically generatedX-bar is the sample mean
* Mu is the population mean
* Sigma is the population standard deviation
* n is the sample size.

If we plug in the numbers from our previous example, we get Z = -2.82

A diagram of a normal distribution

Description automatically generatedLet's check out where the z-score, negative 2.82, lies on the distribution. It's far to the left, almost three standard deviations below the mean. For a normal distribution, the probability of getting a value less than your z-score of -2.82 is calculated by taking the area under the 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. Again, your p-value is the probability of observing a test statistic as or more extreme than that observed when the null hypothesis is true.

Your alternative hypothesis states that the mean delivery time decreased based on your sample data. That's why we're interested in the probability of getting any value lower than your z-score of -2.82. In a different testing scenario, your test statistic might be positive 2.45, and you might be interested in values higher than the z-score 2.45. In that case, your p-value would be located on the right tail of the distribution, and you'll be conducting a right-tailed test.

If you calculate the **p-value**, you'll find that it's 0.0023, so your p-value is 0.0023 or 0.23 percent. This means there's a 0.23 percent probability that the difference in mean delivery time would be 2 minutes or greater if the null hypothesis is true.

**In other words, it's highly unlikely that their difference is due to chance.**

**Draw a conclusion**

* If the p-value < significance level: **reject** null hypothesis
* If the p-value > significance level: **fail to reject** the null hypothesis

**0.23% < 5%**

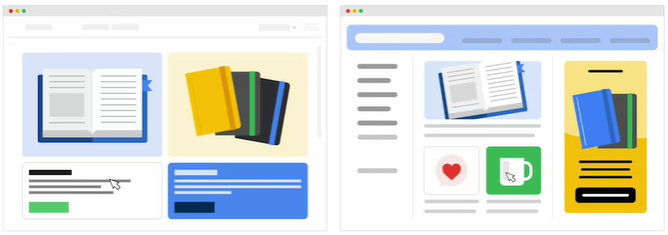
**We reject the null hypothesis and conclude that there is a statistically significant difference in mean delivery time.**

## Video 5.3: Two-sample test: Means

While a one-sample test determines whether a population mean is equal to a specific value, a two-sample test determines whether two population means are equal to each other.

Two sample tests are usually used in A/B testing

**Example**: For example, let's say an online retail store is considering changing the landing page for its reward called members who are the most loyal customers.



The metric that matters most to the company is the average time users spend on the landing page per session. First, I'd set up an experiment for two groups of users group A uses the default landing page and group B uses a redesigned version of the landing page. Then I'd use a t-test to compare the average time spent on each landing page and determine if the difference between the two-sample means is statistically significant. In other words, if group B spends more time on the landing page than group A, the t-test will help determine if that's due to chance, or to the new design of the landing page.

A diagram of a normal distribution

Description automatically generated

**Example**: Imagine you're a data professional who works for a cosmetics company. The company is researching the amount of time customers spend on its website. Your team leader asks you to conduct an A/B test to determine if changing the background color of the landing page from gray to green has any effect on the average time spent on the page.

A screenshot of a computer

Description automatically generated

You randomly select two groups of users. The first group visits the gray landing page named version A. The second group visits the green landing page named version B. You collect the following data from the A/B test.

**Version A**

* **Sample size** = 40
* **Sample mean** = 300 seconds
* **Sample standard deviation** = 18.5 seconds

**Version B**

* **Sample size** = 38
* **Sample mean** = 305 seconds
* **Sample standard deviation** = 16.7 seconds

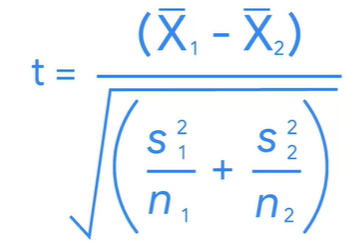
There is an observed 5 second difference between means. We decide to conduct a two-sample mean to analyze the data.

Let’s review the steps for conducting hypothesis testing

A white background with blue text

Description automatically generated

Significance level = 5%

P-value:

* X1 bar and X2 bar are the sample means of your two groups
* n1 and n2 are the sample sizes of your two groups
* S1 and S2 are the sample standard deviations of your two groups.

If you enter the numbers in the formula, and do the calculation, you get a test statistic t of -**1.2508**. For a t-test, the test statistic follows a t-distribution under the null hypothesis.

A diagram of a normal distribution

Description automatically generated

If you calculate the value, you get 21.48%.

**21.8% > 5%**

**This means that there is a 21.48% probability that the absolute difference between this mean time spent on version A and Version B would be five seconds or greater if the null hypothesis is true.**

A blue text on a white background

Description automatically generated

In other words, you fail to reject the null hypothesis that there is no difference in the mean time spent on version A and version B. Your p-value of 0.2148 or 21.48%, is greater than the significance level of 0.05 or 5%. **So, you failed to reject the null hypothesis and conclude that there is not a statistically significant difference between the mean time spent on version A and version B.** In other words, the observed difference in mean time spent is likely due to chance.

## Video 5.4: Two Sample test-Proportions

**Use a two-sample z-test to compare the proportion of**

* Defects among manufacturing products on two assembly lines.
* Side effects to a new medicine for two trial groups
* Support for a new law among voters in two districts

A globe with buildings and a world map

Description automatically generated with medium confidence**Example**: Imagine you're a data professional working for an international construction company. The company has offices in London and Beijing. The human resources team would like to determine whether there is a difference in the level of employee satisfaction between the Beijing office and the London office.

The team surveys a random sample of 50 employees in each office to discover if they are satisfied with their current job. They ask you to find out if there's a statistically significant difference in the proportion of satisfied employees in London and Beijing. If so, the HR team will devote resources to investigating why employees at one office are more satisfied at work.

According to the survey:

A screenshot of a graph

Description automatically generated

There is an observed 10% difference between job satisfaction. We decide to conduct a two-sample z-test to analyze the data.

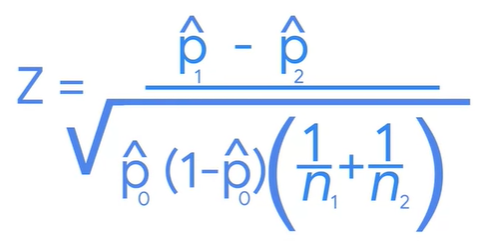
Recall the steps for conducting a hypothesis test.

A white background with blue text

Description automatically generated

Significance level = 5%

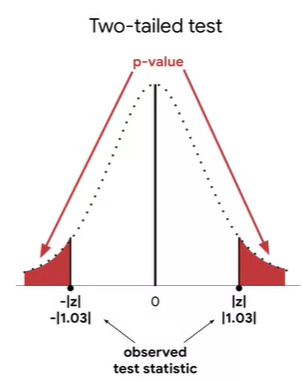
P-value:

 A number with numbers and symbols

Description automatically generated with medium confidence

* p\_1 hat and p\_2 hat are the sample proportions for your first and second group
* n\_1 and n\_2 are the sample sizes for your first and second group
* p\_0 hat is the pooled proportion. The pooled proportion is a weighted average of the proportions from your two samples.

If you plug in the numbers, you get **Z = 1.03**



If you find a statistically significant difference between the two proportions, either less than or greater than the observed difference of 10 percentage points, you will reject the null hypothesis. Because you're interested in values in both directions, either less than or greater than your test statistic, your p-value is the probability of getting a value less than the z-score of negative 1.03 or greater than the z-score positive 1.03. Your p-value corresponds to the area under the curve on both the left tail and the right tail of the distribution. This is a two-tailed test.

P-value : 30.3%

This means that there's a 30.3 percent probability that the absolute difference between the proportion of satisfied employees in London and Beijing would be 10 percent or greater if the null hypothesis is true.

**A blue text on a white background

Description automatically generated**

**30.3% > 5%**

**So, you fail to reject the null hypothesis and conclude that there is not a statistically significant difference between the proportion of satisfied employees in the London office and the Beijing office. In other words, the observed difference in proportions is likely due to chance.**

## Glossary Terms from Module 5

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

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

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

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

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

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

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

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

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

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

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

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

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

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

# Module 6: Course 4 end-of-course project

In this section of the course, you will produce a tangible artifact that you can add to your professional portfolio and present to future employers. For this project, you will use your knowledge of statistics to conduct a statistical test that’s based on a workplace scenario.

## Video 6.1: Introduction to the end-end-of-course portfolio project

**Course Overview:**

* Learn fundamental concepts of statistics.
* Cover descriptive and inferential statistics, basic probability, probability distributions, sampling, confidence intervals, and hypothesis testing.

**Portfolio Project:**

* Apply learned skills to simulate an A/B test for a specific company's product.
* Utilize statistical methods to analyze data and interpret findings.
* Develop storytelling skills acquired in a previous course to present the results effectively.

**Objective:**

* Make an argument about the validity of a product based on statistical analysis.
* Recommend whether to implement a new version of the product based on the A/B test results.

**Role of a Data Professional:**

* Emphasize the importance of data professionals in making informed business decisions.
* Highlight the use of stats to guide investments in company products or services.

**Future Learning:**

* Tease upcoming sections covering advanced techniques like regression analysis and machine learning.
* Emphasize continuous development to excel in the field of data analytics.

**Career Impact:**

* Demonstrate the role of A/B testing in helping employers or clients make informed decisions.
* Showcase the potential value as a data professional by providing actionable insights.

## 6.2: Explore your Corse 4 Workplace TikTok Scenario

A black text on a white background

Description automatically generated

**Project goal:**

The TikTok data team is developing a machine learning model for classifying claims made in videos submitted to the platform.

**Background:**

TikTok is the leading destination for short-form mobile video. The platform is built to help imaginations thrive. TikTok's mission is to create a place for inclusive, joyful, and authentic content–where people can safely discover, create, and connect.

**Scenario:**

The TikTok data team has successfully completed exploratory data analysis on the data for the claims classification project. The team is ready to begin the process of hypothesis testing. You’ve been asked to investigate TikTok's user claim dataset to determine which hypothesis testing method best serves the data and the claims classification project.

**Course 4 tasks:**

* Import relevant packages and TikTok data
* Explore the project data
* Implement a hypothesis test
* Communicate insights with stakeholders within TikTok

**Note*:*** The story, all names, characters, and incidents portrayed in this project are fictitious. No identification with actual persons (living or deceased) is intended or should be inferred. And, the data shared in this project has been created for pedagogical purposes.

## 6.3: Course 4 end-of-course portfolio project overview: TikTok

**Background on the TikTok scenario**

At TikTok, our mission is to inspire creativity and bring joy. Our employees lead with curiosity and move at the speed of culture. Combined with our company's flat structure, you'll be given dynamic opportunities to make a real impact on a rapidly expanding company and grow your career.

TikTok users have the ability to submit reports that identify videos and comments that contain user claims. These reports identify content that needs to be reviewed by moderators. The process generates a large number of user reports that are challenging to consider in a timely manner.

TikTok is working on the development of a predictive model that can determine whether a video contains a claim or offers an opinion. With a successful prediction model, TikTok can reduce the backlog of user reports and prioritize them more efficiently.

**Project background**

* Explore the project data
* Implement a hypothesis test
* Communicate insights with stakeholders

**Assignment**

Conduct hypothesis testing on the data for the claims classification data. You’ve been asked to investigate TikTok's user claim dataset to determine which hypothesis testing method best serves the data and the claims classification project.

**Team members at TikTok**

**Data team roles**

* Willow Jaffey- Data Science Lead
* Rosie Mae Bradshaw- Data Science Manager
* Orion Rainier- Data Scientist

The members of the data team at TikTok are well versed in data analysis and data science. Messages to these more technical coworkers should be concise and specific.

**Cross-functional team members**

* Mary Joanna Rodgers- Project Management Officer
* Margery Adebowale- Finance Lead, Americas
* Maika Abadi- Operations Lead

Your TikTok team includes several managers, who oversee operations. It is important to adjust your general correspondence appropriately to their roles, given that their responsibilities are less technical in nature.

**Specific project deliverables**

With this end-of-course project, you will gain valuable practice and apply your new skills as you complete the following:

* Course 4 PACE Strategy Document to consider questions, details, and action items for each stage of the project scenario
* Answer the questions in the Jupyter notebook project file
* Consider the different groups of data represented in the dataset
* Implement a hypothesis test
* Create an executive summary to share your results