Mathematical Foundations for Data Science

PRESENTED BY TANYA KHANNA

EMAIL: TK759@SCARLETMAIL.RUTGERS.EDU

COURSE MATERIALS: GITHUB LINK - HTTPS://GITHUB.COM/TANYA-KHANNA/DATASCIENCEWORKSHOP_2024_NBL

WORKSHOPS SCHEDULE

Introduction to Python Programming	February 1, 2024; 4 - 5:30 PM
Mastering Data Analysis: Pandas & Numpy	February 8, 2024; 4 - 5:30 PM
Python for Visualization & Exploration	February 15, 2024; 4 - 5:30 PM
Mathematical Foundations of Data Science	February 29, 2024; 4 - 5:30 PM
Introduction to Machine Learning: Supervised	March 7, 2024; 4 - 5:30 PM
Introduction to Machine Learning: Unsupervised	March 21, 2024; 4 - 5:30 PM
Introduction to Deep Learning	March 28, 2024; 4 - 5:30 PM
Deep Dive into Natural Language Processing	April 4, 2024; 4 - 5:30 PM
Large Language Models and Chat GPT	April 11, 2024; 4 - 5:30 PM

https://libcal.rutgers.edu/calendar/nblworkshops?cid=4537&t=d&d=0000-00-00&cal=4537&inc=0

TABLE OF CONTENTS

- WHAT IS DATA SCIENCE?
- WHAT IS MACHINE LEARNING?
- INTRODUCTION TO LINEAR REGRESSION
- FROM LINEAR EQUATIONS TO LINEAR REGRESSION
- STATISTICS FOR DATA SCIENCE
- PROBABILITY FOR DATA SCIENCE
- CALCULUS FOR LINEAR REGRESSION
- HYPOTHESIS TESTING

What is Data Science?

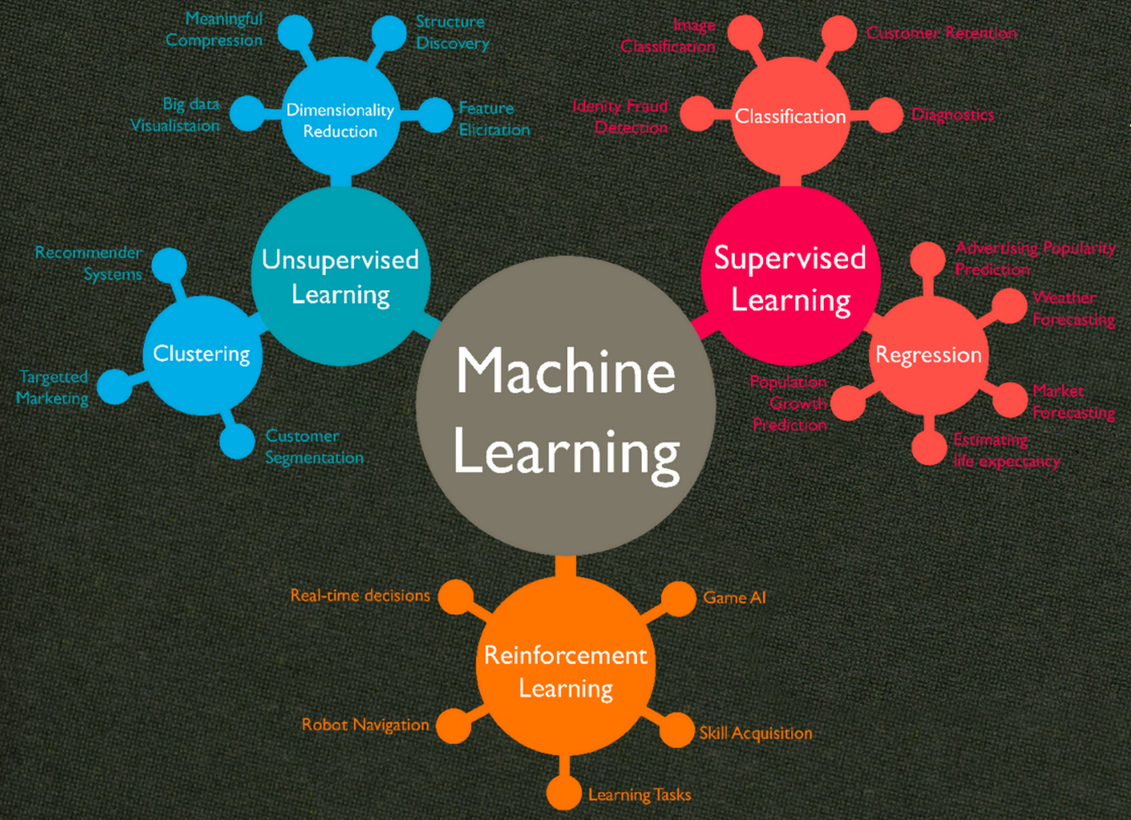
Think of Data Science as a way of uncovering hidden patterns and valuable information from data. Like being a detective, but instead of solving mysteries, you're finding insights from data to help solve real-world problems, make smarter decisions, and understand the world better.

Thus, Data Science is an interdisciplinary field that uses scientific methods, processes, algorithms, and systems to extract knowledge and insights from structured and unstructured data. It applies techniques from mathematics, statistics, computer science, and information science to analyze, interpret, and visualize data, aiming to make informed decisions.

Examples:

- Predicting which movies you'll enjoy based on what you've watched before.
- Companies understanding customer behavior to improve products.
- Health professionals using patient data to predict health outcomes and tailor treatments.

What is Machine Learning?



Imagine teaching your computer to make decisions and predictions, like a student learning from books. But instead of books, the computer learns from data. Over time, just like a student gets better with study, the computer gets better at making predictions and decisions based on the data it has learned from.

Machine Learning is a subset of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. It focuses on the development of computer programs that can access data and use it to learn for themselves.

Examples:

- Email filtering systems learning to block spam.
- Recommendation systems on shopping sites suggesting products you might like.
- Autonomous vehicles learning to navigate traffic.

Supervised vs. Unsupervised Learning

SUPERVISED LEARNING

- It's like teaching a child with flashcards. Each card has a picture and a name (label), and over time, the child learns to associate the picture with its correct name.
- In supervised learning, the computer is given data with labels to learn from, so it can make predictions or decisions based on new, unseen data.

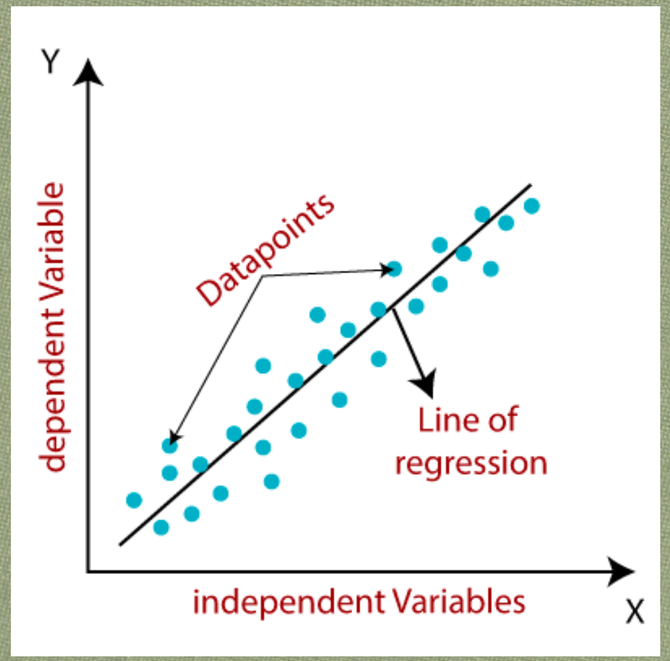
Example: Email spam filters. By being taught what spam looks like (emails labeled as spam), the filter learns to identify and separate future spam emails from your inbox.

UNSUPERVISED LEARNING

- This is like giving a child a box of mixed toys and asking them to sort them out without any instructions.
- The child might decide to group them by color, size, or type.
- In unsupervised learning, the computer tries to find patterns and organize the data on its own, without any labels or guidance.

Example: Customer segmentation in marketing. A business collects data on its customers' shopping habits without knowing in advance how to categorize them. Unsupervised learning algorithms can identify patterns in the data, grouping customers into segments with similar behaviors for targeted marketing campaigns.

Introduction to Linear Regression



Linear Regression is a fundamental statistical and machine learning method used to predict a dependent variable (outcome) based on one or more independent variables (predictors), assuming a linear relationship between them. It calculates the best-fit line or regression equation to describe the relationship between the variables.

Imagine you're trying to predict your final grade based on the number of hours you study. If you notice that the more hours you study, the higher your grade, you're observing a linear relationship. Linear Regression is like drawing the best straight line through all your study hours and grades to predict future grades based on how much you plan to study.

Examples:

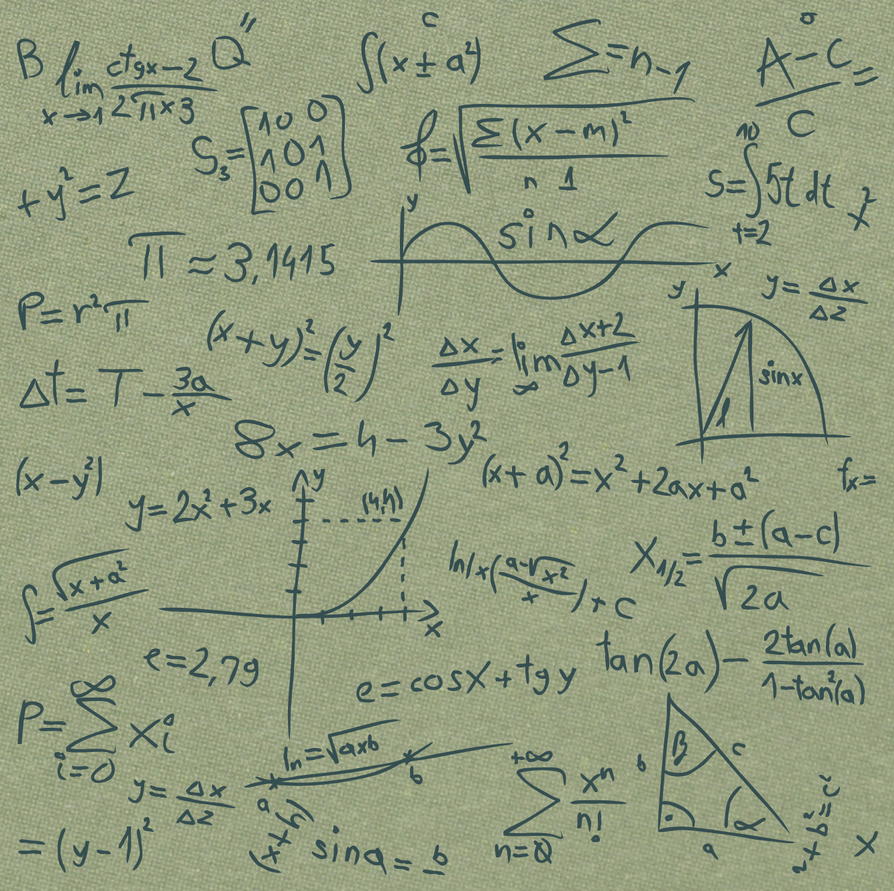
- Predicting house prices based on their size, location, and other features.
- Estimating a person's salary based on their education level and years of experience.
- Forecasting sales for a company based on advertising spend.

Essentials of Algebra

Algebra is a branch of mathematics that deals with **relationships** between **numbers** and **unknown values**, represented by letters called **variables**. It's like a language for expressing and solving problems in a general way, instead of just dealing with specific numbers.

Some key points about algebra:

- **Variables**: These are letters like x, y, and z that represent unknown values we're trying to find or that can take on different values.
- **Expressions**: These combine numbers, variables, and operations like addition, subtraction, multiplication, and division, to form a single mathematical statement.
- **Equations**: These are statements that two expressions are equal, often used to represent relationships between variables. Solving an equation involves manipulating it to find the value of the variable that makes the equation true.
- **Applications**: Algebra is used in countless fields, from science and engineering to economics and finance. It's a fundamental tool for modeling and analyzing real-world problems.



EXAMPLE 1: BASIC ALGEBRAIC EQUATION

EQUATION: $X+3=7$
OBJECTIVE: SOLVE FOR X.

SOLUTION:
TO FIND THE VALUE OF X, YOU SUBTRACT 3 FROM BOTH SIDES OF THE EQUATION TO ISOLATE X ON ONE SIDE:

$X+3-3=7-3$
$X=4$

SO, THE VARIABLE X IS EQUAL TO 4.

EXAMPLE 2: USING VARIABLES IN AN EQUATION

SCENARIO: SUPPOSE YOU ARE SAVING MONEY FOR A NEW BIKE. YOU SAVE $50 EVERY WEEK. HOW MANY WEEKS WILL IT TAKE TO SAVE $400?
OBJECTIVE: SOLVE FOR X.

SOLUTION:
LET W REPRESENT THE NUMBER OF WEEKS. THE TOTAL AMOUNT SAVED CAN BE REPRESENTED AS 50W.
SO, THE EQUATION IS: $50W = 400$.

TO FIND W, DIVIDE BOTH SIDES BY 50:
$50W/50 = 400/50$
$W = 8$

NOTE:

ALGEBRAIC EXPRESSION: $3X + 2$

THIS IS NOT AN EQUATION BECAUSE IT DOES NOT ASSERT THAT TWO THINGS ARE EQUAL. IT SIMPLY REPRESENTS A VALUE DEPENDING ON X.

Slope and Intercept in Linear Equations

The concepts of slope and intercept are fundamental in algebra, especially in the context of linear equations. A linear equation in two variables typically takes the form

$$y = mx + b$$

where

- y is the dependent variable,
- x is the independent variable,
- m is the slope of the line, and
- b is the y-intercept.

How to convert standard form to slope intercept form

It's very simple to convert between these 2 forms. Here's an example:

Let's take **3x+5y=15**

We need to convert it to y=mx+b, so first we need to isolate the y. How do we do this? By moving the x to the other side

Our equation now looks like this:
5y=15-3x

How to convert standard form to slope intercept form

To get y without the coefficient, we need to divide both sides per 5. That way, the Y will end up alone

Our equation is now **5y/5=(15-3x)/5**

That simplifies to **y=3-(3/5)x**, which if we rewrite it as **y=(-3/5)x+3**, it matches slope intercept form

If you want to convert an equation from slope intercept to standard, just do the steps we did, but in reverse!

What is a slope?

In mathematics, the slope of a line describes two important aspects: steepness and direction. It tells you:

How much the line slants: A positive slope indicates the line slants upwards from left to right, while a negative slope indicates it slants downwards. A slope of zero means the line is horizontal, and a slope of undefined means the line is vertical.
The amount of change in the y-value for every unit change in the x-value: This essentially captures the rate of change along the line.

Calculating Slope:

The slope (m) of a line can be calculated using the following formula:

$$m = \frac{y_2 - y_1}{x_2 - x_1}$$

y1 and x1 represent the coordinates of one point on the line.
y2 and x2 represent the coordinates of another point on the line.

Consider a line that passes through the points (2, 4) and (5, 7).

Using the formula:

m = (7 - 4) / (5 - 2)
m = 3 / 3
m = 1
This means the slope (m) is 1. In simpler terms, for every 1 unit you move up (positive y-direction), you also move 1 unit to the right (positive x-direction).

Understanding the slope allows you to interpret the direction and steepness of a line, which is crucial in various applications like:

Modeling real-world phenomena involving rates of change (e.g., velocity, cost vs. quantity)
Graphing lines with specific characteristics
Analyzing relationships between variables

What is an intercept?

The y-intercept of a line is the point where the line crosses the y-axis. In other words, it's the point on the line where the x-coordinate is zero.
Here are some key points about the y-intercept:

- Location: It always lies on the y-axis because the x-coordinate at this point is 0.
- Notation: It is often denoted by the letter b in the slope-intercept form of a linear equation: $y = mx + b$.
- Interpretation: The y-intercept tells you the initial y-value when the x-value is zero.

The y-intercept is a crucial element in understanding and manipulating linear equations. It helps you:

- Graph the line: Knowing the y-intercept (a point on the line) along with the slope allows you to easily graph the line.
- Interpret the equation: The y-intercept provides insight into the initial value of the dependent variable (y) when the independent variable (x) is zero.

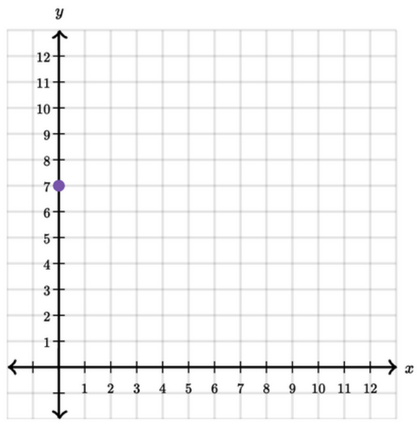
Given the equation $y=4x-5$:

- The slope (m) is 4, indicating a steep incline.
- The y-intercept (b) is -5, meaning the line crosses the y-axis at $y=-5$.

Graphing lines from slope-intercept form

For example, imagine we're given the equation y = 2x + 7 and asked to graph it.

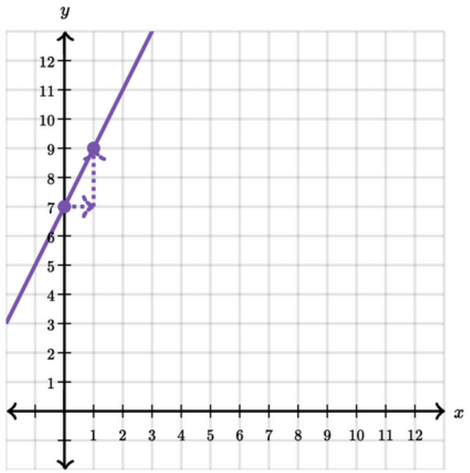
Straight from the equation, we know the y-intercept is 7.



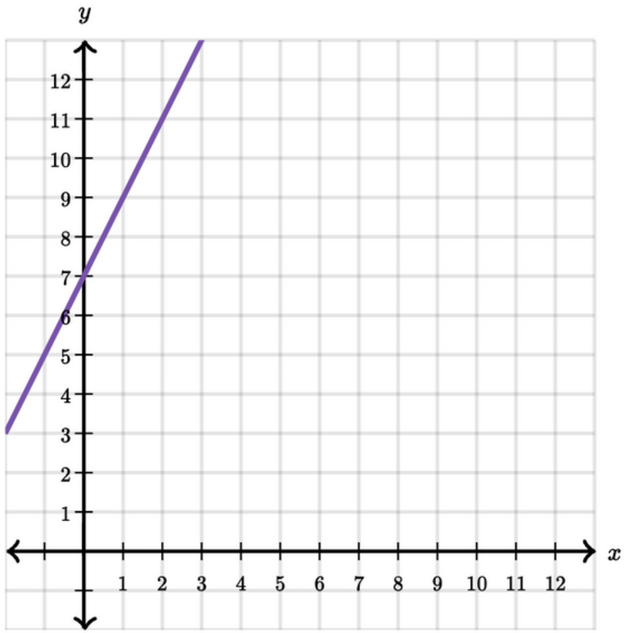
And we know the slope is 2.

$$\text{Slope} = \frac{\Delta y}{\Delta x} = \frac{2}{1} = 2$$

So, for every one unit we go to the right, we must go up two units:



Here is our final graph:



Scalars, Vectors, and Matrices

SCALAR

- A scalar is simply a single number.
- It's the most basic quantity we deal with in mathematics.
- You can think of a scalar as something that gives you information about 'how much' or 'how many'.
- It doesn't have any direction, just a magnitude.

Example of a Scalar:

Temperature: 72 degrees
Money: $100
Apples: 5 apples

VECTOR

- A vector is like an arrow. It has both a magnitude (how long the arrow is) and a direction (where the arrow points).
- In the world of data, vectors can be a list of scalars lined up in a row or a column.
- These scalars could represent different related measurements. For instance, if you're tracking your speed and direction while hiking, you might represent this as a vector: one number for speed, one for the direction you're moving in.

Example of a Vector:

- Speed and direction: 5 mph North-East
- In mathematical terms, a 2-dimensional vector for this could be written as $\begin{bmatrix} 5 \\ 45 \end{bmatrix}$, where '5' is the speed and '45' is the angle in degrees representing the North-East direction.

MATRICES

- Now, imagine you have several vectors, each representing a different hiker's speed and direction.
- If you wanted to keep all this information organized, you could arrange each hiker's vector in rows to form a grid.
- This grid is known as a matrix.
- Matrices are like spreadsheets for numbers, where each row could represent a different hiker and each column a different aspect of their journey, such as speed and direction.

Hiker data: We have three hikers, and we're tracking their speed and direction.

$$\begin{bmatrix} 5 & 45 \\ 3 & 90 \\ 4 & 135 \end{bmatrix}$$

Here, each row corresponds to a different hiker, the first number in each row is the speed, and the second number is the direction in degrees.

Matrix Operations

ADDITION AND SUBTRACTION

Matrices can be added or subtracted element by element if they have the same dimensions.

Example of Matrix Addition:

$$\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} + \begin{bmatrix} 5 & 6 \\ 7 & 8 \end{bmatrix} = \begin{bmatrix} 1+5 & 2+6 \\ 3+7 & 4+8 \end{bmatrix} = \begin{bmatrix} 6 & 8 \\ 10 & 12 \end{bmatrix}$$

MULTIPLICATION

The product of two matrices is not element by element. Instead, the element in the i-th row and j-th column of the product is obtained by multiplying each element of the i-th row of the first matrix by the corresponding element of the j-th column of the second matrix and then summing these products.

Example of Matrix Multiplication:

$$\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} \times \begin{bmatrix} 5 & 6 \\ 7 & 8 \end{bmatrix} = \begin{bmatrix} (1 \times 5 + 2 \times 7) & (1 \times 6 + 2 \times 8) \\ (3 \times 5 + 4 \times 7) & (3 \times 6 + 4 \times 8) \end{bmatrix} = \begin{bmatrix} 19 & 22 \\ 43 & 50 \end{bmatrix}$$

Inverse of a Matrix

The inverse of a matrix is like the reciprocal of a number. When a matrix is multiplied by its inverse, the result is the identity matrix. Not all matrices have inverses. A matrix must be square (same number of rows and columns) and its determinant must not be zero.

Example of Inverse Calculation for a 2×2 Matrix:

Given a 2×2 matrix:

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

The inverse is calculated as:

$$A^{-1} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$$

Transpose of a Matrix

$$\begin{bmatrix} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \\ 9 & 10 & 11 & 12 \end{bmatrix} = \begin{bmatrix} 1 & 5 & 9 \\ 2 & 6 & 10 \\ 3 & 7 & 11 \\ 4 & 8 & 12 \end{bmatrix}^{\mathrm{T}}$$

Linear Regression & Line of Best Fit

- Linear Regression is a statistical method that allows us to study and model the relationship between two continuous variables.
- Imagine you want to predict your score in an exam based on the number of hours you study.
- Linear regression would help us understand how your study hours might influence your score.
- If there's a direct relationship, as you increase your study hours, your expected score might also increase.

MATHEMATICALLY, LINEAR REGRESSION CAN BE REPRESENTED BY THE EQUATION OF A STRAIGHT LINE:

$$y = mx + b$$

where:

Y is the dependent variable we're trying to predict,
X is the independent variable we're using for the prediction,
m is the slope of the line, which represents the change in Y for a change in X,
b is the y-intercept, which is the value of Y when X is 0.

Let's say we have data on the number of hours 10 students studied for an exam and their respective scores:

In this example, we would use linear regression to predict the exam score based on the number of hours studied.
By finding the best-fitting line through the data points, we can model the relationship between study hours and exam scores.

The Line of Best Fit, also known as the regression line, is the straight line that best represents the data on a scatter plot. This line minimizes the distance between itself and all the points in the dataset, typically using a method called "least squares."

The equation of the line of best fit is similar to any linear equation,

Y=mX+b,
where m & b are calculated so that the line is as close as possible to all the data points.

The Line of Best Fit helps us understand the trend within the data. For instance, if we're looking at our study hours versus exam scores example, the line of best fit would show us how scores generally increase with more study hours. It helps in making predictions about the exam score for any given amount of study time.

Formulas for Slope and Intercept

The slope (a) and intercept (b) of the line of best fit can be calculated using the following formulas:

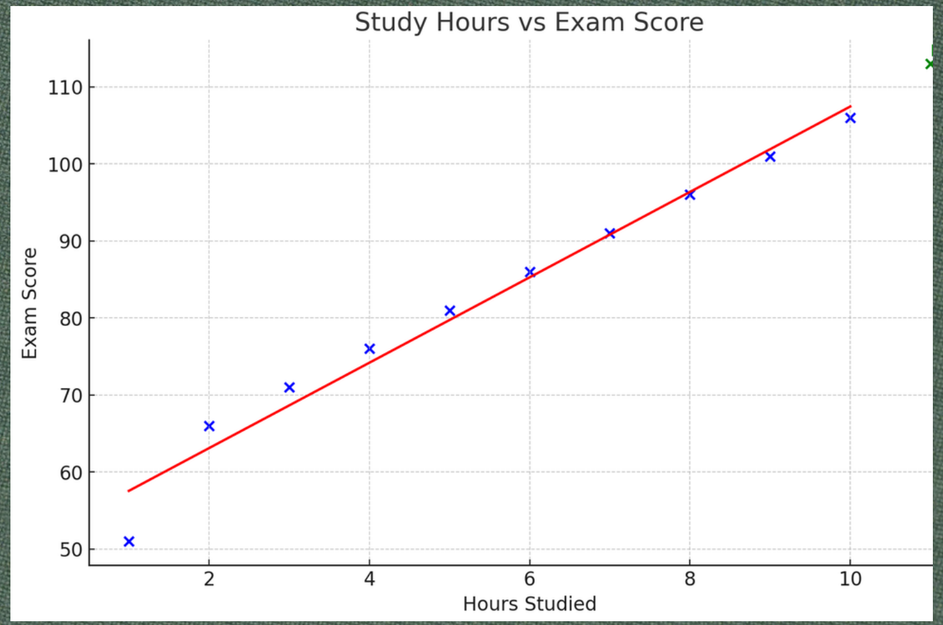
1. **Slope (a) Formula**:
   $a = \frac{N(\sum xy)-(\sum x)(\sum y)}{N(\sum x^2)-(\sum x)^2}$
2. **Intercept (b) Formula**:
   $b = \frac{\sum y - a(\sum x)}{N}$

where:

- N is the number of data points,
- $\sum xy$ is the sum of the product of each pair of x and y values,
- $\sum x$ and $\sum y$ are the sums of x and y values, respectively,
- $\sum x^2$ is the sum of the squares of x values.

These formulas help us compute the exact position and angle of the line that best fits our data, minimizing the overall distance between the data points and the line itself.

Hours Studied (X)	Exam Score (Y)
1	51
2	66
3	71
4	76
5	81
6	86
7	91
8	96
9	101
10	106



Based on the calculations:

The slope (a) of the line of best fit is 495.
The intercept (b) of the line is -2640.
This means the equation of our line of best fit,
based on the study hours and exam scores example, is:

$$Y = 495X - 2640$$

Statistics & Probability for Data Science

WHAT IS STATISTICS?

The science of learning from data; involves collection, analysis, interpretation, and presentation.

FRAMEWORK FOR UNDERSTANDING DATA

Identifies patterns, structures, and relationships.

INFORMED DECISIONS AND PREDICTIONS

Based on empirical evidence and statistical analysis.

ROLE IN DATA SCIENCE

Derives meaningful insights from complex dataset, and take Data-Driven Decisions.

Foundation for Algorithms

APPLICATIONS ACROSS FIELDS

Marketing: Market research, customer segmentation.
Healthcare: Clinical trials, epidemiological studies.

AS DATA'S ROLE GROWS, STATISTICS' IMPORTANCE IN EXTRACTING VALUE AND INSIGHTS WILL CONTINUE TO EXPAND.

Types of Data: Categorical vs. Numerical

CATEGORICAL	NOMINAL	BLOOD TYPES (A, B, AB, O)
	ORDINAL	MOVIE RATINGS
NUMERICAL	DISCRETE	# OF STUDENTS IN A CLASS
	CONTINUOUS	HEIGHT, WEIGHT

Data can be broadly classified into two types: categorical (qualitative) and numerical (quantitative).

CATEGORICAL (QUALITATIVE) DATA

Categorical data represents characteristics or attributes that can be separated into different categories that are distinct.

Nominal: Data that can be categorized but not ranked.

Ordinal: Data that can be categorized and ranked.

NUMERICAL (QUANTITATIVE) DATA

Numerical data represents quantities and can be measured.

Discrete: Countable, distinct values.

Continuous: Measurable quantities that can take any value within a range.

DESCRIPTIVE STATISTICS

A branch of statistics that summarizes or describes the main features of a dataset quantitatively. It provides simple summaries about the sample and the measures.

KEY COMPONENTS

- Measures of Central Tendency: Mean, Median, Mode.
Provide information about the dataset's average or most common values.
- Measures of Variability: Range, Variance, Standard Deviation.
Describe the spread or dispersion of the data points.
- Measures of Shape: Skewness, Kurtosis.
Indicate the asymmetry and peakedness of the data distribution.

PURPOSE

- Simplify large amounts of data in a sensible way. Facilitates data understanding and interpretation.
- Provide a basic understanding of the dataset's characteristics. Helps in identifying patterns, trends, and anomalies in data.
- Serve as the foundation for inferential statistics and hypothesis testing.
- Essential for making informed decisions based on data insights.

Measures of Central Tendency

DATASET: 7, 3, 4, 1, 7, 6

MEAN

Calculated by summing all the values and dividing by the total number of values.

(7 + 3 + 4 + 1 + 7 + 6) / 6

28 / 6

MEDIAN

The middle value in a dataset when it is ordered from smallest to largest. If there is an even number of observations, the median is the average of the two middle numbers.

1, 3, 4, 6, 7, 7

4+6 / 2 = 5

MODE

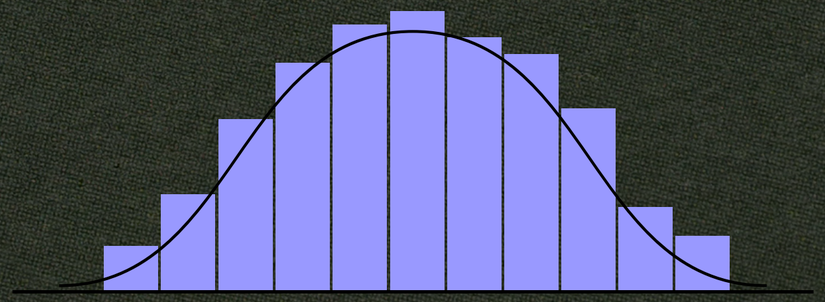
The most frequently occurring value(s) in a dataset.

7

What is Probability?

Probability is a branch of mathematics that deals with calculating the likelihood of a given event's occurrence, which is expressed as a number between 1 and 0.

- - It quantifies uncertainty.
- - Fundamental in making predictions.
- - Used to model and analyze the real world.



Random Variable

Imagine you're flipping a coin. Before you flip it, you're not sure if it will land heads or tails. A random variable is a way to give a number to each possible outcome of this uncertain event. For the coin, you could say heads is 1 and tails is 0. If you were rolling a dice, the random variable could be the number that comes up, which could be any number between 1 and 6. So, a random variable is basically a label we use to describe the outcome of something that can happen in different ways.

Probability Distribution

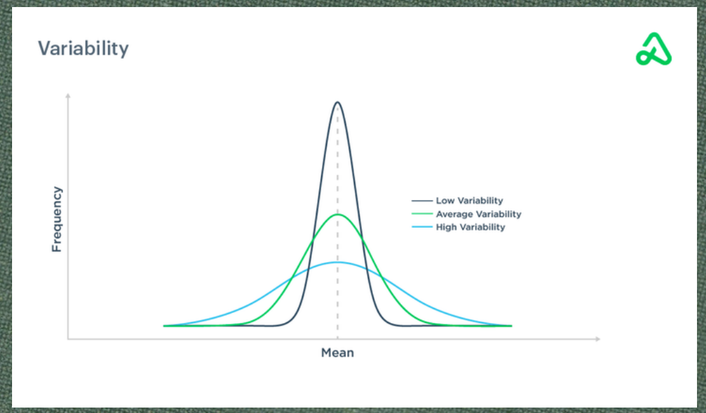
Now, let's say you're flipping the coin many times, or rolling the dice many times, and you start keeping track of what happens. The probability distribution is a way to show how likely each outcome is based on your tracking. If you flip a fair coin many times, you'd expect heads about half the time and tails about half the time. The probability distribution would show that both outcomes have a 50% chance. It's like a summary of how likely each possible outcome is.

Normal Distribution

The normal distribution is a specific type of probability distribution that looks like a bell curve when you graph it. It's a way to describe how data points, like people's heights, are spread out. Most people are of average height, so the top of the bell (the highest point) represents the average height. The curve slopes down on either side, showing that fewer people are much taller or much shorter than average. This distribution is everywhere in nature and human-made phenomena because many things tend to cluster around an average with fewer extremes.

DATA SCIENTISTS USE THEM TO UNDERSTAND THE BEHAVIOR OF DATA AND MAKE INFERENCES ABOUT POPULATIONS FROM SAMPLES.

Measure of Variability



These measures describe the spread or dispersion of data points in a dataset.

- Range: The difference between the highest and lowest values in a dataset.
- Variance: A measure of how much the values in a dataset deviate from the mean. It is calculated as the average of the squared differences from the Mean.

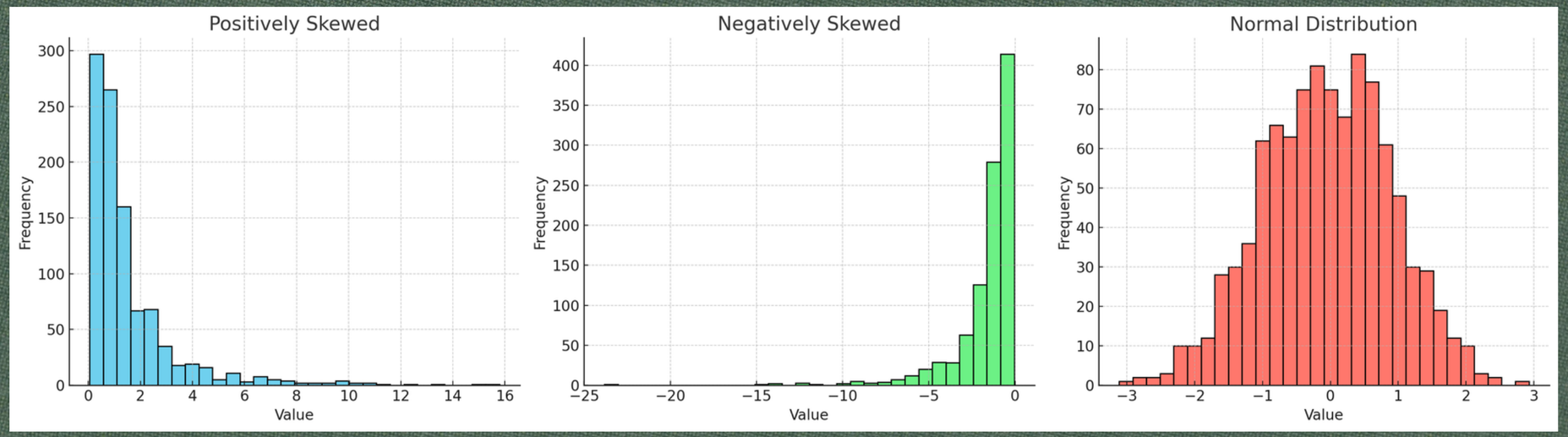
$$\sigma^2 = \frac{\sum(x_i - \mu)^2}{n}$$

- Standard Deviation: The square root of the variance, providing a measure of spread that is in the same units as the data.

Skewness

Skewness measures the asymmetry of the distribution of values in a dataset.

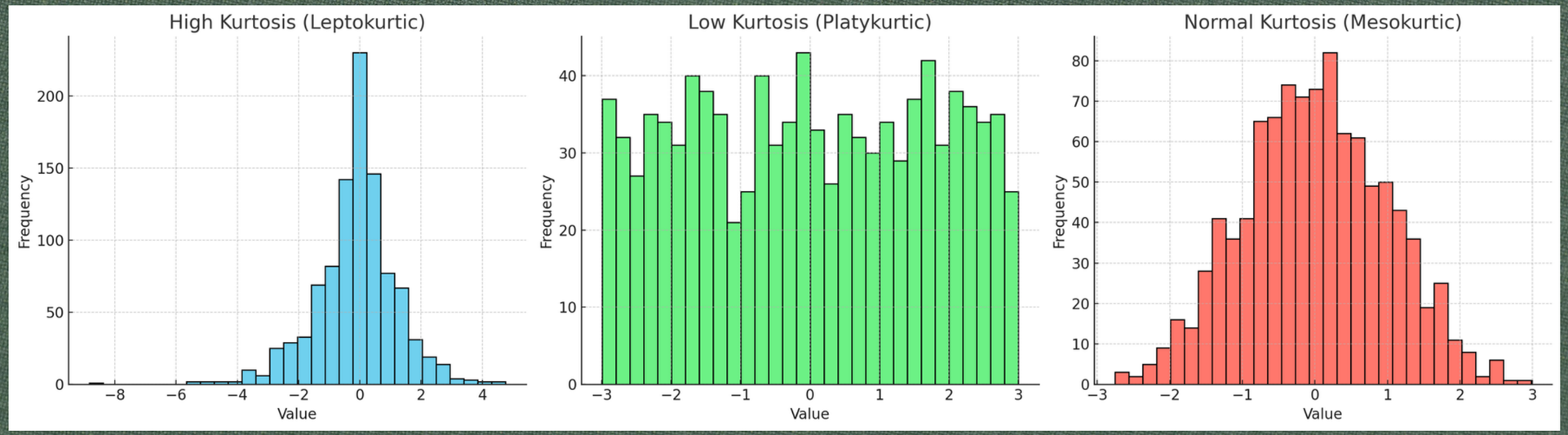
- A distribution is symmetrical if skewness is around 0.
- A distribution is skewed to the right (positive skewness) if the tail on the right side is longer or fatter than the left side.
- A distribution is skewed to the left (negative skewness) if the tail on the left side is longer or fatter than the right side.



Kurtosis

Kurtosis measures the "tailedness" of the distribution of values in a dataset, indicating the presence of outliers.

- High Kurtosis (Leptokurtic): sharper peak and fatter tails, indicating more data in the tails and a higher likelihood of outliers.
- Low Kurtosis (Platykurtic): flatter peak and thinner tails, suggesting less data in the tails and fewer extreme outliers.
- Normal Kurtosis (Mesokurtic): kurtosis value similar to that of a normal distribution, with a moderate peak and tails, indicating an average likelihood of outliers.



Basics of Calculus

$$\frac{P(x)}{Q(x)} = G(x) + \frac{R(x)}{Q(x)}$$

WHAT IS CALCULUS?

Imagine you are driving a car, and you want to know how fast you're going at every moment, not just your average speed over the trip.
Calculus helps you figure out your speed at each instant; it's the mathematical tool for studying things that change smoothly.
It also helps in figuring out the total distance traveled, even if you speed up or slow down along the way, by adding up all those little distances you covered every instant.

Calculus is the mathematical study of continuous change. It is divided into two related fields:

- Differential Calculus focuses on the concept of the derivative, which represents the rate of change of a function with respect to its variable. It's fundamental for understanding how quantities change over time or space.
- Integral Calculus involves the integral, which represents accumulation of quantities and can be considered as the area under a curve of a function. It's used to calculate things like areas, volumes, and totals from infinitesimally small pieces.

Limits & Derivatives

WHAT ARE DERIVATIVES?

The derivative of a function at a point is the rate at which the function's value changes with respect to changes in its input value. It is defined as the limit of the average rate of change of the function over a tiny interval as the interval approaches zero.

If calculus is about understanding change, think of a derivative as a snapshot of that change at a specific moment. For example, if you're watching a balloon deflate, the derivative tells you how fast the air is leaving the balloon at any given time.

The notation "lim" stands for "limit" in mathematics, which is a fundamental concept in calculus. A limit describes the value that a function or sequence "approaches" as the input or index approaches some value. Limits are essential for defining derivatives and integrals, the core concepts of calculus.

Consider the function $f(x) = x^2$, which describes a parabola. The derivative of this function, denoted as $f'(x)$ or $\frac{d}{dx}x^2$, tells us how $f(x)$ changes as x changes.

To find the derivative of $f(x) = x^2$, we use the definition of the derivative:

$$f'(x) = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h}$$

Plugging $f(x) = x^2$ into the formula gives:

$$f'(x) = \lim_{h \to 0} \frac{(x+h)^2 - x^2}{h} = \lim_{h \to 0} \frac{2xh + h^2}{h} = \lim_{h \to 0}(2x + h) = 2x$$

So, the derivative of $f(x) = x^2$ is $f'(x) = 2x$. This means that, at any point x on the graph of $f(x)$, the rate at which $f(x)$ changes is $2x$. For example, at $x = 3$, the rate of change of $f(x)$ is $f'(3) = 2(3) = 6$. This tells us how steep the curve is at that point.

Step-by-Step example:

Step 1: Start with the function

We have $f(x) = x^2$. This is our starting point, describing how $f(x)$ (our "output") changes with x (our "input").

Step 2: Make a tiny change to x

Let's change x by a tiny amount, call it h. So instead of looking at just x, we're now looking at $x + h$.

Step 3: Calculate the function for $x + h$

Now, we calculate what $f(x + h)$ would be. Since $f(x) = x^2$, we have $f(x + h) = (x + h)^2$.

Step 4: Find the change in $f(x)$

To find out how much $f(x)$ changes, we subtract the original $f(x)$ from the new $f(x + h)$. So, we calculate $(x + h)^2 - x^2$.

Step 5: Simplify the expression

$(x + h)^2 - x^2$ simplifies to $2xh + h^2$, because $(x + h)^2$ expands to $x^2 + 2xh + h^2$, and then you subtract x^2.

Step 6: Find the average rate of change

We're interested in the rate of change per unit of h, so we divide by h: $\frac{2xh+h^2}{h}$. Simplifying this gives us $2x + h$.

Step 7: Make h really, really small

The derivative tells us what happens as h gets closer and closer to zero. So, we take the limit of our expression as h approaches zero. This means we're looking at what $2x + h$ becomes when h is infinitesimally small.

Step 8: Conclude the derivative

As h approaches zero, the h in $2x + h$ becomes negligible, leaving us with just $2x$. So, the derivative of $f(x) = x^2$, denoted as $f'(x)$ or $\frac{d}{dx}x^2$, is $2x$.

Derivative Formulas

Power Formula

$$\frac{d}{dx}(x^n) = nx^{n-1}$$

Sum Formula

$$\frac{d}{dx}[f(x)+g(x)] = f'(x) + g'(x)$$

Difference Formula

$$\frac{d}{dx}[f(x)-g(x)] = f'(x) - g'(x)$$

Chain Formula

$$\frac{d}{dx}f(g(x)) = f'(g(x))g'(x)$$

Product Formula

$$\frac{d}{dx}[f(x)g(x)] = f(x)g'(x) + g(x)f'(x)$$

Constant Multiple Formula

$$\frac{d}{dx}[cf(x)] = cf'(x)$$

Quotient Formula

$$\frac{d}{dx}\left[\frac{f(x)}{g(x)}\right] = \left[\frac{g(x)f'(x)-f(x)g'(x)}{[g(x)]^2}\right]$$

Gradient Descent in Linear Regression: Finding the Best Fit Line

Linear regression aims to find the best-fit line, or equation, to model the relationship between a dependent variable (y) and an independent variable (x) based on a set of data points. But how do we quantify how "good" a fit is? This is where the cost function comes in. Also, finding the best-fitting line for a given dataset requires an optimization technique. This is where gradient descent comes in.

COST FUNCTION

Imagine you're trying to fit a line through a bunch of scattered data points. The perfect fit would minimize the distance between the data points and the line. In linear regression, we quantify this distance using a cost function, often the mean squared error (MSE):

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - (mx_i + b))^2$$

where n is the number of observations, y_i is the actual value, x_i is the feature value, m is the slope, and b is the y-intercept.

What is Gradient Descent?

The Guiding Light: Gradient

Now, imagine yourself standing on a hilly landscape blindfolded. You want to reach the lowest point (the minimum). The gradient comes to your rescue. It's a mathematical concept that tells you the steepest direction of change at any point on the landscape.
In the context of linear regression, the gradient points towards the direction of steepest increase in the cost function (MSE) with respect to the model's parameters. These parameters are typically the slope (m) and the intercept (c) of the linear equation.

The Descent Begins: Gradient Descent Algorithm

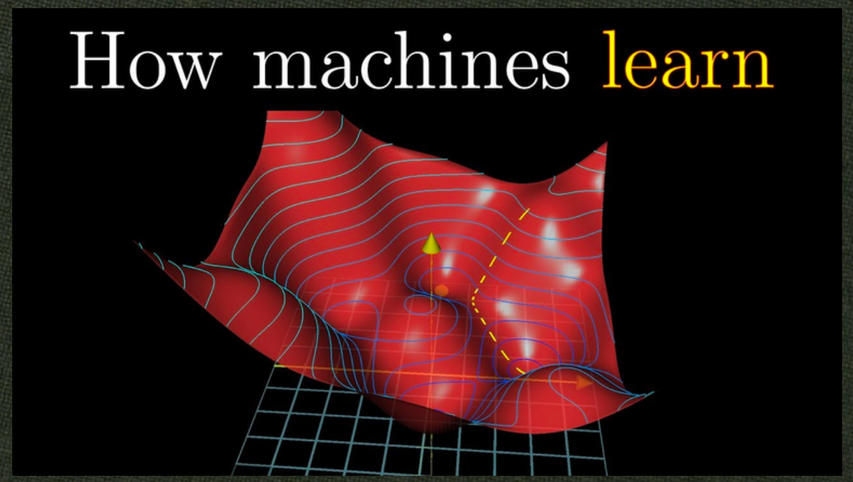
Gradient descent is an iterative optimization algorithm that leverages the gradient to find the minimum of the cost function. Here's how it works:

1. Start with initial values for the slope (m) and intercept (c). These can be random guesses.
2. Calculate the gradient with respect to both m and c. The gradient tells you how much and in which direction to adjust m and c to reduce the cost function.
3. Update the parameters using a learning rate (α), which controls the step size:

m_new = m_old - α * (∂MSE / ∂m)
c_new = c_old - α * (∂MSE / ∂c)

- ∂MSE / ∂m: partial derivative of MSE with respect to m (slope)
- ∂MSE / ∂c: partial derivative of MSE with respect to c (intercept)

4. Repeat steps 2 and 3 until the change in the cost function becomes negligible (convergence) or a maximum number of iterations is reached.



Finding the Line of Best Fit: y = mx + b

Step 1: Define the Data Points

Let's use the following 5 data points (x, y): $(1, 2), (2, 3), (3, 5), (4, 7), (5, 9)$.

Step 2: Initialize Parameters

We need to initialize the parameters m and b. Let's start with $m = 0$ and $b = 0$.

Step 3: Choose a Learning Rate

The learning rate determines the size of the steps we take during each iteration of gradient descent. A typical starting point could be $\alpha = 0.01$.

Step 4: Calculate the Gradient

The cost function (Mean Squared Error, MSE) for linear regression is given by:

$$J(m, b) = \frac{1}{N} \sum_{i=1}^{N} (y_i - (mx_i + b))^2$$

where N is the number of data points.

Partial Derivative with Respect to m:

To find $\frac{\partial J}{\partial m}$, we differentiate $J(m, b)$ with respect to m, treating b and x_i as constants.

$$\frac{\partial J}{\partial m} = \frac{\partial}{\partial m} \left(\frac{1}{N} \sum_{i=1}^{N} (y_i - (mx_i + b))^2 \right)$$

Applying the chain rule, we get:

$$= \frac{1}{N} \sum_{i=1}^{N} 2(y_i - (mx_i + b)) \cdot \frac{\partial}{\partial m}(y_i - (mx_i + b))$$

$$= \frac{1}{N} \sum_{i=1}^{N} 2(y_i - mx_i - b) \cdot (-x_i)$$

$$= \frac{-2}{N} \sum_{i=1}^{N} x_i(y_i - mx_i - b)$$

Partial Derivative with Respect to b:

Similarly, to find $\frac{\partial J}{\partial b}$, we differentiate $J(m, b)$ with respect to b, treating m and x_i as constants.

$$\frac{\partial J}{\partial b} = \frac{\partial}{\partial b} \left(\frac{1}{N} \sum_{i=1}^{N} (y_i - (mx_i + b))^2 \right)$$

Again, applying the chain rule:

$$= \frac{1}{N} \sum_{i=1}^{N} 2(y_i - (mx_i + b)) \cdot \frac{\partial}{\partial b}(y_i - (mx_i + b))$$

$$= \frac{1}{N} \sum_{i=1}^{N} 2(y_i - mx_i - b) \cdot (-1)$$

$$= \frac{-2}{N} \sum_{i=1}^{N} (y_i - mx_i - b)$$

With $m = 0$ and $b = 0$, these simplify to:

$$\frac{\partial J}{\partial m} = \frac{-2}{5}[(1)(2-0) + (2)(3-0) + (3)(5-0) + (4)(7-0) + (5)(9-0)]$$

$$\frac{\partial J}{\partial b} = \frac{-2}{5}[(2-0) + (3-0) + (5-0) + (7-0) + (9-0)]$$

Let's manually compute these values.

For $\frac{\partial J}{\partial m}$:

$$\frac{\partial J}{\partial m} = \frac{-2}{5}[1 \cdot (2-0) + 2 \cdot (3-0) + 3 \cdot (5-0) + 4 \cdot (7-0) + 5 \cdot (9-0)]$$

$$= \frac{-2}{5}[2 + 6 + 15 + 28 + 45]$$

$$= \frac{-2}{5}[96]$$

$$= \frac{-192}{5}$$

$$= -38.4$$

For $\frac{\partial J}{\partial b}$:

$$\frac{\partial J}{\partial b} = \frac{-2}{5}[(2-0) + (3-0) + (5-0) + (7-0) + (9-0)]$$

$$= \frac{-2}{5}[2 + 3 + 5 + 7 + 9]$$

$$= \frac{-2}{5}[26]$$

$$= \frac{-52}{5}$$

$$= -10.4$$

So, the partial derivatives calculated manually for our initial values of m and b are:

- $\frac{\partial J}{\partial m} = -38.4$
- $\frac{\partial J}{\partial b} = -10.4$

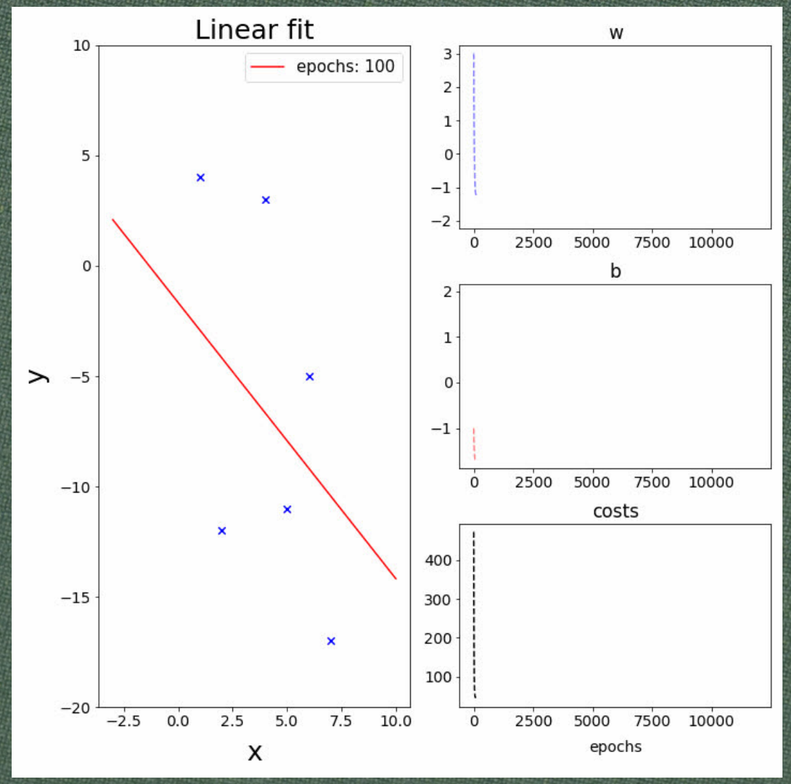
Step 5: Update the Parameters

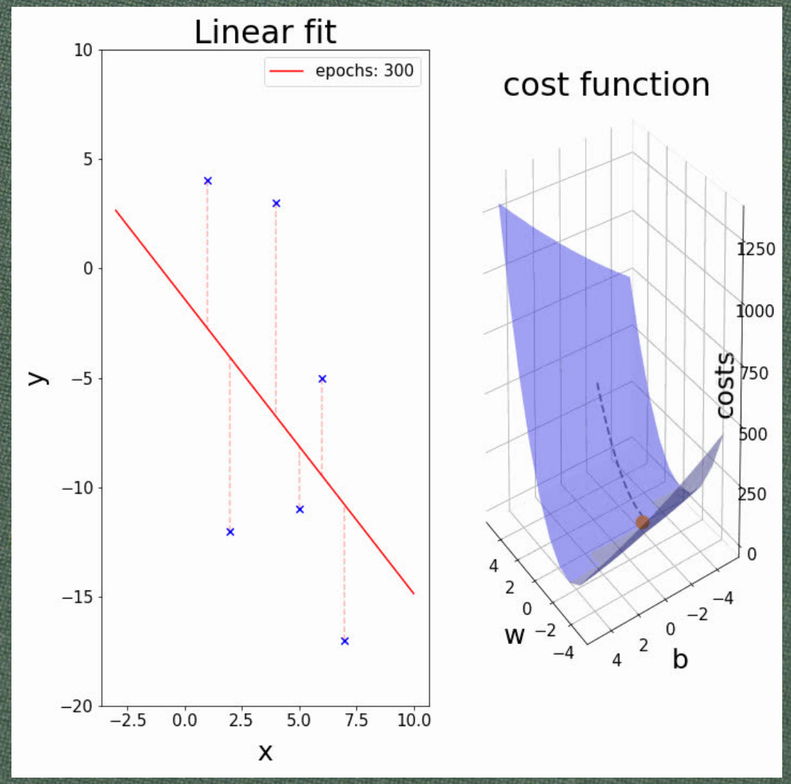
To update m and b, we use the following formulas:

$$m := m - \alpha \cdot \frac{\partial J}{\partial m}$$
$$b := b - \alpha \cdot \frac{\partial J}{\partial b}$$

With our data points fixed, the cost function is only dependent on the parameters w and b. We aim to adjust our parameters until the cost function reaches its minimum.

- When training our model, our objective is to repeat the following for each epoch until we reach convergence.
- Metaphorically speaking, the cost function can be imagined as some mountainous terrain, where beginning from a certain starting point, we want to head downhill until we reach a valley. Analogously, the gradient is giving us the 'multidimensional slope' i.e. the direction of where 'uphill' is on the mountain surface. That's why we intend to constantly adjust our parameters in the opposite direction of the gradient.



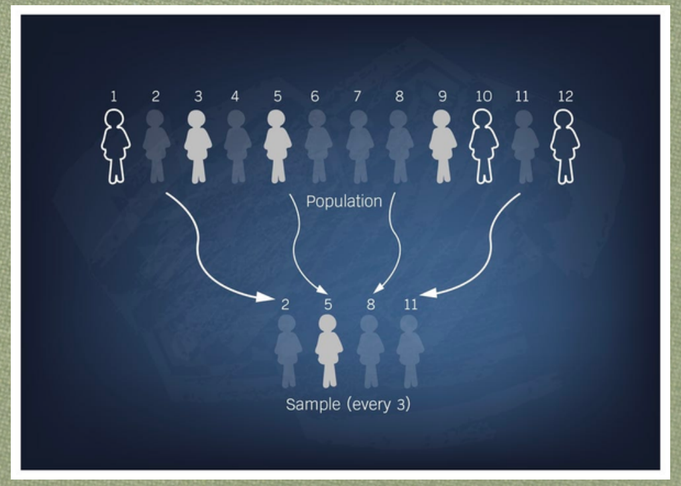


Hypothesis Testing

WHAT IS IT?

Hypothesis testing is like a courtroom trial for statistics. Just like in a trial, there's an initial assumption (the defendant is innocent until proven guilty), in hypothesis testing, we start with an assumption about a population parameter. We then collect evidence (data) to support or reject this initial assumption.

Hypothesis testing is a statistical method that uses sample data to evaluate a hypothesis about a population parameter. The process involves making an initial assumption (the null hypothesis), then determining whether the observed sample data provides sufficient evidence to reject this initial assumption in favor of an alternative hypothesis.



Uses of Hypothesis Testing

HOW IS IT USED IN DATA SCIENCE?

- Testing the effectiveness of new drugs.
- Determining if changes to a website lead to more conversions.
- Evaluating the impact of policy changes in economics.
- Quality control in manufacturing.
- Environmental science to assess impact changes.
- Market research to test consumer preferences.

HYPOTHESIS

A hypothesis is a tentative statement or explanation that proposes a possible relationship between variables and is subject to testing and verification.

OVERVIEW

A hypothesis is a testable prediction that seeks to explain a phenomenon or make a statement about the relationship between variables. It is formulated based on prior knowledge, observations, or theories. Hypotheses serve as a starting point for scientific investigation, and their validity is assessed through empirical evidence.

EXAMPLES

Null Hypothesis: The null hypothesis is a statement that suggests no significant relationship or effect between variables.

A new diet has no effect on weight loss.

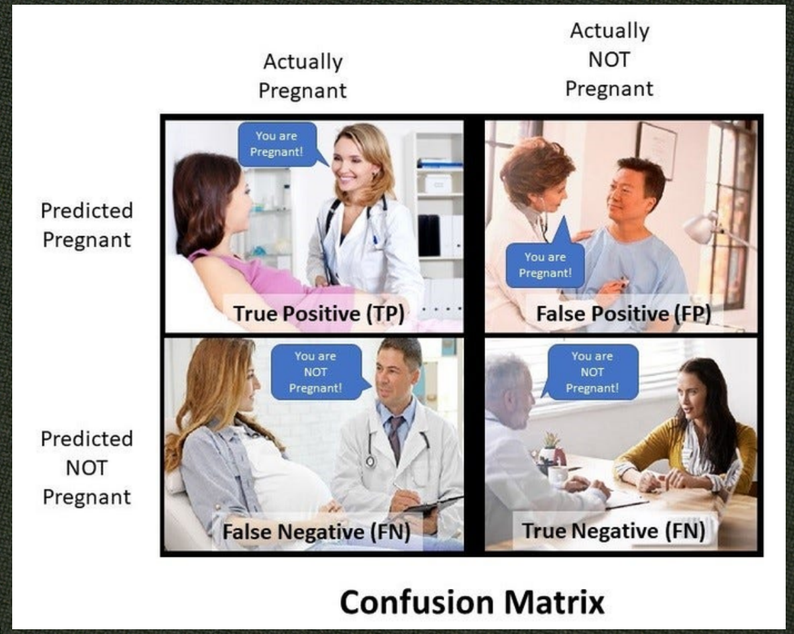
Alternative Hypothesis: The alternative hypothesis is a statement that proposes a specific relationship or effect between variables, challenging the null hypothesis.

The new diet leads to weight loss.

Errors in Hypothesis Testing

Definitions: Types of Errors

- **Type I error**: The null hypothesis is rejected when it is true.
- **Type II error**: The null hypothesis is not rejected when it is false.
- There is always a chance of making one of these errors. We'll want to minimize the chance of doing so!



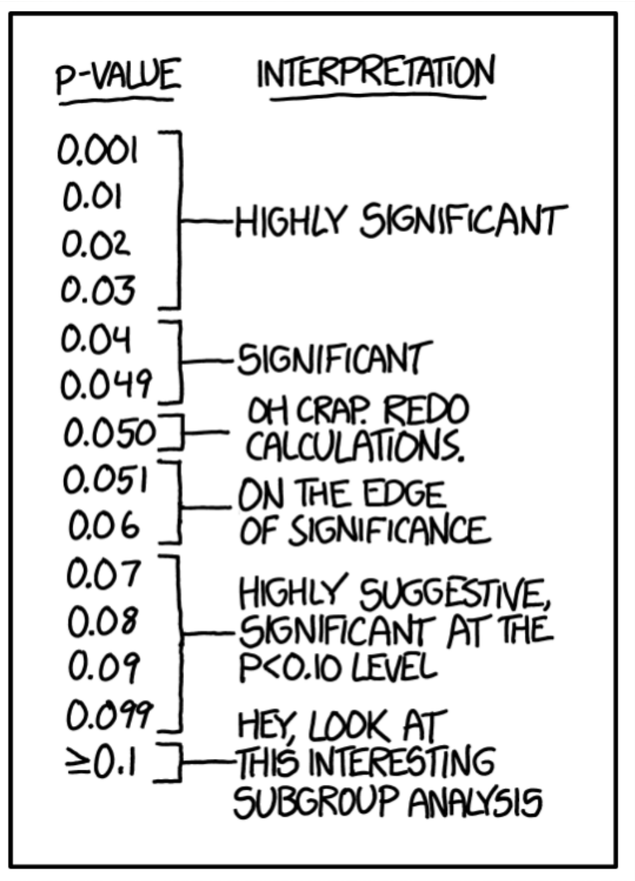
Type 1 Error (False Positive): It's the mistake of assuming something has happened (a "positive" result) when it hasn't, based on the test data.

Difference Between
Type-1 and Type-2 Error

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

Type 2 Error (False Negative): This error represents a missed opportunity to identify an actual effect or difference, mistakenly concluding a "negative" result when in fact there was something to be noted.

P-Values and Significance Level



Imagine you're a teacher who suspects that a new teaching method may improve students' test scores. To scientifically test this hypothesis, you decide to conduct an experiment.

Setting Up the Experiment
Population: You have 100 students.
Current Method: All students currently use the traditional teaching method.
New Method: You want to test a new teaching method.

Step 1: Formulate Hypotheses

- **Null Hypothesis (H0):** The new teaching method does not improve test scores.
- **Alternative Hypothesis (H1):** The new teaching method improves test scores.

Step 2: Choose Alpha (α)

You decide on an alpha (α) level of 0.05, which means you're willing to accept a 5% chance of incorrectly rejecting the null hypothesis (a Type I error).

Step 3: Conduct the Experiment

You randomly select 50 of your 100 students to undergo the new teaching method for a semester, while the other 50 continue with the traditional method. At the end of the semester, all students take a standardized test.

Step 4: Calculate the P-Value

After the test, you calculate the average scores for both groups and perform a statistical test to compare them. This test gives you a p-value, which is the probability of observing the test results, or more extreme results, if the null hypothesis were true.

Example Outcome

- **Average Score with Traditional Method:** 70
- **Average Score with New Method:** 75
- **P-Value:** 0.03

Step 5: Compare P-Value and Alpha

- **P-Value (0.03) < Alpha (0.05):** The p-value is less than your alpha level.

Interpretation

Since the p-value is less than alpha, you reject the null hypothesis. This means there is statistically significant evidence, with a 5% risk of being wrong, that the new teaching method improves test scores compared to the traditional method.

Think of the p-value as an indicator of how surprising your results are under the assumption that the new teaching method has no effect:

Low P-Value (e.g., 0.03): Your results are surprising under the null hypothesis, suggesting the new method does have an effect.
High P-Value (e.g., 0.5): Your results are not surprising under the null hypothesis, suggesting the new method does not have a significant effect.
The alpha level is like a threshold for how surprised you need to be to decide that the new method really makes a difference. Since the p-value in this example is below the threshold (alpha), you conclude that the new teaching method likely caused the improvement in test scores.

Multiple Linear Regression

Imagine you're trying to predict the price of a house based on various features like its size, age, and number of bedrooms.
Each of these features can influence the price.
Multiple linear regression allows us to understand how each feature affects the price while considering the influence of all other features.
It's like finding the best recipe for a cake where the ingredients must be mixed in just the right amounts to get the perfect flavor, where the "perfect flavor" is the most accurate prediction of the house price.

Mathematical Explanation:

In multiple linear regression, the relationship between the dependent variable Y and multiple independent variables $X_1, X_2, ..., X_n$ is modeled as:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n + \epsilon$$

Here:

- Y is the dependent variable we're trying to predict.
- $X_1, X_2, ..., X_n$ are the independent variables (or features) that influence Y.
- β_0 is the intercept, the value of Y when all X are 0.
- $\beta_1, \beta_2, ..., \beta_n$ are the coefficients that represent the change in Y for a one-unit change in X_i, holding all other variables constant. These coefficients are different from just slope and intercept in simple linear regression because each coefficient is associated with a different independent variable, providing a multi-dimensional slope.
- ϵ is the error term, representing the difference between the observed values and the values predicted by the model. It accounts for the randomness or the variation in Y that cannot be explained by the X variables.

Coefficients in Multiple Linear Regression: Significance Testing Experiment

What are Coefficients?

In a multiple linear regression model, coefficients represent the magnitude and direction of the linear relationship between each independent variable (predictor) and the dependent variable (response).

Why are we doing this?

Understanding the significance of coefficients is crucial because:

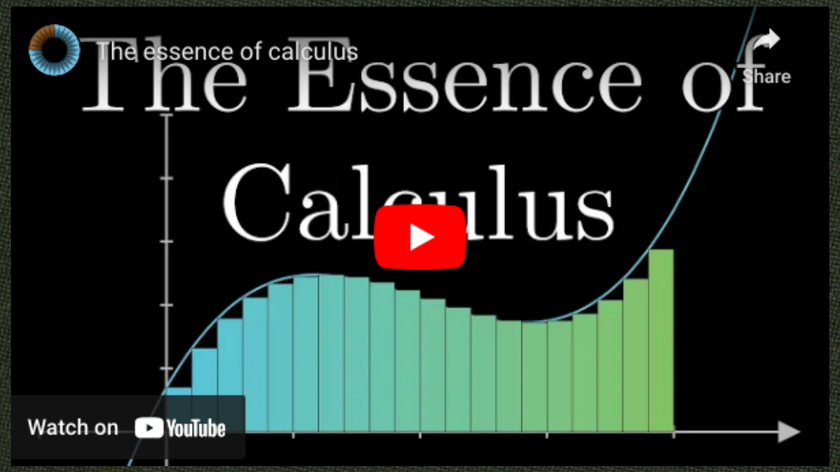
Non-significant coefficients: May indicate a variable doesn't have a statistically significant linear relationship with the response variable. Including such a variable can unnecessarily complicate the model and potentially reduce its interpretability.
Significant coefficients: Help us identify the most important factors influencing the response variable and quantify their individual impact.

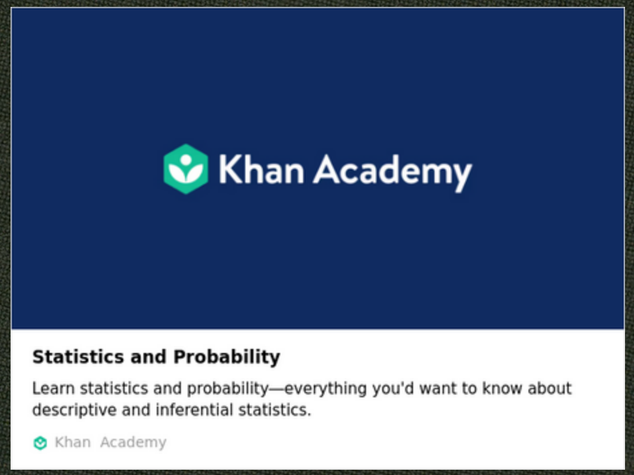
Example

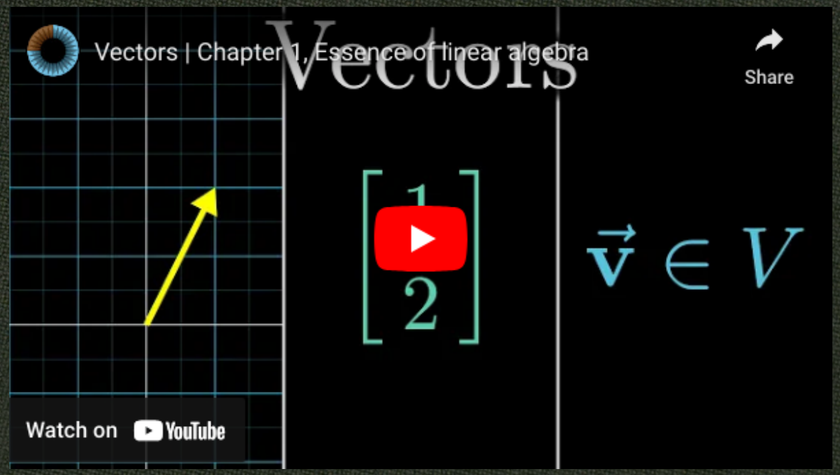
Let's say you're studying factors that affect house prices. Your dependent variable is the price of the house, and your independent variables might include square footage, number of bedrooms, age of the house, and proximity to the city center.

1. **Hypotheses**:
   - H0: The coefficients for square footage, number of bedrooms, age, and proximity to the city center are all equal to zero.
   - H1: At least one of the coefficients is not equal to zero.
2. **Data Collection**: You collect data on 200 houses, including their prices, square footage, number of bedrooms, age, and distance from the city center.
3. **Significance Level**: Alpha = 0.05.
4. **Model Fitting**: You use software to fit a multiple linear regression model and estimate the coefficients.
5. **Hypothesis Testing**: Suppose the p-values for the coefficients are as follows:
   - Square footage: $p = 0.001$
   - Number of bedrooms: $p = 0.04$
   - Age of the house: $p = 0.2$
   - Proximity to the city center: $p = 0.03$
6. **Interpretation**:
   - For square footage, number of bedrooms, and proximity to the city center, the p-values are less than 0.05, so you reject the null hypothesis for these variables. This suggests they have a statistically significant relationship with house prices.
   - For the age of the house, the p-value is greater than 0.05, so you fail to reject the null hypothesis, indicating it does not have a statistically significant effect on house prices in your model.

Thank You!


The essence of calculus
Share
The Essence of Calculus
Watch on YouTube


Khan Academy
Statistics and Probability
Learn statistics and probability—everything you'd want to know about descriptive and inferential statistics.
Khan Academy


Vectors | Chapter 1, Essence of linear algebra
Share
Vectors
$\vec{v} \in V$
Watch on YouTube