Welcome to Data Science Online Bootcamp

Day 7

dφ

Democratizing Data Science Learning

Learning Objectives

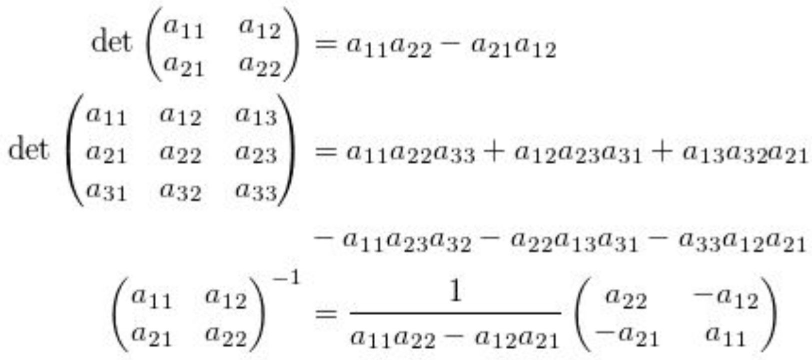
1. Linear Algebra

2. Matrix

3. Basic Statistics

What is Linear Algebra?

Often the first acquaintance with linear algebra looks something like this:



$$\det\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} = a_{11}a_{22} - a_{21}a_{12}$$

$$\det\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} = a_{11}a_{22}a_{33} + a_{12}a_{23}a_{31} + a_{13}a_{32}a_{21}$$

$$- a_{11}a_{23}a_{32} - a_{22}a_{13}a_{31} - a_{33}a_{12}a_{21}$$

$$\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}^{-1} = \frac{1}{a_{11}a_{22} - a_{12}a_{21}} \begin{pmatrix} a_{22} & -a_{12} \\ -a_{21} & a_{11} \end{pmatrix}$$

Not very inspiring, right? Two questions immediately arise: where did all this come from and why is it needed.

Basic definition:

Linear algebra is a sub-field of mathematics concerned with vectors, matrices, and linear transforms.

Do You Really Need Linear Algebra?

- It depends on your goal.

- If you just want to use tools from AI and machine learning as a black box, you arguably just need enough math to figure out if your problem fits the models premise.

- But, if you want to develop new ideas, Linear Algebra is your must-learn thing. I don't mean you need to learn everything concerning math. Doing so you will be stuck at everything and lose motivation towards other more important things like calculus/stats.

- Mathematics in data science and machine learning is not about crunching numbers, but about what is happening, why it's happening, and how we can play around with different things to obtain the results we want.

Linear Algebra

- Let’s start with a simple problem.
  - Condition 1: Imagine price of 2 chocolates and 1 apple is 100 units
  - Condition 2: Similarly imagine, price of 1 chocolate and 2 apples is 100 units.

  Now, we want to find the price of a chocolate and an apple

- Suppose the price of a chocolate is $ ‘x’ and the price of an apple is $ ‘y’. Values of ‘x’ and ‘y’ can be anything depending on the situation i.e. ‘x’ and ‘y’ are variables.
- Translating the above information in mathematical form:

  $2x + y = 100$ ---------- (1)

  Similarly, for the second condition,

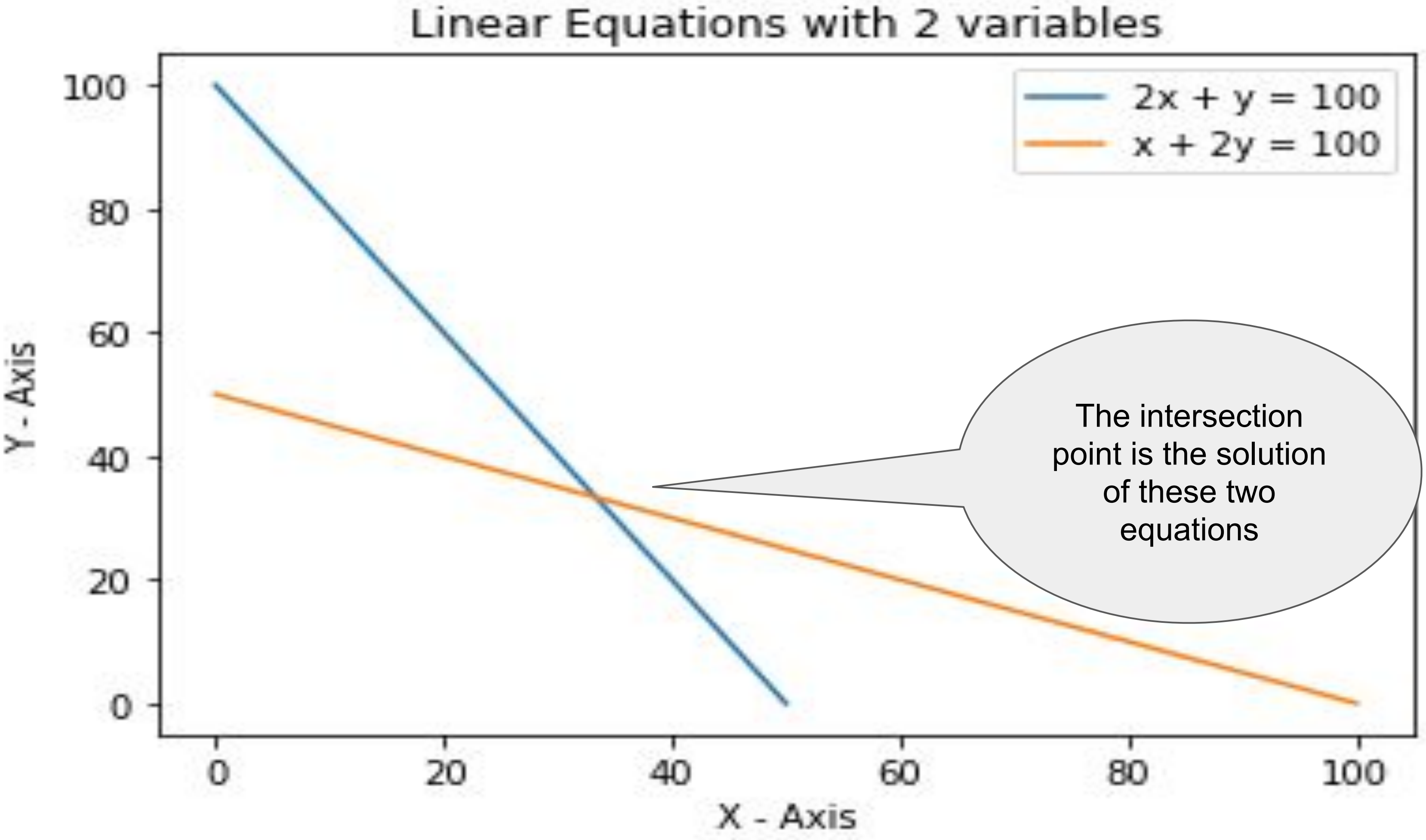
  $x + 2y = 100$ ---------- (2)

Linear Algebra

- To find the prices of chocolate and apple, we need the values of ‘x’ and ‘y’ such that it satisfies both the equations.

- The basic problem of linear algebra is to find these values of ‘x’ and ‘y’ which is nothing but solution of set of linear equation.

Graphical Representation of two Equations



The intersection point is the solution of these two equations

Complicating the Problem

- In the earlier example, we had two variables ‘x’ representing the price of a chocolate and ‘y’ representing the price of a apple.

- Now, suppose you are given a set of three conditions with three variables, say ‘x’, ‘y’ and ‘z’, and asked to find the value of three variables.

- The three conditions are given as:

  $x + y + z = 1$ ---------- (1)

  $2x + y = 1$ ---------- (2)

  $5x + 3y + 2z = 4$ ---------- (3)

- By solving the above three equations we can get the values for ‘x’, ‘y’ and ‘z’.

- In Linear Algebra, data is represented in the form of linear equations. These linear equations are in turn represented in the form of matrices and vectors.

Matrices

- Matrix is a form of **representing data in the form of rows and columns.** It is a very natural approach of organizing data.

- A real life example,
  Consider a reactor which needs to be controlled using multiple attributes from various sensors like Pressure (P), Temperature (T), Density (d), etc.

  In the matrix given,

  Columns:

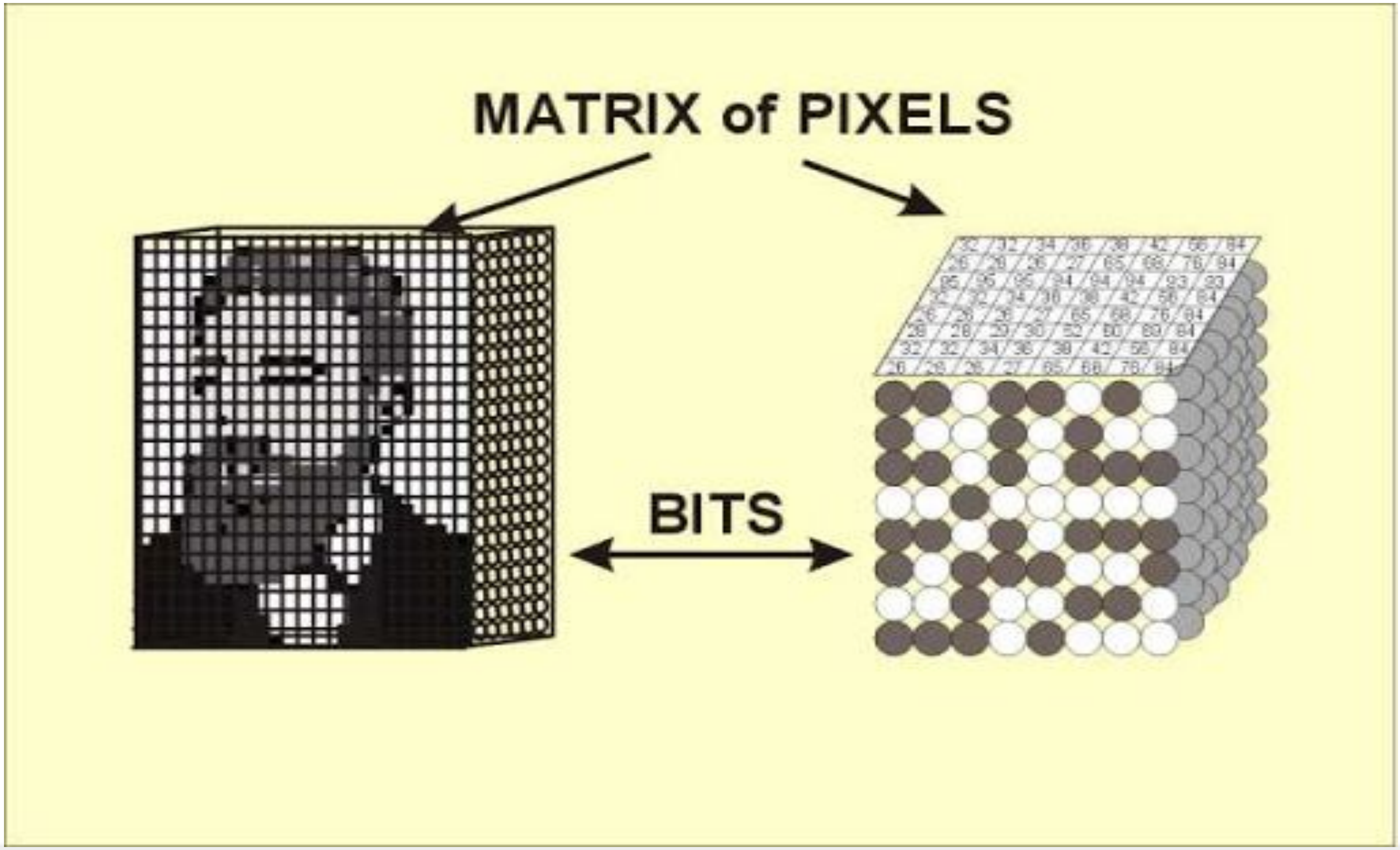
  - First column represents Pressure (P)
  - Second column represents Temperature (T)
  - Third column represent Density (d)

  Row:
  - Each row corresponds to one sample.

$$\begin{array}{ccc} \mathbf{P} & \mathbf{T} & \mathbf{d} \\ 300 & 300 & 1000 \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ 500 & 1000 & 5000 \end{array}$$

Matrices Continued...

- The image stored in the machine is nothing but a large matrix of pixel values across the image.



Democratizing Data Science Learning

Linear Equations in Matrix Form

- Earlier we had two linear equations as:
  2x + y = 100 ---------- (1), and
  x + 2y = 100 ---------- (2)

- The above two linear equations can be represented in the matrix form as:

$$\begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 100 \\ 100 \end{bmatrix}$$

- We can get the above two linear equations from the shown matrix equation just by multiplying the two matrices on left hand side and equating the corresponding value to right hand side.

- **Here is nice resource on Matrices for further reading:**
  https://www.statisticshowto.com/matrices-and-matrix-algebra/

Additional Resources

Interested to learn more about linear algebra operations in Machine Learning?

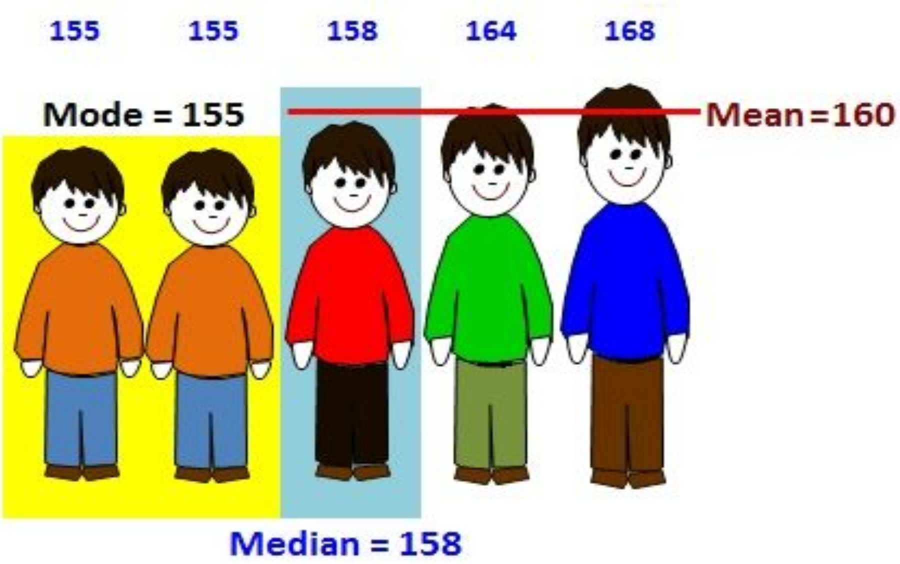
- Check out this helpful resource: https://machinelearningmastery.com/linear-algebra-cheat-sheet-for-machine-learning/

PS. Matrix operations like Addition, Multiplication and Transpose are commonly used in ML.

Statistics

What is Statistics?

A branch of mathematics that takes and transform the data into some useful information which in turn is used to make some decisions.



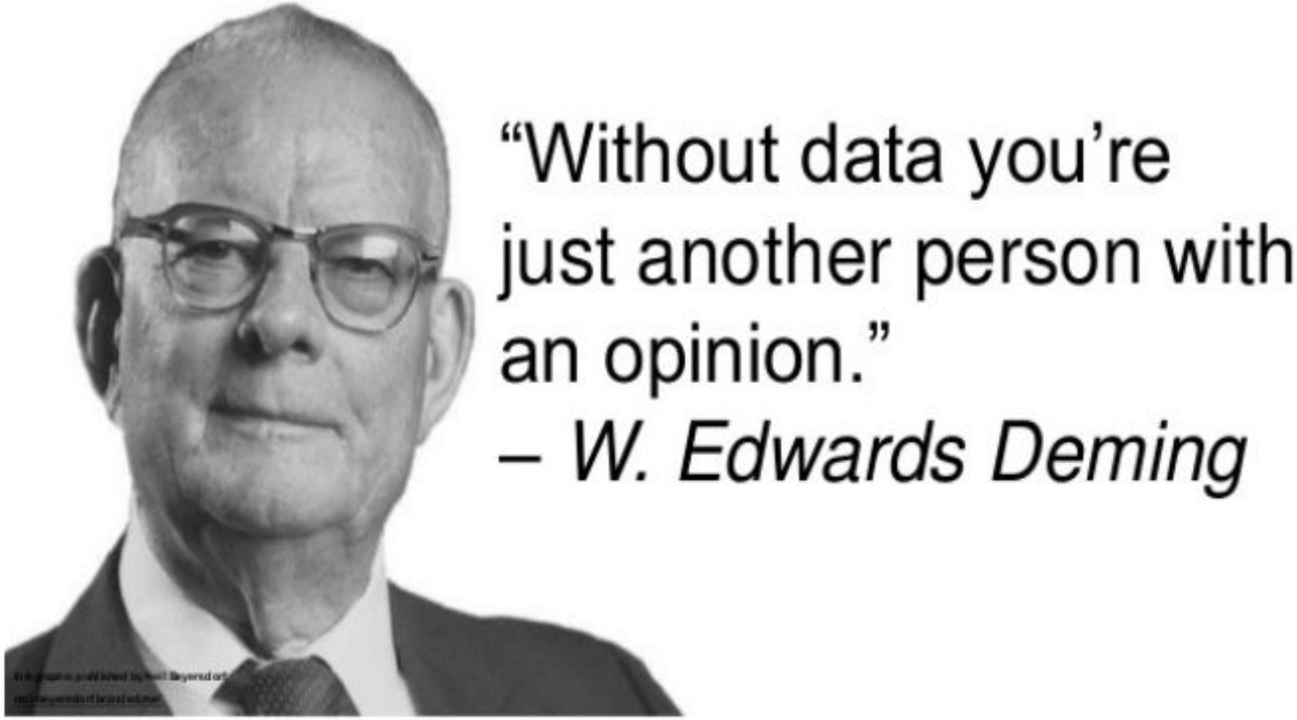
Statistics is concerned with

- Processing and analyzing data
- Collecting, presenting and transforming data to assist decision maker

Data

What is Data?

Data are facts and statistics collected together for reference or analysis.



"Without data you're just another person with an opinion."
– *W. Edwards Deming*

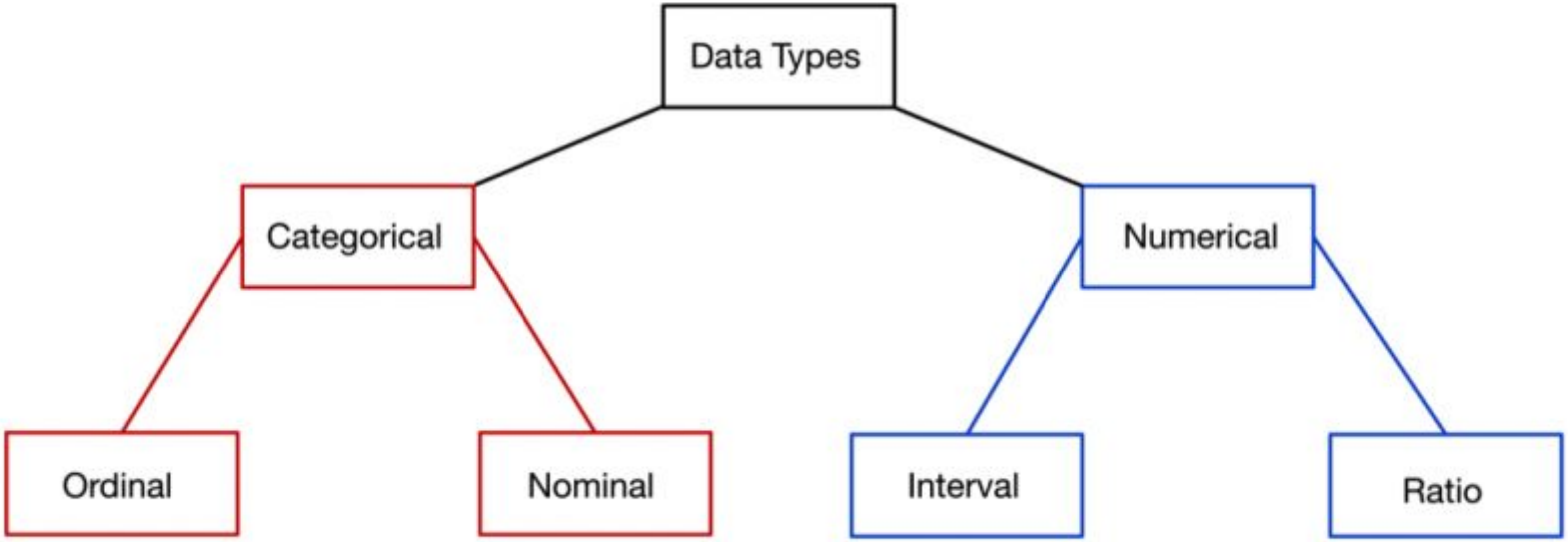
Data Types in Statistics

- **Categorical**
  - **Nominal:** Doesn’t have an order. For example: Gender of a person (male or female)
  - **Ordinal:** Has some order in place. For example:
    Grades of students (first division, second division and third division)

- **Numerical**
  - **Discrete:** Discrete Data can only take certain values. They are distinct and separate.
    Example: the number of students in a class. We can't have half a Student!
  - **Continuous:** Continuous Data can take any value (within a range).
    A person's height: could be any value (within the range of human heights), not just certain fixed heights.

!!! we will get used to these terms soon, no need worry too much about it. Read this article for additional information:
https://builtin.com/data-science/data-types-statistics

Data Types in Statistics



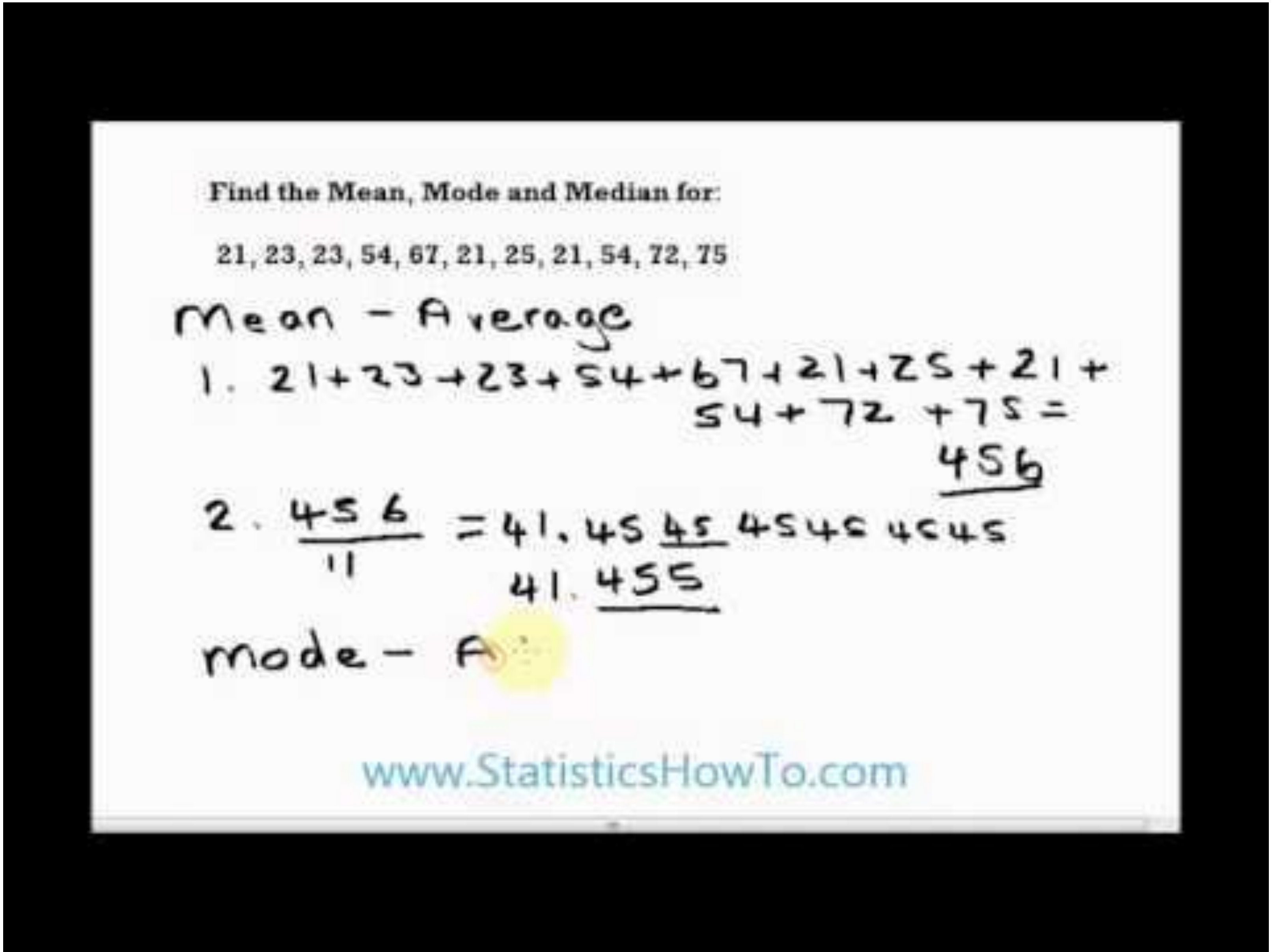
Measures of Central Tendencies

- **Mean:** The mean is the average of a data set. For example, take a list of numbers: 10, 20, 40, 10, 70
  Mean = (10 + 20 + 40 + 10 + 70) / 5 = 30

- **Median:** The median is the middle of the set of numbers.
  To find the median, first we sort the list of numbers:
  10, 10, 20, 40, 70
  The exact middle number i.e. 20 is the median.

- **Mode:** The mode is the most common number in a data set.
  In above list of numbers, 10 has occurred 2 times while other three numbers are occurred one time each.
  So, the mode is 10 here.

Explanation of Mean, Median, Mode



Find the Mean, Mode and Median for:

21, 23, 23, 54, 67, 21, 25, 21, 54, 72, 75

Mean - Average

1. 21+23+23+54+67+21+25+21+
54+72+75 =
456

2. $\frac{456}{11}$ = 41.45 45 4545 4545
41.455

mode - A

www.StatisticsHowTo.com

Applications of Central Tendencies

MEAN: When you watch a baseball game and you see the player's batting average, that number represents the total number of hits divided by the number of times at bat. In other words, that number is the mean. In school, the final grade you get in a course is usually a mean. This mean represents the total number of points you scored in the class divided by the number of possible points. This is the classic type of average – when your overall performance on many items is evaluated with a single number.

MEDIAN: Although the mean is the most common type of average, the median can also be used to express the average of a group. You may hear about the median salary for a country or city. When the average income for a country is discussed, the median is most often used because it represents the middle of a group. Mean allows very high or very low numbers to sway the outcome but median is an excellent measure of the center of a group of data.

Applications of Central Tendencies

MODE: Imagine that you live in a small town where most of the people are employed by a factory and earn minimum wage. One of the factory owners lives in the town and his salary is in the millions of dollars. If you use a measure like the average to try to compare salaries in the town as a whole, the owner's income would severely throw off the numbers. This is where the measure of mode can be useful in the real world. It tells you what most of the pieces of data are doing within a set of information.

That’s it for the day. Thank you!

Feel free to post any queries in the #help channel on Slack