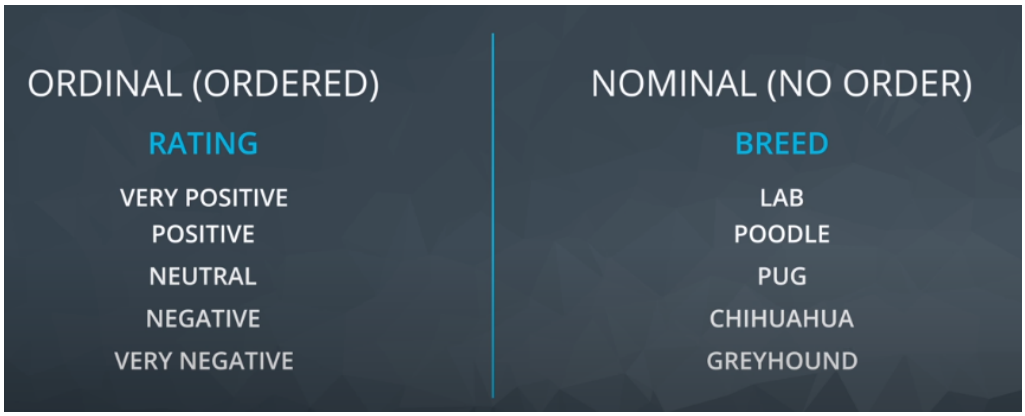
Statistics

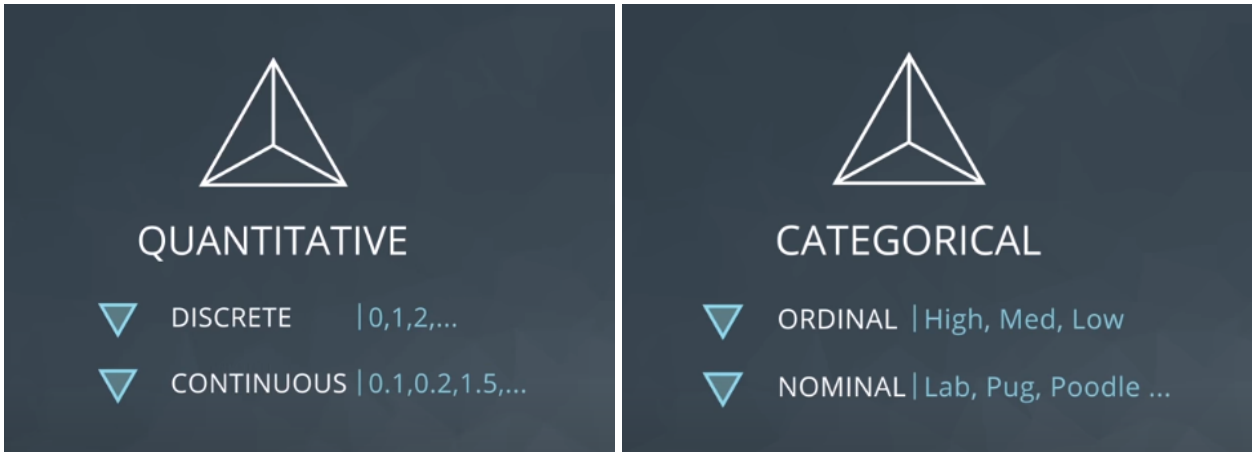
Descriptive Statistics 1.

Data Types:

- Quantitative: Takes on numeric values that allow mathematical operations. (EX: number of dogs), it can be divided into:
  - Continuous: Values that can be split into smaller values.(EX: Age of a dog), can take on any numeric value (decimals, floats or negatives)
    - 

  - Discrete : Values That are countable. (EX: number of dogs)
- Categorical: Labels a group or a set of items (EX: breeds of dogs that pass), it can be divided into :
  - Ordinal (Ordered): Values That are ranked.
  - Nominal (Unordered): Values that don't have a ranked order.
  - 

RECAP:



Analysing quantitative data:

Has 4 main aspects: Center, Spread, Shape, Outliers.

Notation: Common math language used to communicate regardless of spoken language. (Essential to communicating ideas regarding data)

Variables:

- Random: Notated by a capital letter (They have many different values)
- Observed: Notated by a lowercase letter with a subscript (signify a specific value)

Measures:

- Measure of Centre: Gives an idea of the Average (EX: average completion time of a course.), there are 3 widely accepted measure of centre:
  - Mean: The Average of all Values. (Sum of all values / Number of values).
  - Median: The middle value of the data set. (Half of the data is larger, the other half is smaller):
    - first values are ordered then depending on whether the data size is even or odd, we calculate:
      - EVEN : we take the median of the middle 2 value (EX: 8 , we take the mean of 4th&5th / 2)
      - ODD : we take the middle value as the median (EX: 7, we take the 4th value)
  - Mode: The most frequent value in a data set.
- Measure of Spread: Gives an idea of how data is spread (EX: spread of completion of a course.)

Descriptive Statistics II.

Analysing quantitative data:

Histograms:

The most common visual for quantitative data.

Quantitative data Has 4 main aspects: Center, Spread, Shape, Outliers.

Measure of spread:

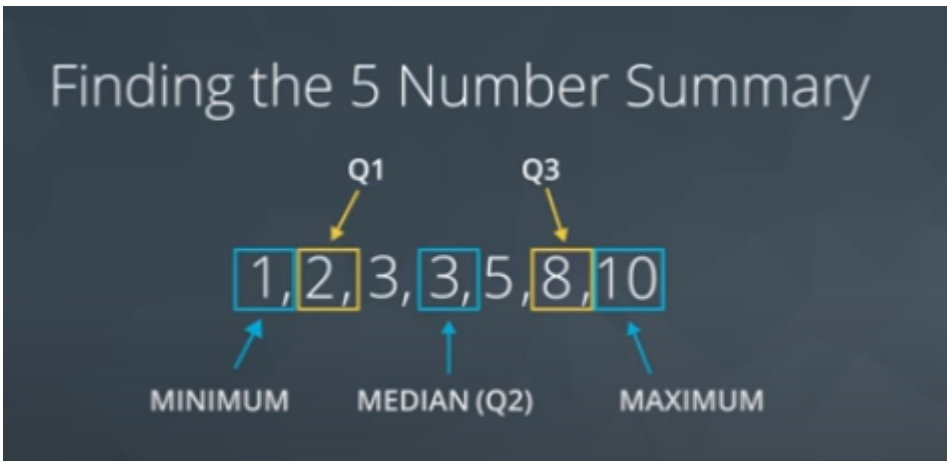
One of the most common ways to measure the spread of data is the 5-Numbers-Summary

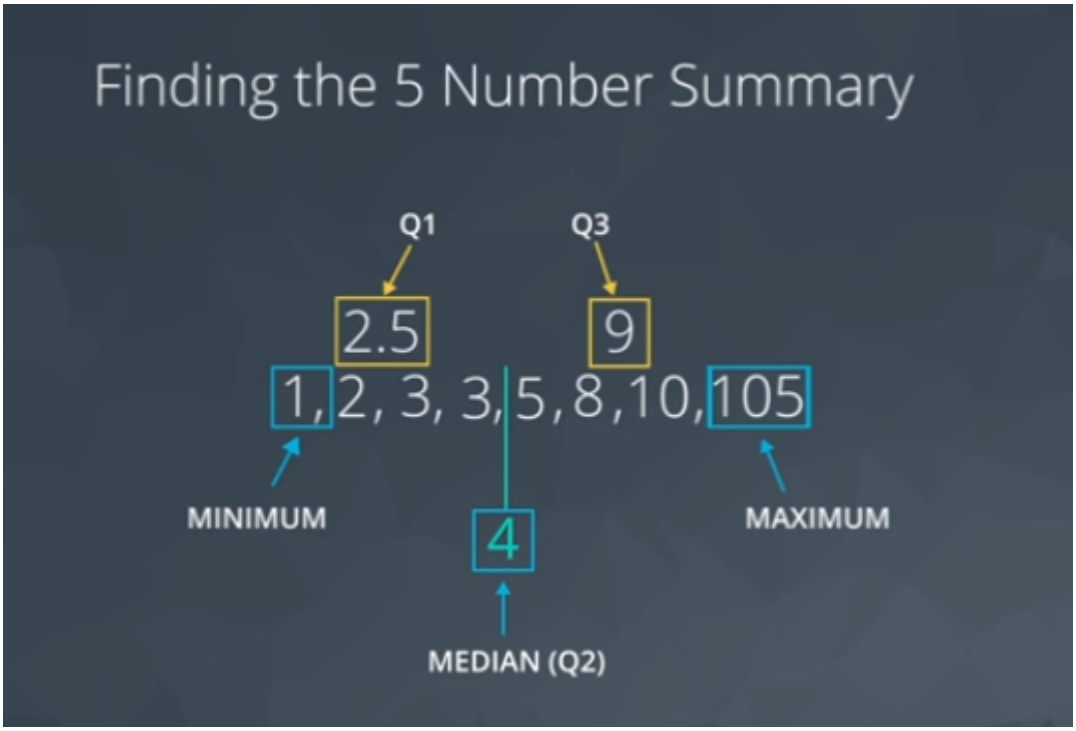
5-Numbers-Summary: gives values for calculating the range and interquartile range for a **ordered** dataset

It consists of 5 values:

- maximum : the biggest value in the data set
- third quartile : the median of the values between the maximum and second quartile (75% of the data falls below it)

- second quartile (median) : the median of the values
- first quartile : the median of the values between the minimum and second quartile (25% of the data falls below it)
- minimum : the smallest value in the data set

- odd set of values EX:
  - 

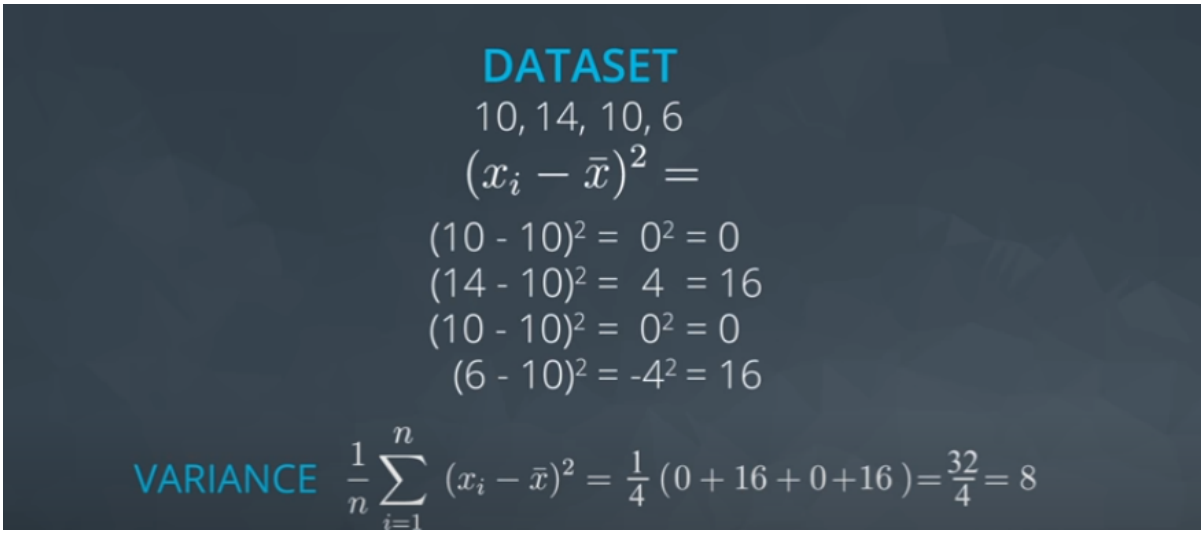
- Even set of values EX:
  - 

The range is calculated: by subtracting the maximum from the minimum.
The interquartile range is calculated: by subtracting the values of the 3rd & 1st quartiles.
The spread of data is measured most commonly using a single value is with Standard deviation or with Variance.
Standard Deviation: How much each point on average varies from the mean of the points (EX: how much on average the distance of each of the employees of a company differs from the average distance all employees are from work).
(IT IS THE SQRT OF VARIANCE)

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

Calculating the standard deviation:

- get the mean ($\bar{x}$)
- square the difference between each value of the data set and the mean ($x_i - \bar{x}$)
- get the average squared distance of each observation of the mean (variance)
- square root the ending value and we get the standard deviation

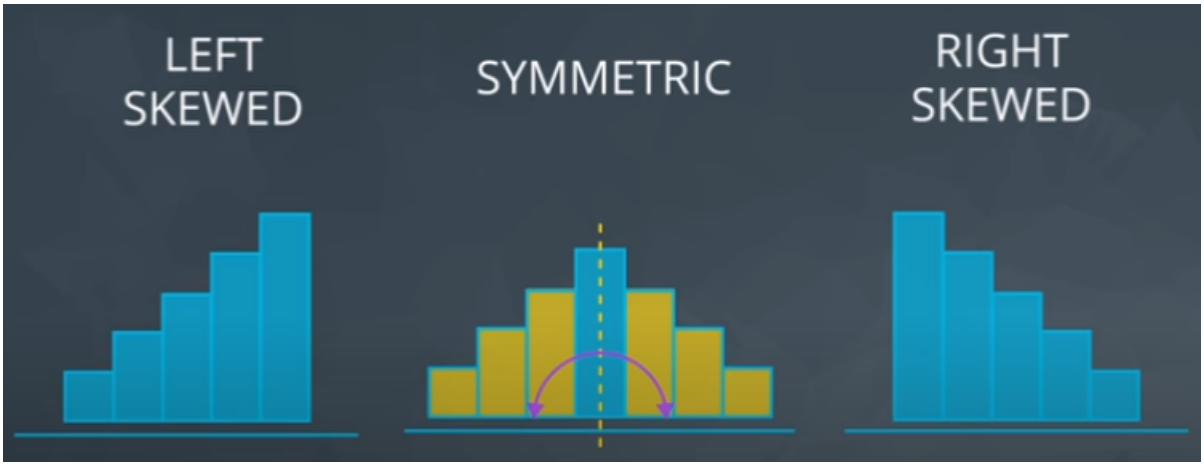
- EX:
  - 

  - The standard deviation is the sqrt of the variance
  - The higher the mean value is the lower the standard deviation and variance are

---

Measures Of Shape:

Shape : is how to use histograms to determine the shape associated with data. here are 3 examples of histogram shapes:

- Left skewed: the left most bin is smaller than the right most bin
- Right Skewed : the right most bin is smaller than the left most bin
- Symmetrical : you can draw a line down the middle and have both sides mirroring
- 

Outliers:

Data points that falls very far from the rest of the data values in out dataset.
with outliers you should:

- Note the impact they have on summary
- Remove / Fix them if they're typos
- Understand why they exist and their impact on questions we're trying to answer
- be careful when reporting and ask the right questions

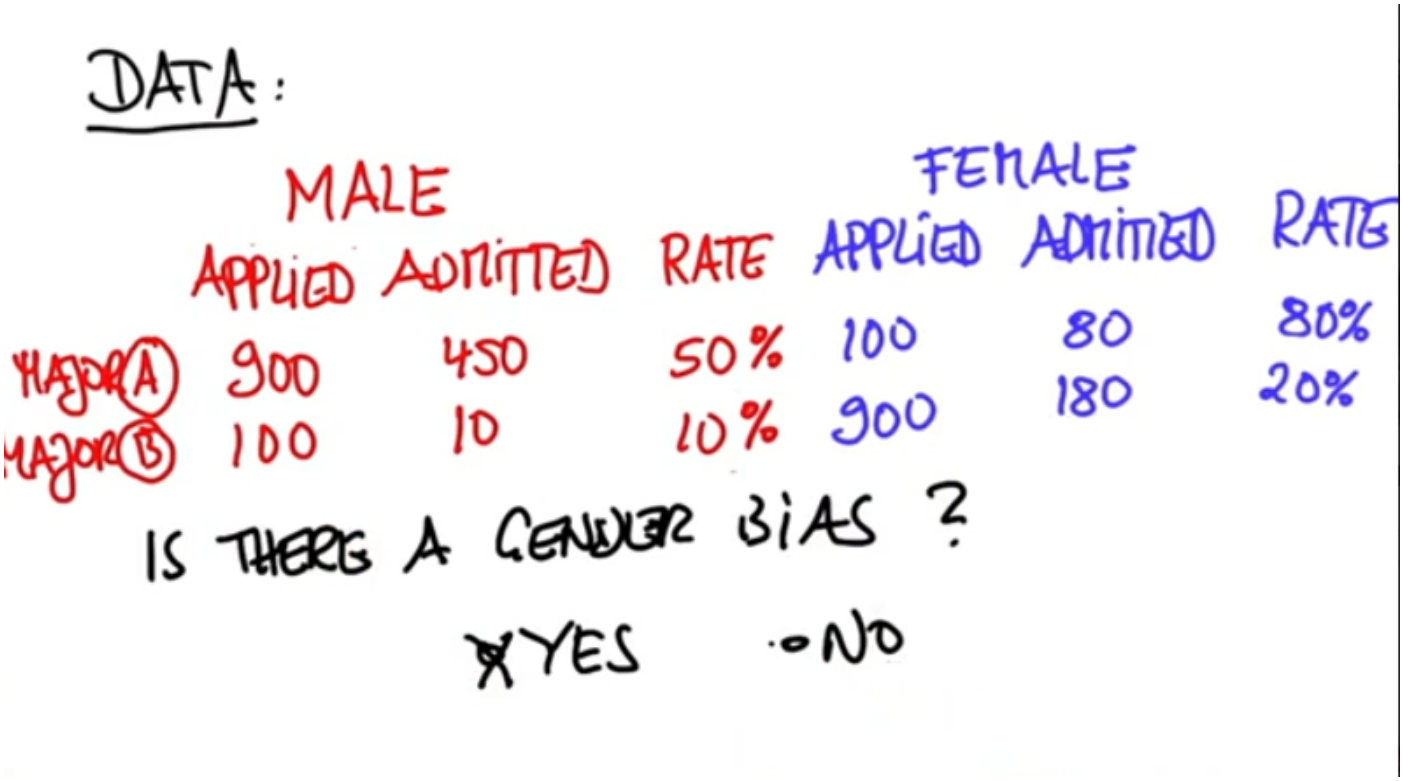
(When outliers are present it's better to use the 5-number-system instead of the mean or median)

Advanced Statistics

Simpsons Paradox:

A phenomenon in probability and statistics where a trend appears in several data groups but disappears or reverses when the groups are combined.
(It shows how different data groupings can lead to very different conclusions)

EX:

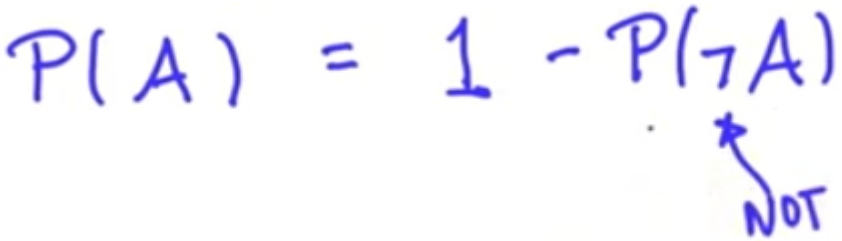


A couple years a go UC Berkeley did a test to see if there was any gender bias in their acceptance rates.(the example above follows what UC Berkeley did but with made-up numbers).
In the example if we look at the admittance rates for both males and females separated by major

we find out that the males have a lower admittance rate instead of the females in both majors
we reach a conclusion that : there is female bias in admittance rates!.
however if we look at both majors combined we'll find that for males, out of the applied 1000, 460 got accepted leading to an admittance rate of 46%.
as for females, out of the 1000 applied only 260 got accepted leading to an admittance rate of only 26%.
by looking at such data we conclude that there is severe male bias in overall acceptance.

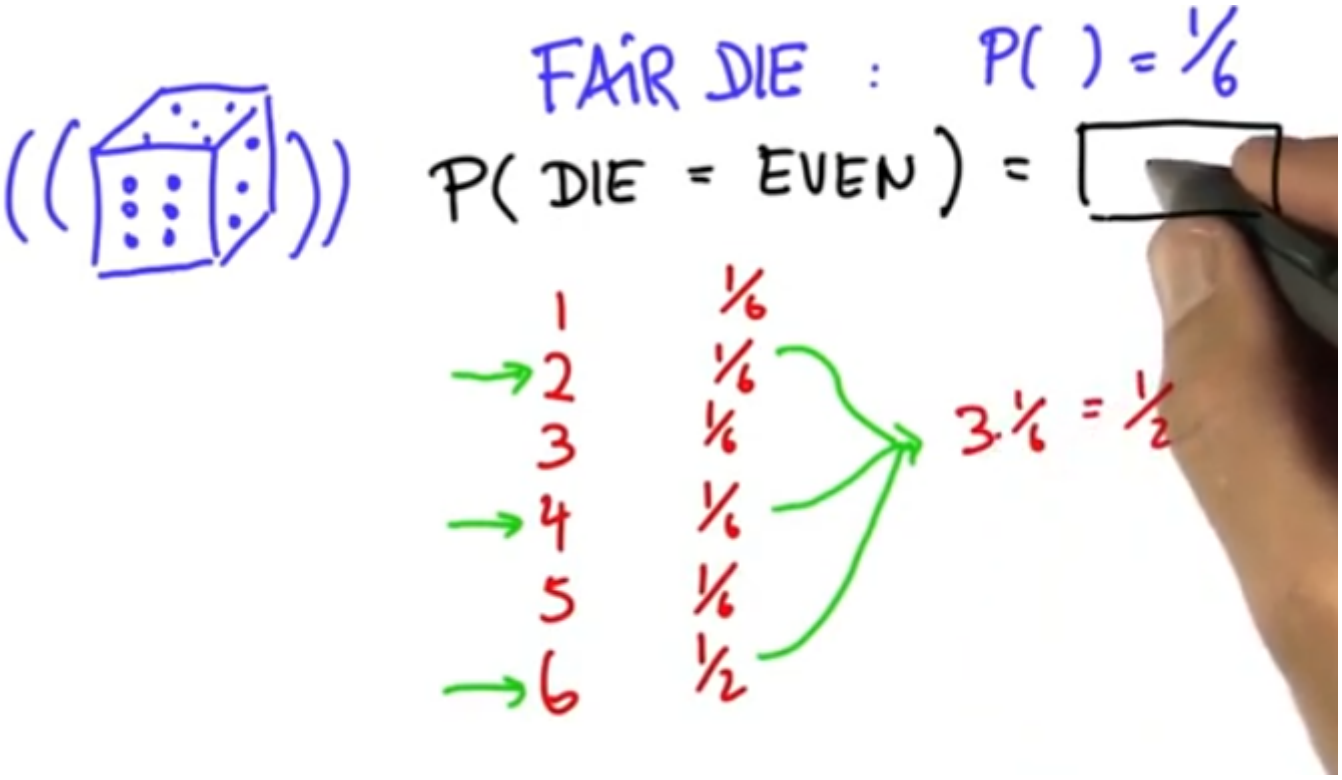
Probability

It is the opposite of statistics as in statistics we use analyse data, in probability we predict data using assumptions we make about it.
Basic probability law : The probability of an event is 1 - the probability of opposite event

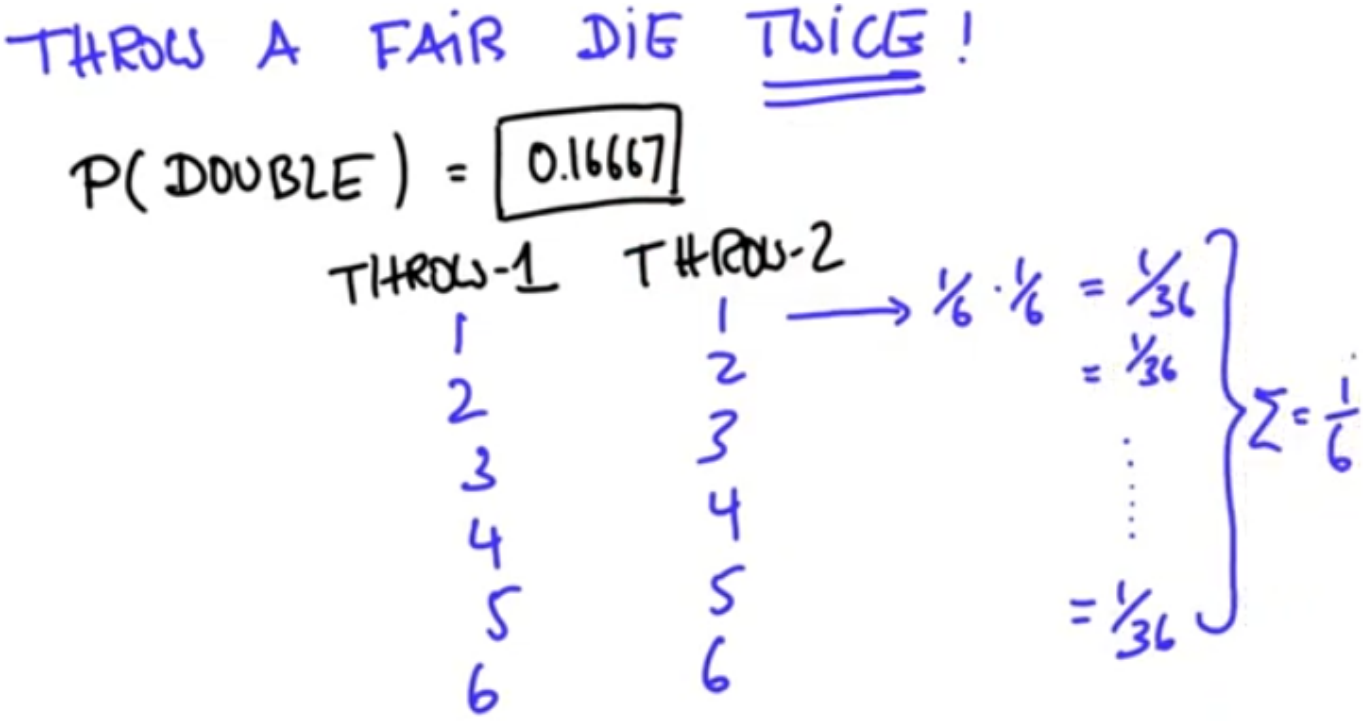


You can get the probability of a composite event which is the probability (p) times how many event wanted:
EX:
how many times can you flip a coin and get a tails :
solution is the P(tails) * P(tails)

EX: (with solution) how many times can you get an even number on a die flip: the outcome is 0.5

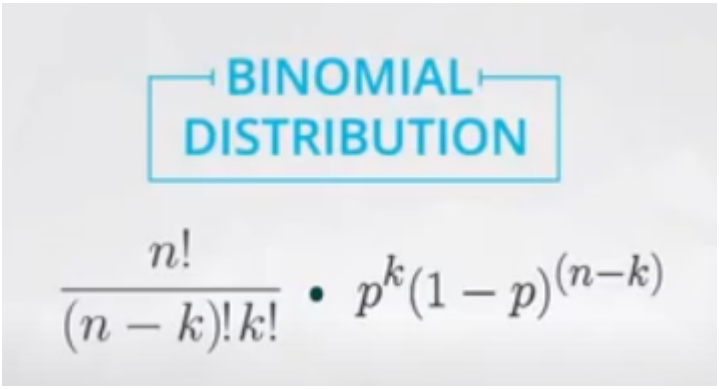


EX2: how many times can you get a double (same number) in a fair dice thrown twice



Binominal Distribution

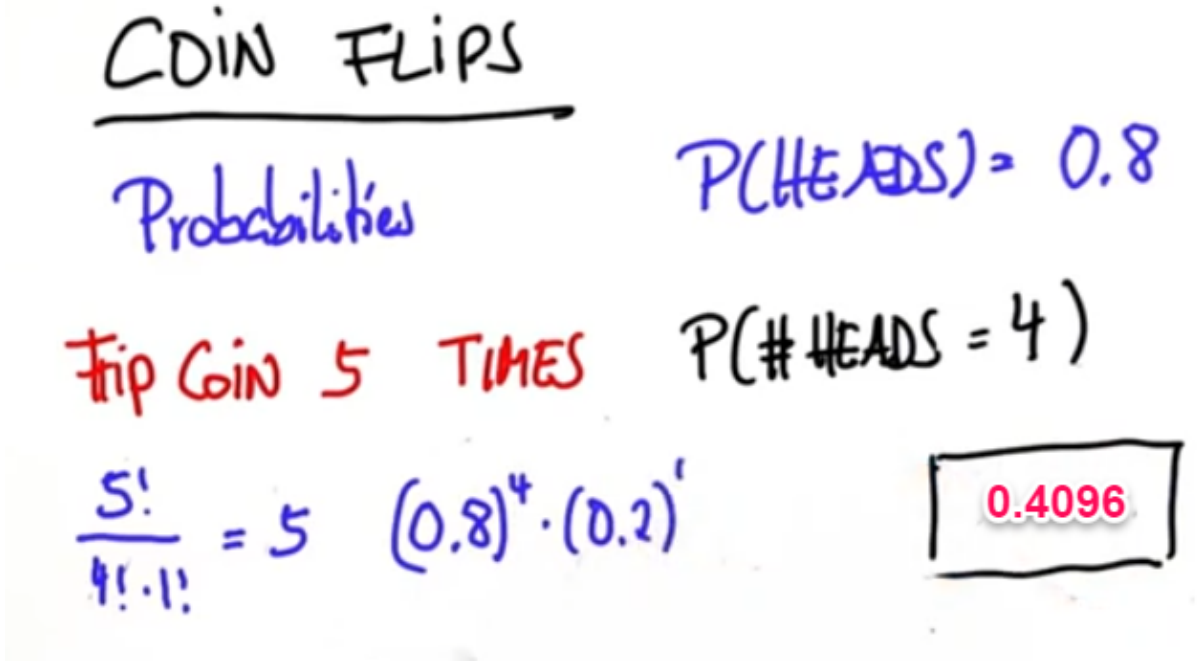
We here continue over the last lesson but we have a mathematical formula of possible ways to get a side of a coin depending on probability



EX:

in the previous example if we flipped a coin (n) times and the we wanted to see how many heads appeared (k)
the formula will use the binominal formula as stated in the example
Examples with solutions:

Solution:



In the previous example, we flipped the coin 5 times and wanted to check how many times we get 4 heads

by **substituting** in the binominal formula we get the **number of times the condition appears**

multiplying the condition by the probability of said condition gets us the **Probability**