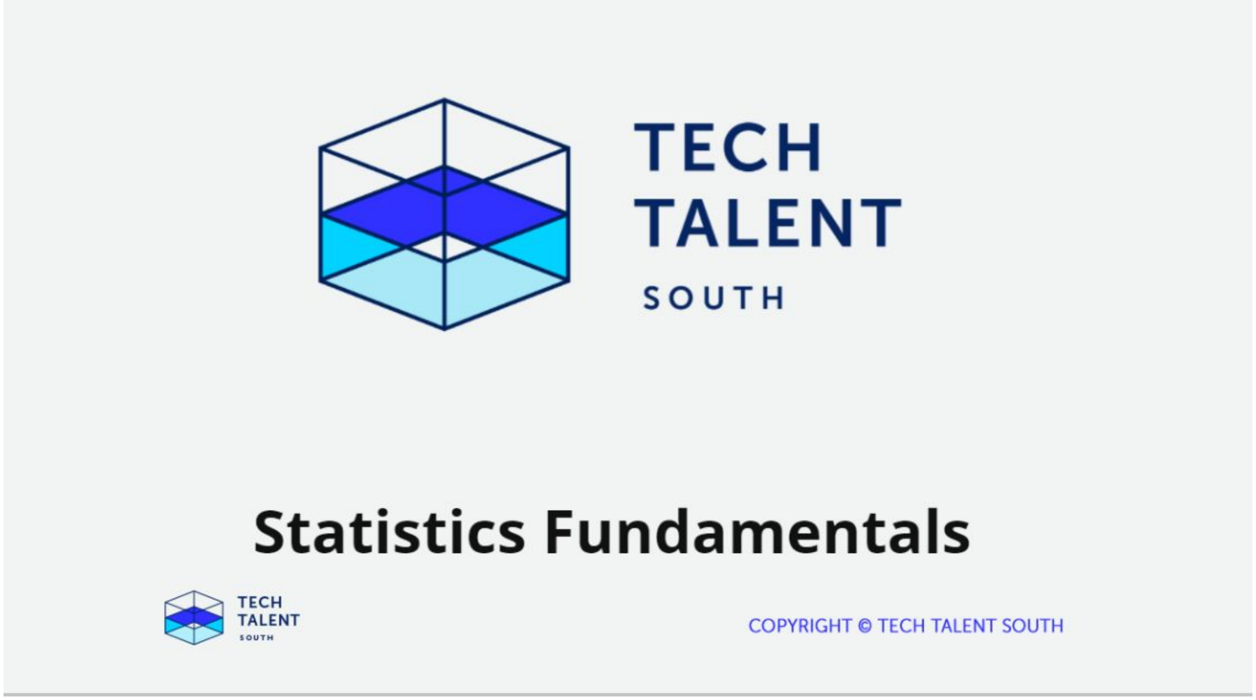
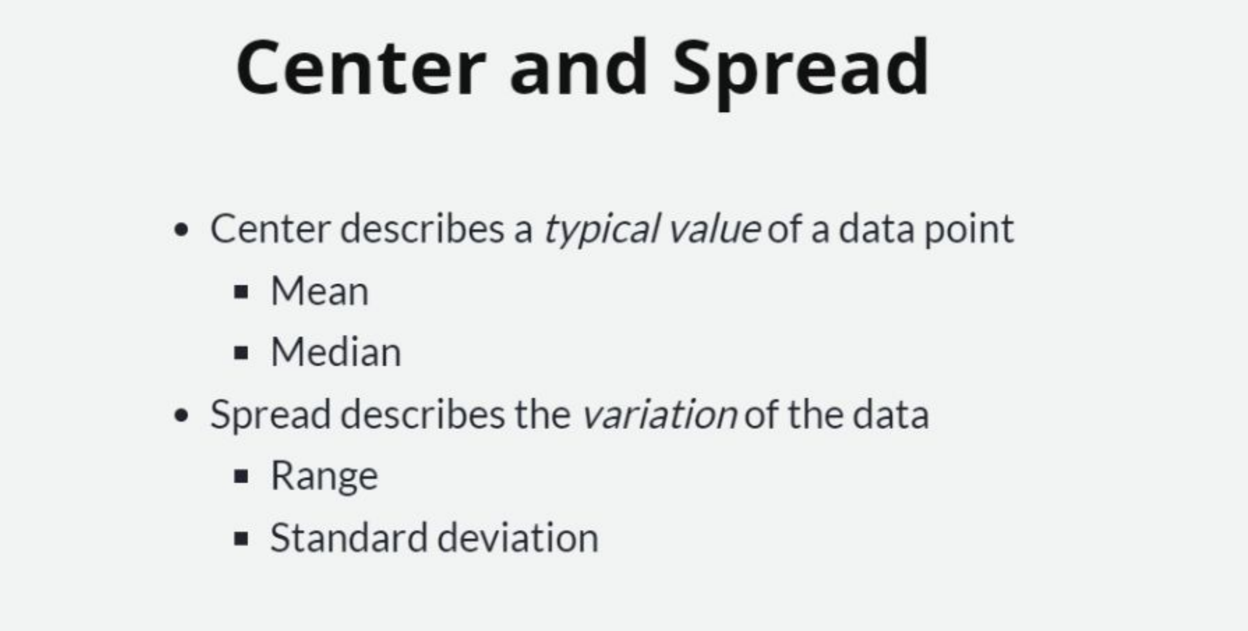


TECH TALENT SOUTH

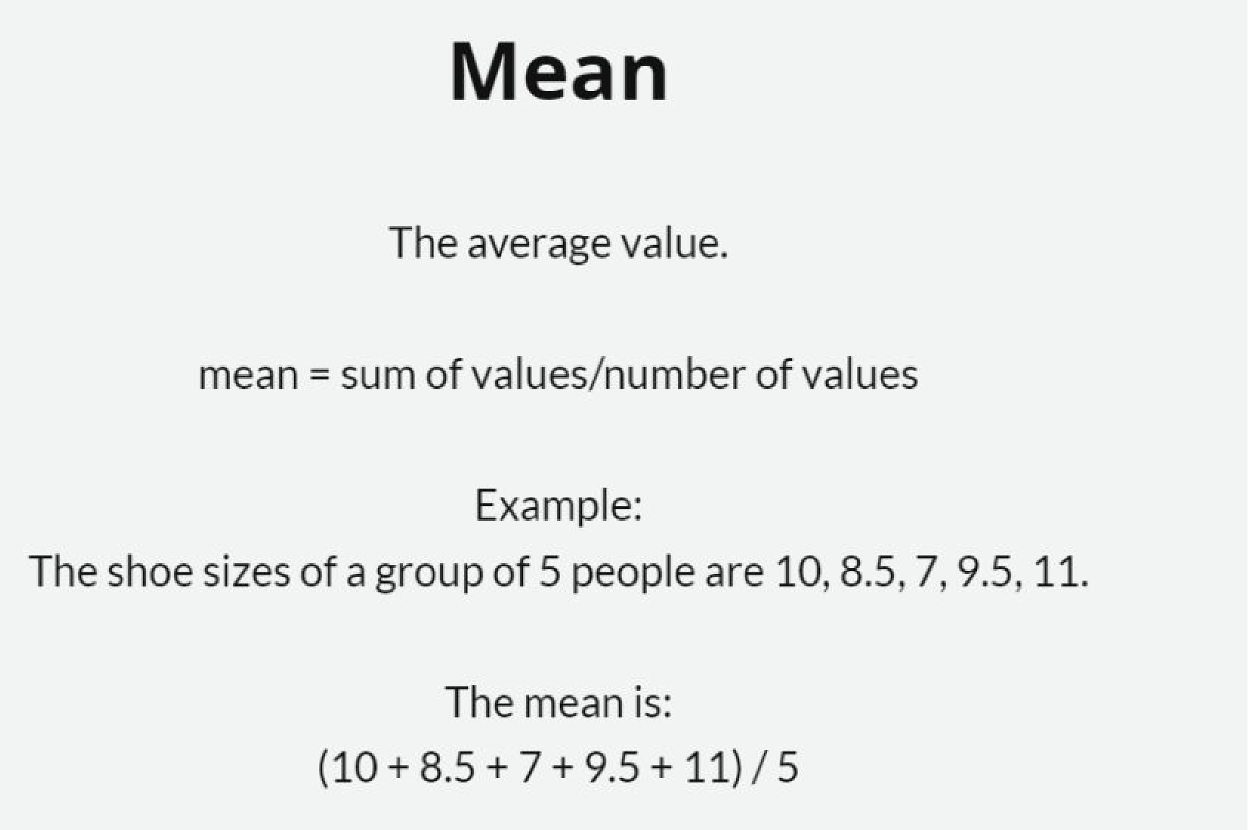
Statistics Fundamentals

TECH TALENT SOUTH



Center and Spread

- Center describes a *typical value* of a data point
  - Mean
  - Median
- Spread describes the *variation* of the data
  - Range
  - Standard deviation



Mean

The average value.

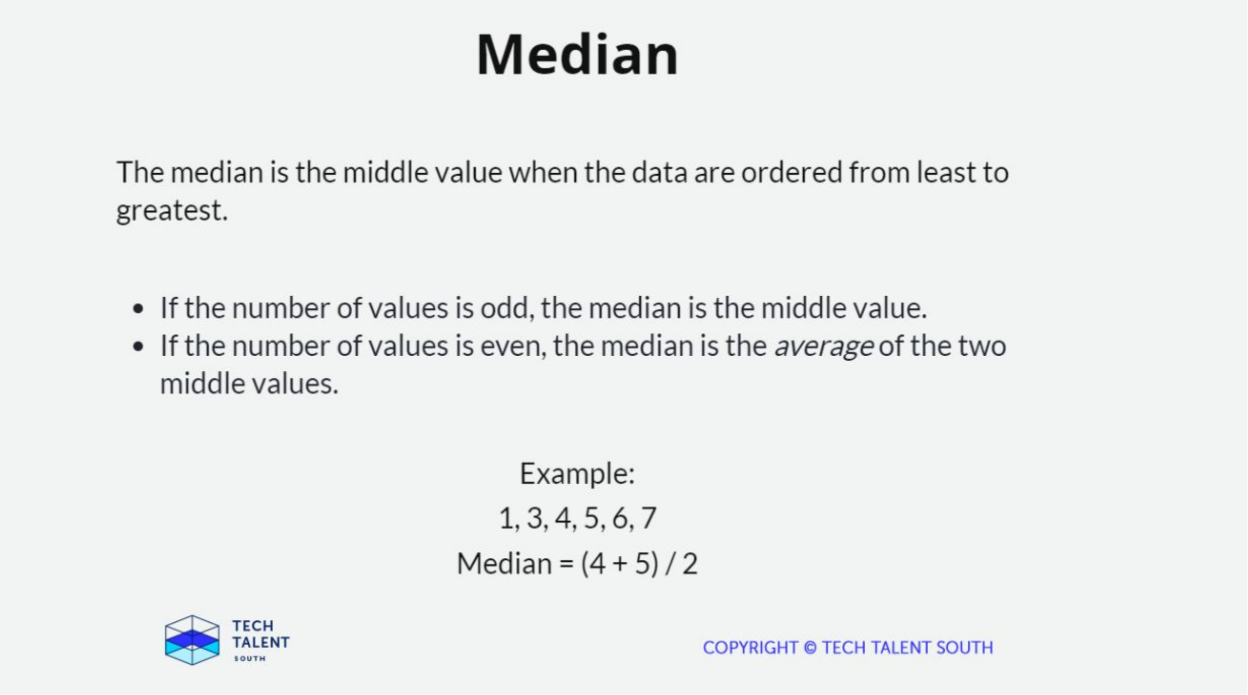
mean = sum of values/number of values

Example:

The shoe sizes of a group of 5 people are 10, 8.5, 7, 9.5, 11.

The mean is:

(10 + 8.5 + 7 + 9.5 + 11) / 5



Median

The median is the middle value when the data are ordered from least to greatest.

- If the number of values is odd, the median is the middle value.
- If the number of values is even, the median is the *average* of the two middle values.

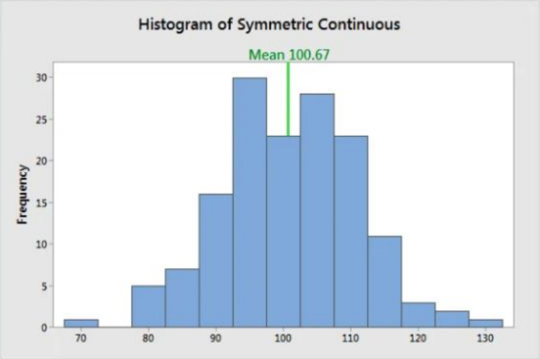
Example:

1, 3, 4, 5, 6, 7

Median = (4 + 5) / 2

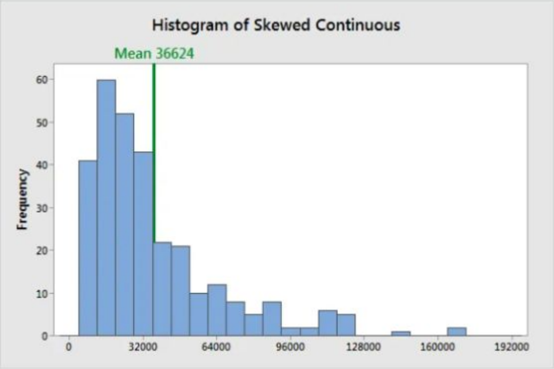
Mean vs Median

The mean is heavily affected by outliers. In a symmetrical distribution, it will locate a value close to the center of the data.



Mean vs Median

However, in a skewed distribution, where the data has many values concentrated on one side with a few outliers on the other, we can see that the mean misses the mark.



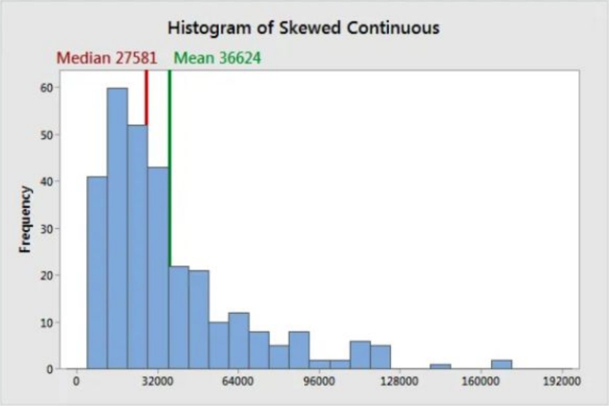
Mean vs Median

Because this, the median can be very useful. It splits the data in half and picks the centermost value, mitigating the issues that skewed data can cause.

For example, when analyzing household incomes, taking the mean will result in a higher value than expected. The outliers who earn lots of money will pull the mean up. Meanwhile, the median will lie closer to the income level one would expect from the average person in the area.

Mean vs Median

This graph showcases this using data from U.S. household income for 2006.

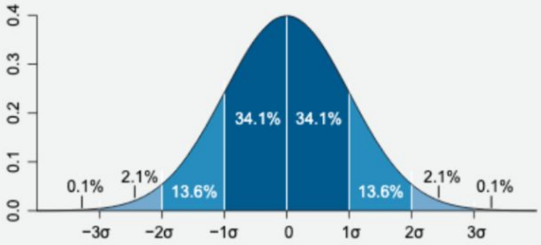


TECH TALENT SOUTH

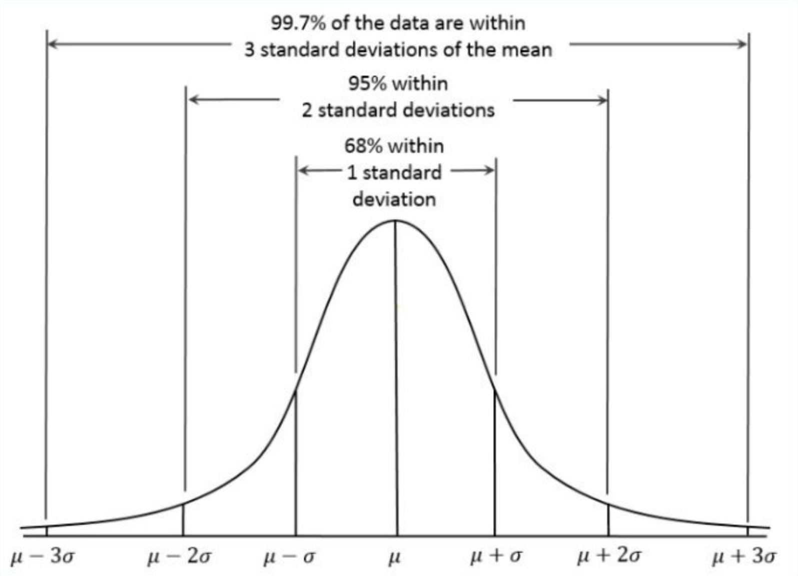
Standard Deviation

Standard deviation is a measure of the spread of the distribution of data.

A low standard deviation indicates that much of the data lies close to the mean. A high standard deviation means the data is spread out.



Standard Deviation



Random Sampling

We spoke a little bit about sampling the other day. Let's take another look at random sampling.

A good sample is *representative* and *random*.

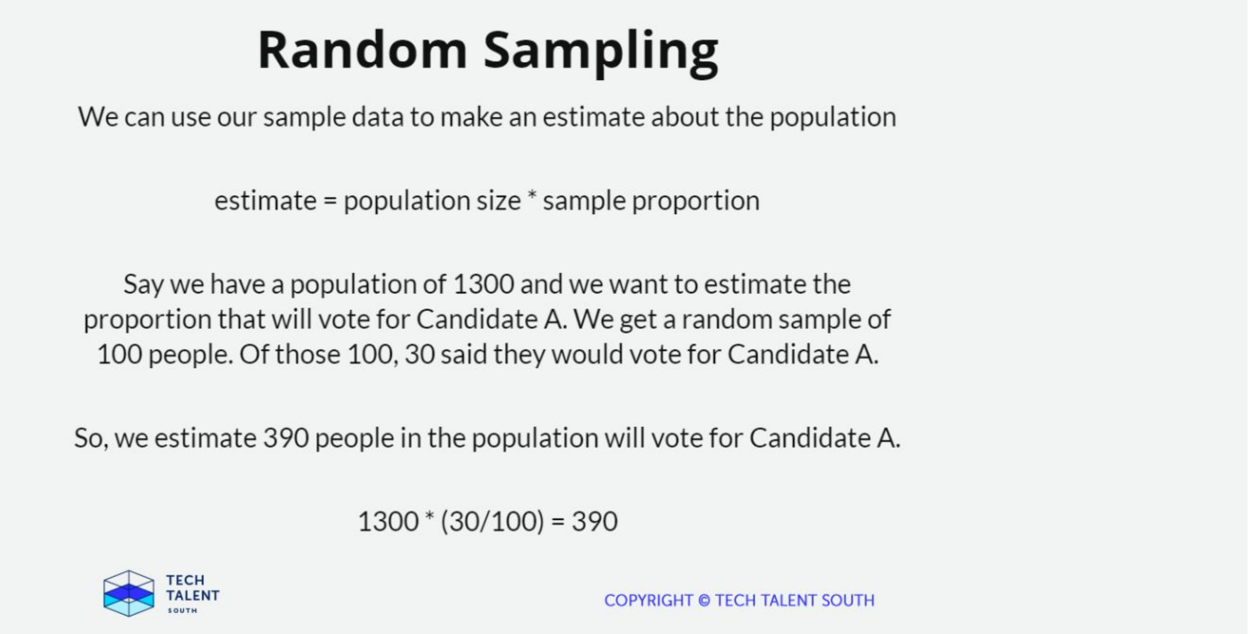
Random Sampling

Representative

- Only members of the population you want are being studied

Random

- Every member of the population has an equal chance to be sampled



Random Sampling

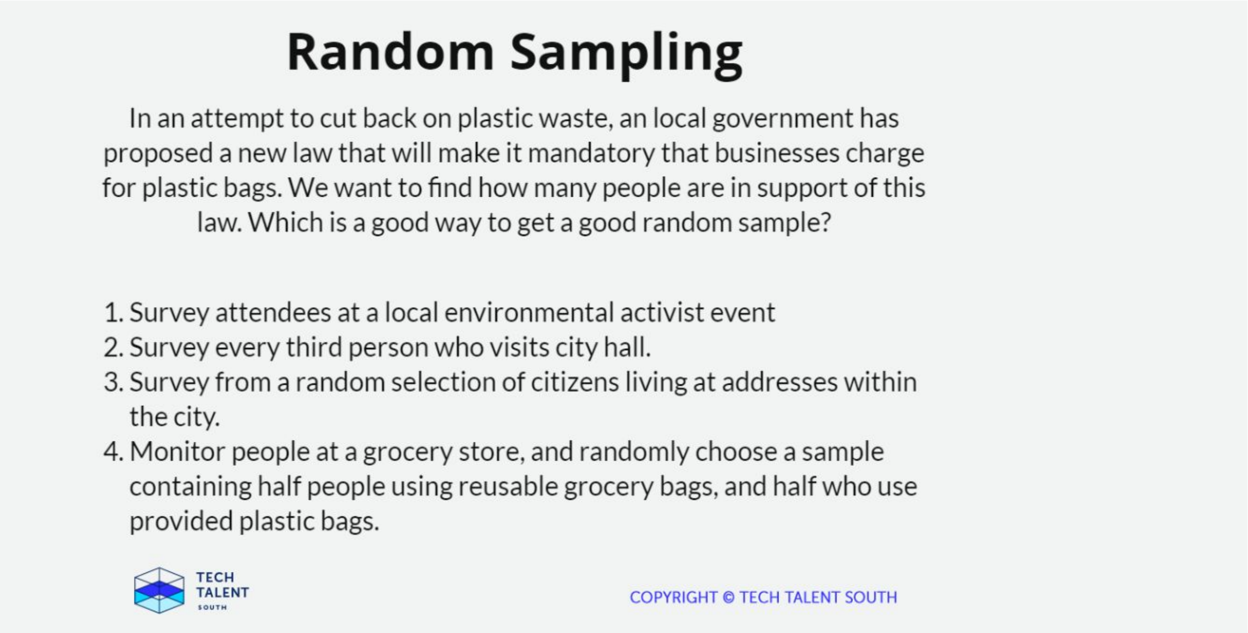
We can use our sample data to make an estimate about the population

estimate = population size * sample proportion

Say we have a population of 1300 and we want to estimate the proportion that will vote for Candidate A. We get a random sample of 100 people. Of those 100, 30 said they would vote for Candidate A.

So, we estimate 390 people in the population will vote for Candidate A.

1300 * (30/100) = 390



Random Sampling

In an attempt to cut back on plastic waste, an local government has proposed a new law that will make it mandatory that businesses charge for plastic bags. We want to find how many people are in support of this law. Which is a good way to get a good random sample?

1. Survey attendees at a local environmental activist event
2. Survey every third person who visits city hall.
3. Survey from a random selection of citizens living at addresses within the city.
4. Monitor people at a grocery store, and randomly choose a sample containing half people using reusable grocery bags, and half who use provided plastic bags.

Random Sampling

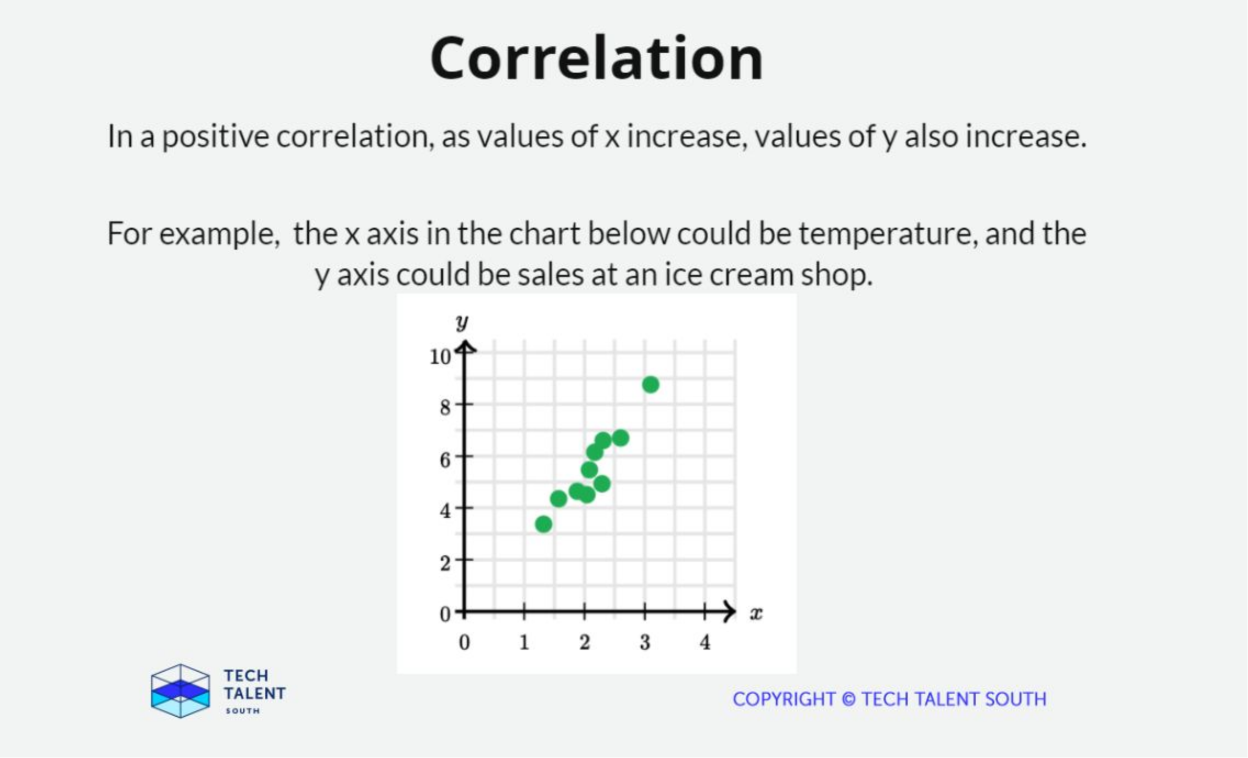
1. Survey attendees at a local environmental activist event
2. Survey every third person who visits city hall.
3. **Survey from a random selection of citizens living at addresses within the city.**
4. Monitor people at a grocery store, and randomly choose a sample containing half people using reusable grocery bags, and half who use provided plastic bags.

The correct answer is 3. Why is this the best option, and why may the other options produce biased results?

Correlation

Correlation is the dependence of two variables on each other. It describes how two variables change in relation to each other.

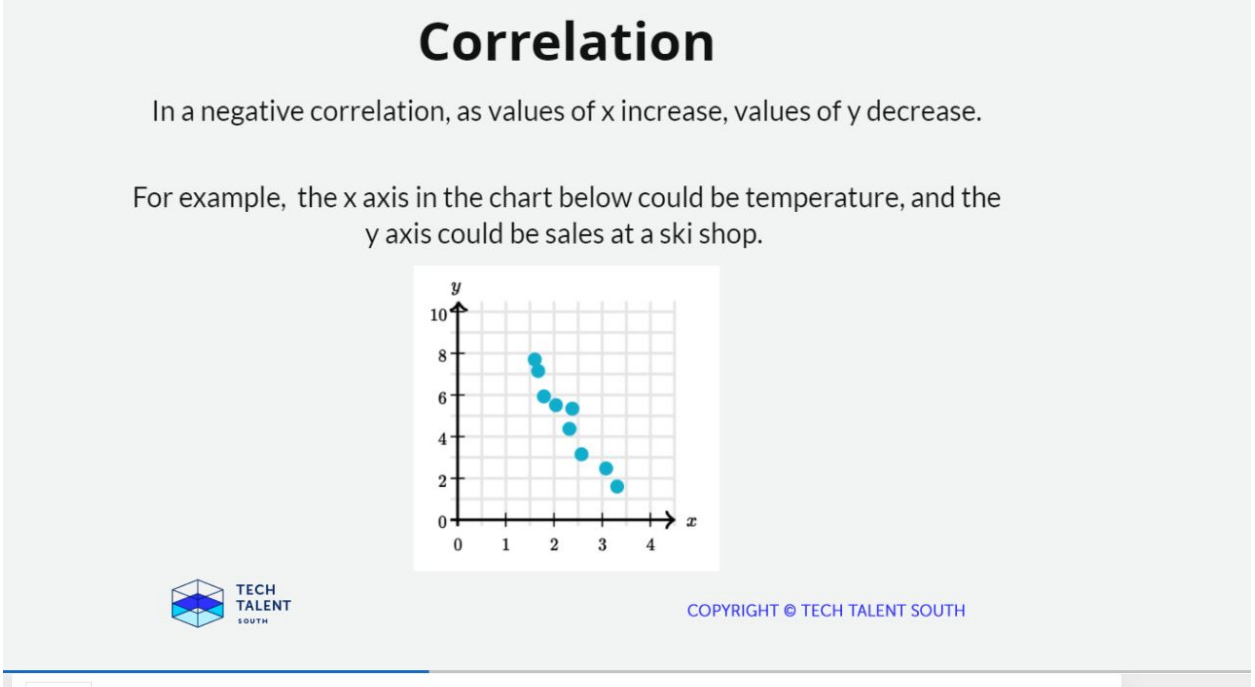
A common way to showcase correlation is through scatterplots.



Correlation

In a positive correlation, as values of x increase, values of y also increase.

For example, the x axis in the chart below could be temperature, and the y axis could be sales at an ice cream shop.



Correlation

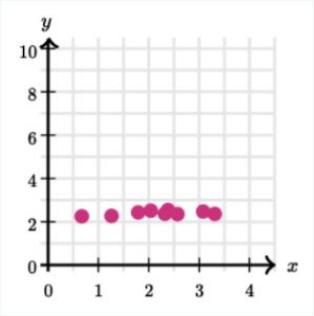
In a negative correlation, as values of x increase, values of y decrease.

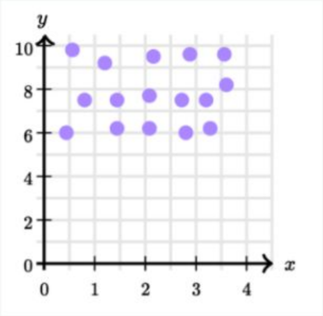
For example, the x axis in the chart below could be temperature, and the y axis could be sales at a ski shop.

Correlation

If there is no correlation, the values of y either stay relatively the same or have no clear pattern as x changes.

What are some examples you can think of for either case?





TECH TALENT SOUTH

Probability

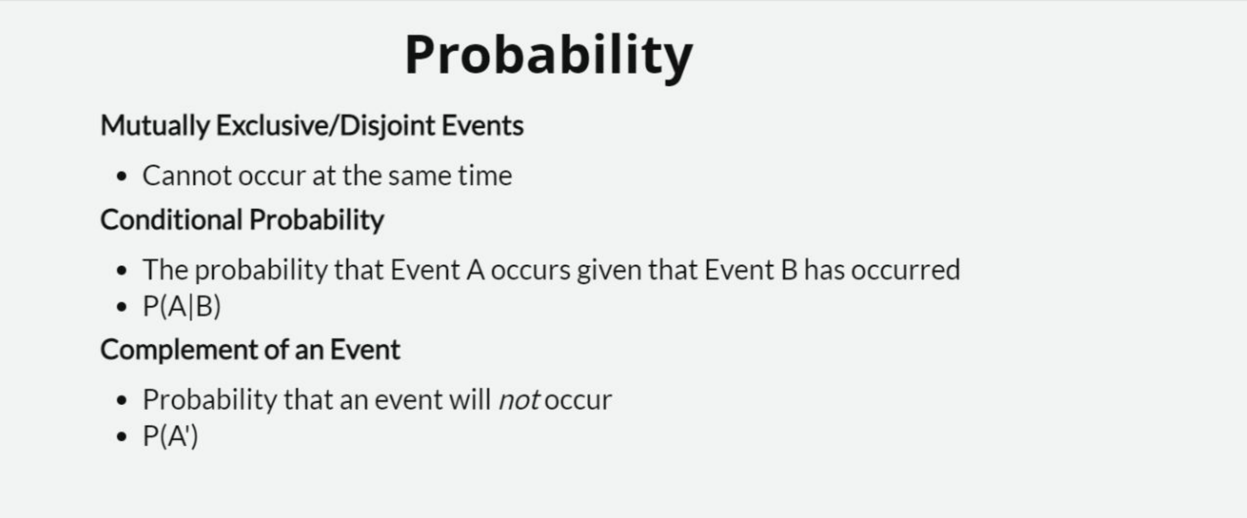
When analyzing data, we'll often want to determine Probability. Probability is the likelihood an event will occur.

The probability of an event E is equal to the number of ways it can happen divided by the total number of outcomes.

For example, when flipping a coin, the probability of heads is:

1/2 = 0.5

(one side is heads)/(two possible sides)



Probability

Mutually Exclusive/Disjoint Events

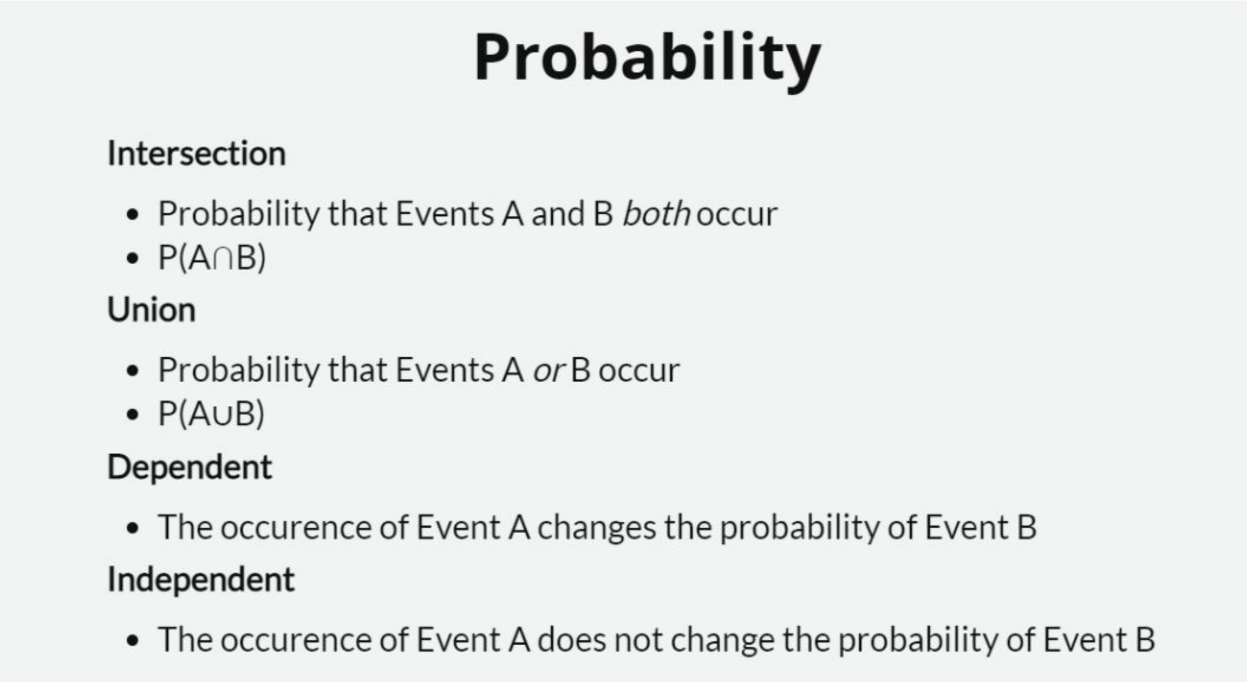
- Cannot occur at the same time

Conditional Probability

- The probability that Event A occurs given that Event B has occurred
- $P(A|B)$

Complement of an Event

- Probability that an event will *not* occur
- $P(A')$



Probability

Intersection

- Probability that Events A and B *both* occur
- $P(A \cap B)$

Union

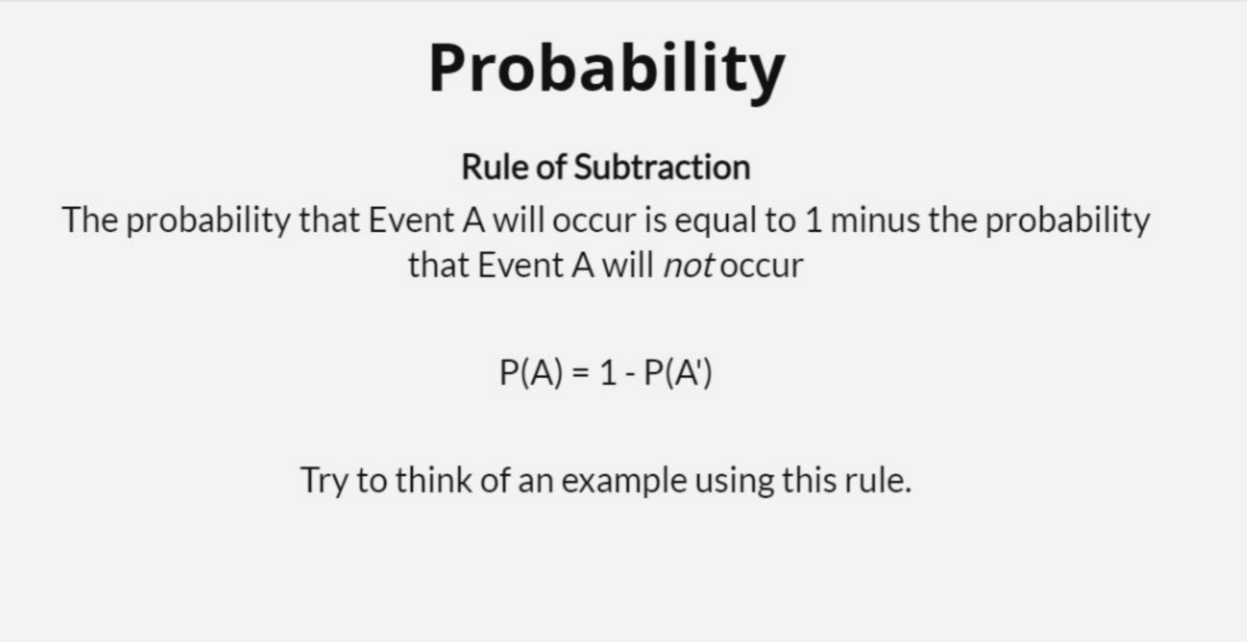
- Probability that Events A *or* B occur
- $P(A \cup B)$

Dependent

- The occurence of Event A changes the probability of Event B

Independent

- The occurence of Event A does not change the probability of Event B



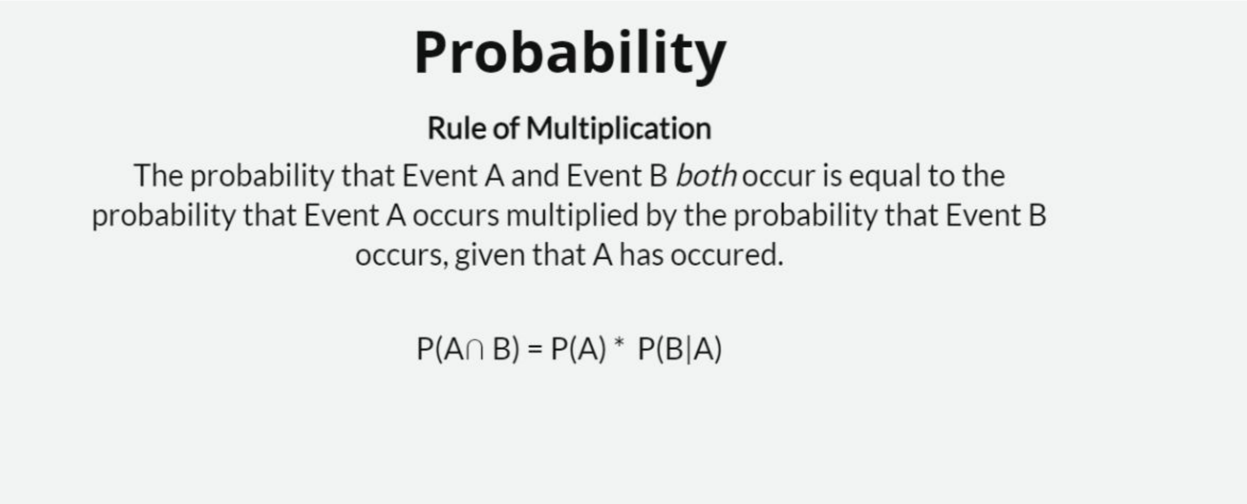
Probability

Rule of Subtraction

The probability that Event A will occur is equal to 1 minus the probability that Event A will *not* occur

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

Try to think of an example using this rule.

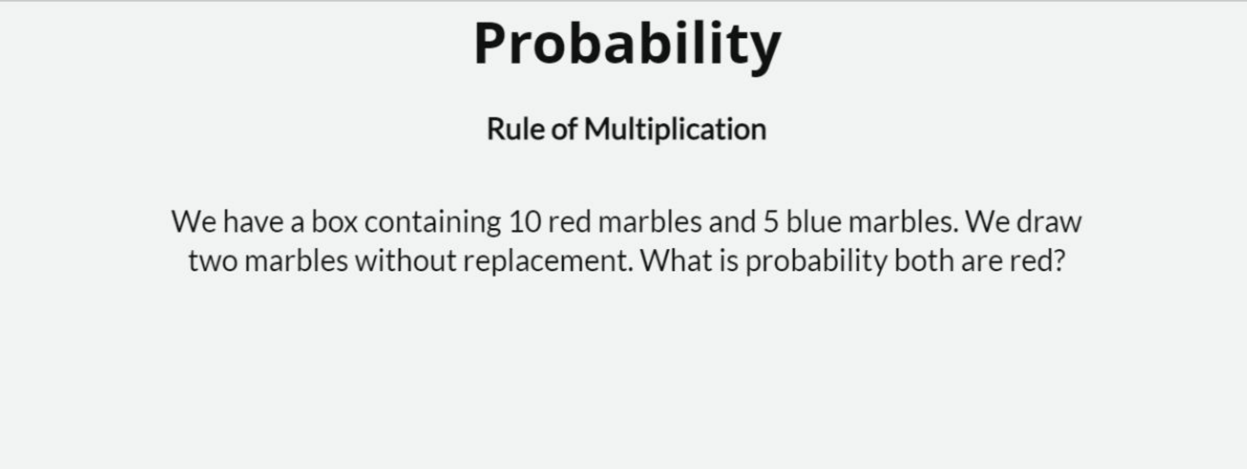


Probability

Rule of Multiplication

The probability that Event A and Event B *both* occur is equal to the probability that Event A occurs multiplied by the probability that Event B occurs, given that A has occured.

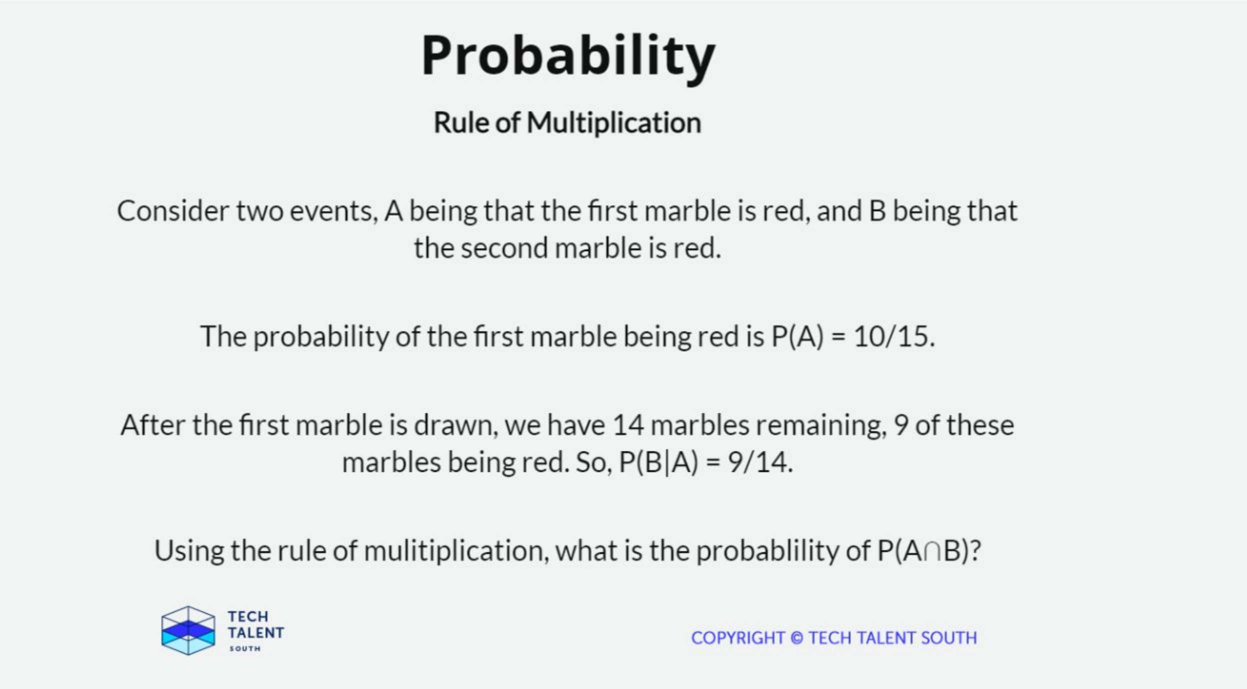
$$P(A \cap B) = P(A) * P(B|A)$$



Probability

Rule of Multiplication

We have a box containing 10 red marbles and 5 blue marbles. We draw two marbles without replacement. What is probability both are red?



Probability

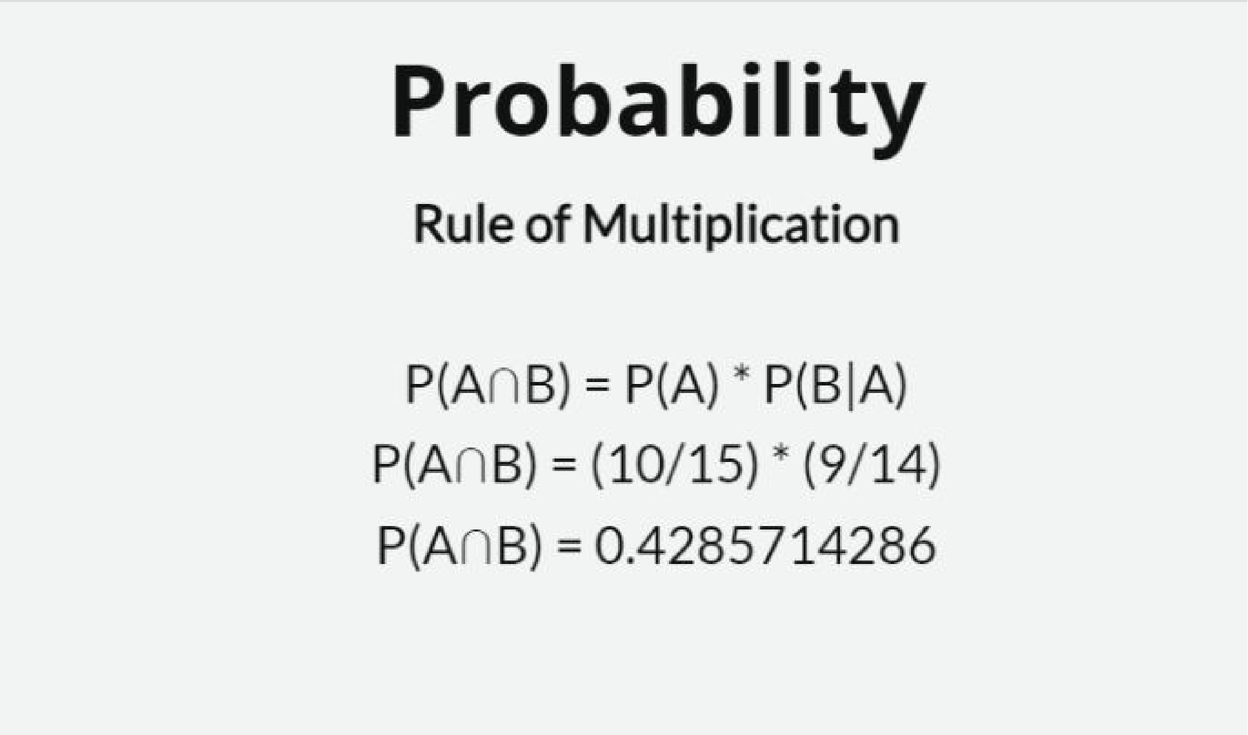
Rule of Multiplication

Consider two events, A being that the first marble is red, and B being that the second marble is red.

The probability of the first marble being red is P(A) = 10/15.

After the first marble is drawn, we have 14 marbles remaining, 9 of these marbles being red. So, P(B|A) = 9/14.

Using the rule of mulitiplication, what is the probablility of P(A∩B)?



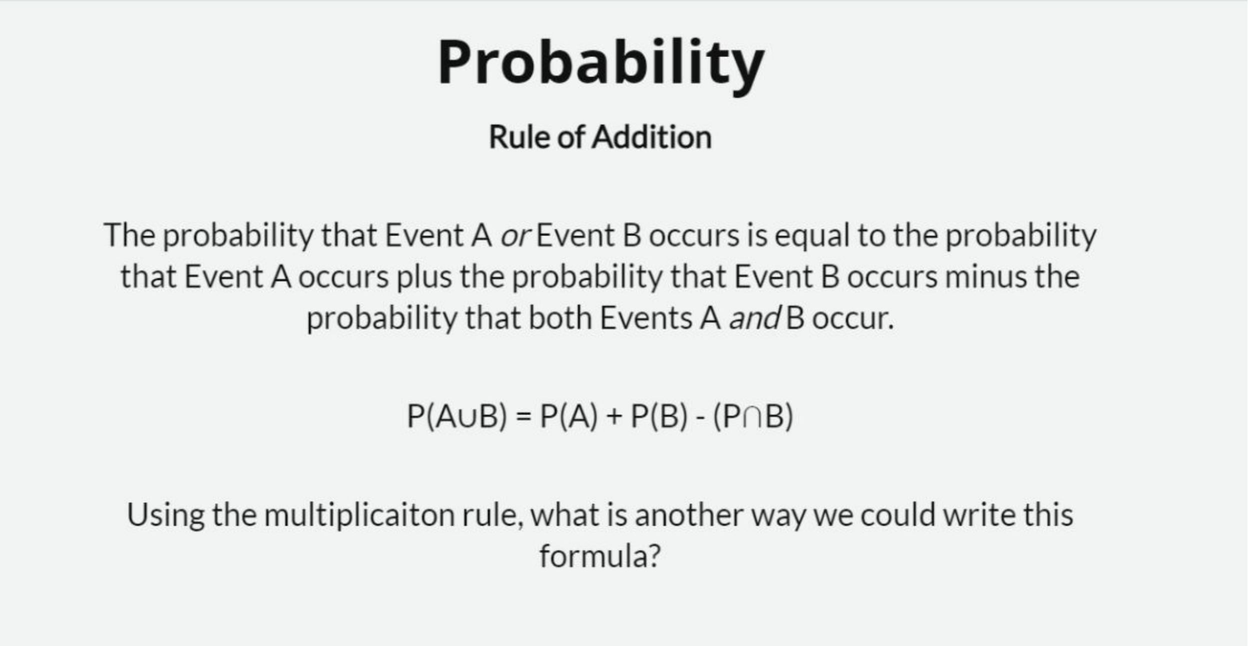
Probability

Rule of Multiplication

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

$$P(A\cap B) = (10/15) * (9/14)$$

$$P(A\cap B) = 0.4285714286$$



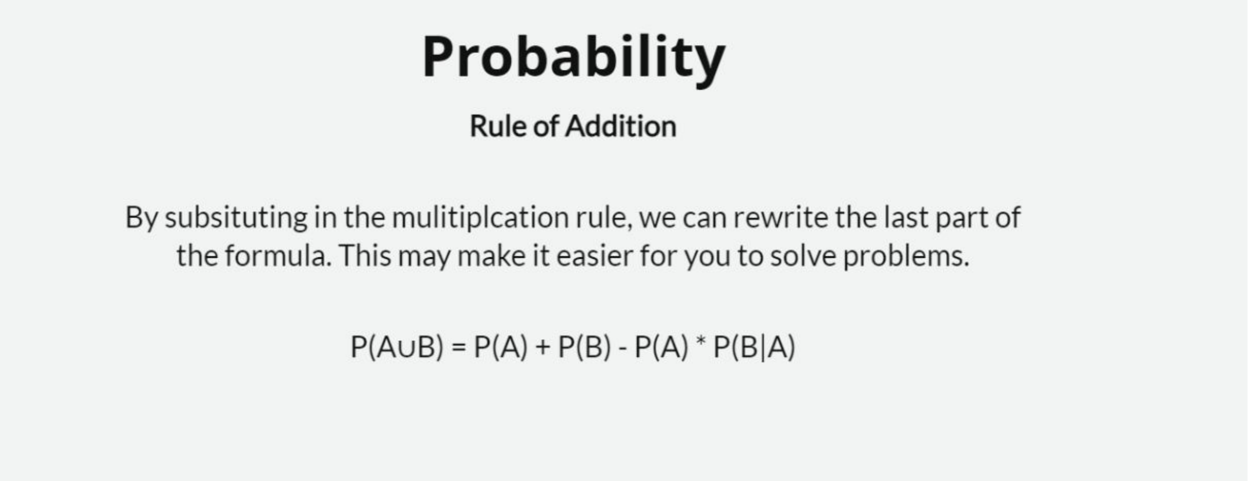
Probability

Rule of Addition

The probability that Event A *or* Event B occurs is equal to the probability that Event A occurs plus the probability that Event B occurs minus the probability that both Events A *and* B occur.

$$P(A\cup B) = P(A) + P(B) - (P\cap B)$$

Using the multiplicaiton rule, what is another way we could write this formula?

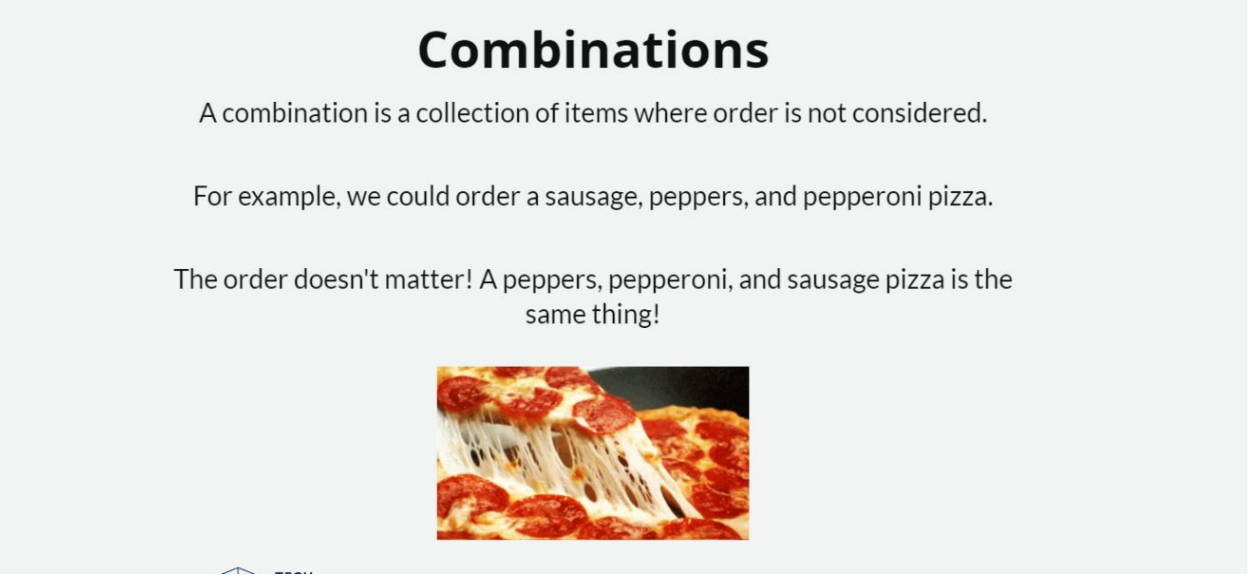


Probability

Rule of Addition

By subsituting in the mulitiplcation rule, we can rewrite the last part of the formula. This may make it easier for you to solve problems.

$$P(A \cup B) = P(A) + P(B) - P(A) * P(B|A)$$

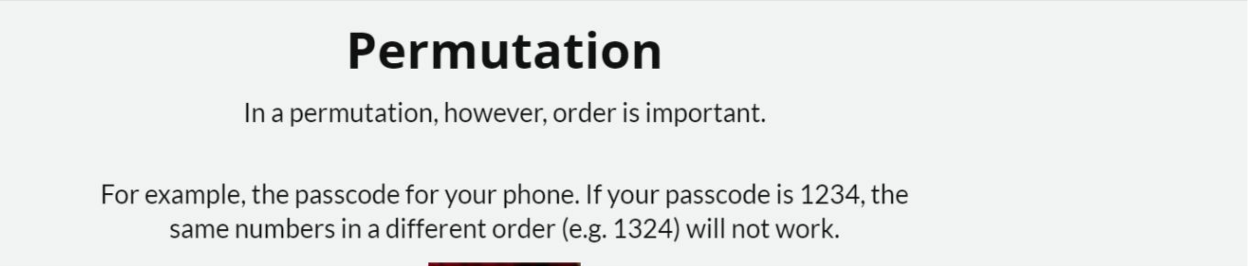


Combinations

A combination is a collection of items where order is not considered.

For example, we could order a sausage, peppers, and pepperoni pizza.

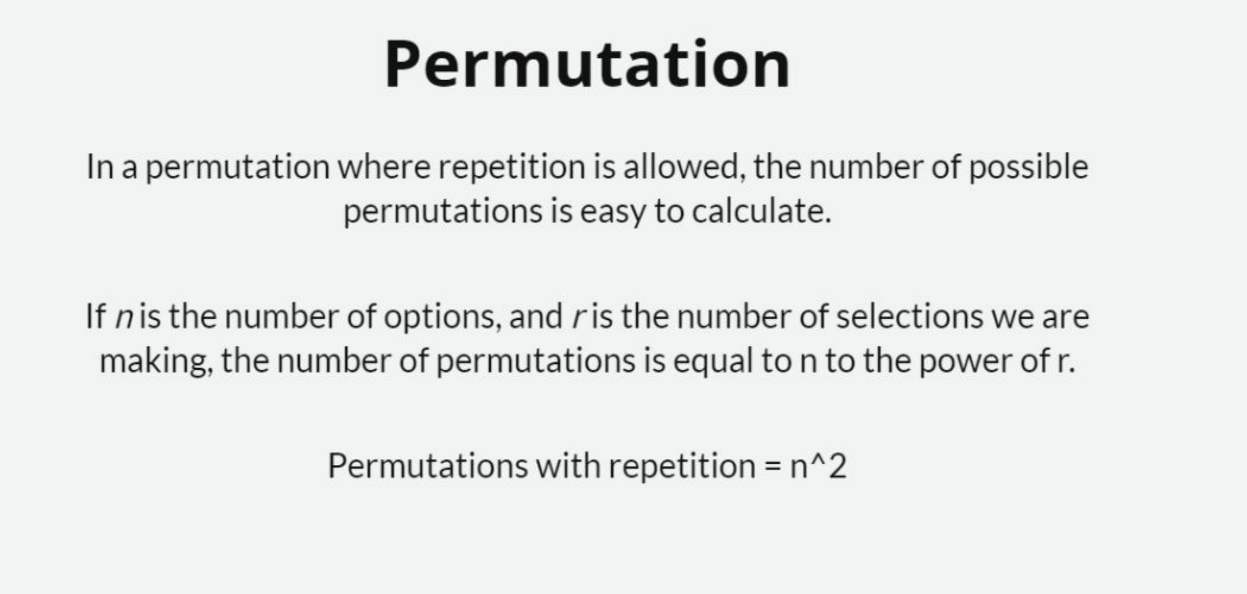
The order doesn't matter! A peppers, pepperoni, and sausage pizza is the same thing!



Permutation

In a permutation, however, order is important.

For example, the passcode for your phone. If your passcode is 1234, the same numbers in a different order (e.g. 1324) will not work.

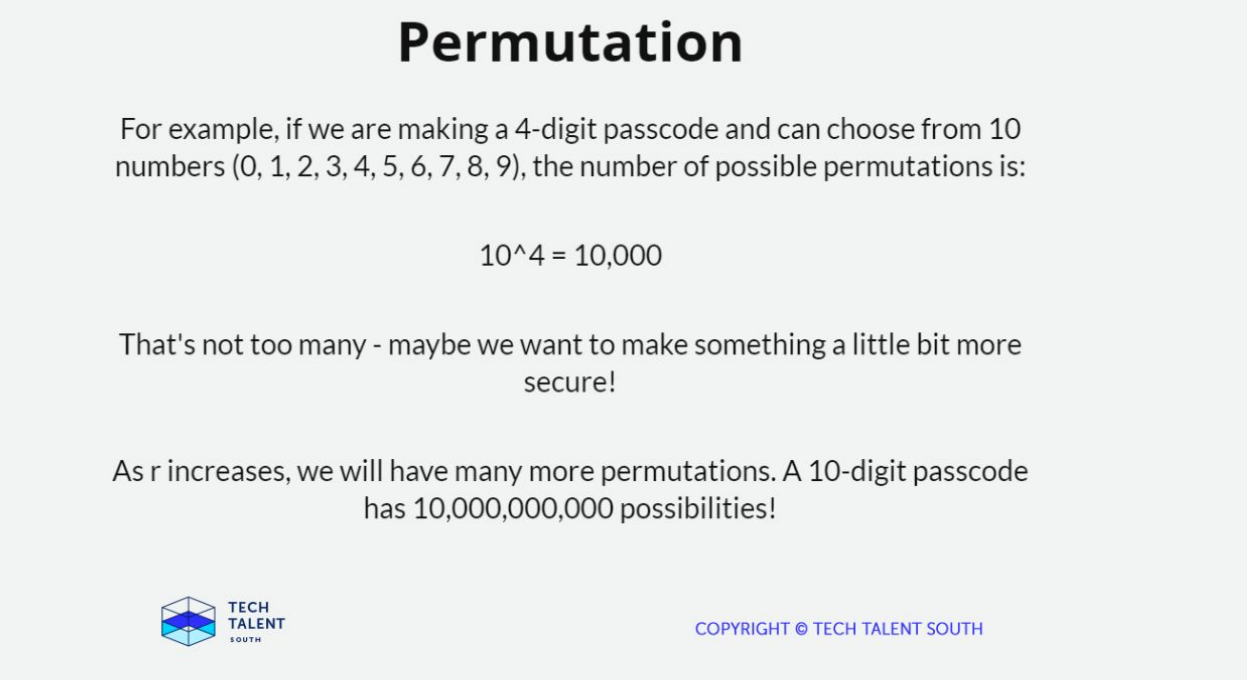


Permutation

In a permutation where repetition is allowed, the number of possible permutations is easy to calculate.

If *n* is the number of options, and *r* is the number of selections we are making, the number of permutations is equal to n to the power of r.

Permutations with repetition = n^2



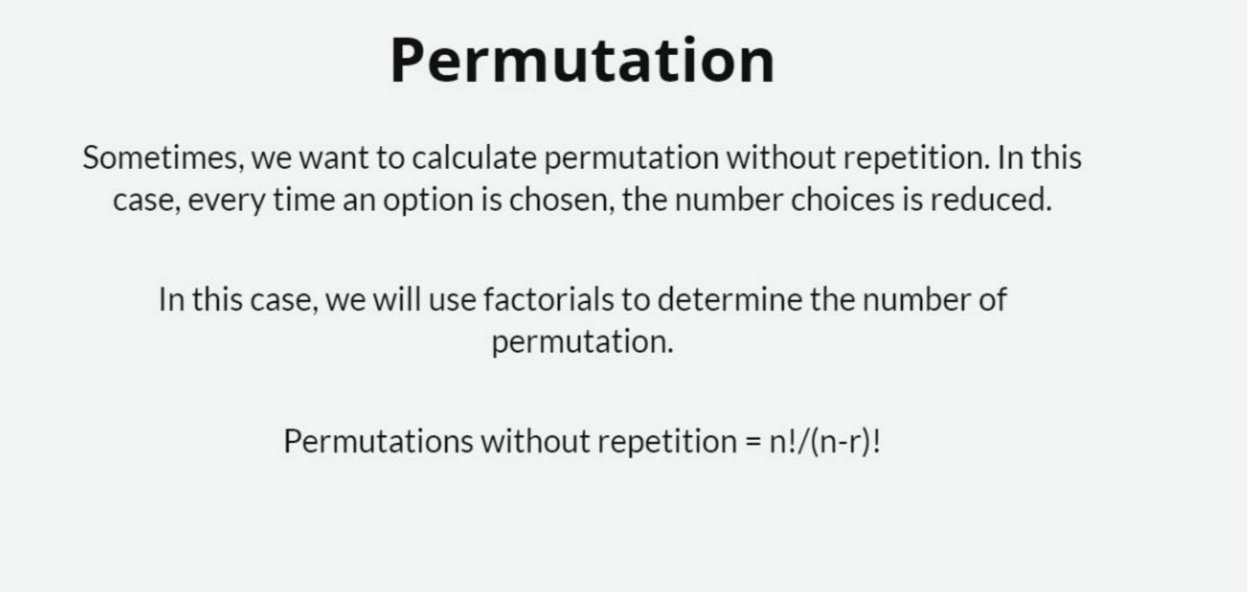
Permutation

For example, if we are making a 4-digit passcode and can choose from 10 numbers (0, 1, 2, 3, 4, 5, 6, 7, 8, 9), the number of possible permutations is:

10^4 = 10,000

That's not too many - maybe we want to make something a little bit more secure!

As r increases, we will have many more permutations. A 10-digit passcode has 10,000,000,000 possibilities!

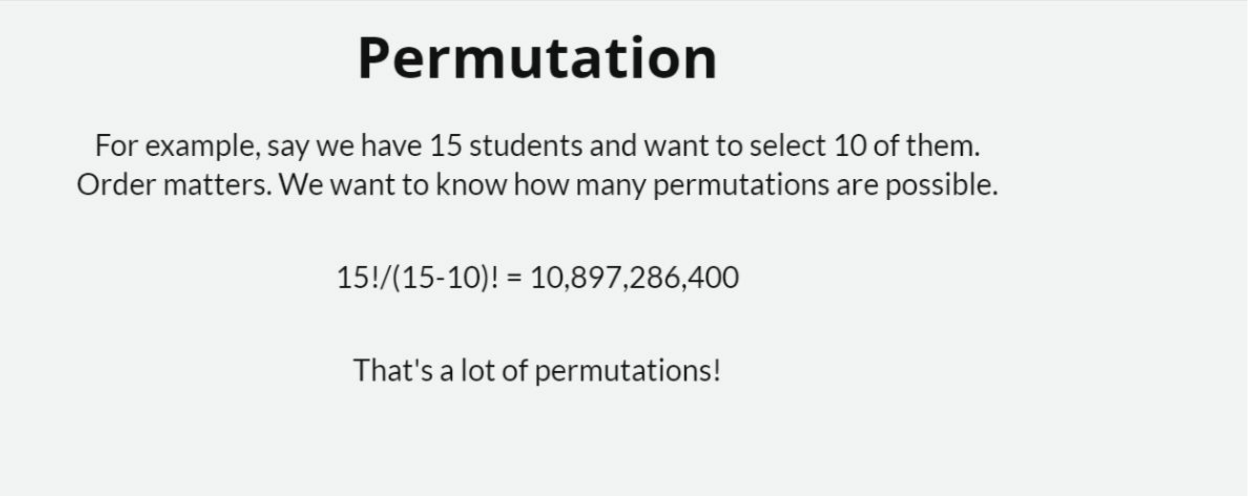


Permutation

Sometimes, we want to calculate permutation without repetition. In this case, every time an option is chosen, the number choices is reduced.

In this case, we will use factorials to determine the number of permutation.

Permutations without repetition = n!/(n-r)!

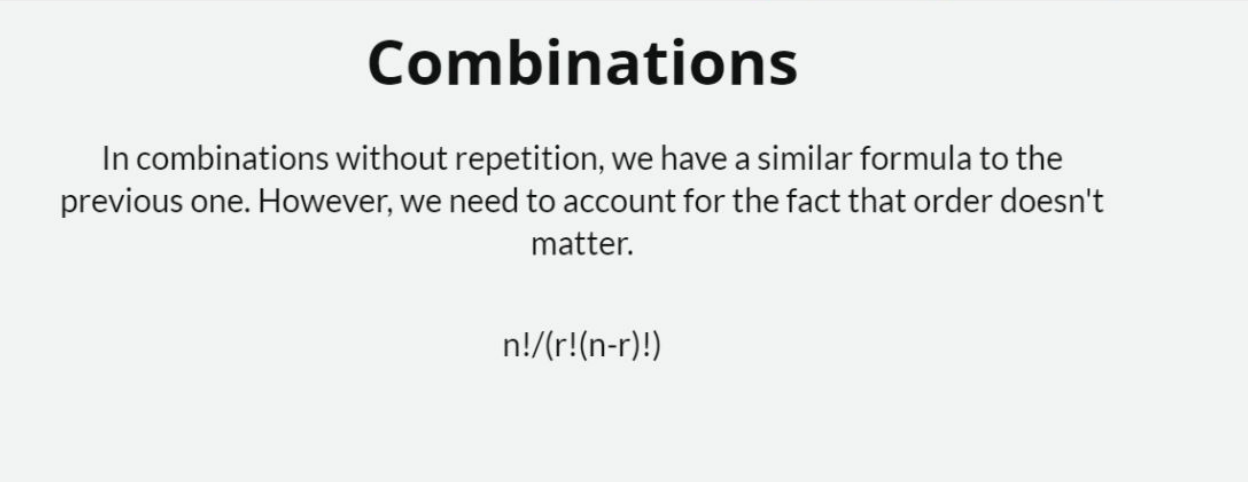


Permutation

For example, say we have 15 students and want to select 10 of them. Order matters. We want to know how many permutations are possible.

15!/(15-10)! = 10,897,286,400

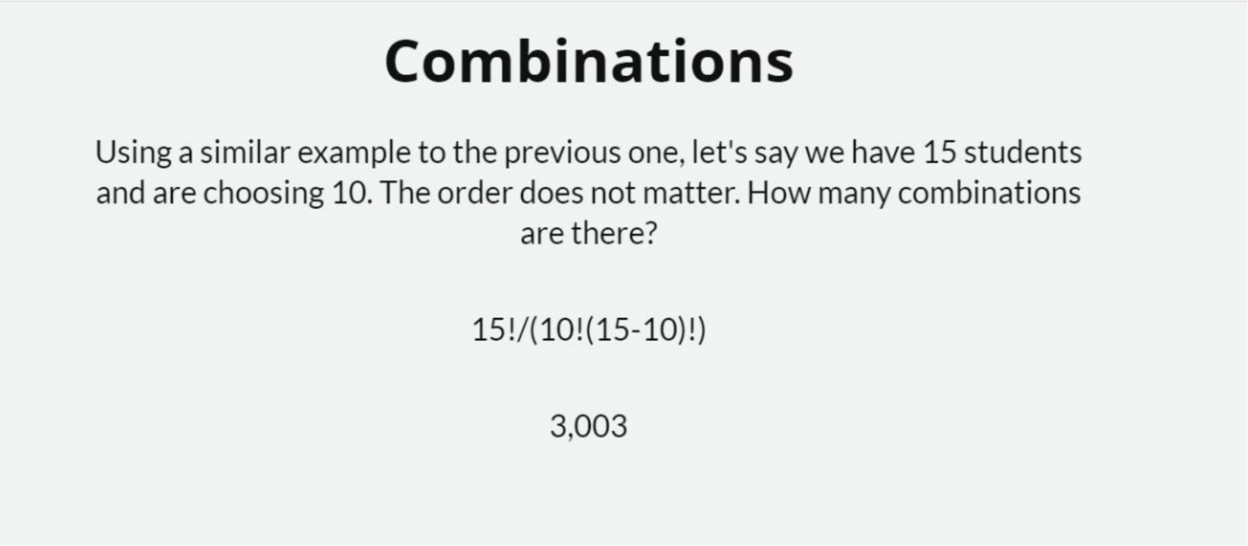
That's a lot of permutations!



Combinations

In combinations without repetition, we have a similar formula to the previous one. However, we need to account for the fact that order doesn't matter.

n!/(r!(n-r)!)

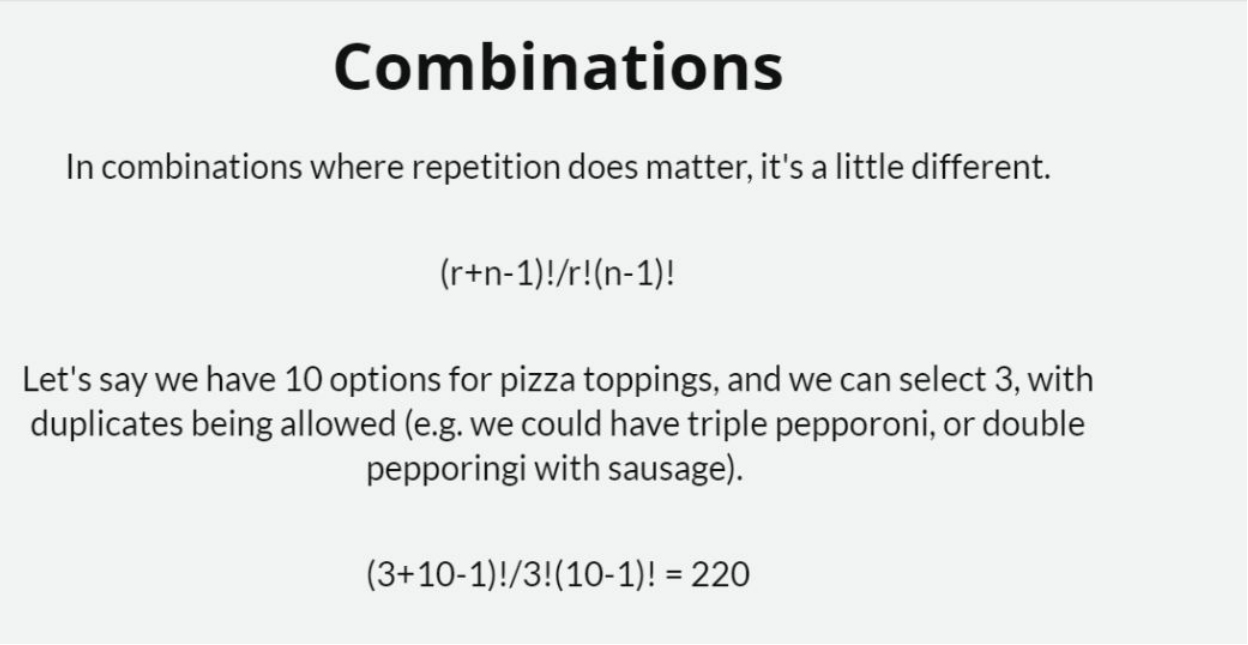


Combinations

Using a similar example to the previous one, let's say we have 15 students and are choosing 10. The order does not matter. How many combinations are there?

$15!/(10!(15-10)!)$

3,003



Combinations

In combinations where repetition does matter, it's a little different.

$(r+n-1)!/r!(n-1)!$

Let's say we have 10 options for pizza toppings, and we can select 3, with duplicates being allowed (e.g. we could have triple pepporoni, or double pepporingi with sausage).

$(3+10-1)!/3!(10-1)! = 220$

https://techtalentsouth.slides.com/techtalentsouth/data-science-4-statistics-fundamentals?token=a1wrNB-3#/33