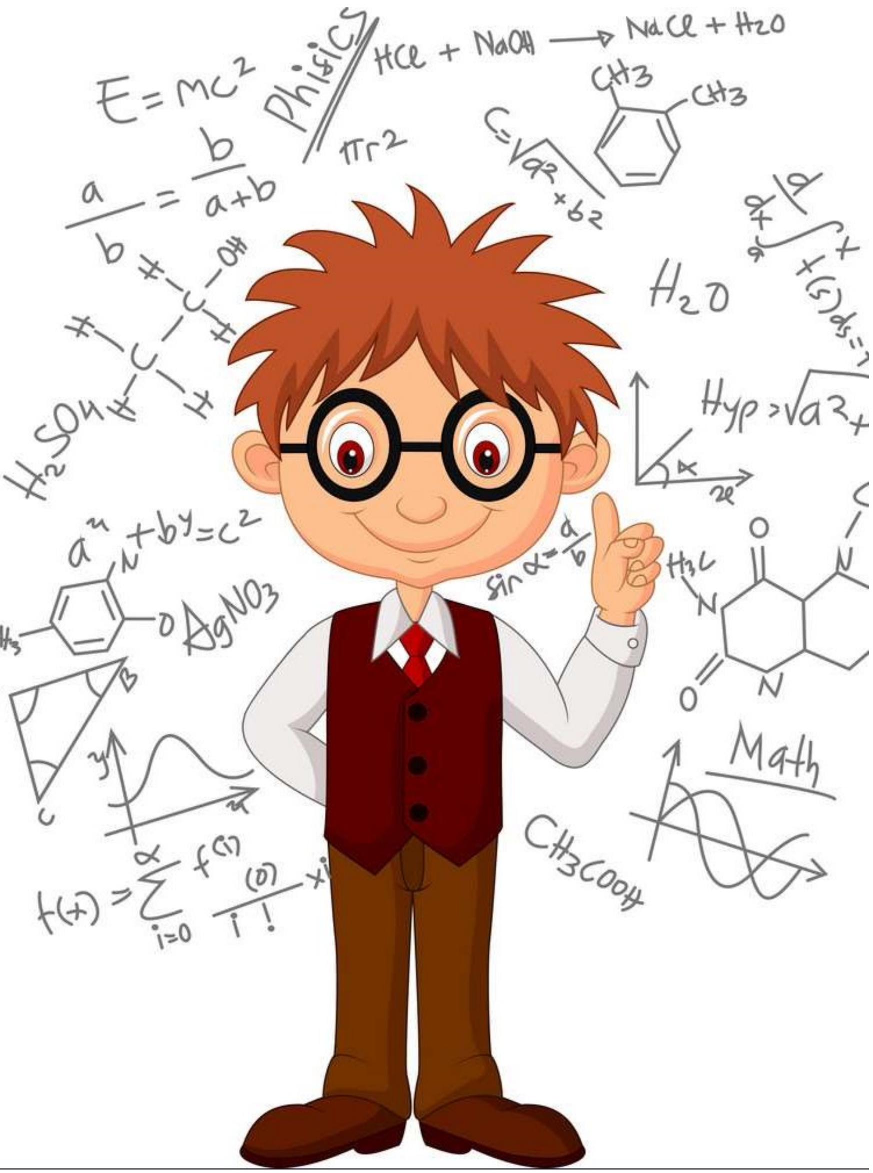
Statistics for Data Analysis

Lecturer: Marina Iantorno

E-mail: miantorno@cct.ie

cct College Dublin

Computing • IT • Business



In today's class we will cover:

- □Probabilities
- □Problem Solving

What is Statistics?

Statistics is a form of mathematical analysis that uses quantified models, representations and synopses for a given set of experimental data or real-life studies.

Statistics include numerical facts and figures.

For example: Compared to women, men are three times more likely to die by suicide.

Statistics is NOT only numbers and figures. The numbers may be right, but the interpretation may be wrong.

For example: The more churches in the city, the more crime there is. Thus, churches lead to crime.

Statistics refer to a range of techniques and procedures for collecting, analysing, interpreting, displaying and making decision based on data.

What is Statistics?

We can find two different kind of events:

- Deterministic event: This is when an experiment could tell us precisely what will happened under certain conditions. For instance, if we put in a jar with one litre of water in it on the fire at 100 degrees Celsius, we know that it will boil after 3 minutes, and if we repeat this experiment many times, we will always get the same result.

- Probabilistic event: This is when an experiment depend on chance, and therefore we could get always different results even though we repeat it under the same conditions. For instance, we can roll a dice several times and although we shake the glass three times and roll it on the same surface, we can get always different results.

Probabilities

We find Statistics every day around us and maybe we do not realise. We are talking about Statistical Events when:

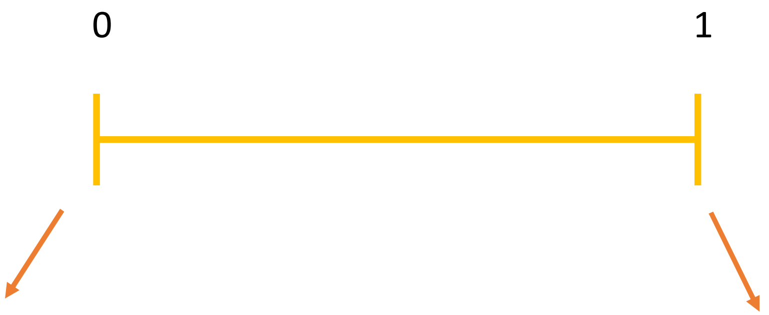
- We check the weather on our phone and it says that today there is a 70% chances of rain/have a sunny day.
- We mention that 9% of the global population live in poverty.
- We play the lottery.
- Choose a card from a deck.
- Discuss about possible results in sports.
- When we discuss about the prices.

Probabilities

Probabilities

Probabilities can take values between 0 and 1.

A probability could be zero and could be one



Null event. This is something that will never happen

Certain event. This is something that for sure will happen

Probabilities

The events are named with a capital letter: A, B, C, D…

For example:

A = picking a white ball from a bag

B = pass an exam

And so on.

Statistic is a linguistic problem! 👇

Colloquial Language	Symbol	Probability	Operation
or	U	Union	Addition (+)
and	∩	Intersection	Multiplication (*)

Multiplication (*) → *We multiply only when the events are independent.*

Probabilities

We need to know the possibility to find some particular events.

- Mutual exclusive events: These are those events that cannot happen at the same time.

For example, if I am an English speaker, I cannot be a non-English speaker.

- Independent events: This happen when the existence of an event does not affect at all the existence of the other one.

For example, the fact that it is sunny in Thailand is not connected to the event of having an accident at home.

Probabilities

This is the basic formula to calculate a probability:

$$P(A) = \frac{\text{Favorable cases of the event A}}{\text{Total cases}}$$

Suppose that in this classroom we have 30 males and 20 females, and we want to randomly choose a person, what is the probability of choosing a male?

A = the person is a male

Number of males in the group

$$P(A) = \frac{30}{50} = \boxed{0.60}$$

Total people in the group

60% of the students are males.

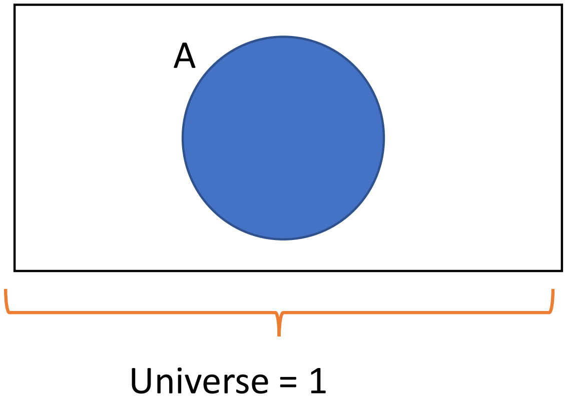
Probabilities

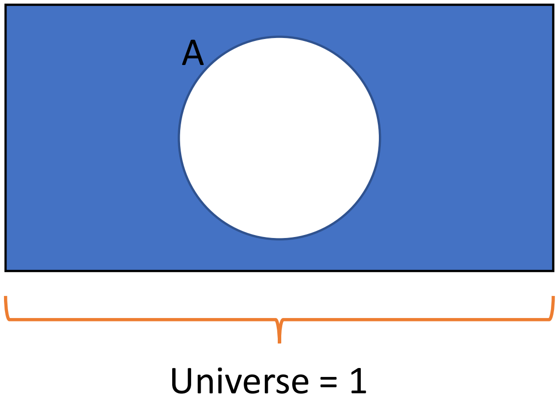
Axioms of Probability

These axioms are auxiliars that help us to solve problems related to probabilities. There are 3 basic Axioms that we will commonly use.

1) Complementary probability.

If we know the probability of A, we could know the probability of $\underline{A}$





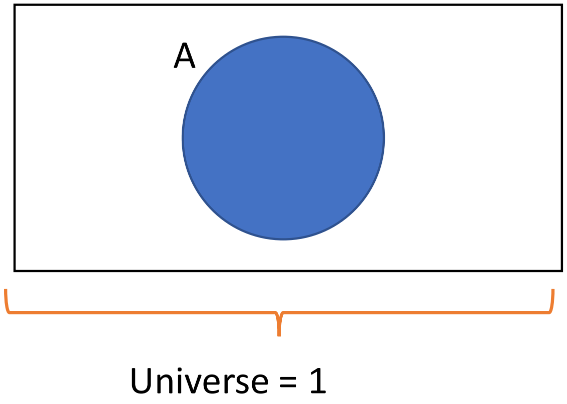
Probabilities

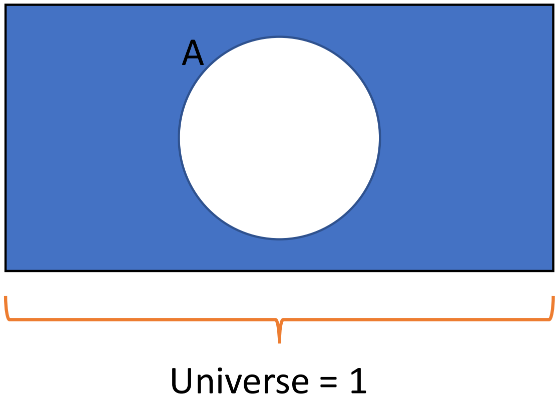
Axioms of Probability

1) Complementary probability:

If the probability of my universe is 1 and I know the probability of A, I can find the probability of $\underline{A}$

$P(\underline{A}) = 1 – P(A)$ And the other way around $P(A) = 1 – P(\underline{A})$



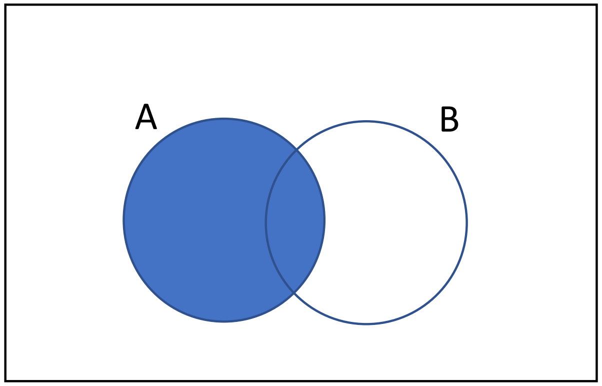


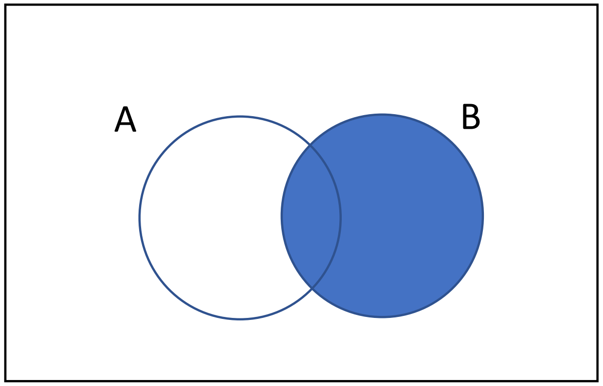
Probabilities

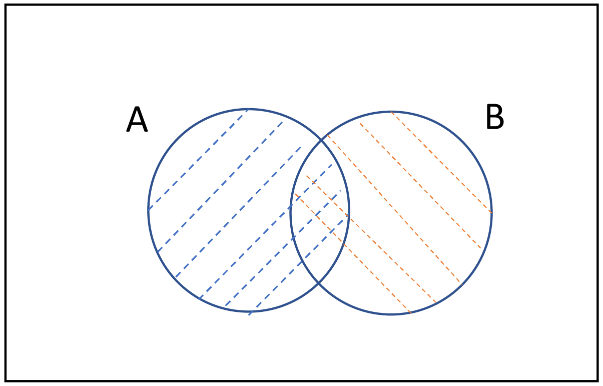
Axioms of Probability

2) Axiom of the addition:

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







Probabilities

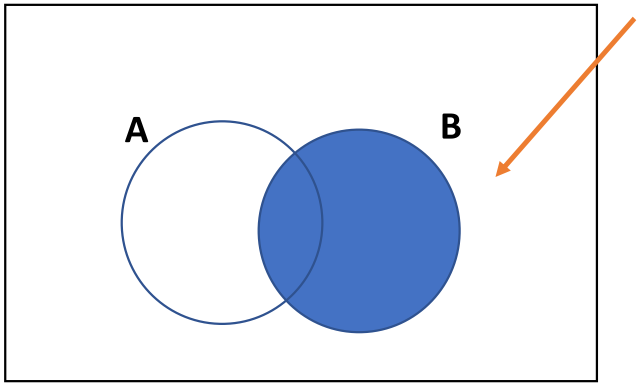
Axioms of Probability

3) Conditional probability:

If we know something about the data, we do not need to use the full group to find our solution.

$$P(A/B) = \frac{P(A \cap B)}{P(B)}$$

The slash means knowing that…
Given that…
Assuming that…



If we know that the observation belongs to B, we do not need to take into consideration what happens with A.

Probabilities

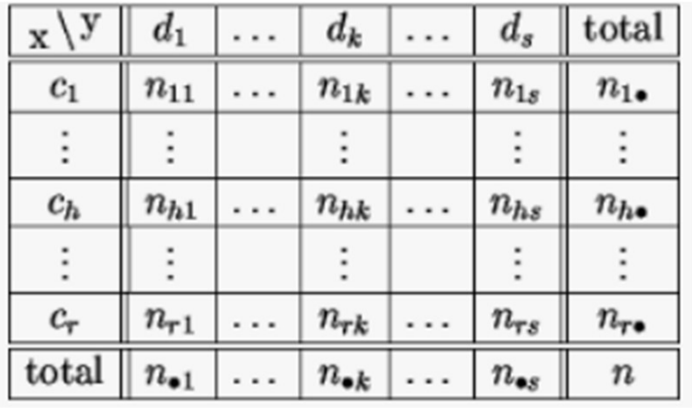
With these information we could solve any probability problem, but there are some ways that will make our journey easier. We will divide them as follow:

Tables

Tree diagrams

Others

Let’s see some examples!



x \ y	d_1	$\dots$	d_k	$\dots$	d_s	total
c_1	n_{11}	$\dots$	n_{1k}	$\dots$	n_{1s}	$n_{1\bullet}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
c_h	n_{h1}	$\dots$	n_{hk}	$\dots$	n_{hs}	$n_{h\bullet}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
c_r	n_{r1}	$\dots$	n_{rk}	$\dots$	n_{rs}	$n_{r\bullet}$
total	$n_{\bullet 1}$	$\dots$	$n_{\bullet k}$	$\dots$	$n_{\bullet s}$	n

Contingency Tables

Probabilities

Contingency Tables: We use them when there is an idea of simultaneity in our data.

You are the owner of a start-up company, and you need to hire people. In order to do so, you post an advertising on a social media and 100 candidates send their CV: 30 of them do not have a third level degree, while the rest do, and 60 candidates speak a foreign language. Also, 10 of the candidates do not have a third level degree but they speak a foreign language. If you randomly choose a candidate:

a) What is the probability of choosing a person who has a third level degree?

b) What is the probability of choosing a person who does not speak a foreign language?

c) What is the probability of choosing a person who has a third level degree or speak a foreign language?

d) If we know that the person speaks a foreign language, what is the probability that also holds a third level degree?

Probabilities

Contingency Tables: We use them when there is an idea of simultaneity in our data.

You are the owner of a start-up company, and you need to hire people. In order to do so, you posted an advertising on a social media and 100 candidates sent their CV: 30 of them do not have a third level degree, while the rest do, and 60 candidates speak a foreign language. Also, 10 of the candidates do not have a third level degree but they speak a foreign language. If you randomly choose a candidate:

D = the person holds a third level degree

F = the person speaks a foreign language

We need to check the information available and build the table. Don't forget to define your events

	D	$\underline{\mathbf{D}}$	**TOTAL**
F		10	60
$\underline{\mathbf{F}}$			
TOTAL		30	100

Probabilities

Contingency Tables: We use them when there is an idea of simultaneity in our data.

You are the owner of a start-up company, and you need to hire people. In order to do so, you posted an advertising on a social media and 100 candidates sent their CV: 30 of them do not have a third level degree, while the rest do, and 60 candidates speak a foreign language. Also, 10 of the candidates do not have a third level degree but they speak a foreign language. If you randomly choose a candidate:

D = the person holds a third level degree

F = the person speaks a foreign language

	D	$\underline{D}$	TOTAL
F	50	10	60
$\underline{F}$	20	20	40
TOTAL	70	30	100

We need to complete the gaps by differences

Probabilities

Contingency Tables: We use them when there is an idea of simultaneity in our data.

a) What is the probability of choosing a person who has a third level degree?

	D	$\underline{D}$	TOTAL
F	50	10	60
$\underline{F}$	20	20	40
TOTAL	70	30	100

P(D) = 70/100 = 0.70

Answer: The probability of choosing a person who has a third level degree is 0.70

**Note: we could also express this as a percentage but it is always good to answer according to the question*

Probabilities

Contingency Tables: We use them when there is an idea of simultaneity in our data.

b) What is the probability of choosing a person who does not speak a foreign language?

	D	$\underline{D}$	TOTAL
F	50	10	60
$\underline{F}$	20	20	40
TOTAL	70	30	100

$P(\underline{F}) = 40/100 = 0.40$

Answer: The probability of choosing a person who doesn’t speak a foreign language is 0.40

Probabilities

Contingency Tables: We use them when there is an idea of simultaneity in our data.

c) What is the probability of choosing a person who has a third level degree or speak a foreign language?

	D	$\underline{D}$	TOTAL
F	50	10	60
$\underline{F}$	20	20	40
TOTAL	70	30	100

Pay attention to the keywords

Remember that earlier we said that Statistics is a linguistic problem. Here we found the first keyword that indicates that we have to use an axiom of probability.

Colloquial Language	Symbol	Probability	Operation
or	U	Union	Addition

We will use the axiom of the sum

Probabilities

Contingency Tables: We use them when there is an idea of simultaneity in our data.

c) What is the probability of choosing a person who has a third level degree or speak a foreign language?

	D	$\underline{D}$	TOTAL
F	50	10	60
$\underline{F}$	20	20	40
TOTAL	70	30	100

Pay attention to the keywords

We remember the Axiom of the sum: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Now we translate this to our exercise

$P(D \cup F) = P(D) + P(F) - P(D \cap F) \rightarrow P(D \cup F) = 70/100 + 60/100 - 50/100 = 0.80$

Answer: The probability of choosing a person who has a third level degree or speak a foreign language is 0.80

Probabilities

Contingency Tables: We use them when there is an idea of simultaneity in our data.

d) If we know that the person speaks a foreign language, what is the probability that also holds a third level degree?

Remember that earlier we said that Statistics is a linguistic problem. Here we found another keywords that indicates that we have to use an axiom of probability.

Pay attention to the keywords

When we know something, we limit our total to that space instead of taking the universe.

The conditional axiom will help us here.

$$P(A/B) = \frac{P(A \cap B)}{P(B)}$$

Probabilities

Contingency Tables: We use them when there is an idea of simultaneity in our data.

d) If we know that the person speaks a foreign language, what is the probability that also holds a third level degree?

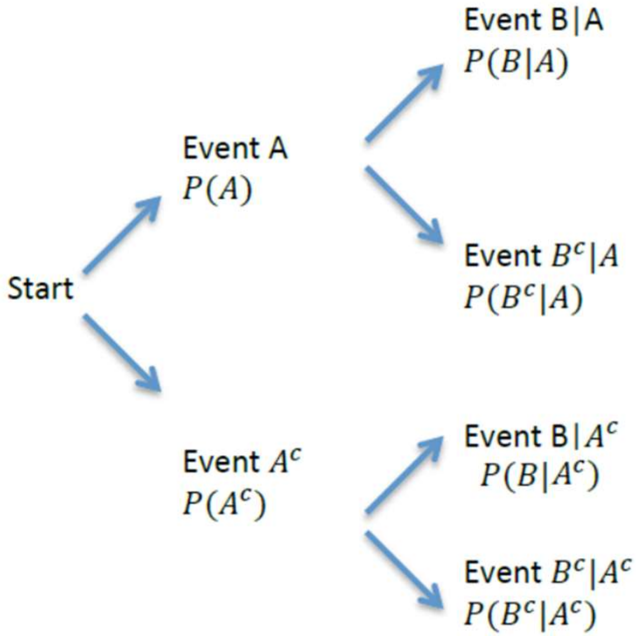
	D	$\underline{D}$	TOTAL
F	50	10	60
$\underline{F}$	20	20	40
TOTAL	70	30	100

We limit our analysis to this row, because we know that the person speaks a foreign language

$$P(D/F) = \frac{P(D \cap F)}{P(F)} = \frac{50}{60} = 0.8333$$

Answer: The probability of choosing a person that holds a third level degree, given that the person speaks a foreign language is 0.8333

Tree Diagrams



Start

Event A
$P(A)$

Event A^c
$P(A^c)$

Event B|A
$P(B|A)$

Event $B^c|A$
$P(B^c|A)$

Event B|A^c
$P(B|A^c)$

Event $B^c|A^c$
$P(B^c|A^c)$

Probabilities

Tree Diagrams: We use them when there is an idea of consecutiveness in our data.

You find two purses. In one of them there are 5 silver coins and 3 bronze coins, and in the second one there are 4 silver coins and 6 bronze coins. If you picked a purse and took a coin from this one, what is the probability of picking a silver coin?

O = This is the purse number one

T = This is the purse number two

S = This is a silver coin

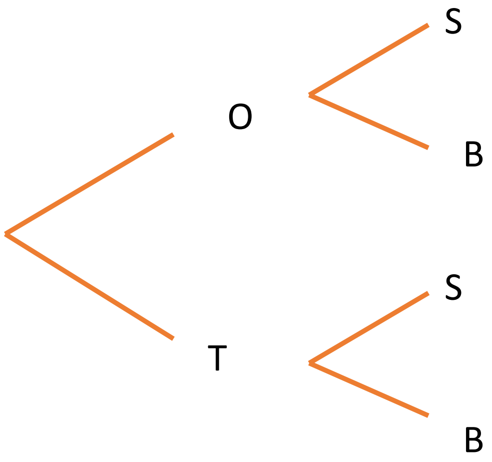
B = This is a bronze coin

Here we clearly see that there is an event that should happen before reaching to the question. First I have to pick a purse and after I have to pick the coin. This is the consecutiveness that we refer when using the trees.

Probabilities

Trees: We use them when there is an idea of consecutiveness in our data.

You find two purses. In one of them there are 5 silver coins and 3 bronze coins, and in the second one there are 4 silver coins and 6 bronze coins. If you picked a purse and took a coin from this one, what is the probability of picking a silver coin?

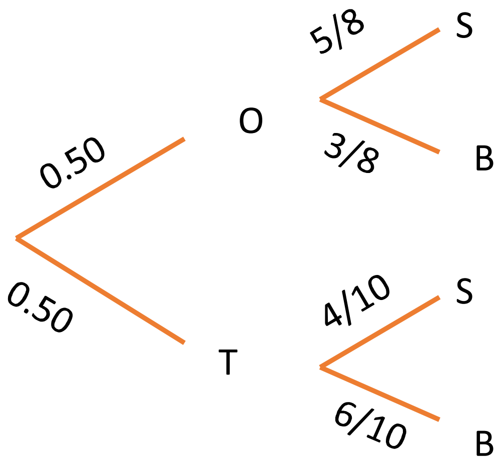


Once we have the tree, we need to add the probabilities on it

Probabilities

Trees: We use them when there is an idea of consecutiveness in our data.

You find two purses. In one of them there are 5 silver coins and 3 bronze coins, and in the second one there are 4 silver coins and 6 bronze coins. If you picked a purse and took a coin from this one, what is the probability of picking a silver coin?

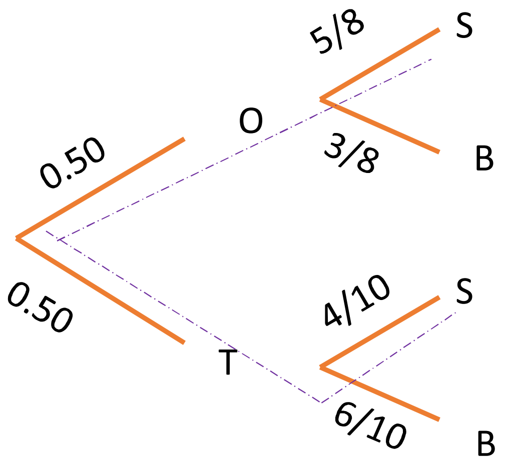


To solve this problem, we need to see where the attribute we are looking for is, and after that, follow the lines of the trees to construct our solution.

Probabilities

Trees: We use them when there is an idea of consecutiveness in our data.

You find two purses. In one of them there are 5 silver coins and 3 bronze coins, and in the second one there are 4 silver coins and 6 bronze coins. If you picked a purse and took a coin from this one, what is the probability of picking a silver coin?



We pick the purse One AND the coin is a Silver coin

OR

We pick the purse Two AND the coin is a Silver coin

$$P(S) = P(O \cap S) + P(T \cap S)$$

$$P(S) = 0.50 \underset{\substack{\uparrow \\ \text{AND}}}{*} 5/8 \underset{\substack{\uparrow \\ \text{OR}}}{+} 0.50 \underset{\substack{\uparrow \\ \text{AND}}}{*} 4/10 = 0.5125$$

Answer: The probability of picking a silver coin is 0.5125

Others

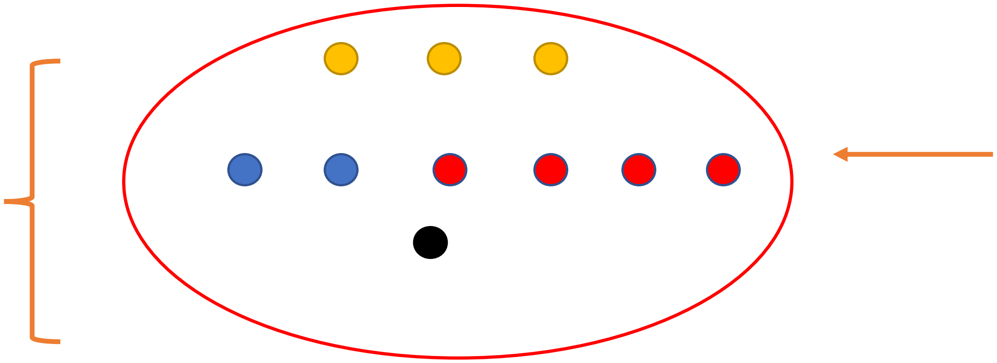
Probabilities

Others: When we reach to this point is because none of the previous methods is applicable and we need to use the logic and our knowledge to solve the problem.

There is a bag with pencils: 3 yellow, 2 blue, 4 red and 1 black. If we pick 3 pencils (one after another one), what is the probability of picking 3 red pencils?

Now, we need to imagine the bag

$P(R) = 4/10$
$P(\underline{R}) = 6/10$



Initially we have 10 pencils:
3 yellow
2 blue
4 red
1 black
10 pencils

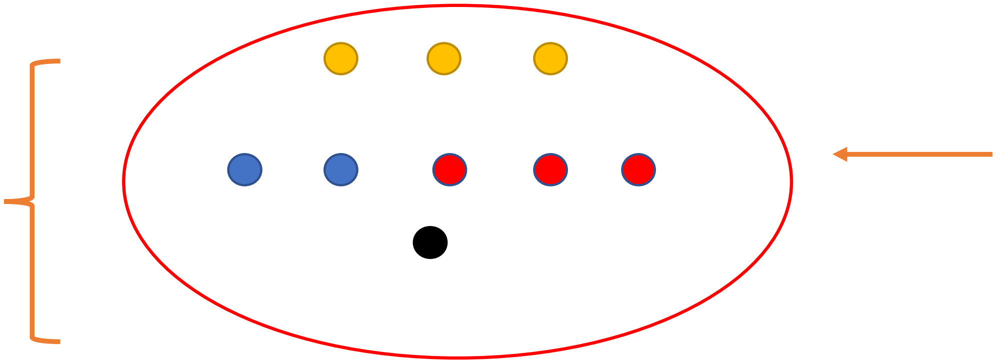
What happens if we take one red pencil?

Probabilities

Others: When we reach to this point is because none of the previous methods is applicable and we need to use the logic and our knowledge to solve the problem.

There is a bag with pencils: 3 yellow, 2 blue, 4 red and 1 black. If we pick 3 pencils (one after another one), what is the probability of picking 3 red pencils?

Now, we need to imagine the bag



P(R) = 3/9
P($\underline{R}$) = 6/9

●

Now we have 9 pencils left in the bag
3 yellow
2 blue
3 red
1 black
9 pencils

What happens if we take another red pencil?

Probabilities

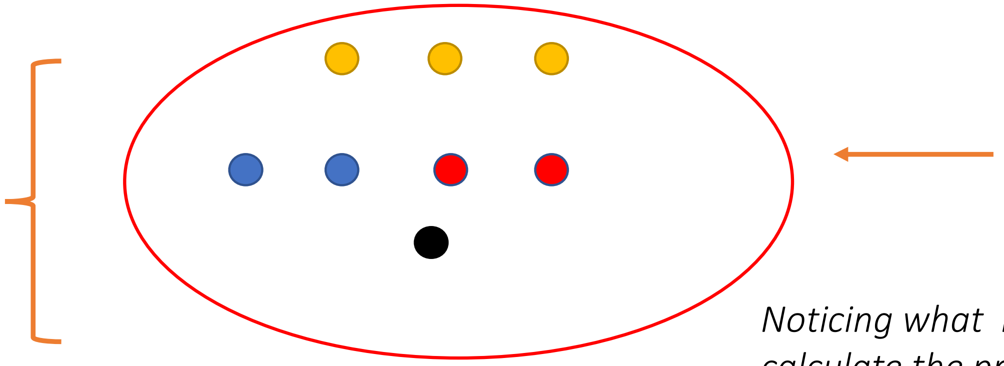
Others: When we reach to this point is because none of the previous methods is applicable and we need to use the logic and our knowledge to solve the problem.

There is a bag with pencils: 3 yellow, 2 blue, 4 red and 1 black. If we pick 3 pencils (one after another one), what is the probability of picking 3 red pencils?

Now, we need to imagine the bag

Now we have 8 pencils left in the bag
3 yellow
2 blue
2 red
1 black
8 pencils

$P(R) = 2/8$
$P(\underline{R}) = 6/8$



Noticing what happens? If we apply the formula to calculate the probability, our numerator and denominator change

Probabilities

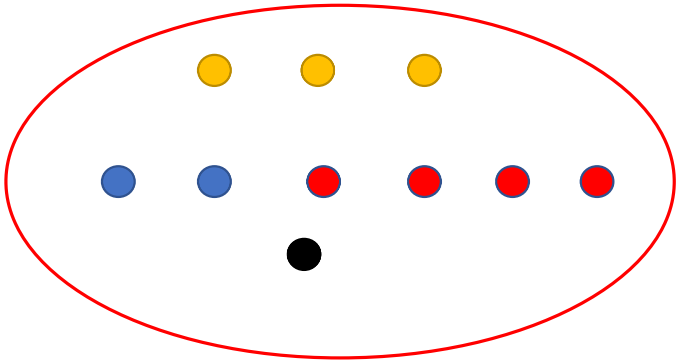
There is a bag with pencils: 3 yellow, 2 blue, 4 red and 1 black. If we pick 3 pencils (one after another one), what is the probability of picking 3 red pencils?

T = 3 pencils are red

We need to pick one red pencil AND another one AND another one

$P(T) = P(R_1) \cap P(R_2) \cap P(R_3)$

$P(T) = 4/10 * 3/9 * 2/8 = \boxed{0.0333}$



Answer: The probability of picking 3 red pencils, one after another one, is 0.0333.

Time To Practice

Probabilities

Exercise 1

A College has 300 students who are pursuing a Data Analytics Degree. For the current semester there are 200 students doing Statistics and 100 are enrolled in Data Preparation. These figures include 30 students who are enrolled in both modules. If we randomly select a student:

a) What is the probability that s/he is not enrolled in any of the module?

b) What is the probability that s/he is enrolled only in Statistics?

c) If s/he is enrolled in Statistics, what is the probability that s/he is also enrolled in Data Preparation?

d) If we select two students, what is the probability that only one of them is enrolled only in Statistics?

HOMEWORK

Homework

Watch the movie “Moneyball” (available on different platforms) and write down some thoughts on the use of Statistics in the game. The movie is based on a true story, and the idea here is for us to analyse different aspects of the events.

Some questions that you could think of:

- Was their plan a good plan? Why yes/not?
- What are the business fields that the movie “touches”?
- Would you do something different?
- What is the importance of Statistics in the plan?
- Did they relate Statistics with another discipline of Data Analysis to accomplish the plan?

THAT’S ALL FOR TODAY

THANK YOU