Content

- Basic Terminologies
  - Experiment
  - Outcomes
  - Sample space
  - Events
    - Mutually exclusive Events (Disjoint Events)
    - Exhaustive Events
    - Joint events
    - Independent events
- Set operations
  - Intersection
  - Union
  - Complement
- Addition Rule
- Cross tab

Basic Terminologies

1. Experiment

- It is basically an activity which I'm trying to do.

> **Let's say I have this mathematical equation**

$a^2 + b^2 + 2ab$

where: a = 3 and b = 4

$3^2 + 4^2 + 2(3)(4)$ = 49

- We are 100% sure that the result of this equation will be 49 only. It cannot be 50 or 48.

This type of experiment is called **Deterministic Experiments** where we can **determine** the exact output, like in this case.

Now, let's see another few more examples:

- Flipping a coin
  - When you flip a coin, there are two possible outcomes: it can land either `heads` or `tails`.

- Rolling a six-sided die
  - When you roll the die, the outcome is uncertain, and the die can land on any of the six faces.
- Cricket Match
  - Suppose there is a match going on between 2 teams, we can't determine the match result.

In all of these above examples, we can notice one common thing.

> **Q. Can we determine the outcome of all these experiments?**

No, because the outcomes are uncertain. These types of experiments are known as **Probabilistic Experiments**.

2. Outcomes

- Suppose we roll a six sided die and we want to know the possible `Outcomes`.
- We know that we could get any digit out of the 6 digits. So, an outcome could be : {1} or {2} or {3} or {4} or {5} or {6}

3. Sample Space

- It is the collection of all the possible outcomes of the experiment.

So the **sample space** for this experiment will be: **{1, 2, 3, 4, 5, 6}**

4. Events

We know that sample space for die is {1,2,3,4,5,6}.

If we say,

> **An Even number is rolled / While rolling a die, an even number has occurred**

- Then the possible outcomes will be: `{2, 4, 6}`

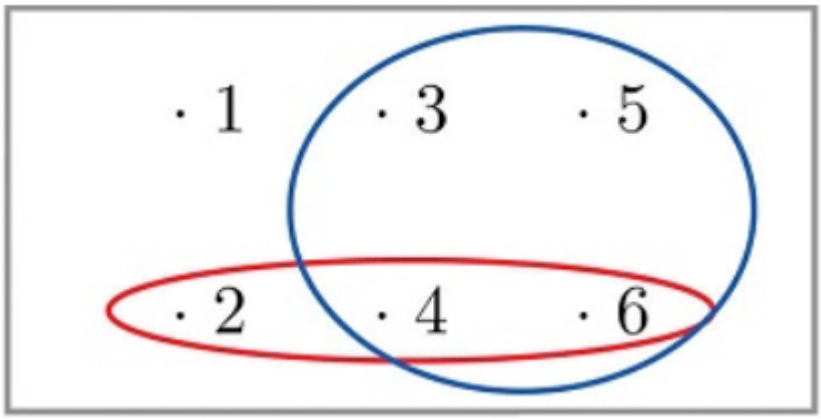
This is known as an **`Event`**.
Any subset of sample space is an event.

- {2, 4, 6} is a subset of sample space.

"**An Even number is rolled**" is an event here and its output is E = {2, 4, 6}, where E denotes an Event.

> **Q1. What are the possible outcomes when a dice is rolled and a number greater than two has occurred?**

- For this Event, outcome will be E = {3, 4, 5, 6}



Here is a graphical representation of a sample space and events

- Here the **sample space** S is represented by a rectangle which is {1, 2, 3, 4, 5, 6}
- **Outcomes** are represented as points within the rectangle which is {1}, {2}, {3}, {4}, {5}, {6}
- **Events** are represented as ovals that enclose the outcomes that compose them.
  - we have two events, $E1$: {2, 4, 6} which is an event for "Even number is rolled"
  - $E2$: {3, 4, 5, 6} which is an event for "A number greater than 2 rolled"

Now let's see few experiments.

Experiment 1: Tossing a single coin

> **Q1. If we toss a single coin then what can be the Possible Outcomes for this experiment?**

- Either we can get **Heads**
- Or we can get **Tails**

Therefore, our outcome becomes: **{H}, {T}**

The **Sample Space** for this experiment will be S **= {H, T}**

> **Based on this sample space, what possible Events can be defined?**

Getting Heads while tossing a coin,

- then our event will be E = {H}

Getting Tails while tossing a coin,

- then our event will be E = {T}

> **Q2. Suppose the given subset is itself {H,T}. Can we define this as an Event or not?**

Yes, It is an event.

- We discussed earlier that any subset of a Sample Space is an Event.
- Also an entire set is a subset of itself so this is a valid event.

> **Q3. So how can we frame this event?**

It is the "**Event of getting Either Heads or Tails**".

> **Q4. Consider the empty set as the given subset denoted by { }. Is it a valid event?**

- We know that, an empty set is a subset of every set. An empty set is therefore a subset of sample space
- It is a valid subset
- So by going with the definition of an Event, we can conclude that this is a valid event.

This can be represented as the **"Event of getting neither Heads nor Tails"**.

> **Q5. Is it possible if we toss a coin and get nothing?**

No, it is not possible.

- Therefore, we will have an **Empty set** here
- As we know an empty set is a subset of sample space, therefore it is an Event.

But, the probability of getting a Null Set (No outcome) is Zero.

As it is not possible to toss the coin and don't get any output. we will either gets a head or a tail.

> **Q6. How many subsets can be formed from the sample space?**

There is one formula to find the number of subsets : 2^N

- where N = number of elements in sample space

For the above experiment, number of elements in the sample sapace is 2 {H,T}, So N = 2

- Therefore the number of subsets will be $2^2 = 4$

- Subsets will be **{ {H}, {T}, {H,T}, { } }**

From this, we can conclude that an empty set is also considered as a valid subset.

Set Operations

Let's recall the experiment "**Rolling a die**" for which the **Sample space** is $\{1, 2, 3, 4, 5, 6\}$

- We can also represent this as a **Universe** or **Universal Set** in context of set operations
- Universal set is the collection of all possible sets

Now let's define some events:

- Mohit bets that he will get an odd number
  - So the outcome of this **Event** will be $A = \{1, 3, 5\}$
- Rakesh bets that he will get either 1, 5 OR 6
  - $B = \{1, 5, 6\}$
- Abhishek bets that he will get an Even number
  - $C = \{2, 4, 6\}$

There are some some questions which can arise

Intersection

> **Q. In which condition, both Mohit and Rakesh will win their bets?**

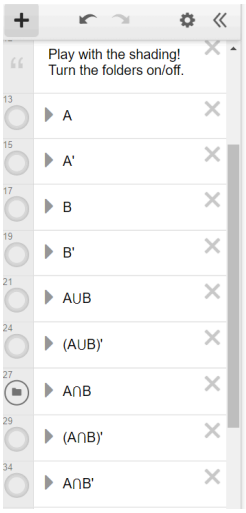
We want a number which occurs in both of their events

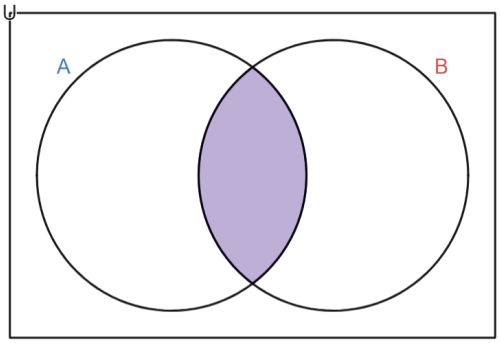
They will win their bets when we get a number 1 or 5 on a die.

- Therefore $\{1, 5\}$ is the possible outcome such that both Mohit and Rakesh will win their bets

This is known as an **Intersection** of two events.

- It is denoted as $A \cap B$
- Intersection means **members belonging to both A `AND` B**
  - So, $A \cap B$ will consists only of the elements present in both events, which in this case are $\{1, 5\}$





Union

Now the next question,

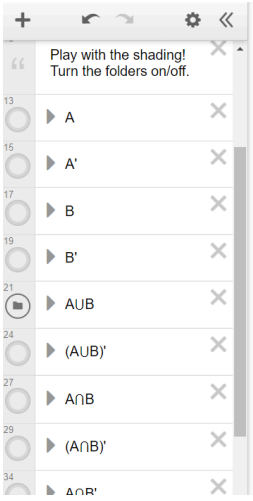
> **Q. When either Mohit or Rakesh will win their bets?**

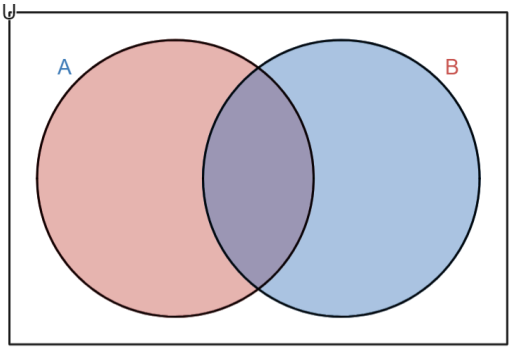
If we get any number out of 1, 3, 5 or 6

- Possible outcomes of this event: $\{1, 3, 5, 6\}$

This is known as **Union** of Two events A and B

- It is denoted by $A \cup B$
- So, **Union** means **members belonging to either A `OR` B**
- So, $A \cup B$ will combine their outcomes, which in this case will be $\{1, 3, 5, 6\}$





Complement

> **Q. When will Mohit lose his bet?**

Mohit will lose his bet if the outcome is $\{2, 4, 6\}$

This is known as **complement** of Event A, denoted by A' or A^c

We can define it as the set that contains all the elements except the elements of A, denoted as $A' = U - A$

While Rakesh will lose if the outcome is $\{2, 3, 4\}$

- Hence $B' = \{2, 3, 4\}$

Mutually Exclusive Events (Disjoint Events)

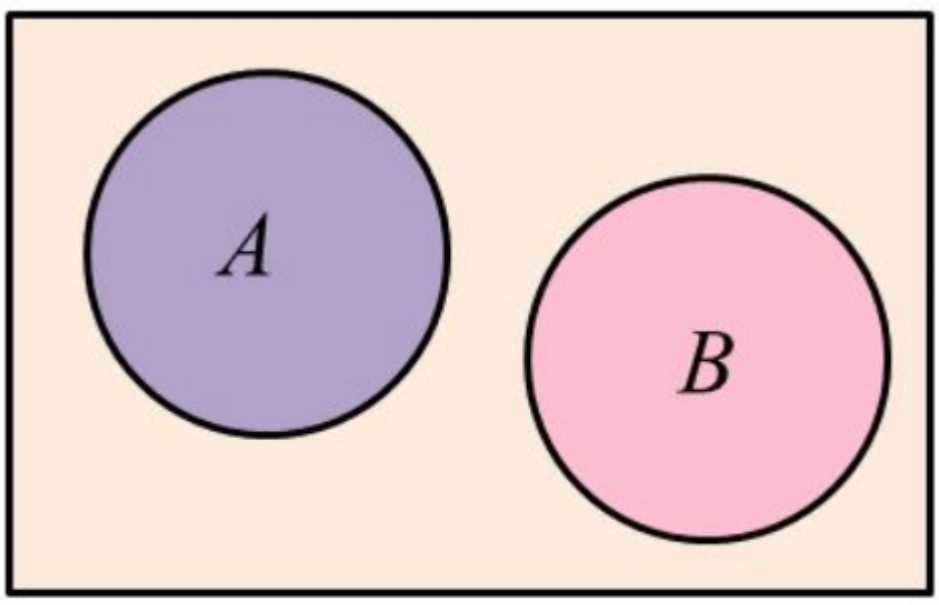
> **Q1. What will be the output of $A \cap C$?**

We will have an empty set $\{\ \}$ which can also be represented by $\emptyset$

Because there are no common elements in Set A and Set C

Or it implies that **both the events can't occur on the same time** means we can't get an **Even number and a Odd number** at the same time on the dice.

- So, when two events cannot occur at the same time or simultaneously then these types of events are known as **`Mutually Exclusive Events`** or `Disjoint Events`



A and B are mutually exclusive

Exhaustive Events

> **Q. What will be the output of $A \cup B \cup C$?**

Our events are:

- $A = \{1, 3, 5\}$, $B = \{1, 5, 6\}$, $C = \{2, 4, 6\}$
  - Therfore $A \cup B \cup C$ = combined elements of Event A, B, C = {1, 2, 3, 4, 5, 6}

This is nothing but the **Sample Space** of our experiment **"Rolling a die"** as these events when combined, giving the all possible outcomes.

- These types of events are known as **`Exhaustive Events`**

Non Mutually Exclusive Events (Joint Events)

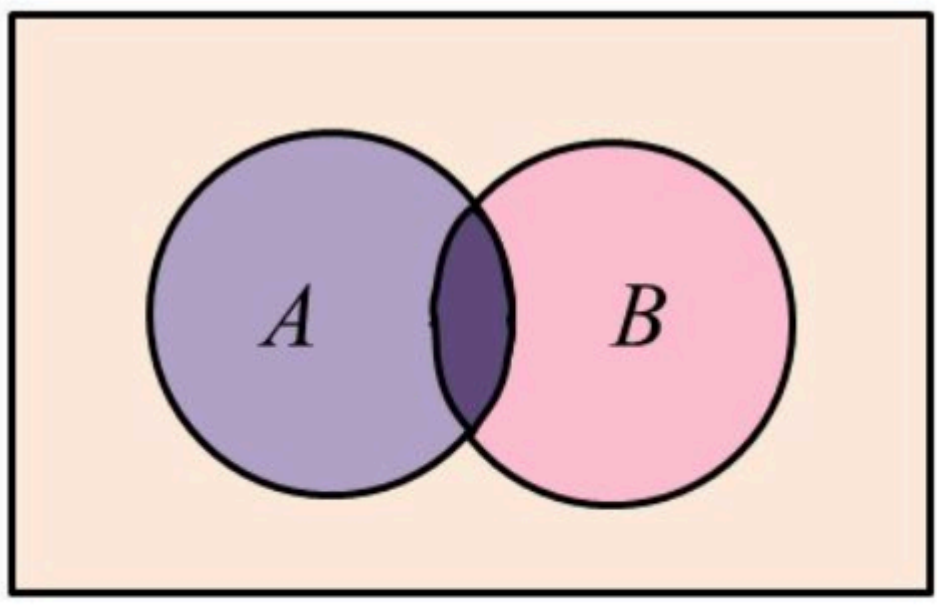
Suppose we define one more Events:

- **Event D**: Rolling a number greater than 3 = (4, 5, or 6).

> **Q. Can we say that Events C (getting even) and D are mutually exclusive?
> **

No, as we can get a number that is both even and greater than 3, which means both **events C and D can occur simultaneously**.

- For instance, if the die shows a 4 or a 6, it fulfills the criteria for both events C and D.
- This type of events are known as **`non-mutually exclusive`** or `joint events`



A and B are not mutually exclusive

Independent Events

While non-mutually exclusive events allow for overlap, where more than one event can occur, independent events focus on how the occurrence of one event **may or may not affect** the likelihood or outcome of another event

Suppose we have 2 two events:

- **Event A**: Rolling an even number (2, 4, or 6)
- **Event B**: Flipping a coin and getting heads

> **Q. Are these two events Independent or not?**

YES, these events are `independent Events` because

- The outcome of rolling the die (**Event A**) `does not affect the outcome` of flipping the coin (**Event B**), and vice versa.

They are unrelated events that are occurring independently.

And if two events A and B are independent, then the probability of happening of both A and B is:

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

In case of Disjoint events, $P(A \cap B)$ **= 0** , as A Intersect B = { }

- **So, if the Events are Independent they cannot be Mutually Exclusive or Disjoint and vice a versa**

In the upcoming lectures, we will see how to derive this formula and also prove this claim.

How to calculate Probability

Now if I want to calculate the Probability of the particular event let's say event A, then we can calculate using this.

$$Probability = \frac{Outcomes\ in\ set\ A}{Total\ Outcomes\ in\ Entire\ Sample\ Space}$$

Now, let's take a random Experiment whose **outcome** could be **{1} or {2} or {3} or {4} or {5} or {6}**, then the **Sample Space** will be **{1, 2, 3, 4, 5, 6}**

Let's define some events:

1. A = **{2, 4, 6}**

> **Q1. What will be the probability of Event A?**

- By looking into the formula = $\frac{Possible\ oucomes}{Total\ outcomes}$
- Possible outcomes of event A = 3 and total Outcome in sample space = 6

So, $P(A) = \frac{3}{6}$

2. B = **{1, 2}**

- Similarly Probability of Event B will be $P(B)$ = $\frac{2}{6}$

3. C = **{1, 4, 5, 6}**

- and Probability of Event C will be $P(C)$ = $\frac{4}{6}$

Addition Rule

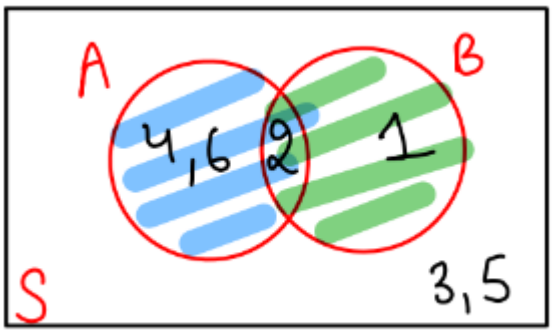
> **Q1. What will be the Probability of $P(A \cup B)$?**

First we need to find $A \cup B$ which is **{1, 2, 4, 6}**

- So by the formula of probability $P(A \cup B)$ will be = $\frac{|A \cup B|}{|S|}$ = $\frac{|\{1,2,4,6\}|}{|\{1,2,3,4,5,6\}|}$ = $\frac{4}{6}$

Where, $|A \cup B|$ = Number of elements(cardinality) of $(A \cup B)$ set,
and $|S|$ = Number of elements in Sample Space

If we want to represent using venn Diagram:



> **Q2. What will be Probability of $P(A \cap B)$?**

$A \cap B$ will be **{2}**

- So by the formula of probability $P(A \cap B)$ will be = $\frac{|\{2\}|}{|\{1,2,3,4,5,6\}|} = \frac{1}{6}$

So by looking into Venn diagram, we observe that AUB means addition of all the elements of **Set A** and **Set B**

- We can also notice in set A we have {2, 4, 6} and in set B we have {1, 2}
- While adding the outcomes of the sets, {2} is occuring twice, which is nothing but $A \cap B$, so we have to subtract it once from our addition, as we want unique outcomes only (Since a set can only have distinct elements).

So the formula for $P(A \cup B)$ can we written as:

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

This is known as `Addition Rule`. This is for Joint Events

In case of **Disjoint Events**

- the intersection of $A \cap B = \{\ \}$ so, $P(A \cap B) = 0$
  - therefore, $P(A \cup B) = P(A) + P(B)$

Experiment 3: Sachin Tendulkar ODI records for India

Problem Statement:

We have a dataset containing Sachin Tendulkar's ODI cricket career stats, including various performance metrics and the outcomes of matches.

In [1]:
```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
```

In [2]:
```
df_sachin = pd.read_csv("Sachin_ODI.csv")
```

In [3]:
```
df_sachin.head()
```

Out[3]:

	runs	NotOut	mins	bf	fours	sixes	sr	Inns	Opp	Ground	Date	Winner
0	13	0	30	15	3	0	86.66	1	New Zealand	Napier	1995-02-16	New Zealand
1	37	0	75	51	3	1	72.54	2	South Africa	Hamilton	1995-02-18	South Africa
2	47	0	65	40	7	0	117.50	2	Australia	Dunedin	1995-02-22	India
3	48	0	37	30	9	1	160.00	2	Bangladesh	Sharjah	1995-04-05	India
4	4	0	13	9	1	0	44.44	2	Pakistan	Sharjah	1995-04-07	Pakistan

Each columns represents different features and each row represents a particular match

In [4]:
```
# shape of the dataset

df_sachin.shape
```

Out[4]:
```
(360, 14)
```

Q1. A match is randomly chosen, what is the probability that India have won that match?

Let's calculate this using the formula of probability, we know:

probability = $\frac{Possible\ Outcomes\ in\ an\ event}{Total\ Outcomes\ in\ an\ Entire\ Sample\ Space}$

Here we want the possible outcomes of India winning a match (WON = True)

Entire sample space will be our entire dataset

In [5]:
```
# find the rows where India have won and store into new dataframe
df_won=df_sachin.loc[df_sachin["Won"]==True]
```

In [6]:
```
# calculate the number of True values which is our possible outcome
df_won.shape[0]
```

Out[6]:
```
184
```

In [7]:
```
# We can also look at the length using len()
len(df_won)
```

Out[7]:
```
184
```

- So, probability

  = $\frac{number\ of\ matches\ won}{total\ number\ of\ matches}$

In [8]:
```
prob_winning=len(df_won)/len(df_sachin)
prob_winning
```

Out[8]:
```
0.5111111111111111
```

Conclusion: :

If a match is randomly chosen, there is **51%** chance that India have won that match.

Q2. A match is chosen at a random, what is the probability that Sachin has scored a Century in that match?

Solution 2:

Let's solve this using value counts function. First let's count the **number of centuries**, Sachin has scored

In [9]:
```
# using value_counts()

df_sachin["century"].value_counts()
```

Out[9]:
```
False    314
True      46
Name: century, dtype: int64
```

Out of 360 matches, Sachin has scored 46 Centuries.

so, probability of Sachin scoring a century will be:

In [10]:
```
46/360
```

Out[10]:
```
0.1277777777777777
```

Conclusion:

If you chose a random match, there is **12.77% chance** that Sachin has scored a century in that match

Cross Tab:

Now,

Let's find out how many matches India have won when Sachin has **`scored a century`** and

How many matches India have won when sachin **`didn't score a century`**.

> **Q. Can we achieve this task and obtain all these values at once?**

In [12]:
```
df_sachin[["century","Won"]].value_counts().T
```

Out[12]:
```
century  Won
False    False    160
         True     154
True     True      30
         False     16
dtype: int64
```

Cross Tab and contingency table

> **Q. Do you remember pivot table from DAV-1 Libraries module?**

- There is a function called `pd.crosstab()` , which accepts parameters **index** and **columns**.

In [13]:
```
pd.crosstab(index=df_sachin["century"],
            columns=df_sachin["Won"],
            margins=True)
```

Out[13]:

century	Won False	True	All
False	160	154	314
True	16	30	46
All	176	184	360

What we did using `.valuecounts()` at above, `pd.crosstab()` did the same thing but converted the output into nice tabular format

- **Century** is taken as the **index** and **Won** is taken as **columns**
- When we do **Margins = True** we get **All**, both in rows and columns,
  - The values of **All** in a ROW represnts the **Total Value** of each columns (False, True, All)
  - The values of **All** in a COLUMN represents the **Total Value** of each rows (False, True, All)

This table is also known as **`Contingency Table`**

We can calculate probabilities using the contingency table.

Q3. A match is chosen at a random. What is the probability that Sachin has scored a century in that match and India have won that match?

In [14]:
```
pd.crosstab(index=df_sachin["century"],
            columns=df_sachin["Won"],
            margins=True)
```

Out[14]:

Won	False	True	All
century			
False	160	154	314
True	16	30	46
All	176	184	360

In [15]:
```
# prob of winning and century
# Won -> True, century -> True

30/360
```

Out[15]:
```
0.08333333333333333
```

Conclusion :

There is **8% chance** that Sachin has scored a century and India have won that match if we choose a random match

This tells us, that **contingency table** is more convenient to calculate probabilities rather than hard coded the every single line

Conclusion of the Problem statement:

Let's have a look how is Sachin's batting can or cannot impact the winning chances of India

1. Out of the **360** matches that Sachin has played, India have **won 184** matches and Loose 176 matches.
2. So, if we choose any match at a random from Sachin's ODI career, there is a **51%** chance that India have won that match.

1. Now, If we choose a random match from Sachin's ODI career, there is **12.77%** chance that Sachin has scored a century in that match.
2. We know if a random match is choosen, there is 12.77% chance that Sachin has scored a century but

   there is **only 8%** chance India have won that match.

   - we can conclude that the chances of India, Winning a match is more when Sachin didn't score a century (what an amazing insight)

Finally,

We can conclude that, if we pick a random match where Sachin played, India's win percentage is 51%. There is 12.77% chance of Sachin scoring a century in that match, and there is only 8% chance that in that match Sachin scores a century as well as India have won that match

Content

- Conditional Probability
- Multiplication Rule
- Marginal and Joint Probability
- Law of Total Probability
- Baye's Theorem
    - Prior, Posterior and Likelihood Probabilities

WhatsApp Autocomplete Example

Conditional probability is a very important concept to understand.

In our daily life, all of you see direct examples of conditional probability. Lets look at one of them.

When typing a message on WhatsApp, we often encounter suggested words after typing a few.

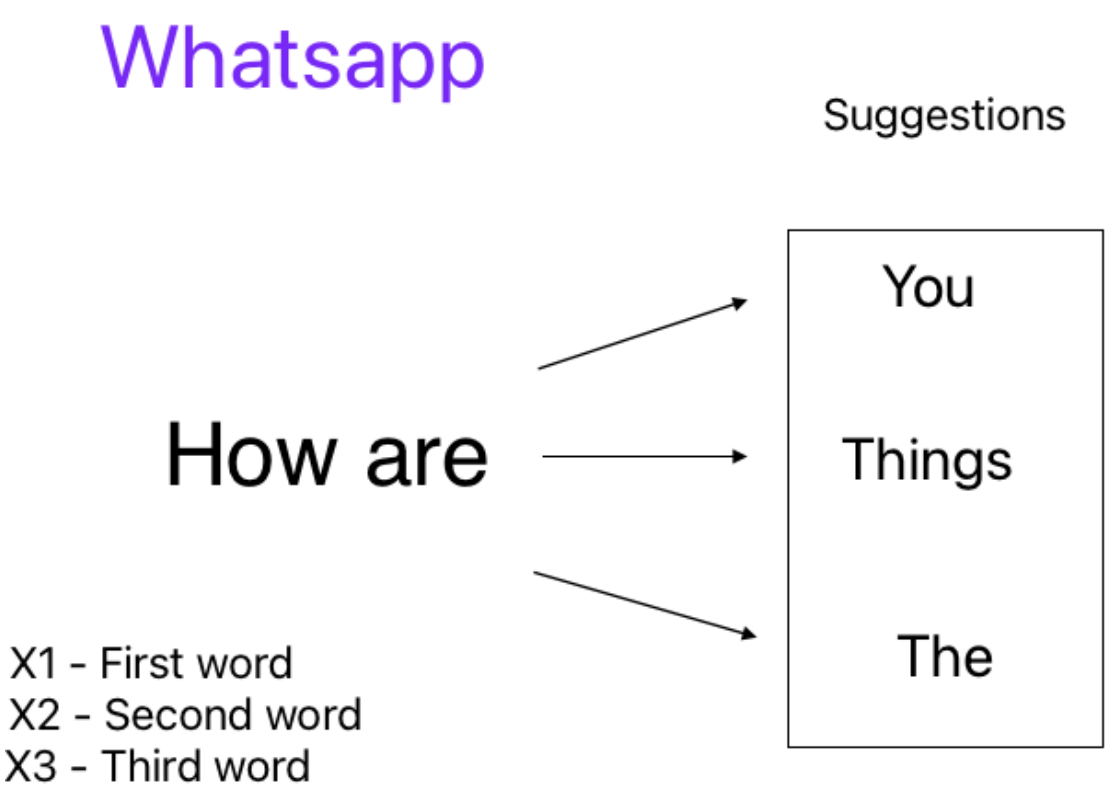
For instance, after typing `"How are"`, we might see suggestions like **"you"**, **"things"**, and **"the"**.

While these suggestions aren't guaranteed to be the next word you'll type but they're highly probable choices.

Is that magic? How did they know which words you may want to use next?

Let's assign a simple notations

- Let x_1 represents the first word
- Let x_2 represents the second word
- Let x_3 represents the third word



Now, you have given the following information to the keyboard:

- x_1 = "How" x_2 = "are"

Now internally, the algorithm needs to compute the probability for a word w that belongs in the dictionary, given the information about words x_1 and x_2.

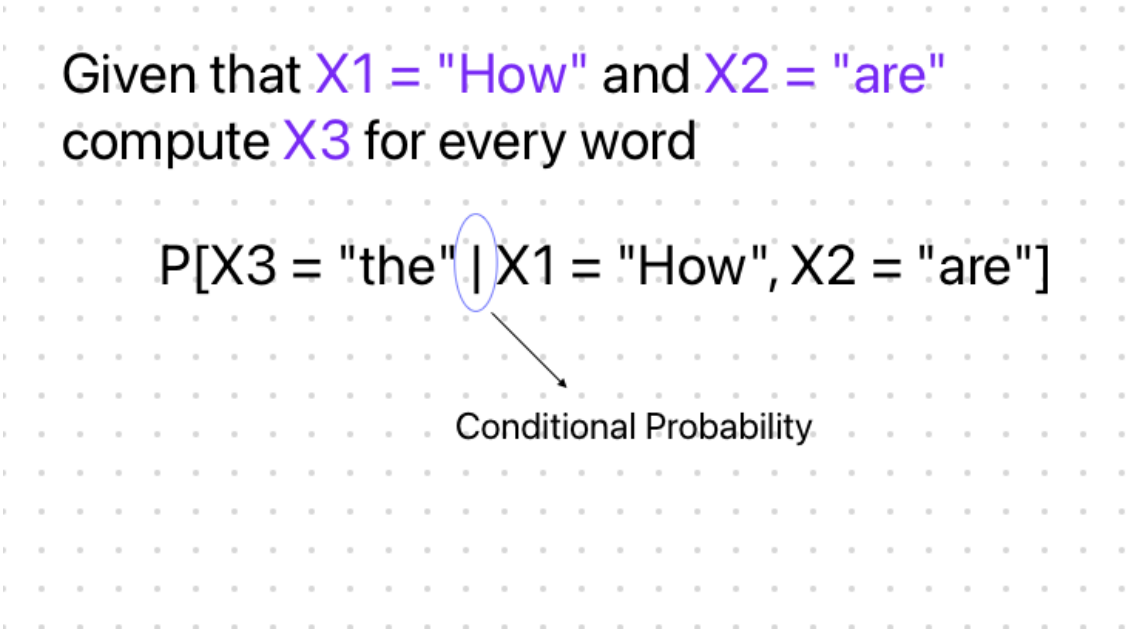
Consider this structure: $P(A|B)$

- Here, A represents the event whose probability we are trying to find
- B represents the events that have already happened / information given to us
- The vertical line `|` represents conditional probability

Therefore, we can represent it as:
$P(x_3 = w|x_1 =" \ How \ " \ \ and \ x_2 =" \ are \ ")$

> Read it as:
>
> - Probability of the word x_3 given that we have seen the words x_1 and x_2.

It then presents its findings, i.e. the words that are most likely to occur (having maximum probability) given that we have seen the words x_1 and x_2.

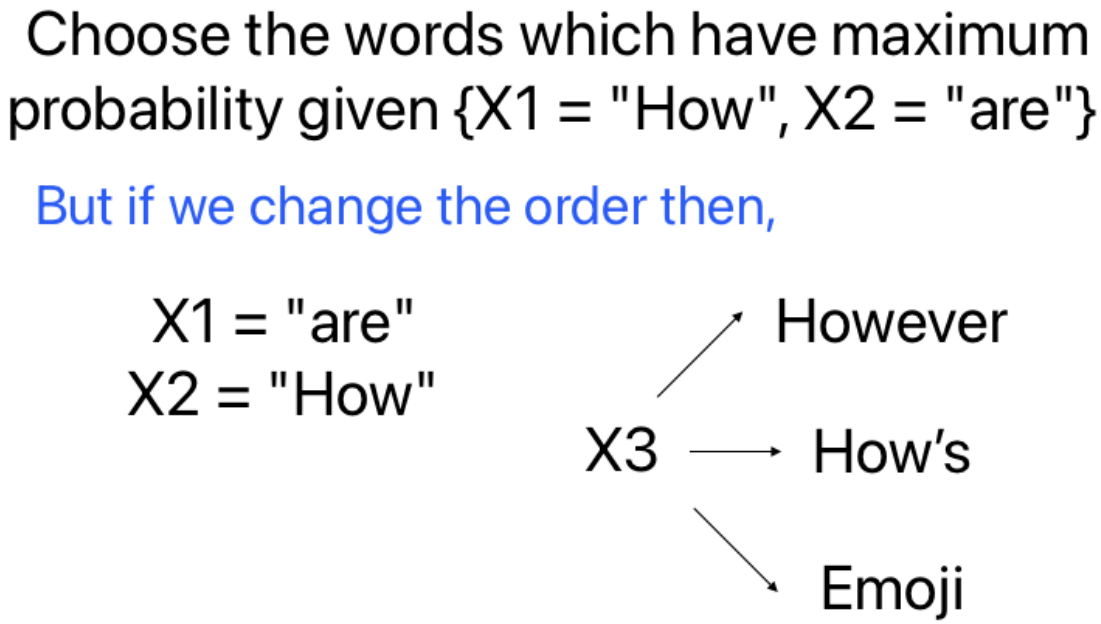


Note:

- The sequence is also important here.
- `you` , `things` , `the` are the top suggestions when x_1 = "How" and x_2 = "are".
- It would suggest different words if the case was x_1 = "are" and x_2 = "How"

Since this is not a sequence of words used very often, it might not give good suggestions here.

Auto complete is another example.



Conditional Probability

Probability of Event A, given Event B has already happened, is equivalent to the probability of A∩B , divided by probability of event B

i.e. $P(A|B) = \frac{P(A \cap B)}{P(B)}$

This equation is known as the **Conditional Probability Formula**

Multiplication Rule

Let's analyse this further,

From the above formula we will get:

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

In probability and statistics, this is known as the **Product / Multiplication Rule**.

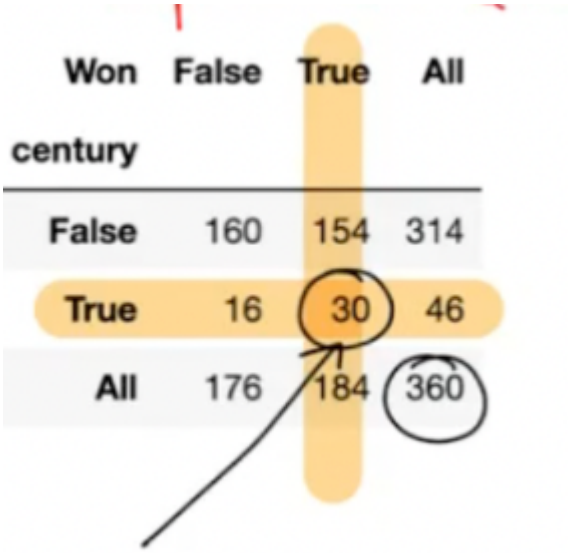
Similarly, we can expand $P(B \cap A) = P(B|A).P(A)$

Marginal and Joint Probabilities

Experiment: Sachin Tendulkar batting for India

Let's define the events happening here:

- W: Sachin's team winning the match
- C: Sachin scoring a century



Won / century	False	True	All
False	160	154	314
True	16	30	46
All	176	184	360

1) Marginal Probability

Let's answer a few questions based on this contingency table

> **Q1.What is the probability that Sachin's team wins the match?**

We need to find $P(W) = \frac{No\ of\ matches\ won\ by\ Sachin}{Total\ no\ of\ matches} = \frac{184}{360}$

> **Q2.What is the probability of Sachin scoring a century?**

$P(C) = \frac{No\ of\ matches\ with\ century}{Total\ no\ of\ matches} = \frac{46}{360}$

Similarly, we can calculate $P(W^C)$ and $P(C^C)$ as well.

All of these probability values are known as **Marginal Probability**

- It is the probability of an event irrespective of the outcome of other variable.
- For instance, consider $P(W)$
  - It denotes the total probability of Sachin's team winning the match, considering both possibilities that Sachin may or may not score a century.
- It is not conditioned on another event. It may be thought of as an **unconditional probability.**
- Other example:
  - Probability that a card drawn is a 4 : P(four)=1/13.
    - This includes the possibility of the 4 being a spades, heart, club or diamond.
  - Probability that a card drawn is spades : P(spades)=1/4.

2) Joint Probability

Now let's look at the second type of probability values, by answering the following questions.

> **Q1.What is the probability that Sachin's team wins AND he scores a century?**

We need to find $P(W \cap C) = \frac{30}{360}$

> **Q2.What is the probability that Sachin scored a century AND his team wins?**

We need to find $P(C \cap W)$

This will be the same as $P(C \cap W) = P(W \cap C) = \frac{30}{360}$

> **Q3.What is the probability that Sachin scores a century AND his team loses?**

$P(W^C \cap C) = \frac{16}{360}$

Similarly, we can find $P(W^C \cap C^C)$ and $P(W \cap C^C)$

Note:

- Here we calculated the likelihood of two events occurring **together** and at the same point in time.
- This type of probability value is known as **Joint Probability**.
- And it is represented as we saw: $P(A \cap B)$
  - Where, A and B are 2 events.
  - It is read as `Probability that event A and B happen at same time`.
- Other Example: the probability that a card is a four and red = P(four and red) = 2/52

The third kind of probability value, we've just studied, i.e. **Conditional Probability**.

Let's answer a few questions on this also

> **Q1.What is the probability that Sachin's team wins the match given that he scored a century?**

Since it is given that he scores a century, our subset reduces to the second row.

Now since we want to find the prob of team winning among these matches, our probability becomes: $P(W|C) = \frac{30}{46}$

> **Q2.What is the probability that Sachin scores a century, given that his team has won the match?**

As per the given extra information, our subset reduces to the second column.

So among these 184 matches, where India won, Sachin scored a century in only 30 matches.

Therefore $P(C|W) = \frac{30}{184}$

Similarly, we can be asked to calculate other conditional probabilities such as: $P(W|C^C)$, $P(W^C|C)$, $P(C|W^C)$, etc.

> **Q.How can we find the values of Marginal Probability?**

Law of Total Probability

- If we re-arrange the formula of conditional probability $P(A|B) = \frac{P(A \cap B)}{P(B)}$, we will get get:

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

This is known as **Law of Total Probability**

Total Probability Law Generic Formula

- Mathematically, The Law of Total Probability is stated as follows:

$P(A) = \sum_{i=1}^{n} P(A \mid B_i)P(B_i)$

Let's have a look into example

```
Example: Email Spam Detection
```

The Law of Total Probability helps combines the information from multiple scenarios or conditions to arrive at a comprehensive probability estimate, making it a valuable tool in various data science and machine learning applications.

Formulas learned so far

1) Conditional Probability:

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

2) Multiplication Rule:

- $P(A \cap B) = P(A \mid B) \cdot P(B)$

3) Law of Total Probability:

- $P(A) = \sum_{i=1}^{n} P(A \mid B_i)P(B_i)$

Let's jump to new concept

Baye's Theorem

- $P(A|B) = \frac{P(B|A).P(A)}{P(B)}$

This equation that we used here is known as the **Bayes Theorem.**

Quick Derivation of Bayes Theorem

From the questions we have solved so far,

> **Q1. Can we say that $P(A \cap B) = P(B \cap A)$?**

We know that $A \cap B$ and $B \cap A$ represent the same subset, i.e. the common elements between A and B.

And, from the Multiplication Rule we can expand them as:

- $P(A \cap B) = P(A|B).P(B)$
- $P(B \cap A) = P(B|A).P(A)$

Since the LHS of both these equations is same, we can equate the RHS also.

$$P(A|B).P(B) = P(B|A).P(A)$$

Dividing both sides by $P(B)$,

$$P(A|B) = \frac{P(B|A).P(A)}{P(B)}$$

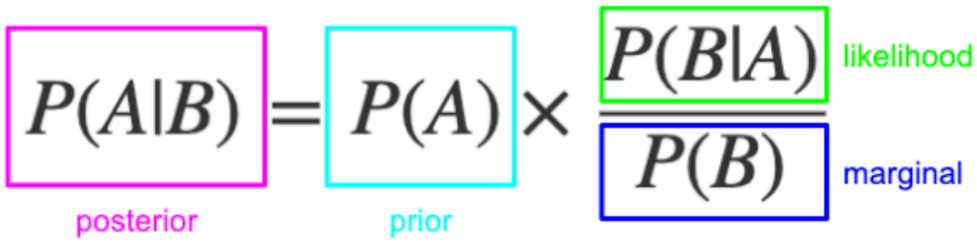
This is exactly the equation of Baye's Theorem.

Let's take a closer look at the Bayes equation.

Prior, Posterior and Likelihood Probabilities

It consists of 4 parts:

- **Posterior probability** (updated probability after the evidence is considered)
- **Prior probability** (the probability before the evidence is considered)
- **Likelihood** (probability of the evidence, given the belief is true)
- **Marginal probability** (probability of the evidence, under any circumstance)



The equation: Posterior = Prior x (Likelihood over Marginal probability)

Let's understand the different terms here.

- **Posterior probability**
  - The Bayes' Theorem lets you calculate the posterior (or "updated") probability.
  - It is the conditional probability of the **hypothesis being true, if the evidence is present.**
  - $P(Hypothesis|Evidence)$
- **Prior Probability**
  - Can be perceived as your **belief in the hypothesis before seeing the new evidence.**
  - Therefore, if we have a strong belief in the hypothesis already, the prior probability will be large.

- $P(Hypothesis)$

- **Likelihood**
  - The prior is multiplied by a fraction.
  - Think of this as the "strength" of the evidence.
  - The posterior probability is greater when the top part (numerator) is big, and the bottom part (denominator) is small.
  - The numerator is the likelihood.
  - It is the conditional probability of the **evidence being present, given the hypothesis is true.**
  - This is not the same as the posterior!!
    $P(Evidence|Hypothesis) \neq P(Hypothesis|Evidence)$

- **Marginal Probability**
  - Notice the denominator of this fraction.
  - It is the marginal probability of the evidence. $P(Evidence)$
  - That is, it is the **probability of the evidence being present, whether the hypothesis is true or false.**
  - We can find it using Total Probability Law
  - The smaller the denominator, the more "convincing" the evidence

In [ ]:

Content

- **Problem Solving**
- **Mini Case Study**

Formulas learnt so far

Let's recall all the formulas that we have learned so far,

1. **Conditional probability:** $P[A|B] = \frac{P[A \cap B]}{P[B]}$
2. From conditional probability we will get,

   $P[A \cap B] = P[A|B] \ * \ P[B]$

   which is known as **Multiplication Rule**

1. **Bayes Theorem:** $P[A|B] = \frac{P[B|A] * P[B]}{P[A]}$
2. **Law of total probability:** $P(A) = \sum_{i=1}^{n} P(A \mid B_i)P(B_i)$
3. **Independent Events:** $P[A \cap B] = P[A] \ * \ P[B]$

Now let's verify one claim.

Claim: If A and B are mutually Exclusive then A and B are not independent.

We know that if A and B are mutually exclusive or Disjoint events:

- $A \cap B = \{\}$
  Note : $A \cap B$ is a null/empty set as A and B can't occur at the same time
- So, $P(A \cap B) = 0$

But in the case of independent events:

- $P(A \cap B) = P(A) * P(B)$ (we just saw above)

In the case of mutually exclusive events $P(A \cap B)$ is not equal to $P(A) * P(B)$, as A and B are not independent.

Therefore, the claim is proven: If A and B are mutually exclusive, then A and B are not independent.

Alternate Method : Using the conditional probability formula:

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

For Disjoint events:

- $P(A \cap B) = 0$
  - So, $P(A|B) = \frac{0}{P(B)} = 0$

For independent Events:

- $P(A \cap B) = P(A) * P(B)$
  - So, $P(A|B) = \frac{P(A) * P(B)}{P(B)} = P(A)$

As we can see in both the events $P(A|B)$ is different

Hence, we can conclude that :

If A and B are mutually Exclusive then A and B are not independent.

Example: 1

```
In a university, 30% of faculty members are females. Of the
female faculty members, 60% have a PHD. Of the male faculty
members, 40% have a PHD
- What is the probability that a randomly chosen faculty
member is a female and has PHD?
- What is the probability that a randomly chosen faculty
member is a male and has PHD?
- What is the probability that a randomly chosen faculty
member has a PHD?
- What is the probability that a randomly chosen PHD holder
is female?
```

Explanation:

Given,

- Female faculty members = 30%
  - Out of this 30% members, 60% have PHD
- Male faculty members = 100 - 30 = 70%
  - Out of this 70% members, 40% have PHD

Let's define probabilities:

- probability that a randomly chosen faculty member is a female i.e. $P(F) = 0.3$
  - Given that faculty member is a Female, the probability that she has a PHD is i.e $P(phd \mid F) = 0.6$
- probability that a randomly chosen faculty member is a Male i.e. $P(M) = 0.7$
  - Given that faculty member is a Male, the probability that he has a PHD is i.e $P(phd \mid M) = 0.4$

Answering questions:

> **Q1. What is the probability that a randomly chosen faculty member is a female and has PHD?**

We know `AND` means intersection, here we want to find $P(phd \cap F)$

- Using the formula of conditional probability,

  $P(phd \mid F) = \frac{P(phd \cap F)}{P(F)}$

  So, $P(phd \cap F) = P(phd \mid F) * P(F)$

  Adding values into the equation

  - $P(phd \cap F) = 0.6 * 0.3 = 0.18$

Conclusion:

The probability that a randomly chosen faculty member is a female and has PHD is **0.18**

Similarly,

> **Q2. What is the probability that a randomly chosen faculty member is a male and has PHD?**

- Using the formula of conditional probability,

  $P(phd \mid M) = \frac{P(phd \cap M)}{P(M)}$

  so, $P(phd \cap M) = P(phd \mid M) * P(M)$

  Adding values into the equation

  - $P(phd \cap M) = 0.4 * 0.7 = 0.28$

Conclusion:

The probability that a randomly chosen faculty member is a male and has PHD is **0.28**

> **Q3. What is the probability that a randomly chosen faculty member has a PHD?**

We have 2 approaches to solve this question.

Approach 1:

- Here, we need to find the probability that If I choose a random person, then he/she have a PHD, no matter whether the person is MALE or FEMALE. i.e. $P(phd)$
- We can add $P(phd \cap F) + P(phd \cap M)$ as it'll give me $P(phd)$
- $P(phd) = P(phd \cap F) + P(phd \cap M)$

  adding values into the equation

  - $P(phd) = 0.18 + 0.28 = 0.46$

Approach 2:

- As we know, we can write $P(phd \cap F)$ **as a** $P(phd \mid F) * P(F)$ because, $P(phd \mid F) = \frac{P(phd \cap F)}{P(F)}$

  Here comes the Law of total probability in picture

- For Male also, we can write $P(phd \cap M)$ **as a** $P(phd \mid M) * P(M)$

Replacing these values in the equation,

- $P(phd) = [P(phd \mid F) * P(F)] + [P(phd \mid M) * P(M)]$
  - $P(phd) = [0.6 * 0.3] + [0.4 * 0.7]$

    = $P(phd) = 0.46$

Conclusion:

The probability that a randomly chosen faculty member has a PHD is **0.46**

> **Q4. What is the probability that a randomly chosen PHD holder is female?**

Here, we are already given that the randomly chosen person is PHD holder and we need to find the probability of this person being Female. We need to find: $P(F \mid phd)$

Using the **formula of conditional probability:**

- $P(F \mid phd) = \frac{P(phd \cap F)}{P(phd)}$

  Replace the $P(phd \cap F)$ with $P(phd \mid F) * P(F)$,

  and $P(phd)$ with $[P(phd \mid F) * P(F)] + [P(phd \mid M) * P(M)]$

Final forumla will be:

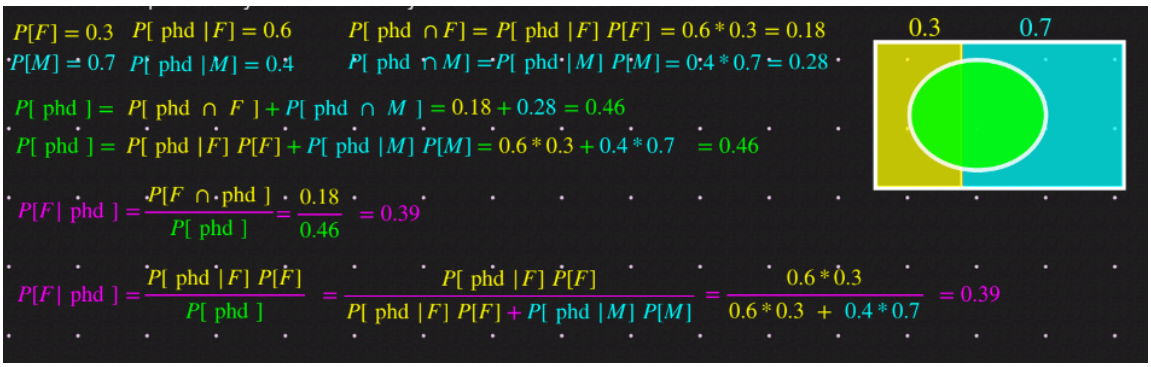
- $P(F \mid phd) = \frac{P(phd \mid F) * P(F)}{[P(phd \mid F) * P(F)] + [P(phd \mid M) * P(M)]}$
  - $P(F \mid phd) = \frac{0.6*0.3}{[0.6*0.3]+[0.4*0.7]}$
  - $P(F \mid phd) = 0.39$

Conclusion:

The probability that a randomly chosen PHD holder is female is **0.39**

There is an alternative approach to solve this question, called **tree based approach**

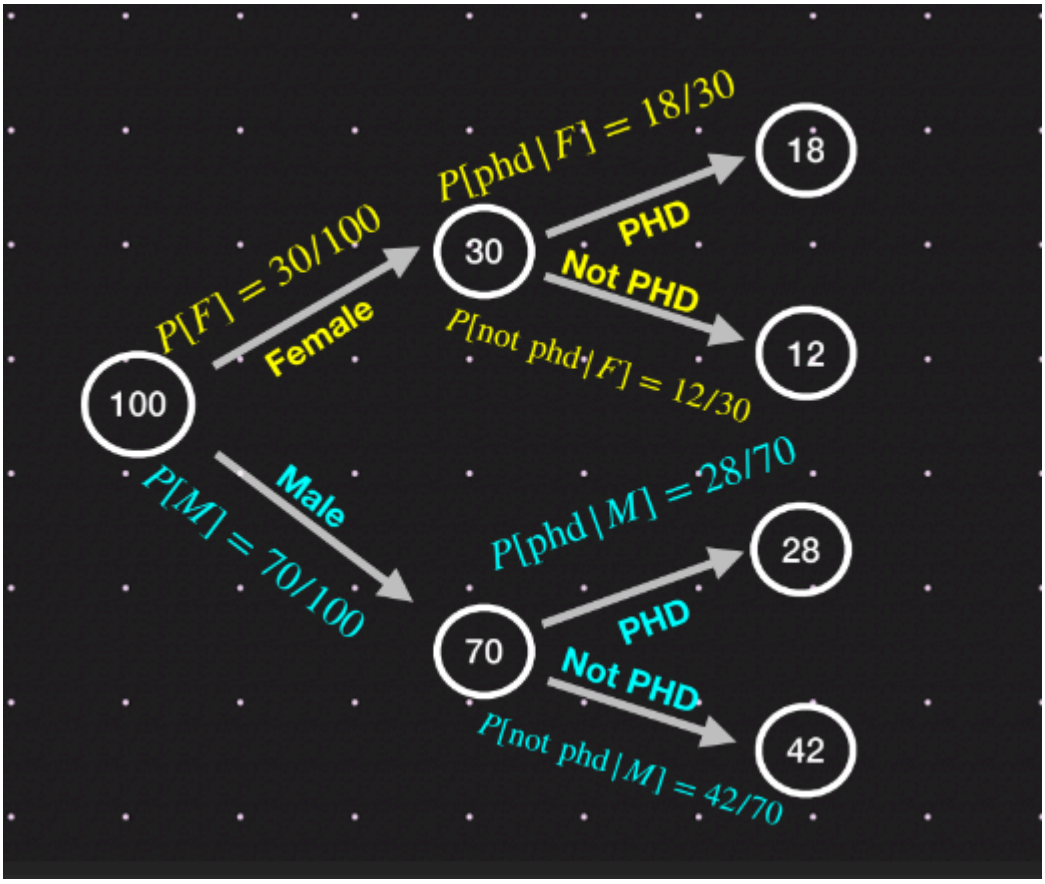
Let's solve this question with tree based approach.



Tree based approach:

Let's assume there are 100 faculty members. Now among these 100 faculty members,

They can be divided into two parts, they can be either male or female.



Explanation of the structure of the Tree:

> **Q1. How many of them are female and how many of them are Male?**

Female : 30% of 100 = 30 (as $P(F) = 0.3$)

We can further seggregate the female part into 2 part:

- Female **AND having** a PHD : 60% of 30 = 18
  - We can represent it as $P(phd \mid F) = 0.6$
- Female **AND NOT having** a PHD : 30 - 18 = 12
  - We can represent it as $P(phd' \mid F) = 1 - P(phd \mid F) = 0.4$

Same for the Male:

Male : 70% of 100 = 70 (as $P(M) = 0.7$)

- Male **AND having** a PHD : 40% of 70 = 28
  - We can represent it as $P(phd \mid M) = 0.4$
- Male **AND NOT having** a PHD : 70 - 28 = 42
  - We can represent it as $P(phd' \mid M) = 1 - P(phd \mid M) = 0.6$

The structure of tree is ready.

Now let's solve the questions

> **Q1. What is the probability that a randomly chosen faculty member is a female and has PHD?**

Let's see how we can easily solve this using tree based approach

We want faculty member and PHD

- From our tree diagram, we can see that there are **18 faculty members who are Female and has PHD**.
  - So $P(F \cap phd)$ = 18/100 = 0.18

    We can observe tha we are getting the same answer but how conviniently we are able to solve this problem with this approach

> **Q2. What is the probability that a randomly chosen faculty member is a male and has PHD?**

Following the same approach as above

- $P(M \cap phd)$ = 28/100 = 0.28

> **Q3. What is the probability that a randomly chosen faculty member has a PHD?**

Here we want to find **total number of faculties having PHD**, it doesn't matter whether the member is male or female

- It will be (18 + 28)/100 = 0.46

> **Q4. What is the probability that a randomly chosen PHD holder is female**?

We have 2 ways to reach the PHD, one through FEMALE and one through MALE

- Now, we need the member **who already has PHD but is a female**.

  It'll be $\frac{18}{18+28}$ = 0.39

> **Q5. What is the probability that a randomly chosen PHD holder is male?**

Following the same approach as above

- $P(M \mid phd) = \frac{28}{18+28} = 0.6$

We can see how conviniently and easily we are able to solve all the questions using this Tree based approach

Kerala Flood Case Study

- The dataset contains the monthly rainfall data from years 1901 to 2018 for the Indian state of Kerala.
- It contains the monthly rainfall index of Kerela and also record weather a flood took place that month or not.

In [1]:
```
# Import libraries
import numpy as np
import pandas as pd
```

In [2]:
```
# Read the data
df = pd.read_csv("kerala.csv")
df.head(10)
```

Out[2]:

	SUBDIVISION	YEAR	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	OCT
0	KERALA	1901	28.7	44.7	51.6	160.0	174.7	824.6	743.0	357.5	197.7	266.9
1	KERALA	1902	6.7	2.6	57.3	83.9	134.5	390.9	1205.0	315.8	491.6	358.4
2	KERALA	1903	3.2	18.6	3.1	83.6	249.7	558.6	1022.5	420.2	341.8	354.1
3	KERALA	1904	23.7	3.0	32.2	71.5	235.7	1098.2	725.5	351.8	222.7	328.1
4	KERALA	1905	1.2	22.3	9.4	105.9	263.3	850.2	520.5	293.6	217.2	383.5
5	KERALA	1906	26.7	7.4	9.9	59.4	160.8	414.9	954.2	442.8	131.2	251.7
6	KERALA	1907	18.8	4.8	55.7	170.8	101.4	770.9	760.4	981.5	225.0	309.7
7	KERALA	1908	8.0	20.8	38.2	102.9	142.6	592.6	902.2	352.9	175.9	253.3
8	KERALA	1909	54.1	11.8	61.3	93.8	473.2	704.7	782.3	258.0	195.4	212.1
9	KERALA	1910	2.7	25.7	23.3	124.5	148.8	680.0	484.1	473.8	248.6	356.6

In [3]:
```
df.shape
```

Out[3]:
```
(118, 16)
```

In [13]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 118 entries, 0 to 117
Data columns (total 16 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   SUBDIVISION      118 non-null    object
 1   YEAR             118 non-null    int64
 2   JAN              118 non-null    float64
 3   FEB              118 non-null    float64
 4   MAR              118 non-null    float64
 5   APR              118 non-null    float64
 6   MAY              118 non-null    float64
 7   JUN              118 non-null    float64
 8   JUL              118 non-null    float64
 9   AUG              118 non-null    float64
 10  SEP              118 non-null    float64
 11  OCT              118 non-null    float64
 12  NOV              118 non-null    float64
 13  DEC              118 non-null    float64
 14  ANNUAL_RAINFALL  118 non-null    float64
 15  FLOODS           118 non-null    object
dtypes: float64(13), int64(1), object(2)
memory usage: 14.9+ KB
```

Let's calculate average rainfall for each month over the years

> **Q. What is the average rainfall for each month over the years**

In [4]:
```
# Calculate the average rainfall for each month
cols = ['JAN', 'FEB', 'MAR', 'APR','MAY','JUN','JUL','AUG','SEP','OCT','NOV

monthly_avg = df[cols].mean()
monthly_avg
```

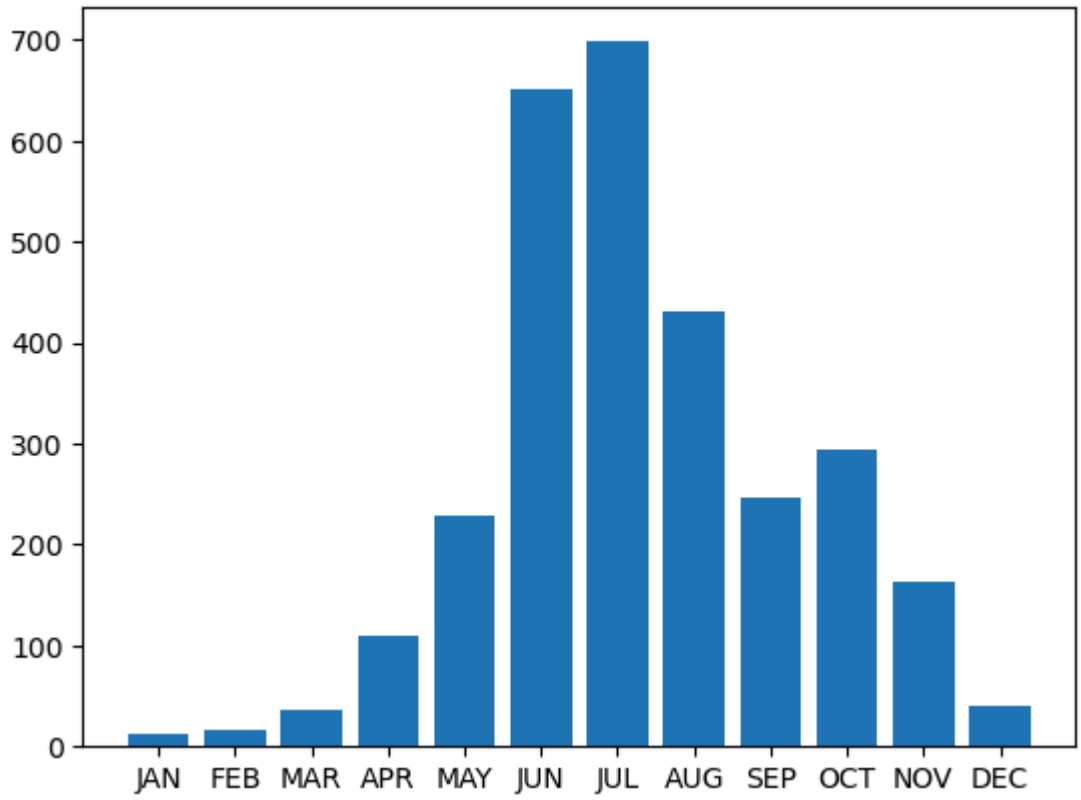
Out[4]:
```
JAN     12.218644
FEB     15.633898
MAR     36.670339
APR    110.330508
MAY    228.644915
JUN    651.617797
JUL    698.220339
AUG    430.369492
SEP    246.207627
OCT    293.207627
NOV    162.311017
DEC     40.009322
dtype: float64
```

In [5]:
```
import matplotlib.pyplot as plt
import seaborn as sns
```

In [6]:
```
x=monthly_avg.index
y=monthly_avg

plt.bar(x,y)
```

Out[6]:
```
<BarContainer object of 12 artists>
```



We can make few **conclusions** here:

- The data reveals significant seasonal variation in rainfall.
  - **June and July** have the **highest average rainfall**, while **January and February** are the driest months
  - The rainfall in **August and September** is still relatively high but begins to decline
  - Surprisingly, **October** has a **higher average rainfall than September**, which may seem counterintuitive.

There are two monsoon seasons in Kerala, **one during Jun-Aug, Other during Oct**.

the important features in this dataset are "JUN", "JUL", "OCT" , "ANNAUL_RAINFALL", "FLOODS"

because in these months only we have seen the peak of the rainfall which can be one of the major source of causing the flood

In [7]:
```
df.columns = [c.replace(' ANNUAL RAINFALL', 'ANNUAL_RAINFALL') for c in df.c

df.head()
```

Out[7]:

	SUBDIVISION	YEAR	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	OCT
0	KERALA	1901	28.7	44.7	51.6	160.0	174.7	824.6	743.0	357.5	197.7	266.9
1	KERALA	1902	6.7	2.6	57.3	83.9	134.5	390.9	1205.0	315.8	491.6	358.4
2	KERALA	1903	3.2	18.6	3.1	83.6	249.7	558.6	1022.5	420.2	341.8	354.1
3	KERALA	1904	23.7	3.0	32.2	71.5	235.7	1098.2	725.5	351.8	222.7	328.1
4	KERALA	1905	1.2	22.3	9.4	105.9	263.3	850.2	520.5	293.6	217.2	383.5

In [8]:
```
df.columns
```

Out[8]:
```
Index(['SUBDIVISION', 'YEAR', 'JAN', 'FEB', 'MAR', 'APR', 'MAY', 'JUN', 'JUL',
       'AUG', 'SEP', 'OCT', 'NOV', 'DEC', 'ANNUAL_RAINFALL', 'FLOODS'],
      dtype='object')
```

In [9]:
```
impactful_columns = ['YEAR', 'JUN', 'JUL', 'OCT', 'ANNUAL_RAINFALL', 'FLOODS

impactful_columns
```

Out[9]:
```
['YEAR', 'JUN', 'JUL', 'OCT', 'ANNUAL_RAINFALL', 'FLOODS']
```

Now, I want to label the months column with 0 and 1

- 0: will represents low rainfall
- 1: will represents heavy rainfall

Similarly for "ANNUAL_RAINFALL" column:

- 0: will represents low rainfall in that particular year
- 1: will represents heavy rainfall in that particular year

> **Q. But how much rainfall index is considered as a heavy ranifall?**

One of the parameter is using the **Median** values of these columns.

If their individual **rainfall index value > median value** then it'll we considered as **heavy rainfall** and vice a versa

In [10]:
```
# new dataset containing only impactful columns

data = df[impactful_columns]

data.head()
```

Out[10]:

	YEAR	JUN	JUL	OCT	ANNUAL_RAINFALL	FLOODS
0	1901	824.6	743.0	266.9	3248.6	YES
1	1902	390.9	1205.0	358.4	3326.6	YES
2	1903	558.6	1022.5	354.1	3271.2	YES
3	1904	1098.2	725.5	328.1	3129.7	YES
4	1905	850.2	520.5	383.5	2741.6	NO

In [14]:
```
# Assuming 'data' is a DataFrame with your specified data
threshold_jun = int(data['JUN'].median())
threshold_jul = int(data['JUL'].median())
threshold_oct = int(data['OCT'].median())
threshold_ar = int(data['ANNUAL_RAINFALL'].median())

threshold_jun, threshold_jul, threshold_oct, threshold_ar
```

Out[14]:
```
(625, 691, 284, 2934)
```

In [15]:
```
thresholds = {
    'JUN': 625,
    'JUL': 691,
    'OCT': 284,
    'ANNUAL_RAINFALL': 2934
}

# Convert columns to binary based on thresholds
for col, threshold in thresholds.items():
    data[col] = (data[col] > threshold).astype(int)

data.head()
```

```
/var/folders/zk/yt14z40j2lb2lz548fqr3v9m0000gn/T/ipykernel_19049/443241069.
py:10: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-doc
s/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  data[col] = (data[col] > threshold).astype(int)
```

Out[15]:

	YEAR	JUN	JUL	OCT	ANNUAL_RAINFALL	FLOODS
0	1901	1	1	0	1	YES
1	1902	0	1	1	1	YES
2	1903	0	1	1	1	YES
3	1904	1	1	1	1	YES
4	1905	1	0	1	0	NO

In [16]:
```
data['FLOODS'].unique()
```

Out[16]:
```
array(['YES', 'NO'], dtype=object)
```

Q1. Calculate the Probability of flood given that rainfall in June is greater than the median june rainfall value (threshold for heavy rainfall)

Question Explanation:

Let A represents : Flood
B represents: heavy rain in June

We need to calculate $P(A|B)$ i.e. $\frac{P(A \cap B)}{P(B)}$

Solution Approach 1:

We can obtain these values using contingency table and put those values into the formula.

Here we need to compare "FLOODS" and "JUN" column.

In [17]:
```
pd.crosstab(data['JUN'],
            data['FLOODS'],
            margins=True,
            margins_name='Total')
```

Out[17]:

FLOODS	NO	YES	Total
JUN			
0	42	16	58
1	16	44	60
Total	58	60	118

Now, $P(A \cap B)$ = Probability of Flood occuring `AND` heavy rainfall in JUNE

As we know in the contingency table, FLOODS = YES represents that flood has occured and JUN = 1 means heavy rainfall.

We need to check value where FLOODS = YES and JUN = 1 which is **44**

Then by the formula of condititonal probability we can feed this data

In [18]:
```
# probability of high rainfall in June P(J)
# P(J) = possible outcomes in june having heavy rainfall / total outcomes

P_J = (16+44)/(42+16+16+44)

# now, P(A and B) (Flood = YES and Jun = 1)

P_F_and_J = 44/(42+16+16+44)

#, so our probability of flood occurring given that the high rainfall occure

P_F_J = P_F_and_J / P_J

print(f'P(J) : {P_J}')
print(f'P(F AND J) : {P_F_and_J}')
print(f'P(F|J): {P_F_J}')
```

```
P(J) : 0.5084745762711864
P(F AND J) : 0.3728813559322034
P(F|J): 0.7333333333333334
```

Approach 2: using normalize attribute

Explanation of Normalize attribute:

Rather putting all the values in the formula and then calculate the probability

We can just pass one more attribute in pd.crosstab() function which will divide all values by the sum of values.

- This is the probability only, as in probability we divide `possible outcome / total outcome (sum of all values)`

Parameter is : **normalize = ' '**

- Without this attribute, the contingency table will show the raw counts of occurrences for each combination of variables.
- It will not be normalized, and the values in the table will represent counts.

Here we can pass these strings in this attribute:

normalize='index' or **normalize='columns'** :

- The normalize attribute specifies how the values in the contingency table should be normalized.
  - When set to **'index'**, it calculates conditional probabilities based on rows, treating each row as a separate condition.
  - When set to **'columns'**, it calculates conditional probabilities based on columns, treating each column as the condition we are focusing on.
- This means that each row in the table is divided by the sum of its row, making each row's values sum up to 1, representing conditional probabilities.

Same with the column

In this case:

By setting **normalize='index'**,

- the code calculates conditional probabilities within each row.
- Each value in the table represents the probability of the corresponding event (FLOODS) given the value of 'JUN' in that row.

The row sums up to 1, ensuring that it reflects the conditional probabilities.

In summary,

setting normalize='index' in pd.crosstab allows you to calculate and visualize conditional probabilities based on the specified row variable ('JUN' in this case),

making it easier to assess the impact of one variable on another.

In [19]:
```
pd.crosstab(index = data['JUN'],
            columns = data['FLOODS'],
            margins=True,
            normalize='index')
```

Out[19]:

FLOODS	NO	YES
JUN		
0	0.724138	0.275862
1	0.266667	0.733333
All	0.491525	0.508475

The values in the table represent the conditional probabilities, where each cell contains the probability of the corresponding outcome (FLOODS) given the condition in June (JUN).

> Then the probability of flood occurring given that the heavy rainfall occured in June will be:

- In the cell at row 1, column 1, the value **0.73333** represents the conditional probability of flooding (FLOODS = YES) given that high rainfall occurred in June (JUN = 1).

Conclusion:

```
  So, there is 73.33% chance of Floods when there is a heavy
rainfall in June
```

As we can see by calculating using formula also, we are getting the same answer as using directly conditional probability using normalize = 'index'

Now, let's jump into the next question

Q2. Given that there is a flooding, calculate the probability that heavy rainfall has occurred in July (more than threshold value)?

Here we want to find $P(July = 1|Flood = YES)$

We are already aware of using formula based approach, so We will solve this using contingency table

Before proceeding,

> **Q. In this question, which string will be passed inside normalize=' ' attribute? 'index' or 'columns'**

In this question, we should normalize the contingency table along the columns

- As we want to find the conditional probability of **high rainfall in July (JUL = 1) given that there was flooding (FLOODS = YES)**,

We want to see how the 'JUL' column behaves when there is flooding.

In [20]:
```
pd.crosstab(index = data['JUL'],
            columns = data['FLOODS'],
            margins=True,
            normalize='columns')
```

Out[20]:

FLOODS	NO	YES	All
JUL			
0	0.655172	0.35	0.5
1	0.344828	0.65	0.5

Conclusion:

The probability that high rainfall occurred in July (JUL = 1) given flooding (FLOODS = YES) is **0.65**.

- **This means that when there is flooding, there is a 65% chance of heavy rainfall in July.**

Q3.Calculate the probability of flood given that june and july rainfall was greater than their median rainfall value?

Solution:

We want to find $P(Flood = Yes|\ june = 1\ and\ Jul = 1)$

Here, we can pass multiple columns in the **pd.crosstab()**

In [21]:
```
pd.crosstab(index = [data['JUN'], data['JUL']],
            columns = data['FLOODS'],
            margins=True,
            normalize='index')
```

Out[21]:

	FLOODS	NO	YES
JUN	JUL		
0	0	0.862069	0.137931
	1	0.586207	0.413793
1	0	0.433333	0.566667
	1	0.100000	0.900000
All		0.491525	0.508475

Conclusion

Frequency (JUN = 1, JUL = 1, FLOODS = YES) = 0.9000000

There is **90%** chance of flood given that heavy rainfall in both june and july

In [ ]: