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}`

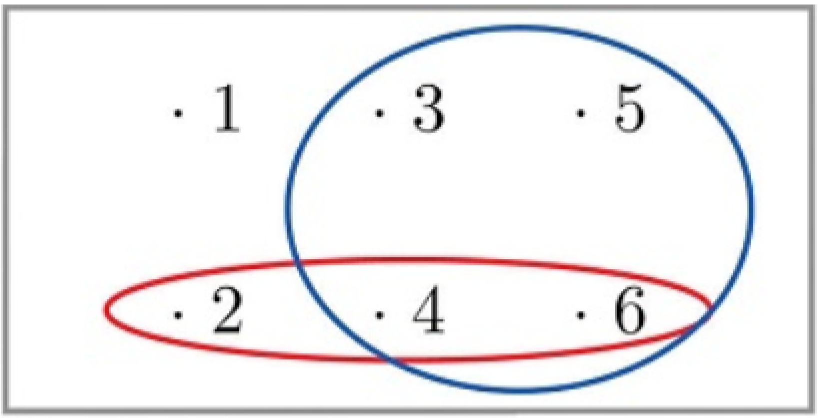
This is known as an **`Event`**.
Any subset of sample space is an event.

- {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?**

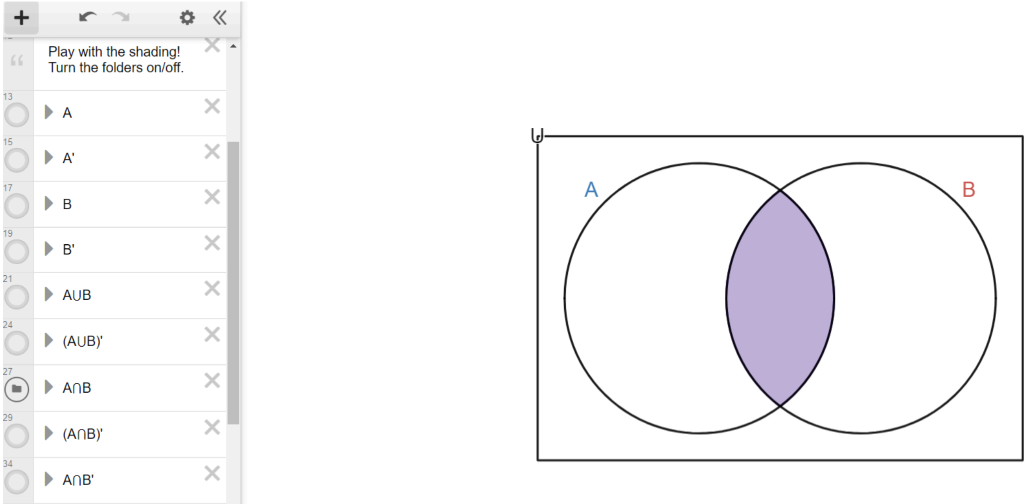
We want a number which occurs in both of their events

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,

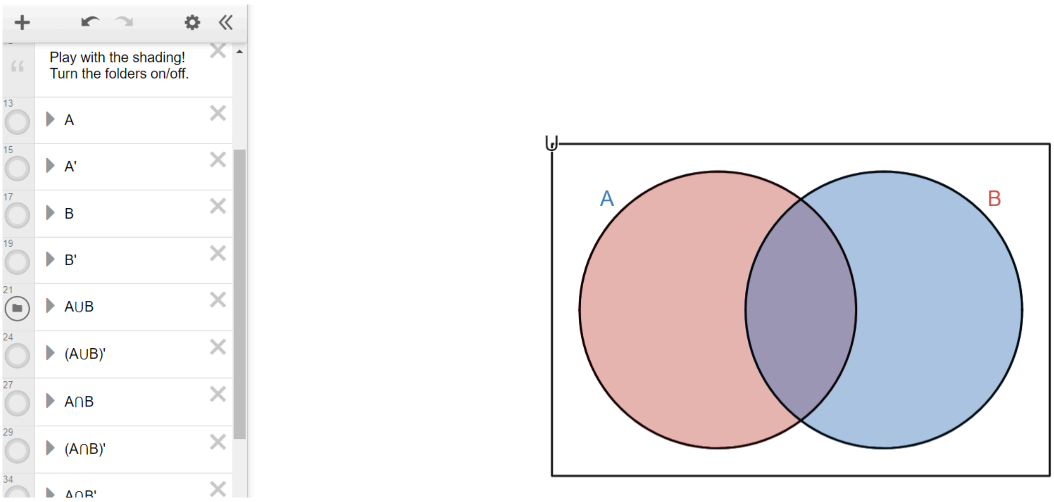
> **Q. When either Mohit or Rakesh will win their bets?**

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)

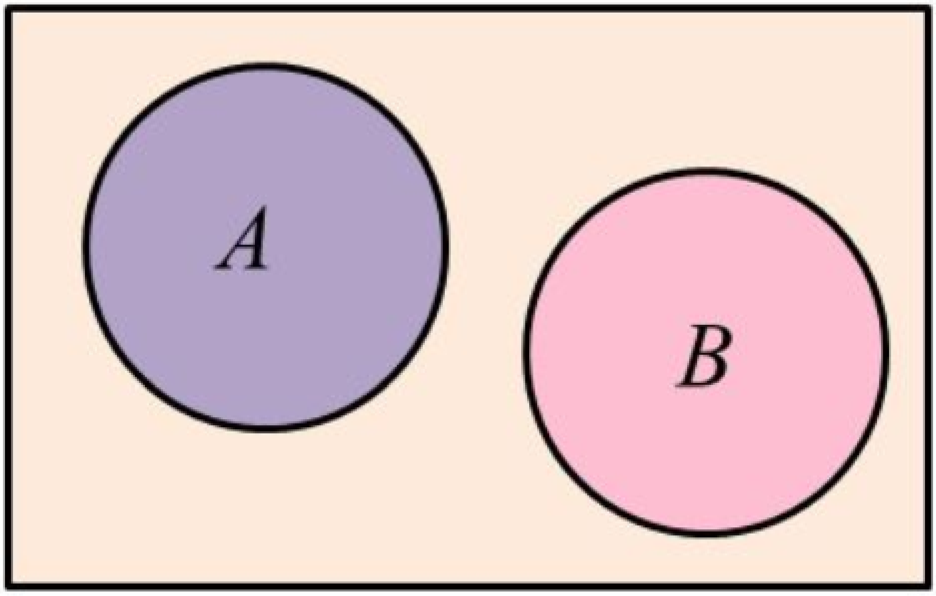
> **Q1. What will be the output of $A \cap C$?**

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)

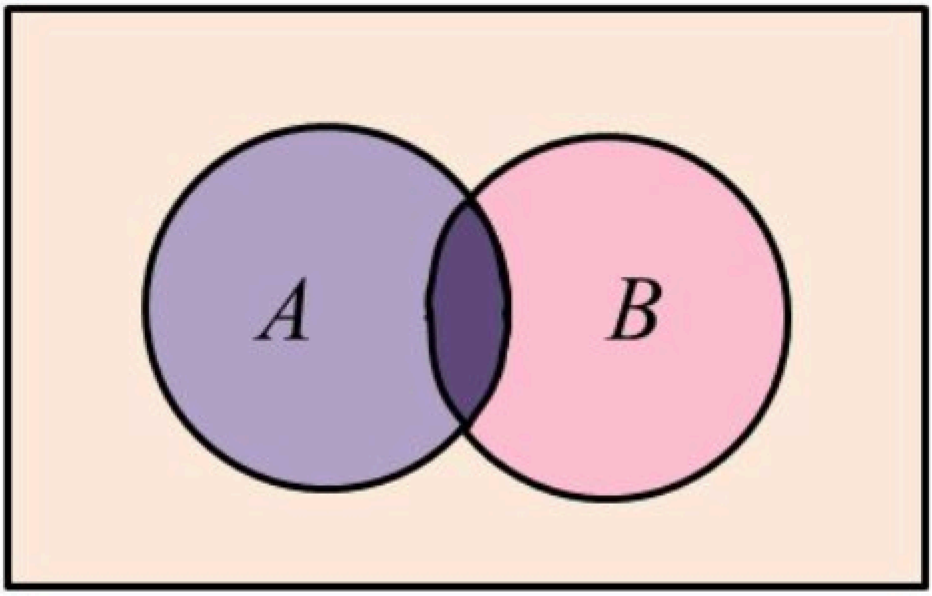
Suppose we define one more Events:

- **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

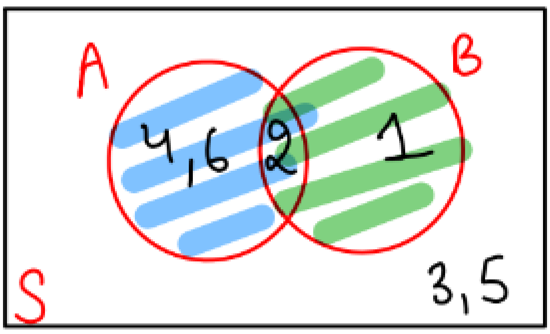
> **Q1. What will be the Probability of $P(A \cup B)$?**

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]:
```
0.12777777777777777
```

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]:

Won / century	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