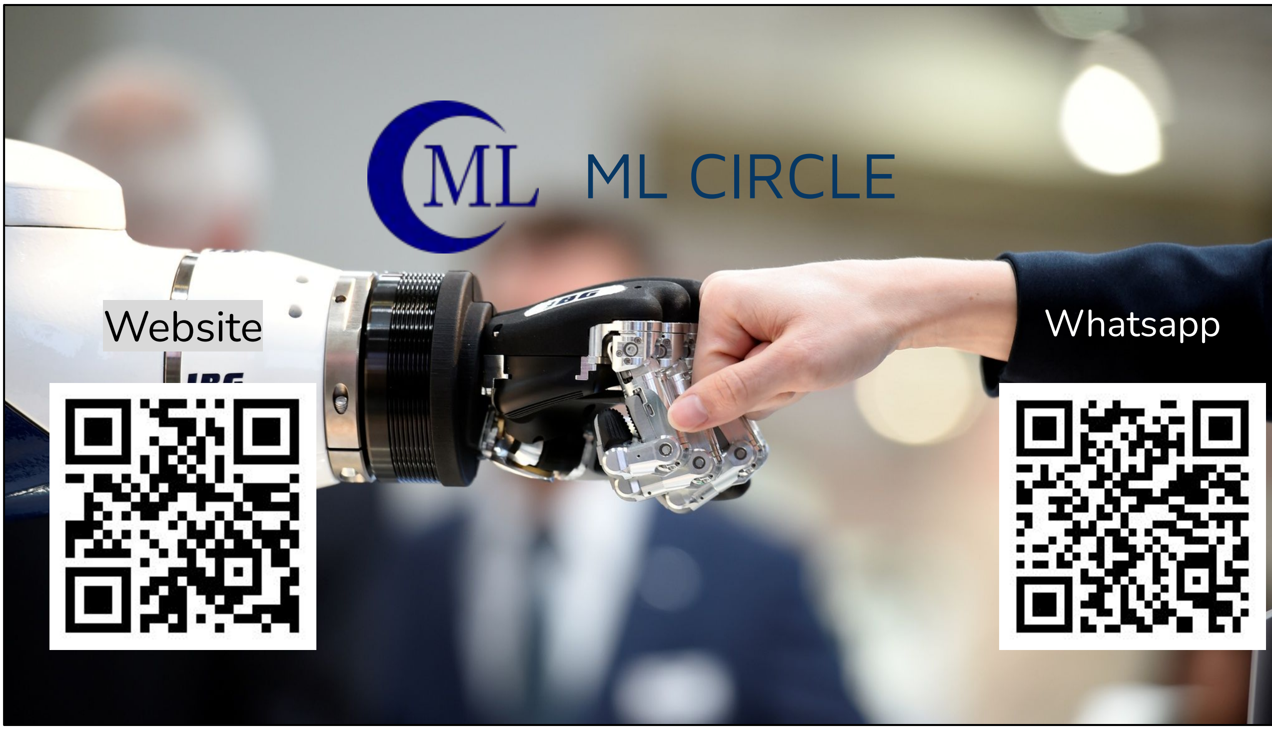

ML CIRCLE
Website
Whatsapp

Hi Everyone! As we start the year 2021 and for some of you, your journey of next 4 years, what better way than to go back to basics?!

Welcome to our first discussion session for probability theory series. Before we begin, I would like to tell you what are my expectation from you so that your expectation for these sessions may be fulfilled. As this is a discussion session and not just a lecture, I would like everyone to participate whenever I put forward a question, or unmute yourself if you have a doubt and also preferably start the video so that I can see you when you talk.

Takeaways

- Understand probability in depth and intuitively
- Understand the concept of Independence and its significance
- Understand what random variables really are and why do we need it
- Understand deeply all the terms/ keywords you may or maynot have known
    - Expectation
    - Correlation
    - Variance, etc

Soft takeaways :

- A push to think on a concept until you deeply understand it

The content I created for this discussion series encapsulates all that I had learnt within the classes, after class discussions, and my own research on the topic from various places.

These are things which I felt would have had helped me better, if I had known them before I began my first course in the domain which according to our curriculum was data analytics.

Many things may challenge you and what you have known until now, but that's what this session is about.

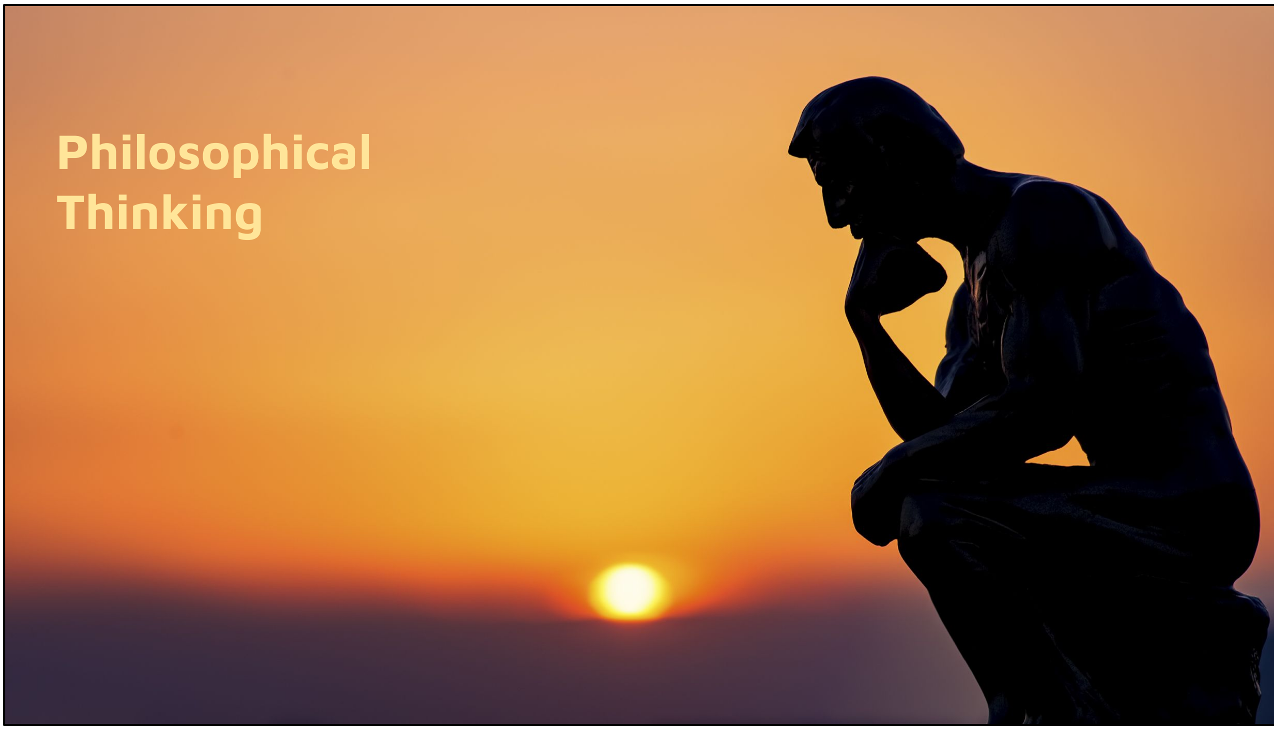
We will discuss anything if it seems wrong and once it's over I encourage you all to search for the queries you have had by yourself too as again finally we all are students.

The takeaways of the discussion series are as you see in this slide.

Probability Theory

Session 1

In our first session we will start right from basics and redefine everything we have known and maybe also our thinking process of how we think about mathematics and approach engineering problems.



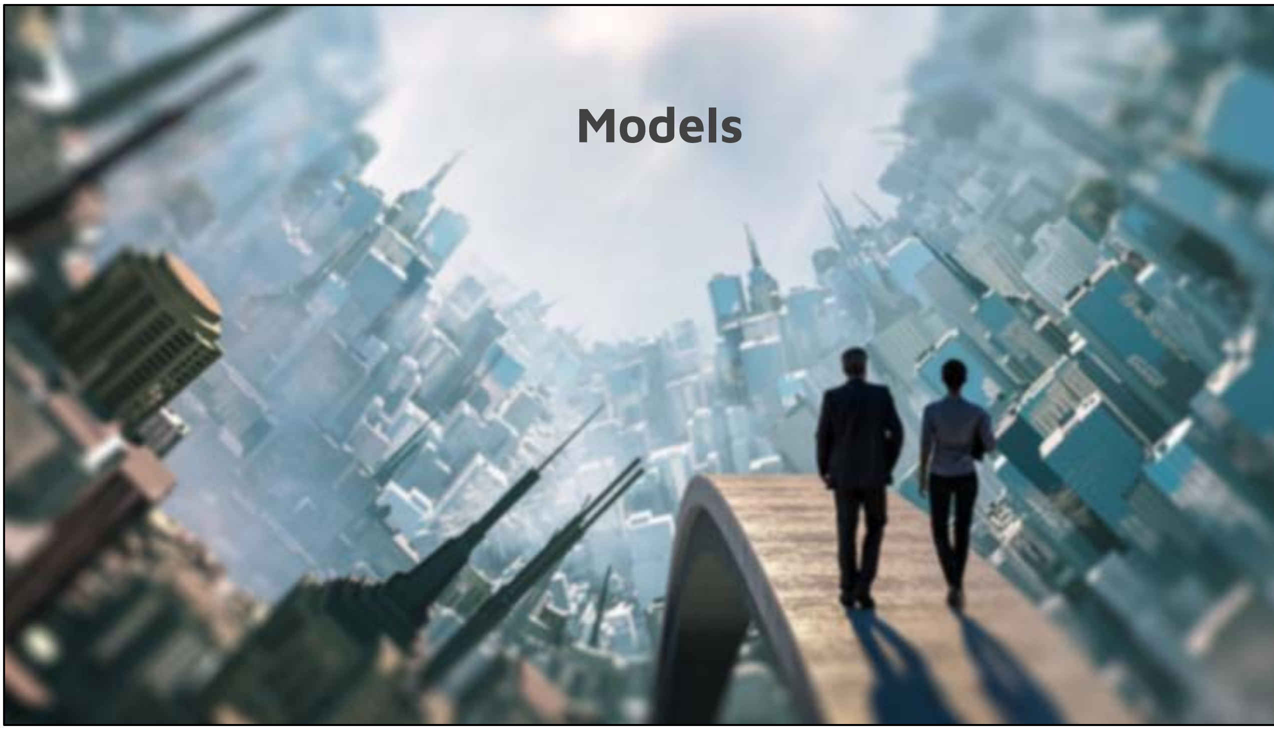
As we proceed through this discussion session, Philosophical Thinking would be the theme I would strongly focus and reflect upon.

I cannot in words put it how important this is to understand the world around you.

Even in education - to understand any subject/topic deeply you need to ask “why”. Why should I do this? Why does this work? And so on..

I am hoping as we move forward this would be another aspect in your personality that you would develop.

I would like you guys prepare your brains to go deep - instead of brushing the surface and saying “it works lets move forward”. But of course currently try to keep it on topic, because its very easy to wander off when thinking such things.



So we are going to prepare this huge structure of machine learning during our bachelors. But as every giant structure, it requires solid foundations and the Theory of Probability makeup this foundations.
We begin our journey to find the blocks to build the theory of probability and we stumble upon our first block.

As a human being residing on planet earth, knowingly or unknowingly you create and refine models everyday, and live your life around the models you create.

So what is a Model? Model is any structure that we create to help capture our reality.

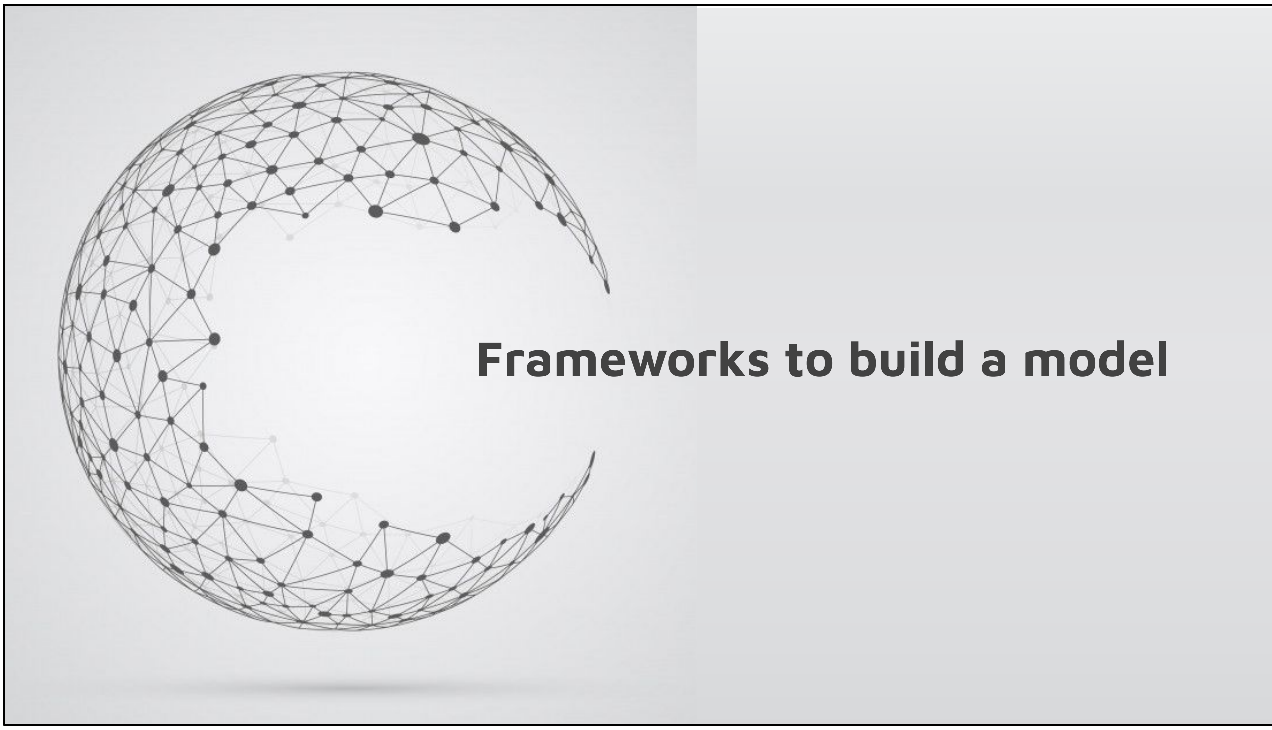
These structures can be hard (something we can touch or feel) e.g. toy model of an airplane, maybe a baking soda volcano,

or soft (something we cannot touch or feel but abstract) e.g. Each one of us have created a model of the personalities of our friends in our mind, we know what they like, we know how they would react to situations, etc. another example would be mathematical equations of motions which captures where an object would be given its current state.

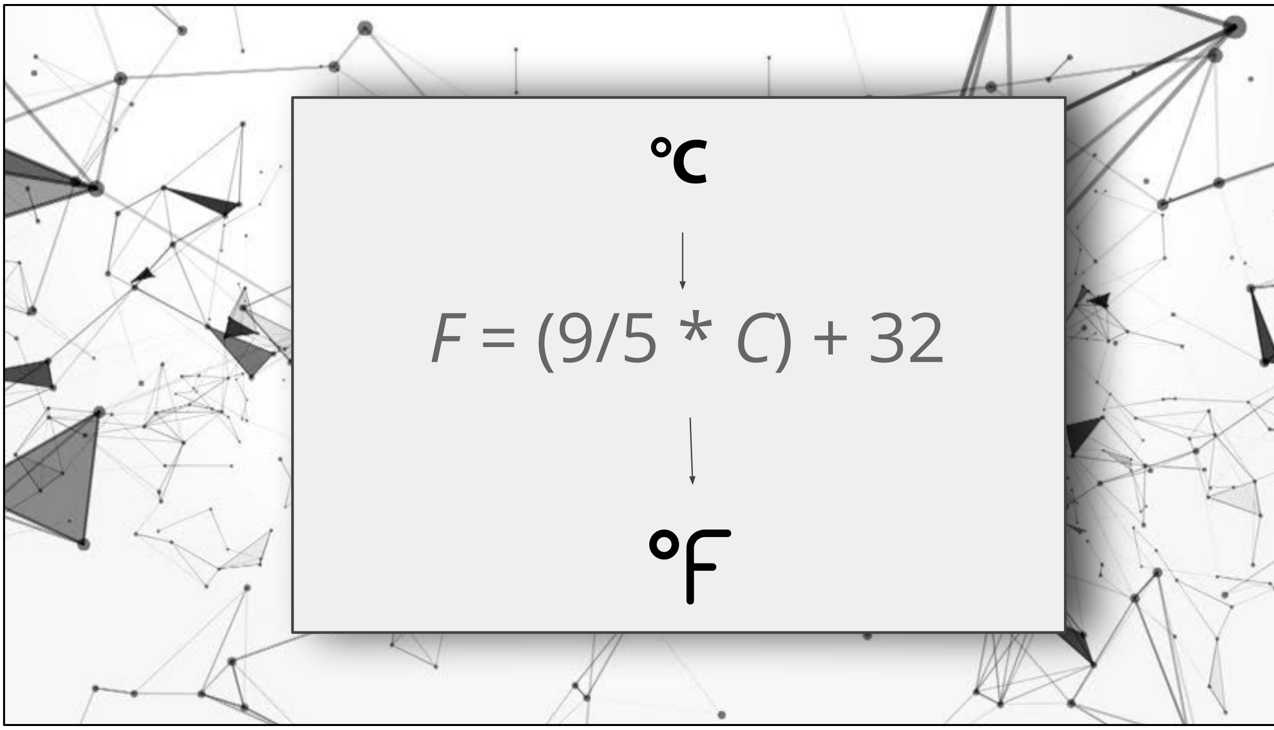
Can anyone else give me an example of a structure which we create to capture our reality?

There are many more models. If you think deeply, what all models are ultimately striving for - is to explain our reality. <pause> People wish for a time machine. But the best time machine we have is right here - our brains. Create an accurate enough model of the reality and you know everything that is coming, well in some sense.

Borrowing the words of Prof. Patrick Winston - Models are used to explain the past, predict the future, understand the subject and control the world
if someone asks us what we do - this is what we do - As data scientists we strive to create accurate models of the world around us using **DATA**.



If building correct models solves our greatest of problems, is there a textbook way to build them? Well, the answer is both no and a yes. No, We don’t have a master equation to the universe. There is no one size fits all. But yes we do have ways to build many different models for different scenarios and it is with experience and knowledge one understand which models to choose, in which situation. The years you have spent in your education you have been doing one of the following 2 things. Either you were learning frameworks to build a model (e.g. Differential Equations, Calculus, Linear Algebra, graph theory, etc) or you were studying models which people before you have already built (e.g. Breadth First Search for the Shortest Path, Using determinants to find roots of quadratic equations, and many more). It is crucial to use right framework in the right situation to give the best models. E.g. Newton represented force using product of mass and double derivatives of distance with time. This was then solved using the frameworks of differential equations which gave us our standard equations of motion. Why do we call it a good model? Because it was a 100% accurate with the real world macro observations. Hence correct frameworks used in the right situations help us make great models.



Next thing which I wanna show you guys are the two major classes, models lies in.

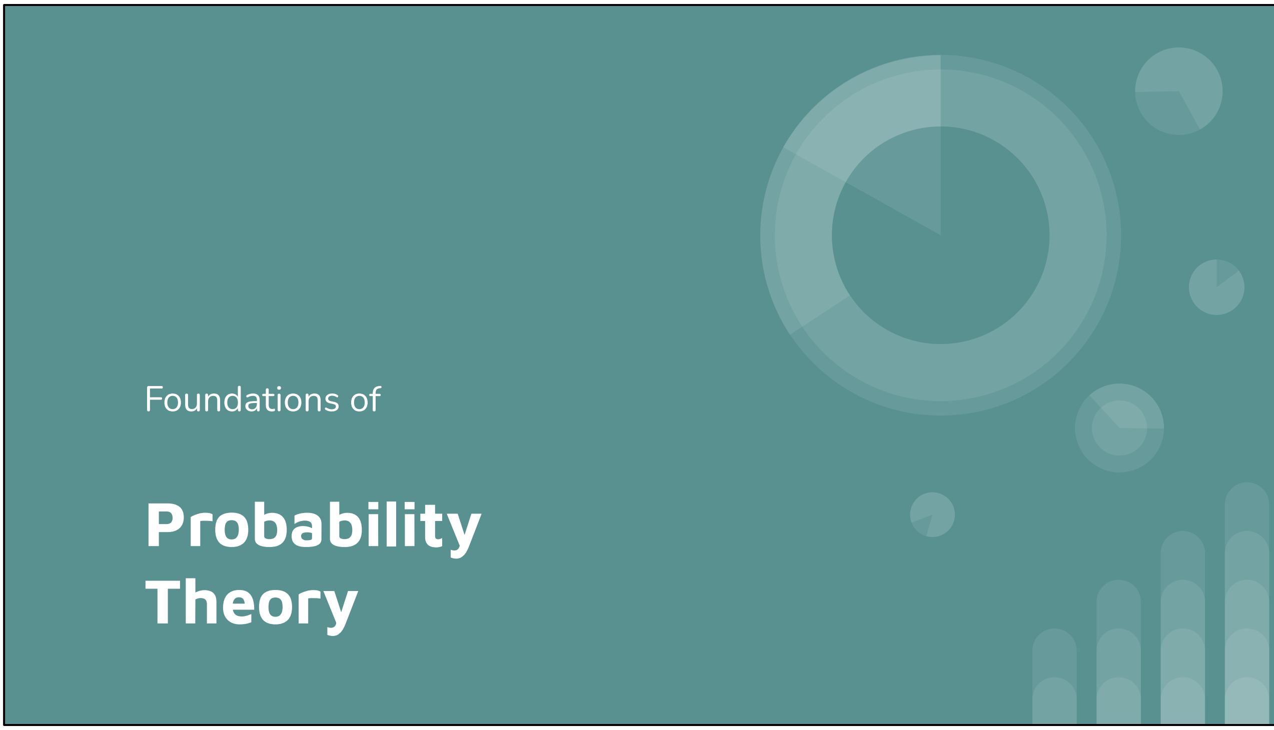
Some things we know for certain. For example, water freezes at 0 degrees Celsius and boils at 100 degrees Celsius. Some relationships, we know for certain as well. If we know the temperature, in degrees Celsius, we can convert that value to the temperature in degrees Fahrenheit using this formula.

This mathematical formula is actually a model of the relationship between two different temperature scales. It is a **deterministic model**, as the relationship between the variables is known exactly. A deterministic model is one in which there is no error in the prediction of one variable from the others. In such cases we are able to “recognize” and “envelop”, all dependencies of an observed phenomenon.

<transition>

But there are certain scenarios where we are unable to either “recognize” all the dependencies, or after having recognized them fail to effectively “envelop” them. This may be due to huge space being taken by the deterministic model or due to thousands of years of computation time to process the results (e.g. integer factorization of large primes). Or it may be even due to our inability to tackle the problem effectively. What I mean by this is - something like the background image you see in this slide, is going on inside our head. The dots or nodes are the concepts which you have learnt and the dependencies of the problem. We have all the nodes but we fail to see which nodes go with which nodes. Its lying scattered like colored specs of rice within biryani inside our head. We don’t know which nodes to connect with which nodes to get desired results. Hence in these cases computations becomes unfeasible.

In such scenarios what we go for is **Probabilistic Models**. These kinds of models don’t give accurate results all the time. They trade accuracy for speed and memory efficiency.



In this Discussion series we won't be studying the probabilistic models that others built but rather we will learn the "framework of probabilistic models" so that we get to create our own such models.

And even more, we will focus on the foundations, rather than upper level stuff, because to build a skyscraper you need equivalently strong foundations, else entire structure would collapse.

The things I am going to tell you would be quite a bit different from what you have had been learning uptill now. We are going to dwell a bit more deeply into mathematics of probability and its going to get a bit abstract and tough. To get the best out of this session, arm your brains to think deeply and out of the traditional box and keep a copy and pen handy for rough work. If you do that I promise you that you will leave this session with a never before grabbed intuition for probability and one which would last.

The classical definition (and its limitations)

$$P(A) = \frac{\text{Number of favourable outcomes}}{\text{Total number of possible outcomes}}$$

- where, all outcomes are equally likely

Because I tell you to open your mind to think about probability differently, it is fitting that I give you the motivation to do it.

This is the statement which I am giving : There are problems in the classical definition of probability. Let me show you what’s wrong with the classical definition of the probability you have been learning.

The definition goes like this : The probability of an event is the number of outcomes favorable to the event, divided by the total number of outcomes, where all outcomes are equally likely.

To those who are not familiar with this definition or the terms used in this definition, this anyways won’t matter to you people.

There are 2 problems here

1) Its restrictive - It only considers experiments with finite number of outcomes. I hope you see why.
2) The definition is actually cyclic - “equally likely” essentially means equally probable and hence we are using probability to define probability.

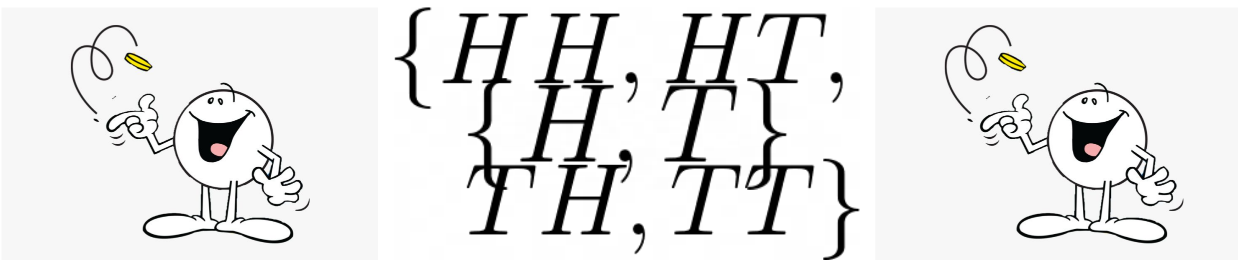
Let me wait for a couple of seconds to let that sink in. <10 seconds>

Though we do get many useful and valid results through this formula because this, in some sense, is the subset of what probability actually is. But we cannot use this itself to define the basis of a mathematical theory of probability.

So let’s see how to define probability in concrete sense.

Ingredients for constructing mathematical theory of Probability

1) Sample Space (Ω): collection of possible outcomes of a random experiment



To build the actual theory of probability we need a few ingredients. The first among them is a Sample Space.

To define sample space I want you to understand the notion of a random experiment first. Its simply an experiment whose outcome you don't know and cannot predict. This maybe due to as I said - you being unable to encapsulate all dependencies, or maybe due to huge computation time required to be able to come up with a meaningful result, etc. An Example of a random experiment is - coin toss.

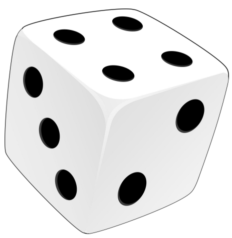
Sample space is the collection of all possible outcomes of a random experiment.

For example if a coin is tossed once, the sample space is Heads, Tails. If the coin is tossed twice what do you think the sample space would be?

To understand if you've understood this, someone give me the sample space when a die is rolled.

Ingredients for constructing mathematical theory of Probability

1) Sample Space (Ω): collection of possible outcomes of a random experiment



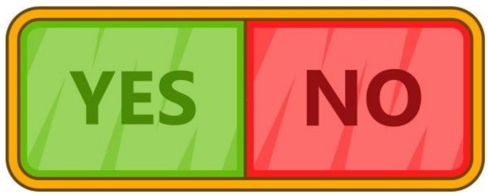
$$\{N \text{ is Even}, \\ N \text{ is Odd}\}$$

Good. Now I want you to think why did you actually say what you said. Because another possible sample space consists of two points, corresponding to the outcomes "N is even" and "N is odd," where N is the result of the dice roll. This is also a valid sample space. Thus different sample space can be associated to the same problem. The nature of the what we are trying to do, determines which sample space you are going to use. E.g. if we are interested, in whether or not Dice roll > 3 in a given performance of the experiment, the second sample space, corresponding to "Dice roll is even" and "Dice roll is odd," will not be useful to us.

In general, the only physical requirement on (the set omega i.e. the sample space) is that, a given performance of the experiment, must produce a result corresponding to exactly one of the points of Ω We have no mathematical requirements on Ω; it is simply a set of points.

Ingredients for constructing mathematical theory of Probability

2) Event : associated with an experiment corresponds to a question about the experiment, that has a yes or no answer.



Next we come to the notion of event.

An "event" associated with an experiment corresponds to a question about the experiment, that has a yes or no answer.

In simpler terms after an experiment is performed you ask a question upon the results of the experiment and if the answer is “YES” then we say the event corresponding to the question has happened.

Mathematically, the subset of the sample space associated with the answer “yes” is called event.

Ingredients for constructing mathematical theory of Probability

$$\boxed{\{HH, HT, TH, TT\}}$$

$$\#heads \leq 1$$

$$A = \{HT, TH, TT\}$$

The event A will occur in a given performance of the experiment if and only if the outcome of the experiment corresponds to one of the points of A.

For example for a coin tossed twice this would be the sample space

You are interested if the number of heads less than or equalto 1 in a given experiment or trial. Trail and an experiment are the same thing hence at times I may use them interchangeably so that you become comfortable with their usage.
<next>

Hence after occurrence of the experiment you ask is the number of heads less than one? If the answer is yes then the event has occurred.

Formally, we go over all the elements in sample space and ask the same question (why because all are possible outcomes of the experiment). We collect all the points whose answer is yes into a set and we call the set the “Event corresponding to number of heads less than or equalto one”.

<next>
Hence our event corresponding to the question number of heads less than one is this specifics subset of the sample space - HT, TH, TT

Does this make sense?

Points belonging to the event A are said to be “favorable to A”. (Its just notations which mathematicians put in to make things more fancy.)

<next>
Mathematical summary of the entire thing is that, the event A will occur in a given performance of the experiment if and only if the outcome of the experiment corresponds to one of the points of A .

Now think about it - if A and B are 2 events is their intersection also an event?

Ingredients for constructing mathematical theory of Probability

$$A = \{HT, TH, TT\}$$
$$B = \{HH, TH, HT\}$$
$$A \cap B = \{TH, HT\}$$
$$A \cup B = \{HH, TT, TH, HT\}$$

Consider the experiment of tossing 2 coins. You already know the sample space.
<next>
Let A be the event that the number of heads less than or equalto 1 in a given experiment or trial.
<next>
Let B be the event that number of Tails less than or equal to 1 in a given trail.

Is A intersection B an event? What is A intersection b?
<next>
To see if its an event check if it answers a yes or no question about the experiment. What is the question to which only the elements of this set say yes to?

A: Number of heads is less than equal to 1 and number of tails is lessthan equal to 1

What about A union B?
<next>
Is it an event ?

A: Number of heads is less than equal to 1 or number of tails is lessthan equal to 1

Okay, Let’s take a complement. Is A complement an event?

As you see unions intersection and complements of events of an experiment are also events of the same experiment. Or In other words are also events in same sample space, because the sample space defines the experiment uniquely.

Ingredients for constructing mathematical theory of Probability

3. $\mathscr{F}$ - Sigma Field of Ω : Collection of subsets of Ω that follow the following rules -

$$\Omega \in \mathcal{F}$$

$$A \in \mathcal{F} \Rightarrow A^c \in \mathcal{F}$$

$$A_1, A_2, A_3, \ldots \in \mathcal{F} \Rightarrow \bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$$

Next comes the sigma field of a set.

This is a special set created out of an existing set. As sigma field of a set is a mouthful I may use only two words “Sigma field” or maybe letter F to denote this set. But always know that this special set is created out of an existing set and is meaningless without the other. So even if I am not mentioning the main set while saying sigma field, as you would see it would be a trivial thing to find out which main set I am talking about.

As we are working with probability I will only use the sets which we have learnt about such as the sample space in this definition. But note that this is true with any general set. Just replace omega the sample space with any general set.

So, mathematically Sigma field of a set is the collection of the subsets of the set that follows the following rules. (Which also means that omega is the universal set while creation of sigma field)

Omega belongs to the sigma field of Omega

If a set belongs to sigma field then its complement belongs to sigma field.

Now what does the upper two rules imply? -<see if anyone answers>- Phi belongs to the sigma field.

Finally for sets A1,A2,A3.. Etc belonging to F, their union also belongs to F.

Now using complement and union rules, you will find intersection of sets in F also belongs to F. Hint, DeMorgan’s Law.

So sigma field is closed under unions, intersections and complements.

If you think about it, in the case of finite sets, the powerset is an example of sigma field.

And There can be many sigma field of a given set.

Ingredients for constructing mathematical theory of Probability

3. $\mathscr{F}$ - Sigma Field of Ω :

Event Space - Set of all possible events generated from a sample space.

Event Space is a sigma field.

Here is a statement which I want you guys to think and which should be obvious if you understood everything I said uptil now.

Let me define event space as set of all possible events of an experiment. E.g. for the 2 coin toss experiment, the event space is the power set of the sample space {HH, HT, TH, TT}

The statement is that Event space is a sigma field of the sample space.

Do you see why?

<pause to see reply>

Ingredients for constructing mathematical theory of Probability

4. P - Probability Measure on $\mathcal{F}$: Mapping of elements (say A) from $\mathcal{F}$ to the real space, which we denote by P(A). This mapping should follow the following :
   - $P(\Omega) = 1$
   - $P(A) \geq 0 \qquad \text{for every } A \in \mathcal{F}$

- $P(A_1 \cup A_2 \cup A_3 \cup \cdots) = P(A_1) + P(A_2) + P(A_3) + \cdots$ $\text{for all disjoint } A_1, A_2, A_3, \cdots \in \mathcal{F}$

The last ingredient is the probability measure on a Sigma field of the sample space. Its a function which maps from elements of the sigma field to Real number. In other words its a function that takes in a event and spits out a real number such that, -

P(Sample Space) is 1

For all sets A in F; P(A) >= 0

Finally function applied on union of disjoint set has the same output as that of the sum of function applied on individual sets.

Definition

A probability Space is a triple (Ω,$\mathscr{F}$,P),

where

- Ω is a set,
- $\mathscr{F}$ is a sigma field of subsets of Ω, and
- P is probability measure of $\mathscr{F}$ i.e. A function from set of events to the real space [0,1] following the constraints given in previous slide.

Equipped with these simple definitions, here is how we define probability space.

Probability space is a triple of - the sample space Omega, F the sigma field of Omega (which is the set of all possible events that we are interested in) and P a function from F to [0,1].

With this definition I am also encouraging you to think of probability as a man made tool rather than a fundamental force of nature. This is a very powerful statement. Because if this is true, then in technical terms, you don't discover probability of events but rather assign probabilities to events.

<next>
For Example in a coin toss it is not the case that inherently coin toss is random and probability of heads is 50% but rather what we are saying is that coin toss is not random, but to correctly know which side is going to land face up there are just too many variables to handle. Hence I won't tell you which side is going to land but only tell you that on next few tries I expect nearly 50% of them are heads and 50% are tails. So I'll assign the probability of 0.5 to the event of heads coming up.

This assignment is in my hands. I may as well have had assigned the probability of 0.7 to heads but I won't because that will not be helpful and does not capture what I think its going to come.
<next>
<next>

Randomness is not a Fundamental Force of Nature.

- Upto debate

Let me bring you back to the statement I made in the last slide.

I stated that probability is not fundamental in nature.

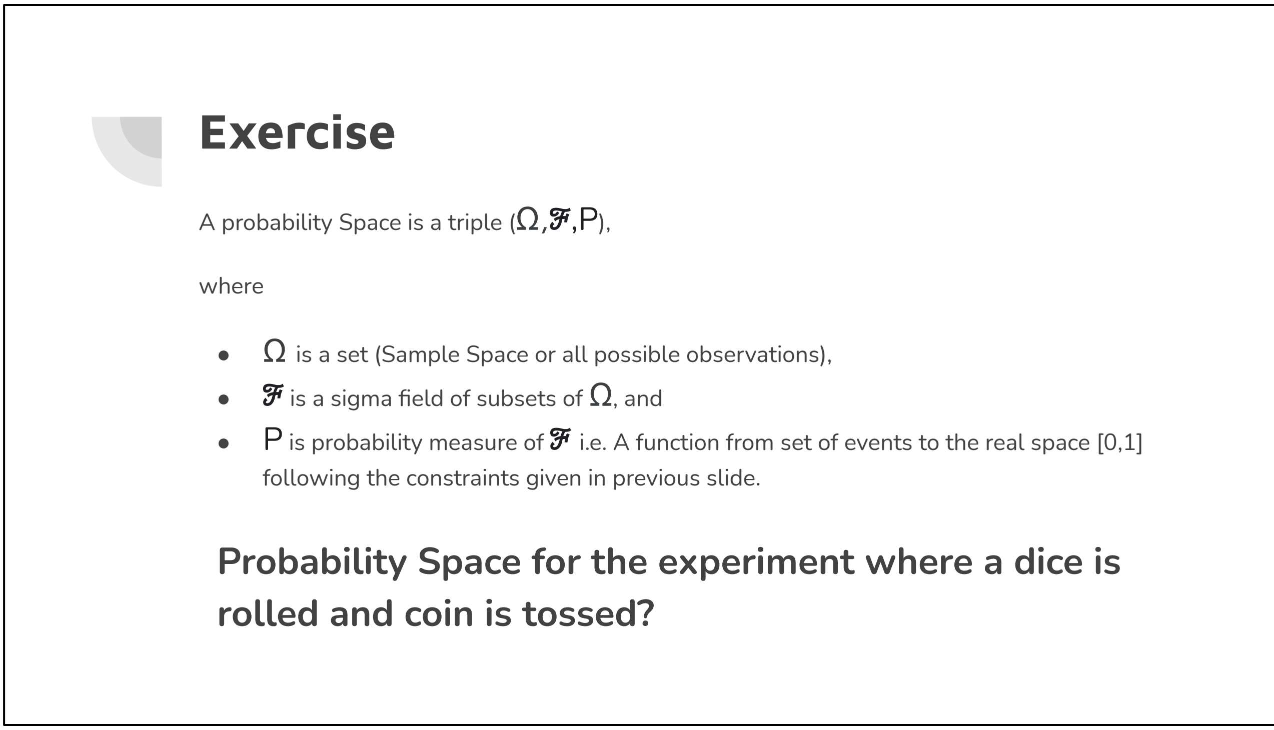
Had randomness been a fundamental force of nature, probability would also be fundamental in nature. Likewise if probability is fundamental in nature, then randomness is fundamental force of nature. Both go hand in hand.

So in turn I am implying that randomness is not a fundamental force of nature but rather our incapacity to understand or effectively model the universe.

<next>
So, before you'all go all Quantum on me,
<next>
allow me to say this. When you look at the origin of Quantum Theory, the double slit electron experiment set it all off. No one expected the same fringe pattern with electrons that was initially observed with light wave. Now as for the detector to illuminate, electrons had to have had hit them. In other words the fringe pattern showed that electrons though they hit the detector, followed wave nature in their existence itself. So the best way to model that observation was to create matter probability waves. These waves model the existence of the electron in a given space. Are they fundamental? Or is it due to some phenomenon which is still yet discovered?

<next>

No one knows. There are arguments from both sides. But majority of physicists lean towards matter probability wave being an inherent property. Regardless, that does not concern our current discussion. For all macro level events, randomness in not fundamental. Hence our definition of probability holds.



Before we proceed, to get a good grip, let's do an exercise. Give the Probability Space for the experiment where you do two things, first you roll a dice and then toss a coin. You can write your answers in notepad and share the screen or any other method you like to share. I'll give you people 2 minutes.

Exercise

A probability Space is a triple (Ω,$\mathscr{F}$,P),

where

- Ω is a set (Sample Space or all possible observations),
- $\mathscr{F}$ is a sigma field of subsets of Ω, and
- P is probability measure of $\mathscr{F}$ i.e. A function from set of events to the real space [0,1] following the constraints given in previous slide.

Probability Space for the experiment where a coin is tossed on pluto? (Roll the dice too because why not)

Okay so now, you go to Pluto and toss a coin and roll the dice. What's the probability space?

Would it be the same or would it be different?

Why did you say it's the same?

Yes you assumed that the gravitational force is uniform and the direction is same as that across all points in the planet. Is this a safe assumption? In some sense yes. You may have some model of pluto right now in your mind which has been shaped from your childhood in such a manner that you think that gravity in pluto behaves same as that on earth. This may be due to science magazines, maybe discovery channel, etc.

I call this a pre-information bias. The first step to mitigate this bias comes with the acknowledgement of the existence of such a thing. And in general this is how science works. You make hypothesis and assumptions. Do experiments based on those assumptions and test out your hypothesis. Then alter your hypothesis and statements accordingly. You may very well give the same assignments to coin toss as that on earth, but correct thing would be to go to pluto and see if your assignments and assumptions were actually correct.

Reflection

Borrowing certain remarks from Robert B Ash, the main problem in the definitions that you have learnt till now is that their approach is somehow try to mathematically prove that probability of heads in tossing a coin is ½. Or probability of getting a 3 on a roll of die is ⅙. But this cannot be done. All we can say is that if a die is rolled and the process is repeated, the ratio of occurrence of face 3 to the total number of times the die is rolled would be ⅙ in accord to our experience. Hence we ASSIGN a probability of ⅙ to the event that a face 3 comes. And the only reason we can do this is because the consequences agree with our experience. If tomorrow some mysterious phenomenon causes 3 to appear half of the times the die is rolled we will simply change this assignment to ½ without disturbing anything else.

We can never use mathematics to prove a specific physical fact. E.g. we cannot prove the existence of force mathematically. What we can do is postulate existence of such entity that satisfies certain differential equation. We then build our results around those equations until some other mathematical theory is built that better explains the observation of what we call “force”. Same is in case of probability theory.

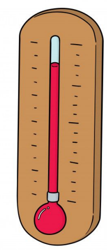
In short, intuition and experience lead us to an assignment of probabilities to events. These assignment change as physical observations change.

This concept that you have learnt forms the frameworks of the theory of probability. And it is yet another tool to help us understand our world better.

Independence

Total number of teeth in a room = 64

A



Temperature outside > 20

B

One of the least intuitively understood concept in probability is the concept of independence of events. Now as we have a good understanding of the base of probability let's try to understand independence.

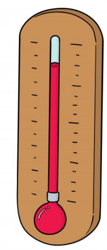
Consider an experiment where we enter a hostel room <next> and count the total number of teeth including both the roommates. Then we go out <next> and record the temperature. We are interested in both the roommates having all 32 teeth i.e. total teeth 64 <next> which we call as event A for convenience, and the temperature outside is above 20C <next> which we call as event B for convenience.

Now as you can see intuitively knowledge about occurrence or the non-occurrence of one should not disturb what we know about the odds of occurance of the other.

Independence

Total number of teeth in a room = 64

P(A) = 0.3



Temperature outside > 20

P(B) = 0.4

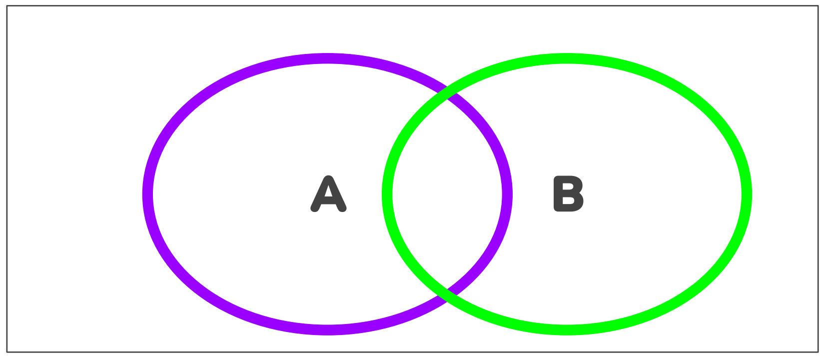
Let us assign some probabilities to these events.
<next>
Let P(A) be 0.3 and

<next>
P(B) be 0.4

What we are saying is irrespective of whether A occurs or not, the probability of occurrence of B is 0.4

Let’s look at this as a venn diagram.

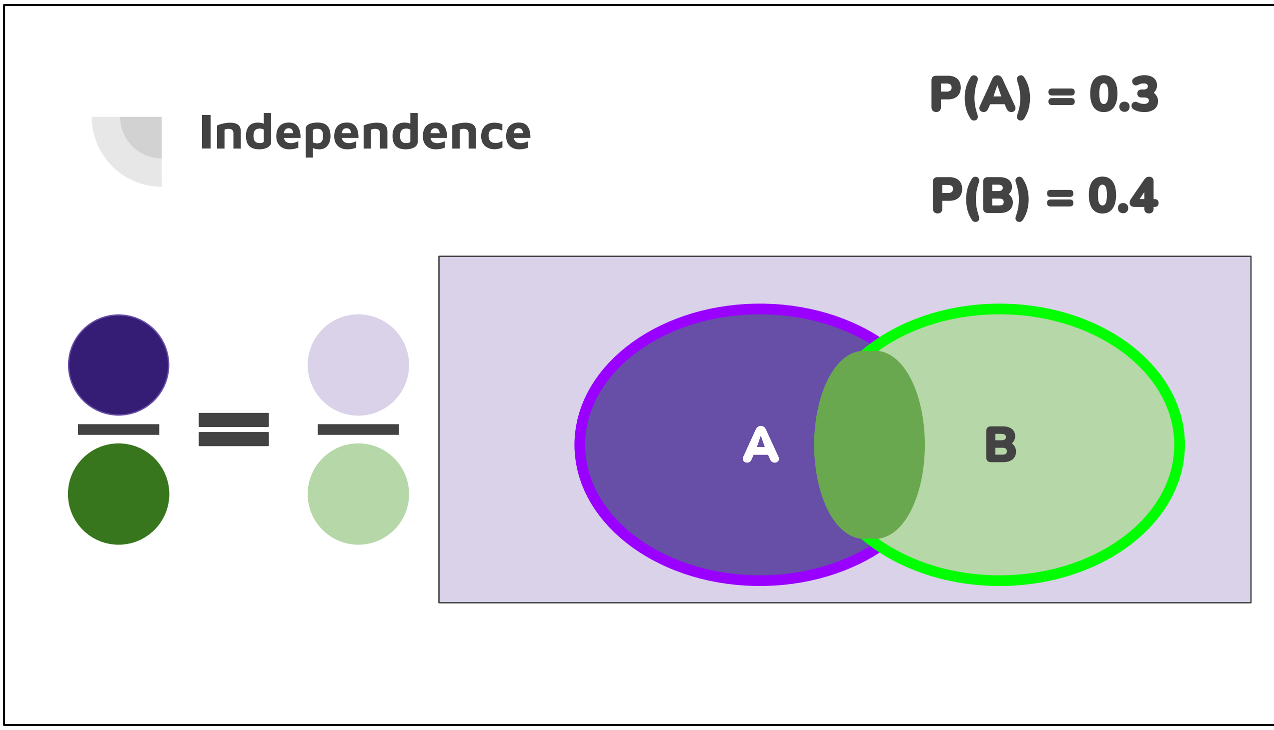
Independence



So the square denotes all possible outcomes of the experiment. In other words square is a graphical substitute for the sample space.

Understand that the outcomes for this experiment are of the form of tuples e.g. (58,21) (64, 19) etc

The purple circle are those elements which are in event A and green circle are those elements which are in event B.



I’m coloring important areas.

As you can figure out events corresponding to

Dark Purple is - A happens but B does not happen

Dark Green is - A happens and B happens

Light green is - B happens but A does not happen

Light purple is - both A and B does not happen.

<next>

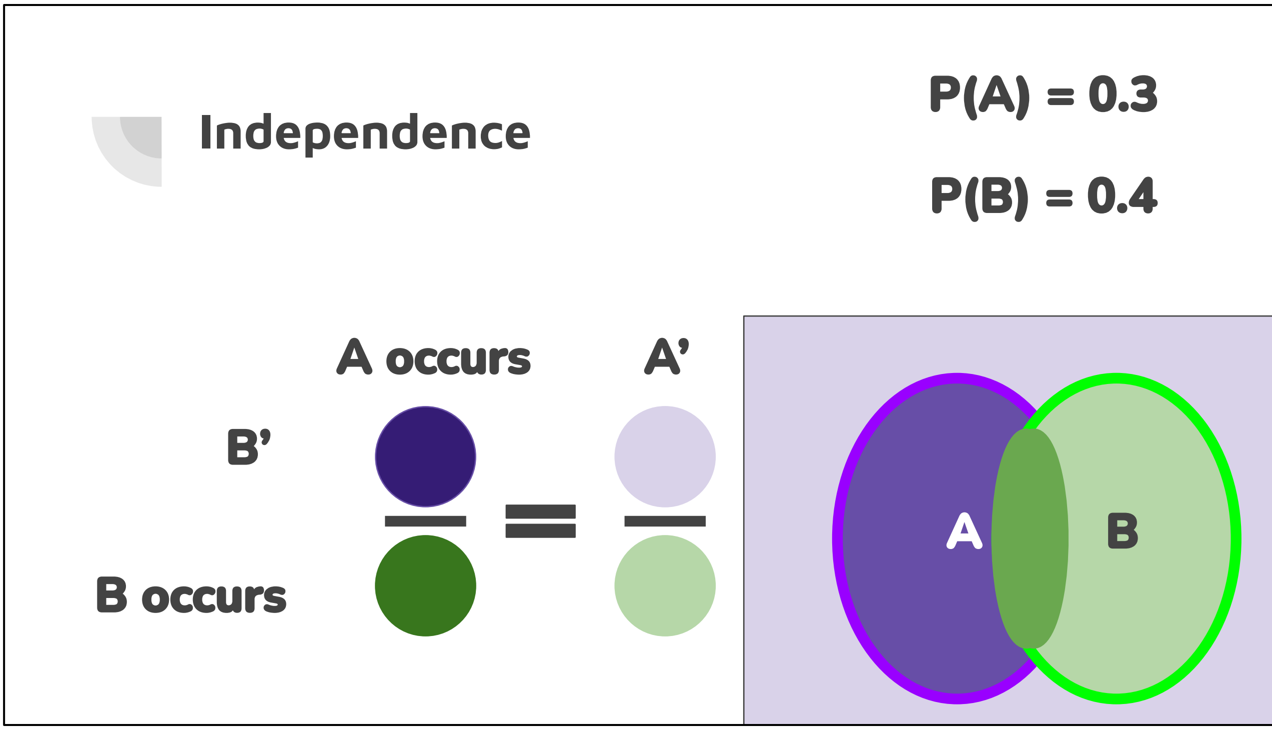
Now we are saying that there is a 30% chance that A occurs. And if that happens there is a 40% chance that B will occur and a 60% chance that B will not occur.

Also there is a 70% chance that A does not occur. And if that happens even then there is a 40% chance that B will occur and 60% chance B will not occur.

Intuitively it means the following in terms of probabilities of shaded area.

<next>
<wait>

<next>



Let me add some text to make this more clear.

<next>

“A” prime is A complement or the event that A does not happen. Same goes for B.

So when we are saying that irrespective of whether event A happens or not the probability of happening of event B is 0.4, we mean that the probabilities assigned to the shaded sets should follow this ratio. Do you see this?

In order to check they are independent we need to see if this ratio is satisfied. Now where are we going to get the probabilities of these shaded sets from?

If you recall we had the function P which took in an input from the sigma field and returned the value which we assigned to those sets while we were creating the probability space.

After checking if this ratio is followed then indeed the events are independent.

Ingredients for constructing mathematical theory of Probability

4. P - Probability Measure on $\mathcal{F}$: Mapping of elements (say A) from $\mathcal{F}$ to the real space, which we denote by P(A). This mapping should follow the following :
   - $P(\Omega) = 1$
   - $P(A) \geq 0 \qquad \text{for every } A \in \mathcal{F}$

- $P(A_1 \cup A_2 \cup A_3 \cup \cdots) = P(A_1) + P(A_2) + P(A_3) + \cdots$
  $\text{for all disjoint } A_1, A_2, A_3, \cdots \in \mathcal{F}$

But wait. We can simplify that ratio even further. Recall from previous slides, that we were not free to assign any arbitrary values, to the sets in sigma algebra. We had to follow these rules while assigning, for it to be called a probability measure.

So if you take these constraints into consideration and work out the math. In order to Follow the ratio, the assignments of probability to sets A and B should follow the famously known equation

Independence

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

P assigned to (A intersection B) is the same as P assigned to (A) times P assigned to (B)

And the math from previous to this is very trivial so I am skipping.

<Doubts?>

Closing Remarks

By now I hope so that you people have gained a very good intuition of probability. Before we close there is one last thing I want to do. I said a line in the start of the presentation wherein I said the classical definition is mere subset of what probability actually is. Can anyone now tell me how?

I hope this session was useful to you all and hopefully you have understood most of the things discussed, and even if not at least gained a good intuition to help in your upcoming courses. Of course the amount of content I covered just barely scratches the surface and as you go deep into the topic you encounter much more difficult topics such as measure theory and other abstract mathematics. If you are interested I encourage you to go through Robert B Ash - Basic Probability theory from which most of the content of the slides have been taken from. Though you will encounter this again in your 7th Semester in the course of Probability theory and would go much deeper into all these stuff and basic proofs.

Also I currently have not made the slides of the next part which would probably be random variables and will do only if you people have actually gained something from this discussion session and think that next part would indeed be useful.

Also we have brilliant professors whom you can ask any doubts you encounter on this journey. Dr. Clint P. George for application of probability in AI. And Dr. Kaushik Majumder for doubts related to theoretical side of probability. But of course they will expect that you have done your research and are coming to them after putting in efforts from your side.