Introduction To Probability (MIT OCW)

★ Sample Space

★ Discrete / Finite sample space :-

→ Imagine that 2 die with 4 faces each are thrown. Then, you would assume that the sample space consists of 16, but that depends on you i.e. if the sequence is important.

→ ~~If only~~ Consider, for example, that only the sum or multiplication of the number on die are important. Then sequence would be irrelevant & sample space size would reduce to 10.

→ This was an example of discrete / finite sample space

★ Continuous sample space :-

→ Consider an example that you have a square dart board such that (x, y) ~~the~~ are co-ordinates of thrown dart where $0 \leq x, y \leq 1$. The co-ordinates (x, y) are measured with infinite precision. Hence, it can't be a finite sample space.

→ You can't directly have probability of any individual point. So, you need probabilities of events, where these events are subsets of sample space. Example:- You can't measure probability of the dart hitting the center if you have infinite precision. But, you could perhaps measure the probability of the dart landing in the upper half of the board.

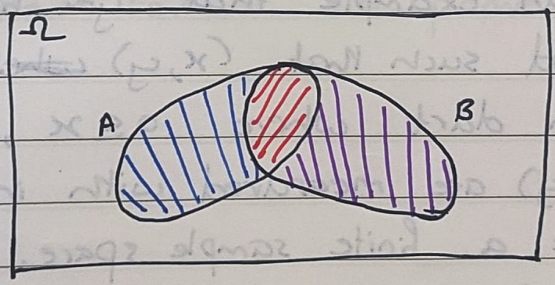
★ Axioms for a probability model:-

- Non-negativity : $\Rightarrow P(A) \geq 0$
- Normalization: $P(\Omega) = 1$ .... (where, Ω = sample space)
- Finite additivity:
  If $A \cap B = \phi$, then $P(A \cup B) = P(A) + P(B)$

★Note:- Consider a set 'A' & its complement ("A^c"). You can call A^c as 'not A'

$A \cup A^c = \Omega$

So, if you have two not-disjoints sets A & B they can be represented as:



The portion shade in blue:- $A \cap B^c$
The portion shaded in red:- $A \cap B$
The portion shaded in purple:- $A^c \cap B$

$\therefore P(A \cup B) = P(A \cap B^c) + P(A \cap B) + P(A^c \cap B)$
.... (by finite additivity)

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

- Union Bound: $P(A \cup B) \leq P(A) + P(B)$

★ Discrete Uniform Law

Consider Ω with n equally likely outcomes.
Consider A, where $A \subset \Omega$ which has k equally likely outcomes.

$\therefore P(A) = k \cdot \frac{1}{n}$

★ Probability of continuous sample space.

- Consider the previous example of dart thrown at a square dart board & precision of measuring the coordinates in infinity.

- Choice of probability law is arbitary. It's up to us how to model a certain situation.

- For the above example, we can assume an uiform probability law, where probability = area of board.

- In above dartboard, co-ordinates are:
  (x, y) such that $0 \leq x, y \leq 1$.
  - ★ So, you can calculate probability of an event which covers a certain area. Example:- $P(\{(x,y) \mid x+y \leq 1/2\}) = 1/8$
  - ★ But, probability of an event which consists of a single point i.e. doesn't cover any area is zero. Ex:- $P(\{(0.5, 0.3)\}) = 0$

★ Validity of probability law

- You must ensure that the given law satisfies all the axioms. You can use modified results to do this.

- For example, consider a sample space of Natural numbers. (this is an infinite discrete sample space).
  You have been given a law (non-uniform) that $P(n) = \frac{1}{2^n}$, $n = 1, 2, 3 \ldots\ldots \infty$ & you want to check its validity.

- It must satisfy the axiom: $P(\Omega) = 1$

  i.e. $\sum_{n=1}^{\infty} \frac{1}{2^n} = 1$

  $$L.H.S. = \sum_{n=1}^{\infty} \frac{1}{2^n} = \frac{1}{2} \sum_{n=1}^{\infty} \frac{1}{2^{n-1}} = \frac{1}{2} \sum_{n=0}^{\infty} \frac{1}{2^n}$$

  $$= \frac{1}{1 - (1/2)} \quad \ldots\ldots \text{(Infinite sum of a G.P.)}$$

  $\therefore L.H.S. = 1 = R.H.S.$

  Hence, the law holds.

★ Countable additivity axiom:-

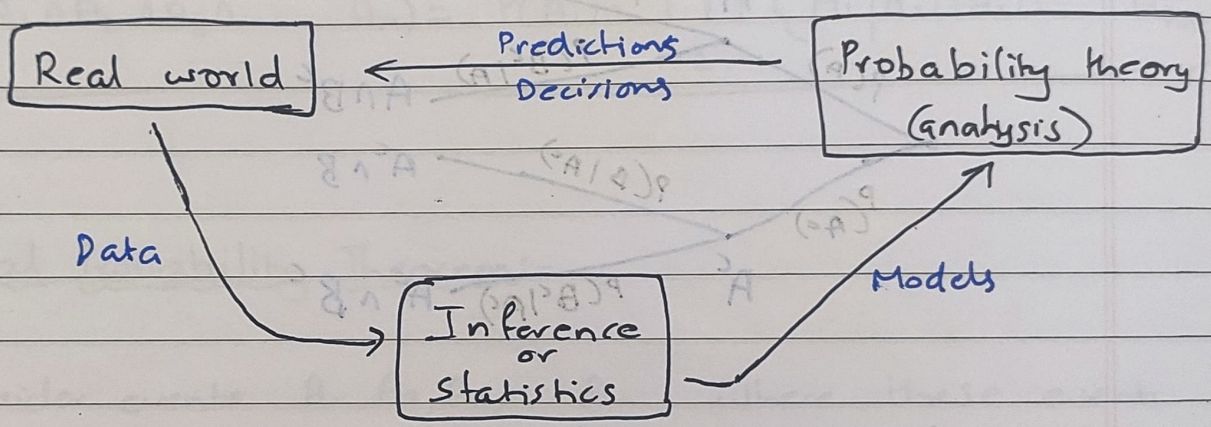
- If $A_1, A_2, A_3 \ldots\ldots$ is an infinite sequence of disjoint events, then $P(A_1 \cup A_2 \cup A_3 \cup \ldots\ldots) = P(A_1) + P(A_2) + P(A_3) + \ldots\ldots$

- There are 2 types of infinite sets; discrete & continuous.

- Additivity only holds for "countable" sequence of events.

- The unit square (similarly, the real line) is not countable (its elements cannot be arranged in a sequence)

- However, 'Area' is a legitimate probability law on the unit square as long as we deal with "nice" subsets of the unit square.

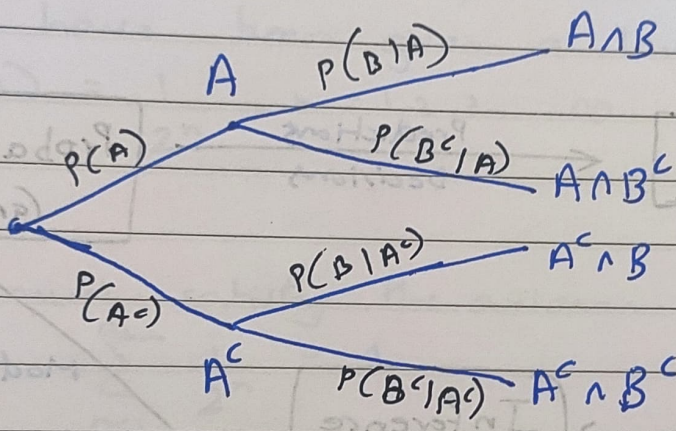


★ Conditional Probability :-

- For two events A & B, the probability of event A happening given that event B has already happened is given by:

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

★NOTE :- It becomes really helpfull to visualize the events & probabilities by drawing a horizontal tree diagram.



As further you add branches to the tree, they become conditional probabilities, because current branch / leaf can occur only if previous branch has occurred.

You can get the probabilities of leaf nodes by multiplying the probabilities of the branches that led to this node.

Ex:- $P(A^c \cap B) = P(B|A^c) \cdot P(A^c)$ ..... (You can prove this)

..... (multiplication rule)

★ Multiplication Rule:-

- $P(A \cap B) = P(B) \cdot P(A|B) = P(A) \cdot P(B|A)$
- $P(A \cap B \cap C) = P(A) \cdot P(B|A) \cdot P(C|A \cap B)$
- We can generalize this rule as follows:-

$$P(A_1 \cap A_2 \cap A_3 \cap \ldots\ldots A_n) = P(A_1) \cdot \prod_{i=2}^{n} P(A_i | A_1 \cap A_2 \cap \ldots\ldots A_{i-1})$$

★ Total Probability Theorem:-

- Consider events $A_1, A_2, A_3, \ldots A_n$ where these events are mutually exclusive & exhaustive.

  i.e. $A_1 \cap A_2 \cap A_3 \cap \ldots\ldots A_n = \phi$

  &

  $A_1 \cup A_2 \cup A_3 \cup \ldots\ldots A_n = \Omega$ (sample space)

- Then total probability for any random event 'B' can be given by:-

$$P(B) = \sum_{i=1}^{n} P(A_i) \cdot P(B|A_i)$$

★ Baye's Rule:-

- Baye's rule is generally associated with inference statistics.
- Consider above example of n sets of event 'A' which are mutually exclusive & exhaustive. These events can be called as 'initial beliefs' that may happen.
- Now, if we consider that a random event 'B' has occured, these 'initial beliefs' should be revised according to changed sample space. Now, ~~can~~ we can calculate the probabilities of these initial beliefs given that an random event has already occured.

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

Now, Using total probability theorem, & multiplication rule,

$$\boxed{P(A_i|B) = \frac{P(A_i) \cdot P(B|A_i)}{\sum_{j=1}^{n} P(A_j) P(B|A_j)}}$$

★ Independence of events:-

- Intuitive definition:- when occurence of 'A' provides no new information about 'B' i.e. ~~else~~ if $P(B|A) = P(B)$

But you can't use this mathematically for $P(A) = 0$, because then, you would have to divide be zero to check if the equality holds.

- **Mathematical definition:-** Events 'A' & 'B' are independent iff $\boxed{P(A \cap B) = P(A) \cdot P(B)}$

This definition also holds for $P(A) = 0$ or $P(B) = 0$ or both 0. This is derived from the intuitive definition.

***NOTE:-** Independence is completely different from disjointness of two events.

Ex:- Take 2 disjoint events A & B. Then considered that A has occured. This tells us that B hasn't occurred. Thus, intuitively & mathematically, A & B are NOT independent.

- Now, intuitively, if A & B are independent, then their complements must also be independent. Mathematically it can be proven that,

> if A & B are independent, then A & B^c are independent & A^c & B are independent & A^c & B^c are independent.

★ Conditional Independence:-

- Conditional independence, given 'C' (i.e. C has occurred), is defined as independence under the probability law $P(\cdot|C)$

- > If event 'C' has occurred, then A & B are independent iff $P(A \cap B | C) = P(A|C) \cdot P(B|C)$

★ NOTE :- Independence of two events does NOT imply their conditional independence w.r.t. some other event.

i.e. Conditioning may affect independence

★ Independence of Collection of events :-

- Intuitive definition :- Information on some collection of events does not change probabilities related to remaining events.

- Mathematical definition :-

Events $A_1, A_2, A_3, \ldots A_n$ events are independent if:
$P(A_i \cap A_j \cap \ldots A_m) = P(A_i)P(A_j)\ldots P(A_m)$ for any distinct $i, j, \ldots m$.

Here, any collection of events chosen from given events must satisfy above condition to be independent.
For example :-
$P(A_1 \cap A_2 \cap A_3)$ must equal $P(A_1) \cdot P(A_2) \cdot P(A_3)$
$P(A_5 \cap A_3)$ must equal $P(A_5) \cdot P(A_3)$
$P(A_i \cap A_j)$ must equal $P(A_i) \cdot P(A_j)$ for any i, j from n.

★ Pairwise Independence:-

If for 3 events A, B & C,
$P(A \cap B) = P(A) \cdot P(B)$
& $P(A \cap C) = P(A) \cdot P(C)$
& $P(B \cap C) = P(B) \cdot P(C)$,
Then events A, B, C are pairwise independent.

★ NOTE:- Pairwise independence doesn't imply collective independence.

★ Pairwise independence vs. Independence of collection:-

- You can observe that pairwise independence $\not\Rightarrow$ collective independence using a simple example:-
- Consider 2 independent fair tosses.
  Event A: First toss is H
  Event B: ~~First toss is~~ Second toss is H
  Event C: Both tosses have same result

You can cross check that A, B, C are pairwise independant. But, $P(A \cap B \cap C) \neq P(A) \cdot P(B) \cdot P(C)$. Thus, they are not collectively independent.

You can understand it intuitively by:
$P(C|A \cap B) = 1 \neq P(C) = 1/2$
i.e. The fact that A & B happened affected probability of C happening.

★ Partitions :-

- Imagine there are $n \geq 1$ distinct items & $r \geq 1$ persons. And we want to give n_i items to i^{th} person.
  Now, $\sum_{i=1}^{r} n_i = n$

- So, the no. of ways in which can partition is $= \frac{n!}{n_1!\, n_2!\, n_3! \ldots n_r!}$

- This is also called the multinomial coefficient.

- One way to imagine this would be this example: Imagine that there are 9 people who want to play a game. In this game, there are 3 teams of 3 people each. Thus, $n = 9$, $r = 3$ (3 teams)
  Then, the no. of unique matches that can be played i.e. ~~no. of combinations set of~~ the no. of unique 3-team combinations (the teams are also obviously unique) is given by the partions formula
  $\therefore$ No. of unique matches $= \frac{9!}{3!\,3!\,3!} = 1680$

- Another example is: There are 4 players & you want to play doubles i.e. there are 2 teams. Now you want to count unique matches.
  That is given by: $\frac{4!}{2!\,2!} = 3$ (which you already know, because you did this 1rd)

• Another way to imagine it is this:
Imagine there are n people & we want to split them into 3 teams of n_1, n_2, n_3 size respectively. Thus, r=3.

Intuitively, we can say that for the 1st team we can choose the people in ${}^{n}C_{n_1}$ ways.
For the 2nd team, we have only $n-n_1$ people left to choose from. And we can do it in ${}^{n-n_1}C_{n_2}$ ways.
Similarly, for 3rd team, only $n-n_1-n_2$ people are available. We can choose them in ${}^{n-n_1-n_2}C_{n_3}$ ways.

If we take their product, we will get the no. of ways n people can be split into 3 groups of n_1, n_2, n_3 size.

$$\therefore \text{No. of ways} = {}^{n}C_{n_1} \cdot {}^{n-n_1}C_{n_2} \cdot {}^{n-n_1-n_2}C_{n_3}$$

$$= \frac{n!}{(n-n_1)!\, n_1!} \cdot \frac{(n-n_1)!}{(n-n_1-n_2)!\, n_2!} \cdot \frac{(n-n_1-n_2)!}{(n-n_1-n_2-n_3)!\, n_3!}$$

$$= \frac{n!}{(n-n_1-n_2-n_3)!\, n_1!\, n_2!\, n_3!}$$

$$= \frac{n!}{0!\, n_1!\, n_2!\, n_3!} \quad \ldots (n = n_1 + n_2 + n_3)$$

$$\therefore \text{No. of ways} = \frac{n!}{n_1!\, n_2!\, n_3!}$$

★ Multinomial probabilities

- Imagine you have balls of 'r' different colours. The colours are: $i = 1, 2, 3, \ldots r$
- The probability of picking a ball of colour 'i' is 'p_i'
- Assume that 'n' balls are drawn independently.
- Now there are also 'r' non-negative no.s given to you which are n_i where $n_1 + n_2 + n_3 + \ldots n_r = n$
- Find P(n_1 balls of colour 1, n_2 balls of colour 2, ... n_r balls of colour r are drawn)
- Special case: example is r = 2; colours: 'heads', 'tails'

→ **Solution:** Let E be the event such that for n balls drawn, n_1 are of colour 1, n_2 of colour 2 .... & n_r of colour r.

The probability of the sequence can be easily be found out as $\prod_{1}^{r} (p_i)^{n_i}$

Now, we just want to count how many of these sequences are possible. You might have guessed it that we could just use the partitions formula.

$$\therefore \boxed{P(E) = \frac{n!}{n_1!\, n_2!\, n_3! \ldots n_r!} \cdot p_1^{n_1} \cdot p_2^{n_2} \cdot \ldots p_r^{n_r}}$$

★ Random variables

- It's a numeric quantity whose value is determined based on the outcome of a probabilistic experiment.
  Ex:- Weight (kg) or Height (m) of a randomly selected student.
- A random variable (r.v.) associates a value to every possible outcome.
- Mathematically, a function from sample space Ω to the real numbers.
- It can have discrete (weight, height) or continuous (measurement of time with infinite precision)
- Notation: r.v. denoted by 'X' (capital letter)
  it's value denoted by 'x'
- You can have multiple random variables defined on a single sample space.
- A function of one or several random variables is also a random variable (BMI from weight & height)

★ Probability Mass function of a Discrete r.v. X :-

- It's the 'probability law' or 'probability distribution' of X. ★★★ Note: P.M.F. is ~~only~~ used when there are r.v.s
- Suppose you want to find out the probability of getting value 'x' for r.v. 'X', you will need to find out the outcomes of the experiment that have these values 'x'.
- For example, consider an experiment with 4 ~~things~~ equally likely outcomes: a, b, c, d.
  Now a & b have value x for r.v. X.

b & c have values y & z respectively for r.v. X. Then, the probability of getting x for X will be $P(a|b) = P(a) + P(b) = 1/2$.

- Now, we can also draw the P.M.F. for r.v. X. Because X can take values x, y, z with unequal probabilities, PMF will be of type $P_X(x)$ vs. x

- Now, $P_X(x) \geq 0$ & $\sum_x P_X(x) = 1$ ... i.e sum of all probabilities of all values of X will be 1. (Since one outcome can only have 1 value of r.v.)

★ Bernoulli random variable: (The simplest r.v.)

- This r.v. takes values of only 1 and 0
- It depends on a parameter 'p' where $p \in [0, 1]$

$$X = \begin{cases} 1, & P_X(1) = p \\ 0, & P_X(0) = 1-p \end{cases}$$

- Bernoulli r.v. show up when there are ~~only two~~ 0-1 experiments like success/failure or Heads/Tails, etc.

- Bernoulli r.v. are used for Indicator r.v. (i.e. indicator r.v. are Bernoulli r.v.)

For example,
Indicator r.v. of event A: $I_A = 1$ iff A occurs
$I_A = 0$ Otherwise

Here, $P_{I_A}(1) = P(I_A = 1) = P(A)$

$P_{I_A}(0) = P(I_A = 0) = P(A^c)$

★ Discrete Uniform Distribution

- In a discrete uniform distribution, the outcomes of an experiment are all discrete & equally likely.
- If there are n outcomes, then Probability of any randomly chosen outcome is $1/n$.

★ Discrete ~~random~~ uniform random variable

- It's where a r.v. follows discrete uniform distribution
- It's defined by 2 parameters: a & b, $a \leq b$
- Sample space for r.v. $\{a, a+1, a+2, \ldots b\}$ which has $(b-a+1)$ values / outcomes
- Random variable X: $X(w) = w$
- All values of r.v. are equally likely.

Special case:- $a = b$, i.e. r.v. is a deterministic r.v. or simply, a constant.

★ Binomial random variable

- It's the random variable that follows binomial distribution.
- It's completely defined by 2 parameters: n (no. of outcomes) & $p : p \in [0, 1]$ (probability of independent event)

- A very good example is n independent tosses of a coin for which $P(\text{Heads}) = p$.
And X: no. of heads observed. (no. of successes)
Then, the PMF can be defined by:
$$P_X(x) = {}^nC_x \, p^x \cdot (1-p)^{n-x} \quad ; \text{where } x = 0, 1, 2 \ldots n$$

- This r.v. is used to model no. of successes in given no. of independent trials.

Geometric random variable

- It is defined by parameter p: $0 < p \leq 1$
or p: $p \in (0, 1]$
- Experiment: Infinitely many tosses of a coin where $P(\text{Heads}) = p$ (i.e. $P(\text{success}) = p$)
- Sample space: Set of infinite sequences of H & T
- Random variable: X: no. of tosses until first Heads (or in general, no. of failures till we arrive at success)
Note:- It includes the toss for which we get Heads

- Model is used for waiting times or no. of trials till we get success at independent repeating events.
- $P_X(k) = P(x=k) = (1-p)^{k-1} p$, where, $k = 1, 2, 3, \ldots$

For $k \to \infty$, $P_X(k) = 0$

- The distribution has form of exponential decay with asymptote at 0.

★ Expectation / Mean of a r.v.

- It is the value of the r.v. that we would get on an average if we repeat the experiment independently ~~in~~ in large amounts.
- It's basically the weighted average

$$E[X] = \sum_x x \cdot P_X(x)$$

~~which~~ (which is analogous to area under the curve for continuous curve)

- In case of an infinite sum, we assume $\sum_x |x| \cdot P_X(x) < \infty$ so that the series converges.

- Expectation of Bernoulli r.v. :-

$$E[X] = 1 \cdot p + 0 \cdot (1-p) = p$$

For indicator r.v. of event A,
$E[I_A] = P(A)$

- Expectation of uniform r.v. :-

If $X = 0, 1, 2, \ldots n$
$E[X] = n/2$

★ Properties of expectations

- If $X \geq 0$, then $E[X] \geq 0$
- If $a \leq X \leq b$, then ~~$E[X] =$~~ $a \leq E[X] \leq b$
- If c is constant, $E[c] = c$

★ The expected value rule for calculating $E(g(X))$

- Let X be a r.v. & $Y = g(X)$

$$\therefore E[Y] = \sum_y y \cdot P_Y(y)$$

&

$$E[Y] = E[g(X)] = \sum_x g(x) P_X(x)$$

- Note:- In general, $E[g(X)] \neq g(E[X])$
  For example, $E[X^2] \neq (E[X])^2$

★ Linearity of Expectation:

- Consider X & Y such that $Y = 2X + 100$
- Then, $E[Y] = E[2X + 100] = 2E[X] + 100$
- This is one of the few exceptions for the the general rule $E[g(X)] \neq g(E[X])$

★ Variance & Standard deviation:

- It's the spread of the PMF or a measure of it's distribution horizantally.
- You would presume that the average deviation from the expected value would give variance, but that value is 0. (check by trying $E[x-\mu]$)
- **Variance:** $\text{var}(x) = E[(x-\mu)^2]$

  where, X is the r.v. with mean $\mu = E[x]$ & $X-\mu$ gives the distance from the mean.
- Thus, by expected value rule,

$$\text{var}(x) = \sum_x (x-\mu)^2 P_x(x)$$

Note that if data has units 'm', then variance has units (m^2).

- **Standard deviation:** $\sigma_x = \sqrt{\text{var}(x)}$

★ Properties of variance:

- $\text{var}(X+b) = \text{var}(x)$

  i.e. if you add a constant to all values & frequency of the r.v., then ~~it's~~ its variance stays same.
  (check by ~~$\sum$~~ var(Y), $Y = X+b$)
- ~~var(aX)~~ $\text{var}(aX+b) = a^2\,\text{var}(x)$

- ~~$\text{var}(X) = E[X^2] =$~~

- $$\boxed{\text{var}(X) = E[X^2] - (E[X])^2}$$

★ Variance of Bernoulli r.v.

For a Bernoulli r.v., remember that

$$X = \begin{cases} 1 \text{ with prob. } p \\ 0 \text{ with prob. } (1-p) \end{cases}$$ & thus, $E[X] = p$

$$\boxed{\text{Var}(X) = \sum_x (x - E[X])^2 P_X(x) = p(1-p)}$$

If you plot var(X) vs. p for Bernoulli r.v., you will get an inverted parabola in the positive quadrant. Its max value is at $p = 0.5 = 1-p$. You can get the intuitive reasoning behind this by again looking at var(X) actually is. Variance measures the uncertainty as well as spread of the PMF. Focusing on the uncertainty part, you could maybe see that for Bernoulli r.v. the outcome of experiment would be most random when $p = 1-p = 0.5$ i.e. equally likely i.e. no outcome is favoured more.

★ Variance of uniform r.v.

$$\boxed{\text{Var}(X) = \frac{1}{12}(b-a)(b-a+2)}$$

where $X = a, a+1, a+2, \ldots b$

★ Conditional PMF & expectation, given an event

- Remember that for any r.v., $\sum_x P_X(x) = 1$ which is obvious, because they are all mutually exclusive & exhaustive.
- Then, let's say the r.v. is conditioned such that an event has already happened and now, only some values of the r.v. are valid that correspond to the happened event.
- So, even then, these changed probabilities must add to zero i.e. $\sum_x P_{X|A}(x) = 1$
- $P_{X|A}(x) = P(X = x|A)$
- $E[X|A] = \sum_x x \cdot P_{X|A}(x)$ (Expectation)
- $E[g(X)|A] = \sum_x g(x) P_{X|A}(x)$ (Expected value rule)
- All ~~not~~ laws that apply to regular PMFs apply to conditional PMFs
- In general, $var(X|A) \gtrless var(X)$ because the uncertainty has decreased after conditioning (obviously)

★ Total Expectation Theorem:

- Recall the Total Probability Theorem:

$P(B) = P(B \cap A_1) + P(B \cap A_2) + \dots P(B \cap A_n)$

OR

$P(B) = P(A_1) P(B|A_1) + P(A_2) \cdot P(B|A_2) + \dots P(A_n) P(B|A_n)$

- We basically replace event B with event $X = x$ such that

$$P_X(x) = P(A_1) \cdot P_{X|A_1}(x) + \dots + P(A_n) \cdot P_{X|A_n}(x)$$

- We take weighted avg of this equation for all values of X to get:

$$E[X] = P(A_1) \cdot E[X|A_1] + \dots + P(A_n) \cdot E[X|A_n]$$

- This is sort of a divide & conquer approach to calculate $E[X]$

★ Conditioning a Geometric r.v.

- Consider X: no. of independent coin tosses until first head
  $P(H) = p$
  The r.v. is defined by PMF $P_X(x) = (1-p)^{x-1} \cdot p$, $x = 1, 2, 3 \dots$

- Assume the condition that first n tosses were tails, how would that affect the PMF? What will be the conditioned PMF?

- Conditioned on $X > n$, $X - n$ is geometric with parameter p. i.e. even after shifting / changing / conditioning, the probabilities relatively remain the same.
  $P_{X-n|X>n}(k) = P_X(k)$

- This is called memorylessness: No. of remaining coint tosses conditioned on Tails in the 1st toss, is Geometric, with parameter p.

- An intuitive way to think about this would be:
  Assume the condition $X > 2$.
  The previous PMF: $P_X(k) = (1-p)^{k-1} p$, $k = 1, 2, 3, \ldots$
  The new PMF: $P_X(k-2) = (1-p)^{k-3} p$, $k = 3, 4, 5 \ldots$
  You can see that it essentially remains the same.

- Another example:
  Consider $P_X(k) = (1-p)^{k-1} p$, $k = 1, 2, 3, \ldots$
  Condition: $X > 2$
  By conditioning, we have removed the first two values & their probabilities of the r.v. from the mix.
  Now, since the probabilities of all remaining values of the r.v. must also sum to 1, we have to scale the probabilities by some factor 'q'.

Thus, for $X > 2$, probabilities start from $(1-p)^2 p$ & so on since k starts from 3 now.
We multiply all these probabilities by 'q' & sum them to 1 to get the scaling factor 'q'.

$$q \cdot (1-p)^2 p + q \cdot (1-p)^3 p + q \cdot (1-p)^4 p + \ldots = 1$$

$$\therefore q \cdot (1-p)^2 \left[p + (1-p)p + (1-p)^2 p + \ldots \right] = 1$$

~~By summation rule~~ This is a geometric series for which, $a = p$ & $r = 1-p < 1$ & we can infinite sum.

$$S_\infty = \frac{a}{1-r} = \frac{p}{1-(1-p)} = 1$$

$$\therefore q \cdot (1-p)^2 = 1$$

$$\therefore q = \frac{1}{(1-p)^2}$$

Now, we will get the new shifted probabilities by multiplying them by 'q'.

$$\therefore P_{X|A}(k) = (1-p)^{k-1} \cdot p \cdot q \quad \text{for } k = 3, 4, 5, \ldots$$

$$\therefore P_{X|A}(k) = \frac{(1-p)^{k-1}}{(1-p)^2} \cdot p \quad \text{for } k = 3, 4, 5 \ldots$$

$$\therefore P_{X|A}(k) = (1-p)^{k-3} \cdot p \quad \text{for } k = 3, 4, 5 \ldots$$

As you can see, the new probabilities are same as the old ones. This is possible because of the geometric nature of the r.v.

☆ Expectation of a geometric r.v.

$$E[X] = \sum_{k=1}^{\infty} k \cdot P_X(k) = \sum_{k=1}^{\infty} k \cdot (1-p)^{k-1} p$$

This is an arithmetico-geometric series.
Solving this gives

$$\boxed{E[X] = \frac{1}{p}}$$

★ Multiple random variables & joint PMFs

$X: p_X$
$Y: p_Y$
↑

Joint PMF: $p_{X,Y}(x,y) = P(X=x \ \& \ Y=y)$

There individual PMFs are called marginal PMFs in context of joint PMFs

- $\sum_x \sum_y p_{X,Y}(x,y) = 1$
- Just like total probability theorem,

$$p_X(x) = \sum_y p_{X,Y}(x,y) \quad \& \quad p_Y(y) = \sum_x p_{X,Y}(x,y)$$

- You can also extend this to multiple random variables:
  - $p_{X,Y,Z}(X=x \ \& \ Y=y \ \& \ Z=z) = p_{X,Y,Z}(x,y,z)$
  - $\sum_x \sum_y \sum_z p_{X,Y,Z}(x,y,z) = 1$
  - $\sum_y \sum_z p_{X,Y,Z}(x,y,z) = p_X(x)$

  &

  $\sum_z p_{X,Y,Z}(x,y,z) = p_{X,Y}(x,y)$

- Expected value rule for function of multiple r.v.

$Z = g(X,Y)$

Its PMF: $p_Z(z) = P(Z=z) = P(g(X,Y)=z)$

$$E[g(X,Y)] = \sum_x \sum_y g(x,y) p_{X,Y}(x,y)$$