Probability & Statistics For Machine learning.

Probability

The measure of how likely an event will occur

→ 10 kids 3 play soccer 7 don't $P(S) = 3/10 = 30\%$

→ 2 coins flipped → {H,T} {HT} → {HH, HT, TH, TT} $P(HH) = 1/4$

→ 3 coins flipped → {HHH, HTH, HTT, HHT, THH, TTH, THT, TTT} = $8 = 2^3$
$P(HHH) = 1/8$

→ Dice rolled $P(5) = 1/6$

→ 2 dice rolled → $6^2 \Rightarrow 36$ total $P(6,6) = 1/36$

Complement of Probability

→ 10 kids soccer, 3 play 7 no. $P(\overline{\text{soccer}}) = 1 - P(\text{soccer}) = 1 - 0.3 = 0.7$

→ coin tossed 3 times $P(\overline{HHH}) = 1 - P(HHH) = 1 - 1/8 = 7/8$

Sum of Probabilities. Disjoint events

→ 1 school kids play 1 sport $P(\text{Soccer}) = 0.3$ $P(\text{Basketball} = 0.4)$
$P(\text{Soccer} \cup \text{Basketball}) = 0.3 + 0.4 = 0.7$

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

→ when throwing a die $P(\text{even})$ or $P(5)$
→ $4/6 = 4/3$

→ roll two fair dice $P(\text{difference of 2})$ or $P(\text{difference of 1})$
→ {(6,4),(5,3),(4,2),(3,1),(6,5),(5,4),(4,3),(3,2),(2,1)} ×2 to rotate dice
→ $18/36 \Rightarrow 1/2$

Sum of Probabilities (Joint Events)

→ $P(\text{rain})$ 60% $P(\text{windy})$ 70% $P(\text{rain or windy}) = 130\%$ WTF!!
rain & windy can happen together they are not independent

→ School kids can play as many sports as they want
$P(S) = 0.6$ $P(B) = 0.5$
$P(S) \cup P(B) = P(S) + P(B) - P(S \cap B)$ ← $P(\text{soccer \& basketball})$
$= 0.6 + 0.5 - P(0.7)$
$= 1.1 -$ cannot infer

Now 6 kids play soccer 5 basketball 3 both.
$P(S) = 6/11$ $P(B) = 5/11$ $P(A \cap B) = 3/11$
8

Disjoint vs Joint events

Mutually exclusive: $P(S \cup B) = P(S) + P(B)$

Non-mutually-exclusive: $P(S \cup B) = P(S) + P(B) - P(S \cap B)$

Ex Sum of 7 or a difference of 1 rolling 2 fair die
sum {(6,1),(5,2),(4,3),(3,4),(2,5),(1,6)} → 6
diff {(6,5),(5,4),(4,3),(3,2),(2,1),(1,2),(2,3),(3,4),(4,5),(5,6)} → 10
(3,4),(4,3) repeat so $\frac{10+6}{36} - 2(36) = \boxed{14/36}$

Independent Event

When one event does not affect another event

→ School + do soccer do not soccer.

Product Rule → $P(A \cap B) = P(A) \cdot P(B)$

Ex coin tossed 5 times heads all 5 times
$P(HHHHH) = P(H) \cdot P(H) \cdot P(H) \cdot P(H) \cdot P(H)$
$= 1/2 \cdot 1/2 \cdot 1/2 \cdot 1/2 \cdot 1/2 = 1/32$

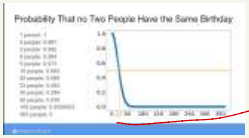
Ex If you roll a fair die 10 times. P(10 sixes)?
→ $P(\text{six})^{10} = \left(\frac{1}{6}\right)^{10}$

Ex 30 friends at a party. what is more likely?
- two people with the same bday or
- that no two of them have the same bday
- Assume 365 days No Feb 29.

Diff Bday

Person 1	Person 2	3	4	5	6	7	8	9
365/365	× 364/365	× 363/365	362/365	361/365	360/365	359/365	358/365	357/365

$= 0.905$ which is pretty high.



This is what the graph looks like

0.9

It starts being less than 0.5 at 23

Conditional Probability

today(rain) = 0.6
but yesterday rainy so today P(rain) changes

Ex Probability of landing heads twice if the first coin flipped is heads?
1/2 $P(HH \mid 1\text{st is } H)$

Product Rule for Independent events

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

Product rule for Conditional Probability

$$P(A \cap B) = P(A) \cdot P(B/A)$$

Single variant

Lemme explain:
role 2 dice, probability that the sum is 10 (B) & the first is 6 (A).
10 Let's take A first → first is 6 narrows down the possibilities to
61 62 63 64 65 66 now implement B to these
sum 10 for 64 so $P(A) = 6/36$ $P(B/A) = 1/6 = \frac{1}{6} \times 1/6 = 1/36$

If you go opposite way it works also
$P(B) \rightarrow$ sum 10 → {6,4, 5,5, 4,6} → ×2
→ 6 total possibilities → $6/36$

$P(B/A) \rightarrow 1/6$
→ $\frac{1}{6} \cdot \frac{1}{6} = 1/36$

Ex $P(S) = 0.4$ $P(\overline{NS}) = 0.6$ $P(\text{Running shoes}) = 0.8$

$P(S \cap R) = P(S)\,P(R/S) = 0.4 \cdot 0.8 = 0.32 = 32$ kids

40 play football 80% of students that play football wear running shoes

Ex $P(R/\overline{NS}) = 0.5$ so $P(\overline{NS} \cap R) = ?$
$P(\overline{NS}) \cdot P(R/\overline{NS})$
→ $0.6 \cdot 0.5 = 0.3 = 30$

Bayes Theorem

Let's say there are 1,000,000 people. Disease affects 1/10,000
you go to doctor (test) accuracy 99%. Diagnosis → sick
what is the probability that you are sick

among 1,000,000 — 999,990 — 100 sick

among these → sick & diagnosed sick / (sick & diagnosed sick + healthy & diagnosed sick)

⇒ sick & diagnosed sick / everyone diagnosed sick

1M
- sick 100
  - diagnosed | sick = 99% — sick & diagnosed sick = 99
  - diagnosed healthy | sick = 1% — " " " healthy = 1
- healthy 999,900
  - diagnosed sick | healthy = 1% — healthy & diagnosed sick = 9999
  - diagnosed healthy | healthy = 99% — " " " healthy = 989,901

P(sick | diagnosed sick) = ?

P(sick) = 0.01%
P(not sick) = 0.99%
P(diagnosed sick | sick) = 99%
P(" | not sick) = 1%

A: sick B: diagnosed sick
P(A|B) = ?

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

$$= \frac{P(\text{sick \& diagnosed sick})}{P(\text{diagnosed sick})}$$

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

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

P(sick & diagnosed sick) = P(sick) · P(diagnosed sick | sick)

$$\frac{P(\text{sick})\cdot P(\text{diagnosed sick} \mid \text{sick})}{P(\text{diagnosed sick})}$$

→ Now P(diagnosed sick)
→ P(Healthy ∩ diagnosed sick) + — 1.1
P(sick ∩ diagnosed sick) — 1.2

Now if you look 1.1 & 1.2 are independent as 1.1 is among the not sick & 1.2 is among the sick.

$$\frac{P(\text{sick})\cdot P(\text{diagnosed sick} \mid \text{sick})}{P(\text{not sick} \cap \text{diagnosed}) + P(\text{sick} \cap \text{diagnosed sick})}$$

P(sick ∩ diagnosed sick) = P(A∩B) = P(sick) · P(diagnosed sick | sick) — 2
P(not sick ∩ diagnosed ") = P(A'∩B) = P(not sick) · P(diag " " | not sick) — 3

$$P(A|B) = \frac{P(A)\cdot P(B|A)}{P(A)\cdot P(B|A) + P(A')\cdot P(B|A')}$$

Bayes Theorem Spam Example

we start with 20% of all emails spam.
Now we notice spam emails tend to have the word lottery.
P(spam | lottery) - ? we notice among 20 spam 14 had lottery
" " " 80 non " 10 " "
total lottery → 14 + 10 = 24
P(if email has lottery is spam)

$$P(\text{spam} \mid \text{lottery}) = \frac{P(A\cap B)}{P(B)} \Rightarrow \frac{14}{24} \Rightarrow \frac{7}{12}$$

recall Baye's → $\frac{P(A)\cdot P(B)}{P(A)\cdot P(B|A) + P(A')\cdot P(B|A')}$ where A' is not A

A: email spam, B: email has lottery   P(spam) = 0.2, P(spam') = 0.8

$$\frac{0.2\times 0.7}{(0.2\times 0.7) + (0.8\times 0.125)} = 0.583 = 7/12$$

Prior → Original probability that you can calculate not knowing anything else the something happens.
Event → gives info about the probability and with that you can calculate the
Posterior →

Prior	Event	Posterior
P(A)	E	P(A|E) – P of A given E

email — spam / ham   Email contains lottery
email — spam — lott — spam & lott
email — spam — No lott
email — ham — lott — ham & lott
email — ham — No lott

What about 2 events.

Prior: $P(\text{spam}) = \frac{\text{spam}}{\text{spam} + \text{ham}}$

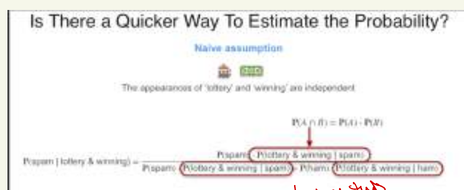
Event:
- Email has 'lottery' — $P(\text{spam} \mid \text{lottery}) = \frac{\text{lottery spam}}{\text{lottery spam} + \text{lottery ham}}$
- Email has 'winning' — $P(\text{spam} \mid \text{winning}) = \frac{\text{winning spam}}{\text{winning spam} + \text{winning ham}}$
- Combine

Email has 'lottery' & winning → P(spam | lottery & winning) = spam (lottery + winning) / spam (- - -) + ham (- - -)

$$P(\text{spam} \mid \text{lottery \& winning}) = \frac{P(\text{spam})\cdot P(\text{lottery \& winning} \mid \text{spam})}{P(\text{spam})\cdot P(\text{lottery \& winning} \mid \text{spam}) + P(\text{ham})\cdot P(\text{lottery \& winning} \mid \text{ham})}$$

This could become very complicated for say 100 variables (events)

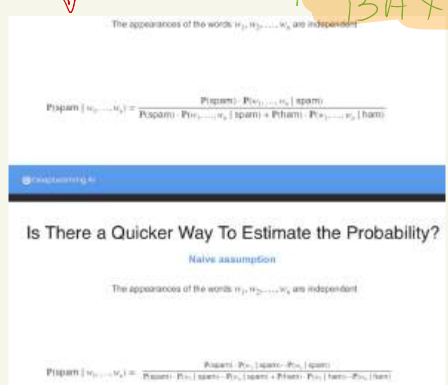
There is a better way. We make a Naive assumption. that the words 'lottery' & 'winning' are independent. which we know is not true words are not independent.



↓ gives this

↓ for n events

NAIVE BAYES



ML use

conditional

You wanna calculate the probability of something based on other factors.

- Spam detection → guess prob of spam based on features.
- Image recognition → calculate prob that there is a cat in the image based on pixels of an image
- Image generation → you want to maximise the prob that a bunch of pixels form a face
- Text generation → Bunch of characters form a word & words form a sentence

Image recognition

- What is the probabity that there is a cat in the image
- P(cat | image) = P(cat | pixel₁, pixel₂, . . . , pixelₙ)

Sentiment analysis

- Is this a happy sentence
- Calculate P(happy | words in the sentence)

Picture (cat) → model → P_{cat}(cat | pixels) = 0.9

Picture car → model → P_{car}(cat | pixels) = 0.1

Image Generation

Generate a group of pixels such that the resulting image looks like a human face

Goal: generate images such that P(face | pixels) is high.

Week 1 Quiz

1) HT, TH 2/4 = 1/2

2) sum=10 {6,4, 4,6, 5,5, 5,5} → 3/36 = 1/12

3) P(no 10) = (1) ten times $1 - 1/6^{10} = \frac{6^{10}-1}{6^{10}}$

4) 100 patients 50 headache 50 fever P(F ∪ H) = P(F) + P(H)

5) A: 4000, B: 5000 bug: 3000

P(Bug ∩ B) = 1500

$P(B|Bug) = \frac{P(B \cap Bug)}{P(Bug)} \Rightarrow \frac{1500}{3000} = \frac{1}{2} = 50\%$

$$P(X|Y) = \frac{P(X \cap Y)}{P(Y)}$$

Random Variables

They can take many values. eg Temp, no. of heads out flipping coins

X = No. of heads

flip a coin

1 heads X=1 P(X=1)=0.5; 0 heads X=0 P(X=0)=0.5

Discrete: 0 1 2 3 4 — countable but can be infinite

Continous: 0.1, 0.01, 0.5 — infinite — takes value on an entire range

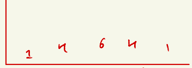
Flip 4 coins $2^4 = 16$: x=0 – 1, x=1 – 4, x=2 – 6, x=3 – 4, x=4 – 1

Probability Distributions

eg X: No. of heads in 3 coins.

$2^3 = 8$ - total outcomes

x=0 – 1
x=1 – 3
x=2 – 3
x=3 – 1



X_3: 5 coin flips.

Now this is $2^5 = 32$ hard to calculate.

so now $P(X_3 = ?)$ we can represent it as $P(X_3 = x)$, x = 0,1,2,3,4,5

This is PMF – Probability Mass Function.

$p_X(x) = P(X_4 = x)$, x = 1, 2, 3, 4, 5.

PMF must satisfy → $p_X(x) \geq 0$, $\sum_x p_X(x) = 1$

X_1, X_2, X_3, X_4, X_5 are very similar

They all represent No. of heads in n experiments

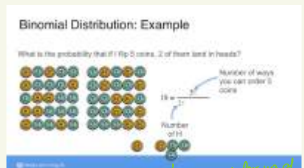
The way the probability distributes along the possible outcomes seems to have a similar pattern.

Could there be a single model to represent this pattern.

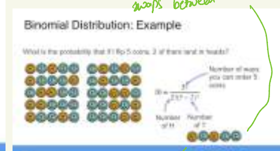
Binomial distribution

let's say you toss a coin 5 times, 2 land on heads?

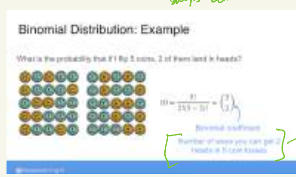
is there some way to find this?



↳ to count for no of swaps between some position



↳ to count for no of swaps between some position



Not probability. Just no of ways to rearrange them.

$\binom{n}{k}$ = counts all the combinations for landing k heads in n coin tosses.

Property: $\binom{n}{k} = \binom{n}{n-k}$

obtaining k heads is the same as n−k tails.

General PMF for X: no. of heads in 5 coin tosses?

Coin P(H) = p (usually 0.5)

Event X = x: x heads in 5 tosses

$p^x (1-p)^{5-x}$

p^x – Probability of seeing x heads; $(1-p)^{5-x}$ – Probability of seeing 5−x tails

This is however just one order.

Now we need no. of ways → Binomial coefficient.

$\binom{5}{x} p^x (1-p)^{5-x}$, x = 0, 1, 2, 3, 4, 5

X follows a binomial distribution

for n coins

$P_X(x) = \binom{n}{x} p^x (1-p)^{n-x}$, x = 0, 1, 2 . . . n

Q) Throw dice 5 times for 3 ones.

P(1) = 1/6
P($\bar{1}$) = 5/6

$P_X(3) = \binom{5}{3} \cdot \left(\frac{1}{6}\right)^3 \left(\frac{5}{6}\right)^2$

$= \frac{5!}{3!\,2!} \cdot \left(\frac{1}{6}\right)^3 \left(\frac{5}{6}\right)^2$

$\frac{5 \times 4 \times 3!}{3! \times 2}$

$10 \cdot \left(\frac{1}{6}\right)^3 \left(\frac{5}{6}\right)^2$

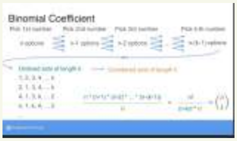
Binomial Coefficient

now this works more generally when you need to pick k out of n numbers

Pick 1st No. — n options; Pick 2nd No. — n−1 options; Pick 3rd No. — n−2; ... Pick k-th number — n − (k−1) options

set of length 1,2,3,4.... k — 2,1,3,4....k — 3,1,3,k,...,2 — 4,1,3,4....2

As these can be taken as independent events.



Bernoulli Distribution

X = No. of heads.

coin → heads: $P(X=1)=0.5$ Success; tails: $P(X=0)=0.5$ failure

X = No. sick patients

Patient → $P(X=1)=P$ Success; $P(X=0)=1-P$ failure

x = No. of 1's

dice → $P(X=1)=1/6$ Success; $P(X=0)=5/6$ failure

Bernoulli	Binomial
Single trial	multiple (n) independent trials.
2 outcomes (success or failure)	No. of successes in n trials
1 parameter P(success)	two parameters n & p(success)
0 to 1	0 to n
tossing a coin once	tossing a coin n times & getting no. of heads

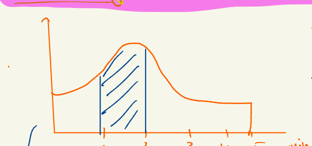
Probability Distributions (continous)

No. of heads → can be listed (0, 1, 2, 3, ...)

amount of time on phone or on the bus (0, 1, 2.1, 2.13) this cannot be listed.

- Discrete → histograms in which sum of heights = 1
- Continuous → Area under the curve = 1

Probability Density Function



P (between 1 & 2 min) = area

Now if we say exactly 2 min that would be a line segment. A line has an area of 0.

This graph

PDF: $f_X(x)$ → tells you the rate you accumulate probability around each point.

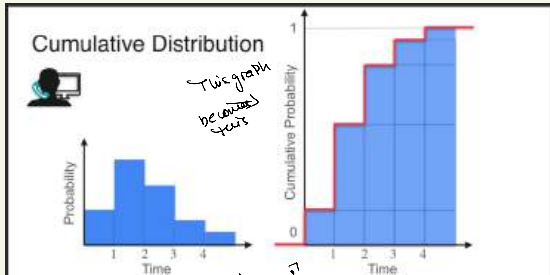
$P(a < X < b)$ = area under $f_X(x)$

$f_X(x)$ needs to satisfy:
- It is defined for all numbers
- $f_X(x) \geq 0$.
- Area under $f_X(x) = 1$

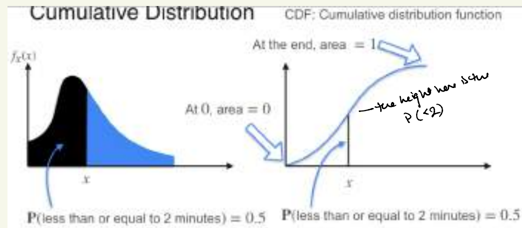
Cumulative Distribution Function

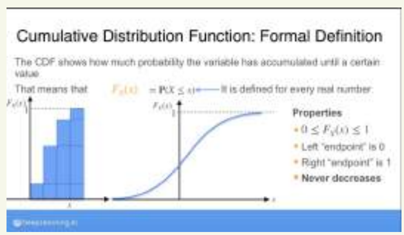
We don't always wanna calculate areas.

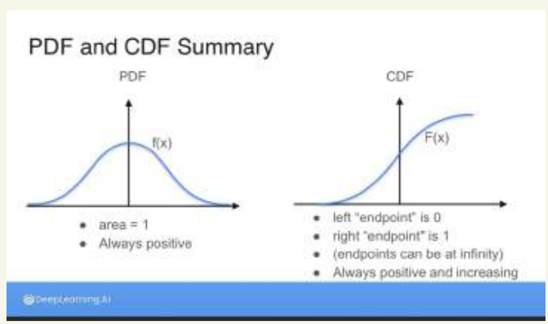
The CDF tells you cumulative P from 0 to a specific n



we are cumulating the probabilities.





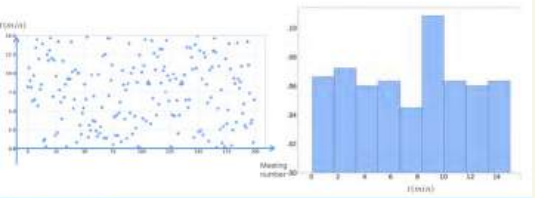


Uniform Distribution

You're calling a tech support line. They can answer any time between zero and 15 minutes and if they don't answer in this time, the line is disconnected.

Uniform Distribution: Motivation

wait time over 200 - calls



Uniform Distribution: Motivation

T: time (in minutes) you have to wait

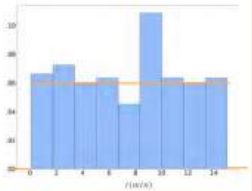
Any value between 0 and 15 minutes must have the same frequency of occurrence.

↓

The pdf must be constant for all values in the interval (0,15)

↓

Which constant? $\rightarrow 15 \times h = 1 \rightarrow h = \frac{1}{15} = 0.06$



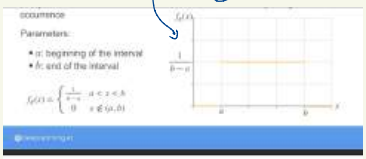
A Continuous random variable can be modelled with a uniform distribution if all possible values lie in an interval & have same frequency of occurrence

Parameters:

a: beginning of interval

b: end of interval

$$f_X(x) = \begin{cases} \frac{1}{b-a} & a < x < b \\ 0 & x \notin (a,b) \end{cases}$$

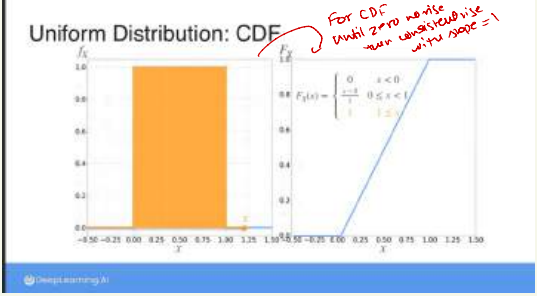


This is what the PDF would look like:

Uniform Distribution: PDF

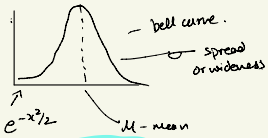
Uniform Distribution: CDF

For CDF must zero monotonic non-decreasing with slope = 1



Normal / Gaussian Distribution

When n is very large the binomial distribution can be approximated by the Gaussian distribution very well.



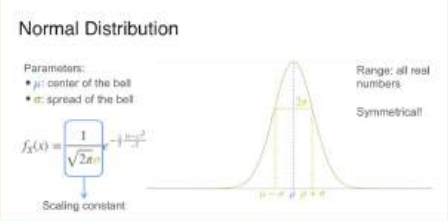
- bell curve.
- spread or wideness
- $e^{-x^2/2}$
- μ - mean

$$\frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

this is to regulate the area under the graph

The idea is $\sigma\sqrt{2\pi}$ - is the area under the original curve but the probability the area should add up to 1. So we divide it by $(\sigma\sqrt{2\pi})$. σ is the standard deviation or spread from the mean. $(x-\mu) \rightarrow$ This tells you to bring x to μ as the bell curve's top is usually at 0 i.e. we adjust by μ to regulate the std deviation or spread.

Normal Distribution



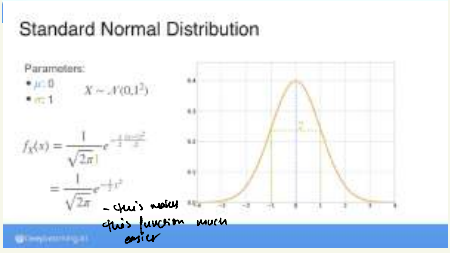
Parameters:
- μ: center of the bell
- σ: spread of the bell

Range: all real numbers

Symmetrical!

Scaling constant

Standard Normal Distribution



Parameters:
- μ: 0
- σ: 1

$X \sim N(0,1^2)$

- this makes this function much easier

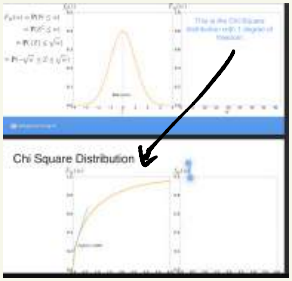
The $\frac{x-\mu}{\sigma} \rightarrow$ I explained earlier is standardization.

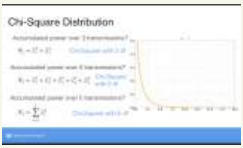
Chi Squared Distribution X^2

One common assumption is that the communication channel has noise with a standard normal distribution.

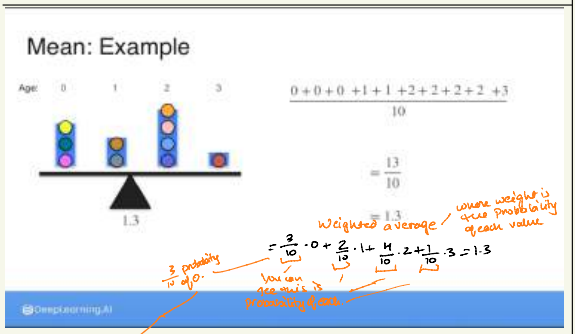
What is the power of the noise -> associated variance & dispersion of noise.

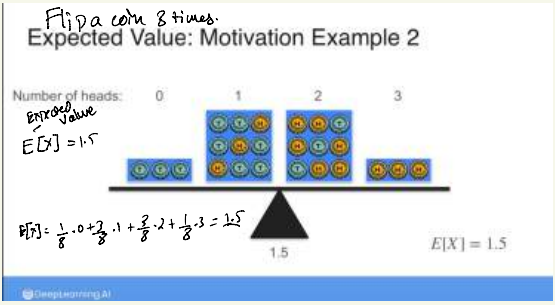
$W = Z^2$





As the number of variables (z) increases it becomes a bell curve





Expected Variable

X is a discrete random variable

PMF of X

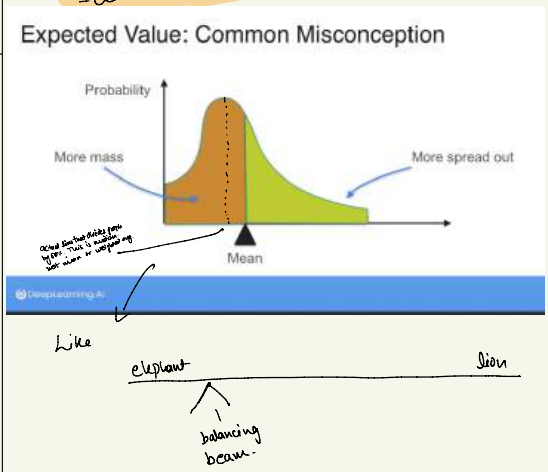
$P_X(x) = P(X = x)$

$$E[X] = \sum_x x\, p_X(x)$$

X is across continous random variables

$$\int_{-\infty}^{+\infty} x \cdot f_X(x)\, dx$$

→ Summing the area of bars that get thinner & thiner.



Expected Value:

- $E[X]$
- Mean / Balancing point
- Defined for discrete & continous random variables
- Weighted Average of the PMF / PDF