Probability & Statistics :-

Random variable :-

ex:- rolling dice → 6 sides = {1, 2, 3, 4, 5, 6} → when rolled → any one of these. } Random Experiment

equal outcome

random variable → X = {1, 2, 3, 4, 5, 6} → Sample Space.

tossing a coin → Y = {H, T}

P(X = 1) = 1/6 P(X = 2) = 1/6 ....

P(X is even) = 3/6 = 1/2

(probability of X being even)

(P(X=2) + P(X=4) + P(X=6) = 1/6 + 1/6 + 1/6)

P(X is odd) = 1/2

P(x = x_i) → P(x_i) same thing diff notations.

Finite set of values → Discrete random value

→ Height of randomly picked student.

y could be 162, 180, 120, 140, ... → infinite values. → Continuous Random Variable

Outliers :-

Y : height of student.

{122.2, 146.4, 132.5, ... 12.2, 156.3, 92.7 ...}

92.7 → could be an outlier

12.2 → outlies → could be human error (or) actual height

→ Outlies can corrupt data.

→ A discrete value is obtained by counting

→ A continuous value is obtained by measuring.

Sample Space :- Set of all possible outcomes of an experiment.

→ A random variable value depends on the outcome of a random phenomenon.

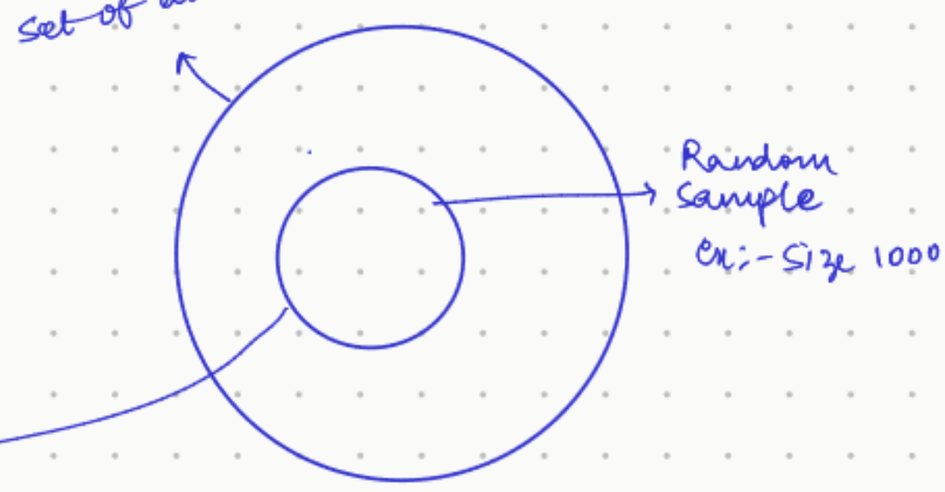
Population & Sample :-

→ Estimating the average height of human

↳ $\mu = \frac{1}{Pop} \sum_{i=1}^{Pop} h_i$ (IMPOSSIBLE)

So we estimate.

often represented by $\bar{x}$ ← $\bar{x} = \frac{1}{1000} \sum_{i=1}^{1600} h_{is}$



→ As sample size increases $\bar{x} = \mu$

$\bar{x}$ → Sample mean

μ → original mean

Sampling is of two types:-

(i) Simple Sampling

(ii) Stratified Sampling ⟶ Unbiased sampling & more accurate results.

ex:- 1000 people ⋯⋯⋯⋯⋯> original sample

- 400 have cars
- 300 have bikes
- 200 have cycles
- 100 have nothing

Sample size ⟶ 250

Simple Random Sampling

250 could have cars (or)

250 could have bikes (or)

100 bikes + 150 cars.

Stratified Random Sampling

- Cars ⟶ 100
- bikes ⟶ 75
- cycles ⟶ 50
- nothing ⟶ 25

} These are random but equal imp to all classes

Gaussian Distribution:- (AKA Normal Distribution)

→ If X is a continuous random variable & X has a PDF curve (⩘), then we say X has a Gaussian distribution.

PDF = Probability Density Function

[Mean, Median & Mode are all same for GD. There are equal number of measurements above & below the mean]

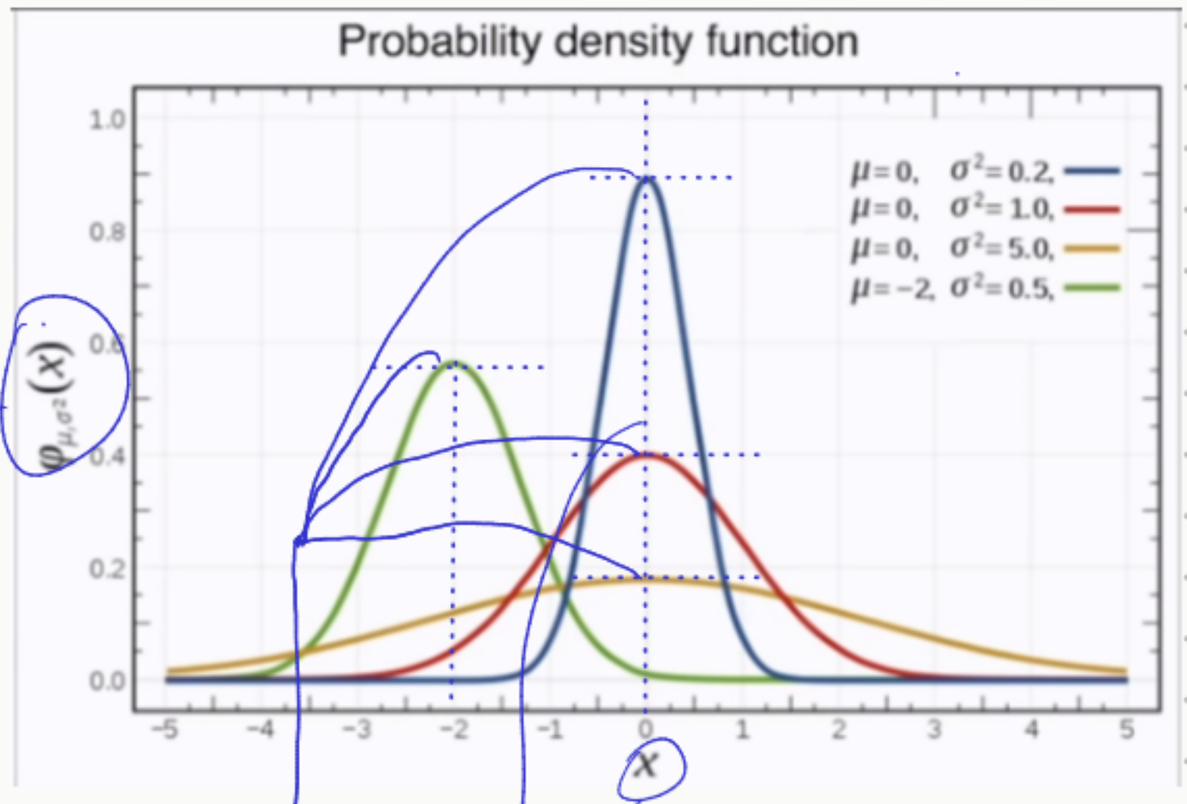
Why learn?

→ Stuff in nature tends to follow GD.

→ heights, weights of people follow it. (Natural Phenomenon)

→ They are simple models that summarize R.V.

ex:-



PDF

μ = mean
σ^2 = variance
} ⟶ parameters

[If this is not known, then we can't]

[If we are given μ, σ^2 & told that X follows GD, we can plot PDF. We don't need the whole data]

Variance is a measure of spread.

Red, Yellow, Blue have $\mu = 0$ but varying variances.

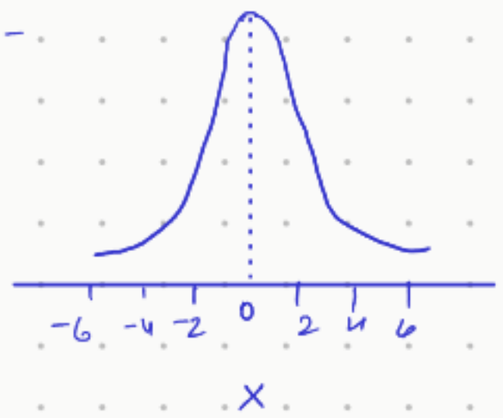
The peak of curve is usually at 'μ'.

X → Continuous Random Variable

→ The parameters of Gaussian Distribution are μ & σ^2

$X \sim N(\mu, \sigma^2)$ ($\Longrightarrow$ X follows Gaussian Distribution with μ & σ^2)

ex:-



→ $X \sim N(0, 2)$

→ $P(X = x) = p(x) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left\{\frac{-(x-\mu)^2}{2\sigma^2}\right\}$

↑ Probability Density at a point (x). PDF at any given point gives the probability density at that point. Probability of getting a single discrete value is 0.

ex:- If $\mu = 0$, $\sigma^2 = 1$, $\sigma = 1$

$f(x) = \frac{1}{\sqrt{2\pi}} \exp\left\{\left(\frac{-1}{2}\right)x^2\right\} = y$

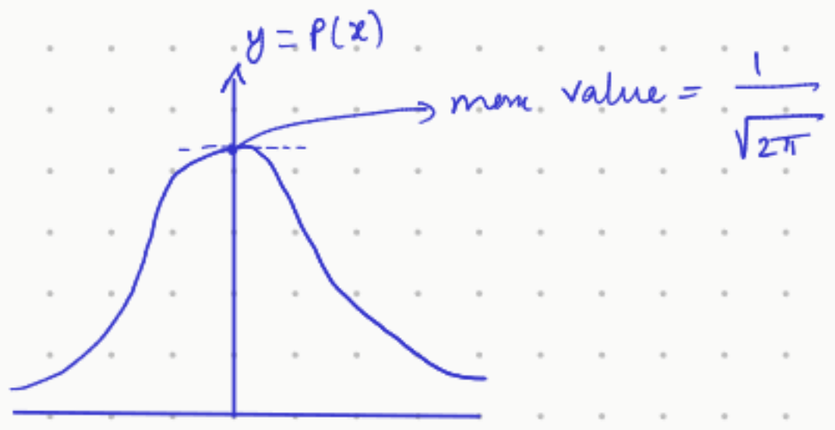
↓ constants

further simplifying

$y = \exp(-x^2)$ → when plotted

as x increases y decreases.

as x decreases y decreases.



conclusions :-

① x moves away from μ, y decreases.
② Graph is symmetric.
③ In this particular graph, y is reducing exponential squared. (e^{-x^2})

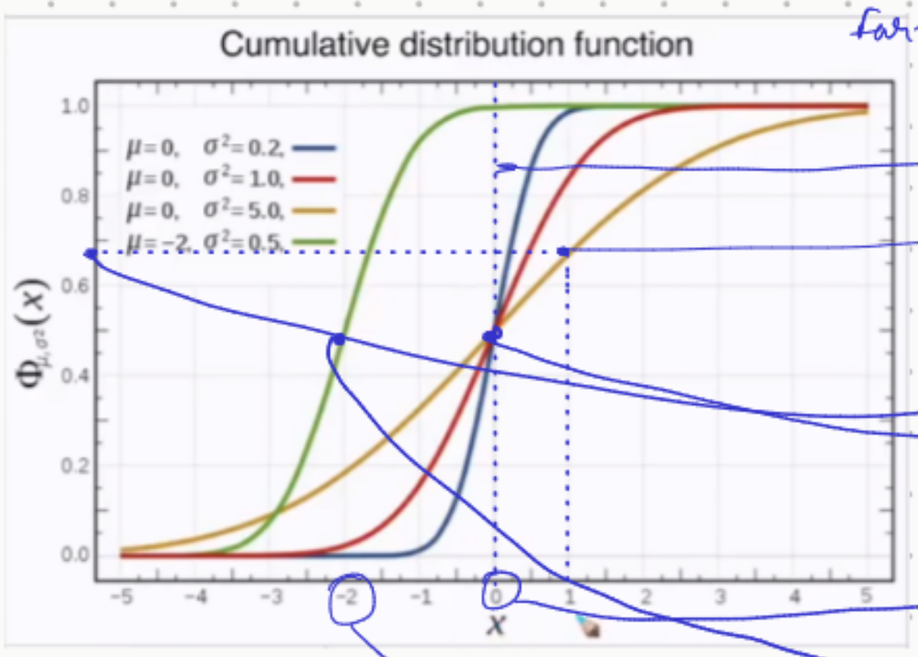
→ If p(x) is the probability density at a point 'x', the probability can be obtained by computing the integral of p(x) over a given interval.

i.e., probability of getting $X \in [a, b]$ is $\int_a^b p(x)\,dx$

Cumulative Distribution Functions (CDF) of Gaussian Distributions/Normal Distributions :-

As σ^2 increases, CDF goes farfromcenter line

CDF of a random variable looks like



→ $P(X \leq 1) = 0.65$

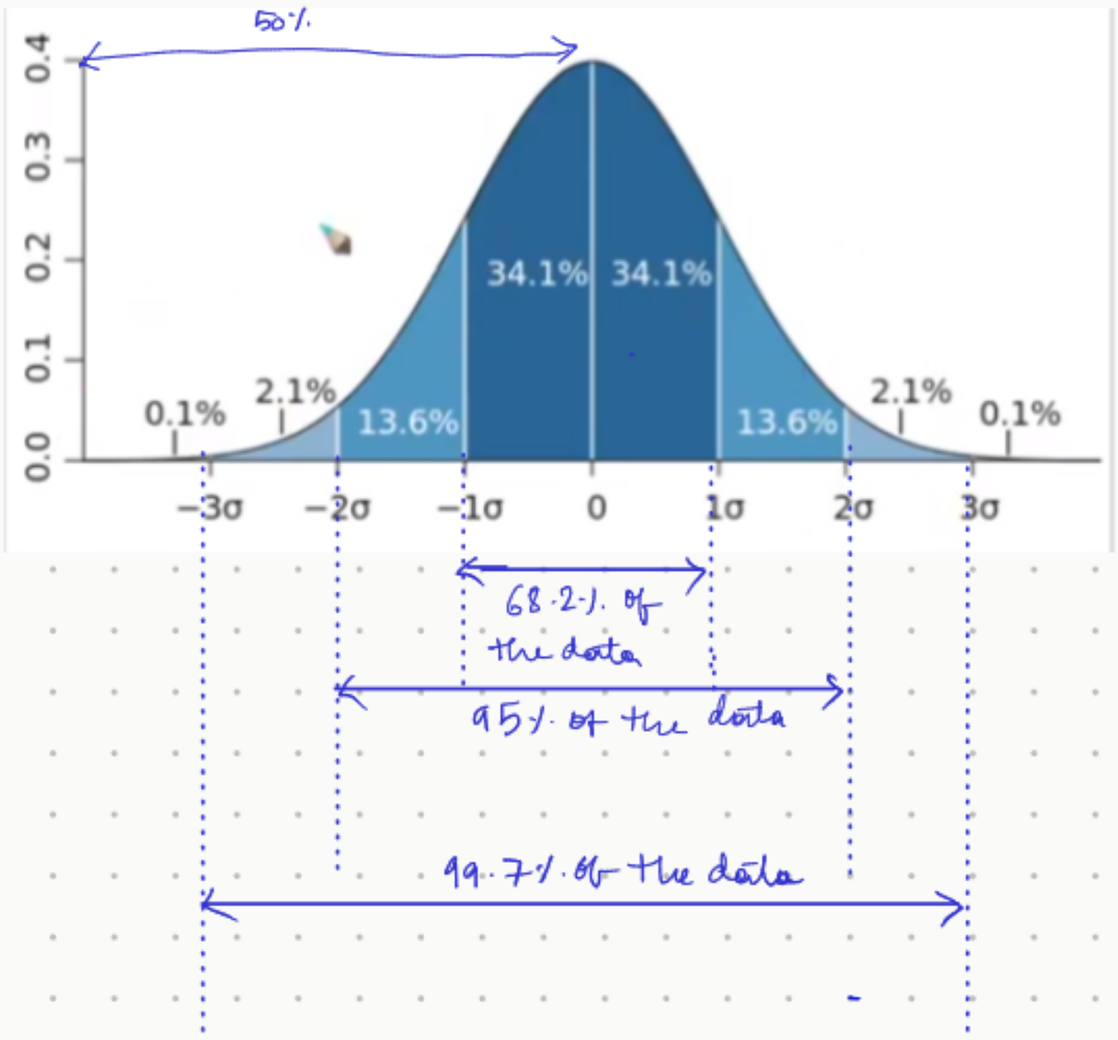
→ $\mu = 0$, center of CDF is at 0.

→ $\mu = -2$, center of CDF is at -2

$$CDF = \frac{1}{2}\left[1 + erf\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right] \longrightarrow \text{No need to memorize}$$

68-95-99.7 rule :-

if $\mu = 0, \sigma^2 = 4 \Rightarrow \sigma = 2$ $\quad X \sim N(0, 4)$



. How is this useful?

ex:- if human population weight

$X \sim N(150, 25)$ — 150 → μ, 25 → σ

$\Rightarrow$ 68.2% of human population lies b/w $(150-25, 150+25)$

95% of people $(150-50, 150+50)$

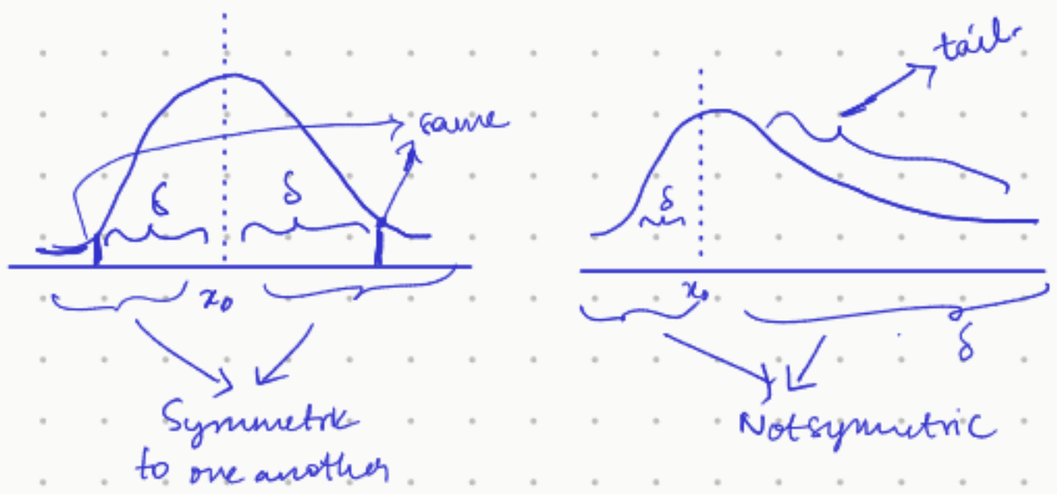
99.7% of people $(150-75, 150+75)$

→ A standard gaussian distribution always has a mean of 0 & variance 1.

If it has other mean & variance, it's a non standard gaussian distribution.

Symmetric Distribution, Skewness & Kurtosis :-

→ They help understand shape of PDF.



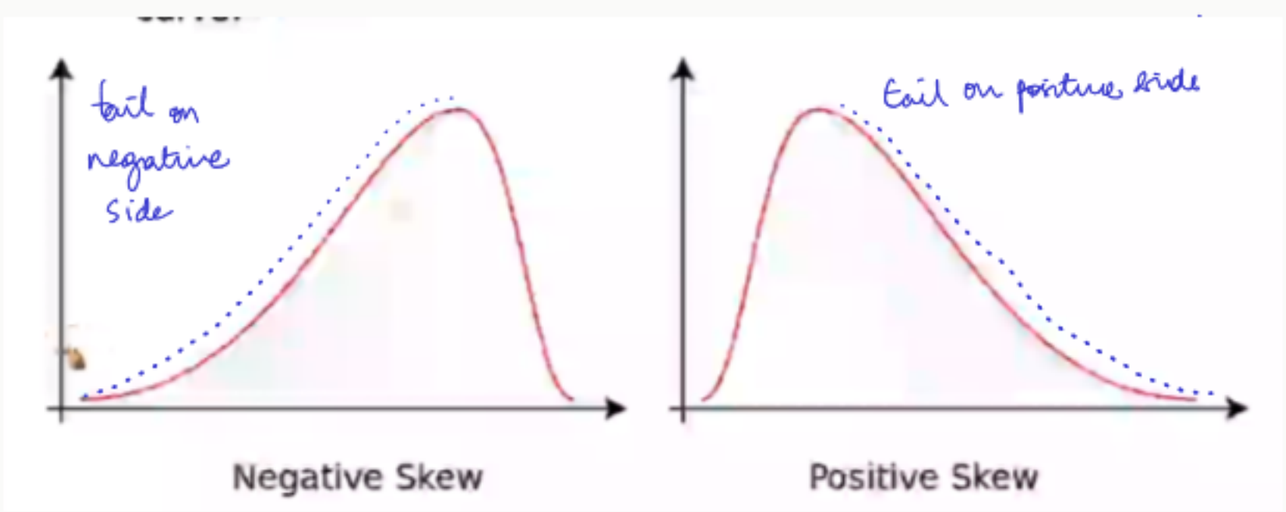
→ A probability distribution is said to be symmetric if and only if there exists a value x_0 such that

$f(x_0 - \delta) = f(x_0 + \delta)$ for all real numbers δ

$f(x)$ is the height of PDF at any point 'x'

Skewness :-

Skewness is a measure of asymmetry -



$$X = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \qquad \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\text{Skewness} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{\left[\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2\right]^{3/2}}$$

③ → if 2 = variance

↳ sample std-deviation

Kurtosis :-

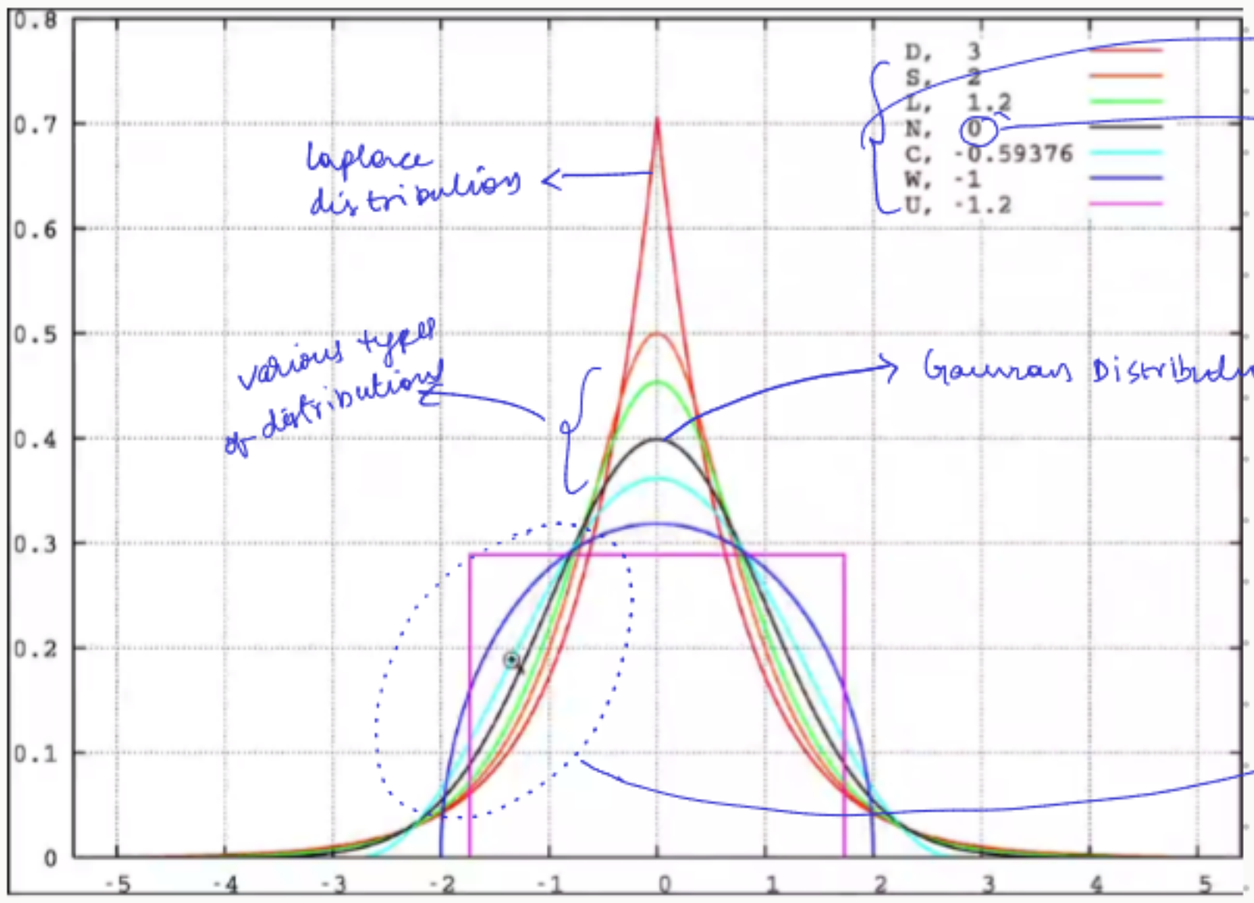
Kurtosis of Gaussian Distribution = 3

→ Measure of tailedness

excess kurtosis = kurtosis − 3

$$\text{excess kurtosis} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4}{\left(\underbrace{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}_{\text{variance}}\right)^2} - 3$$

→ Kurtosis is not a measure of peakedness. Might look like it -



→ excess kurtosis values.

→ Gaussian distribution, excess kurtosis = 0 ⇒ Kurtosis = 3

→ The rate at which the tails touch axis is different

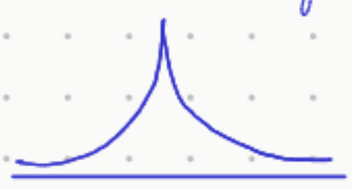
→ Kurtosis tells us if there are outliers are there in the data.

→ Kurtosis is used in analysis of trading, finance etc-)

→ Kurtosis measures how much density/weight is in tails compared to middle part

→ If there are some stocks. Some stocks fall ↓
some go up ↑
some go up enormously ↑

To estimate risk after buying some stocks, we model data - Its generally assumed that they follow gaussian distributions for ease of calculations (Pre 2008 financial crisis). Acc to gaussian distributions chances of losses are less & has good chance of high profits But now, instead of assuming the data is a gaussian distributions, we look at the actual data & calculate the kurtosis

If kurtosis is > 3 ⇝  ⇒ chances of extreme profits are high & chances of losses are also high-

If kurtosis = 3 ⇒ Follows gaussian distribution

If kurtosis < 3 ⇒  ⇒ chances of extreme profits are low & chances of losses are also fairly low.

→ Kurtosis has a degree of 4. So if there is large deviations, then it will lead to higher kurtosis value.

→ High kurtosis is caused by infrequent extreme deviations rather than frequent modestly sized deviations.

→ In order to compare kurtosis b/w two curves, they need to have the same variance.

Standard Normal Variate:- (z)

$Z \sim N(0,1)$

$\mu = 0$

$\sigma^2 = 1$

Let $X \sim N(\mu, \sigma^2)$

$X = [x_1, x_2, x_3, \ldots, x_{50}]$

Standardization:- $x_i' = \frac{x_i - \mu}{\sigma}$

$\Rightarrow x_i' \sim N(0,1)$

↑ Standard Normal Variate.

Given any random variable X, where $X \sim N(\mu, \sigma^2)$

$$z = \frac{x - \mu}{\sigma}$$ → This is also called as basic z-score formula. This basically tells us how many σ is x away from μ.

$\Rightarrow z \sim N(0,1)$

Why?:-

① After standardization, PDF becomes

→ 68.2% of the data

(∵ 68-95-99.7 rule)

② If we are comparing multiple GDs with different μ & σ^2, doing this helps understand better & interpret better

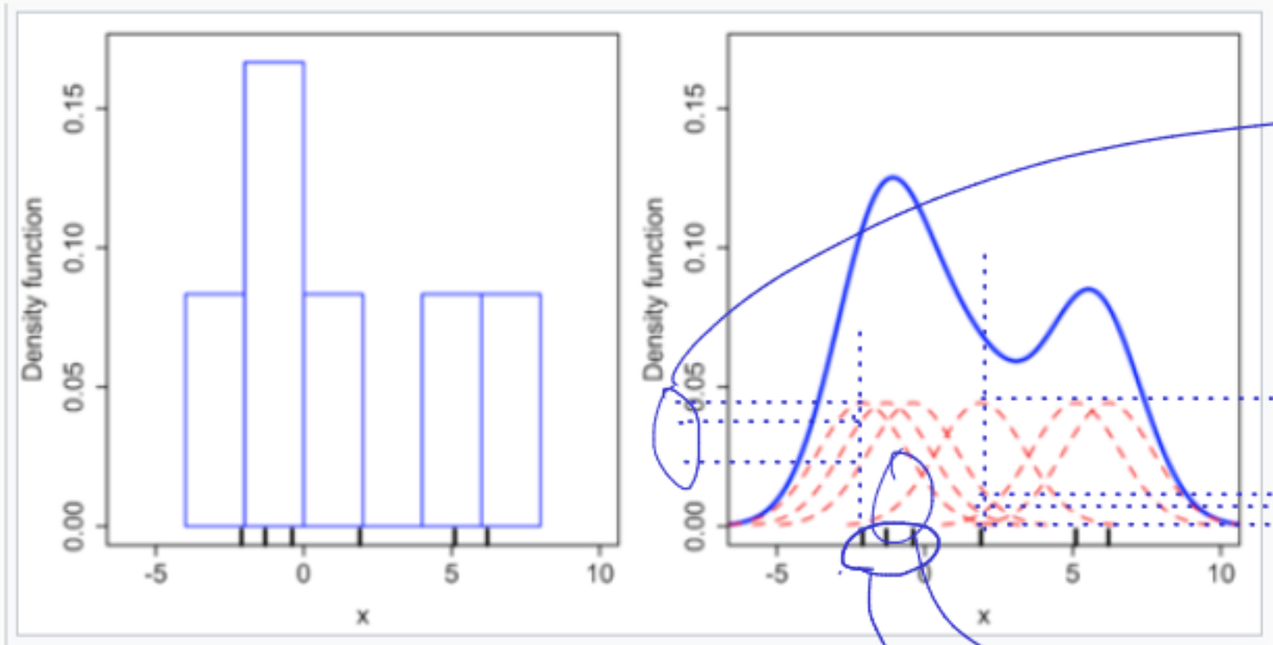
→ We use StandardScaler for standardization & MinMaxScaler for Normalization.

Two techniques of feature scaling.

Normalization → $x_i = \frac{x_i - x_{min}}{(x_{max} - x_{min})}$

Kernel Density Estimation:-

→ Used for smoothing histograms to obtain PDFs.



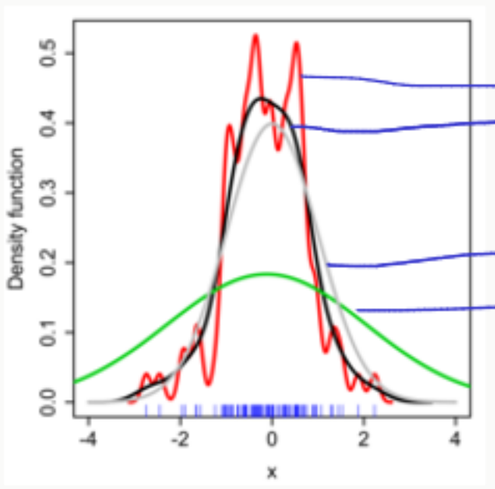
Comparison of the histogram (left) and kernel density estimate (right) constructed using the same data. The six individual kernels are the red dashed curves, the kernel density estimate the blue curves. The data points are the rug plot on the horizontal axis.

→ give height at that point

→ Variance of the gaussian kernel is known as Bandwidth.

→ Gaussian Kernel.

→ Since there are many points here, the number of means will be high. So the PDF will be high. In normal histogram,



→ Narrow Bandwidth

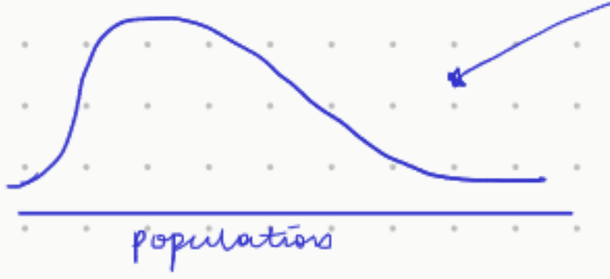
→ Medium Bandwidth

→ Actual Distribution.

→ Wide Bandwidth

Sampling Distributions & Central Limit Theorem :-

Let X be distribution of incomes over populations [Not necessarily Gaussian]



Let S_1 be random sample of size n (Let n = 30) → mean = $\bar{x}_1$ → Sample means.

" S_2 " " " " → mean = $\bar{x}_2$

" " " " " "

" " " " " "

" " " " " → mean = $\bar{x}_m$

These will also have a distribution

The distribution $\bar{x}_i$ = Sampling distributions of sample means.

Central Limit Theorem :- If original distribution 'X' has finite mean (there can be infinite mean ex:- pareto) & variance & 'm' samples are created of size 'n' whose sample means are $\bar{x}_1, \bar{x}_2, \bar{x}_3, \ldots\ldots$ whose distributions is called sampling distributions of sampling mean, central limit theorem states that

$$\bar{x}_i \longrightarrow N\left(\mu, \frac{\sigma^2}{n}\right) \text{ as } n \to \infty$$

(But IRL if $n \geq 30$, then it becomes gaussian. Rule of thumb)

mean is same as original or that of distributions

Gaussian Distributions

Why? → By using just m x n datapoints, we are able to estimate μ & σ^2 of any distributions, if we just know that they are finite.

Quantile Quantile Plot (Q-Q-Plot) :-

Let X be a random variable

$X : x_1, x_2, x_3, \ldots\ldots, x_{500}$

Question :- Is X Gaussian Distributed. Q-Q-Plot helps us identify. Other techniques such as statistical testing (KS testing) etc, also exist.

Q-Q-Plot → graphical method

KS testing → more powerful.

How? :- ① Sort X & calculate percentiles.

$x'_1, x'_2, x'_3, \ldots\ldots, x'_{500}$

$x'_5 \longrightarrow 1^{st}$ percentile $\longrightarrow x^{(1)}$

$x'_{10} \longrightarrow 2^{nd}$ " $\longrightarrow x^{(2)}$

$\vdots$

$x'_{500} \longrightarrow 100^{th}$ percentile $\longrightarrow x^{(100)}$

② take $Y \sim N(0,1)$

$Y : y_1, y_2, y_3, \ldots\ldots, y_{1000}$

$\downarrow$

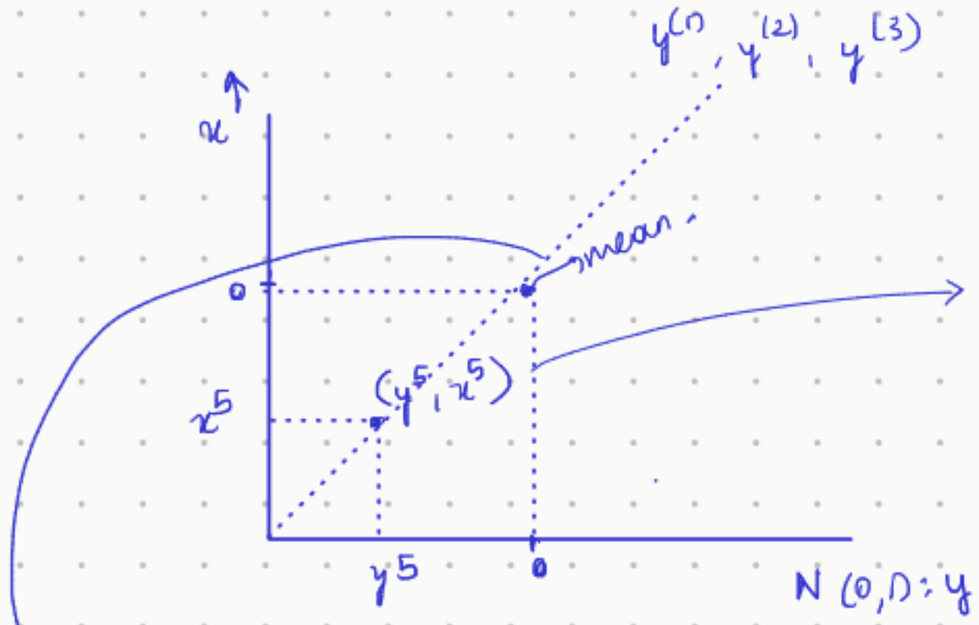
$y'_1, y'_2, y'_3, \ldots\ldots, y'_{1000}$

$\downarrow$

$y^{(1)}, y^{(2)}, y^{(3)}, \ldots\ldots, y^{(100)}$

These are called as Theoretical Quantiles

③ Plot Q-Q-Plot using $x^{(1)}, x^{(2)}, x^{(3)}, \ldots$



we get 100 such pairs (y^1, x^1), (y^2, x^2), ⋮, (y^{100}, x^{100}) — All are plotted.

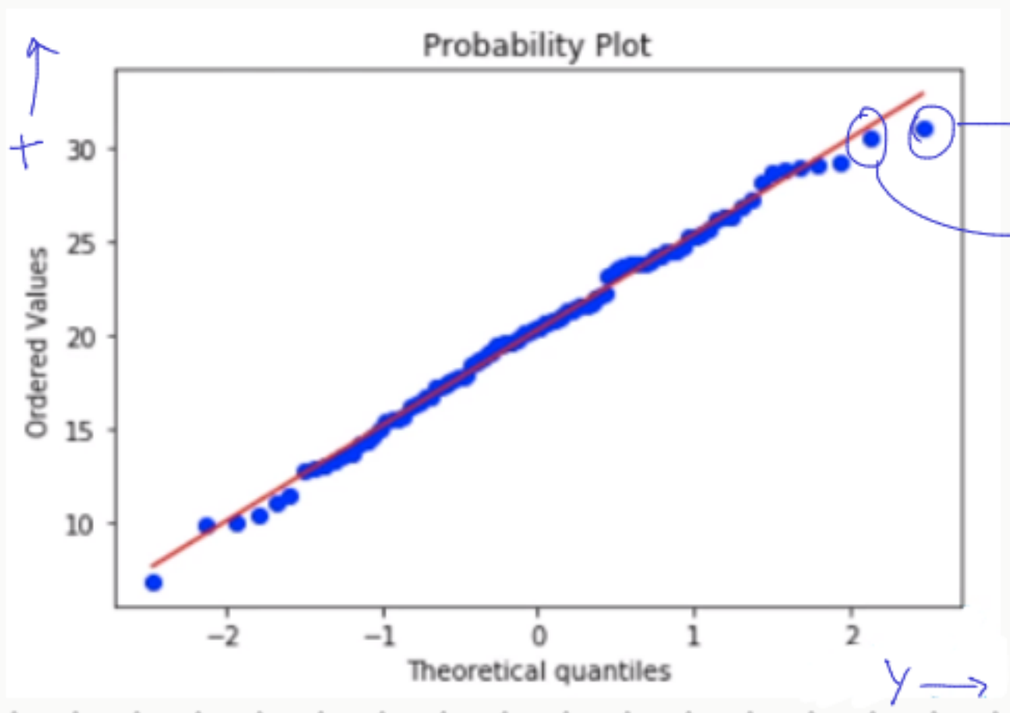
→ If (y^i, x^i) $\forall$ i → 100, lie roughly on a straight line, they y & x have similar distributions.

⇒ X also has gaussian distributions.

→ stats.probplot( )

→ If size of X is small, it is hard to interpret Q-Q plot. Won't be a straight line.

→ Another use of Q-Q plot is: given X, Y, does X & Y have the same distribution.



Generating Y :- (theoretical Quantiles)

```
#Q-Q plot
import numpy as np
import pylab
import scipy.stats as stats

# N(0,1)
std_normal = np.random.normal(loc = 0, scale = 1, size=1000)
```

mean, std-dev

Generating X :-

```
# generate 100 sanples from N(20,5)
measurements = np.random.normal(loc = 20, scale = 5, size=100)
#try size=1000


stats.probplot(measurements, dist="norm", plot=pylab)
pylab.show()
```

How & where to use distributions :-

→ All probability concepts are used for EDA most of the time.

→ If data is Gaussian Distributed, many assumptions can be made by plotting PDF & CDF

Margin of error :- How much deviations from original numbers is allowed in samples.

Chebyshev's Inequality :-

→ If we know that data is Gaussian Distributed, we can apply 68-95-99.7 rule.

→ But if we don't know the distribution, but we know μ, σ can we make assumptions like this ?

(μ: finite; σ: non zero & finite)

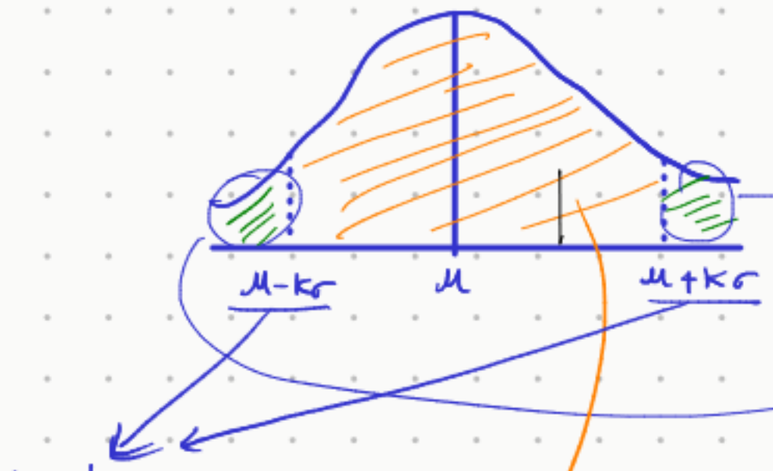
→ Chebyshev's Inequality stats that

'if X is a random variable with finite mean μ and non zero & finite σ

$$P(|X-\mu| \geq k\sigma) \leq \frac{1}{k^2}$$

k std deviations

68 → 1 std dev away
95 → 2 std dev away
99.7 → 3 std dev away



Probability of being in this regions

$$\Rightarrow P(\mu-k\sigma < X < \mu+k\sigma) \geq 1 - \frac{1}{k^2}$$

→ From Chebyshev's inequality we can conclude that 75% of any distributions lies within $[(\mu-2*\sigma)$ & $(\mu+2*\sigma)]$ and 90% within $[(\mu-3*\sigma)$ and $(\mu+3*\sigma)]$

→ In case the range is not symmetric, there are asymmetric variations of chebyshev as well.

Uniform Distributions:-

- Discrete
- Continuous.

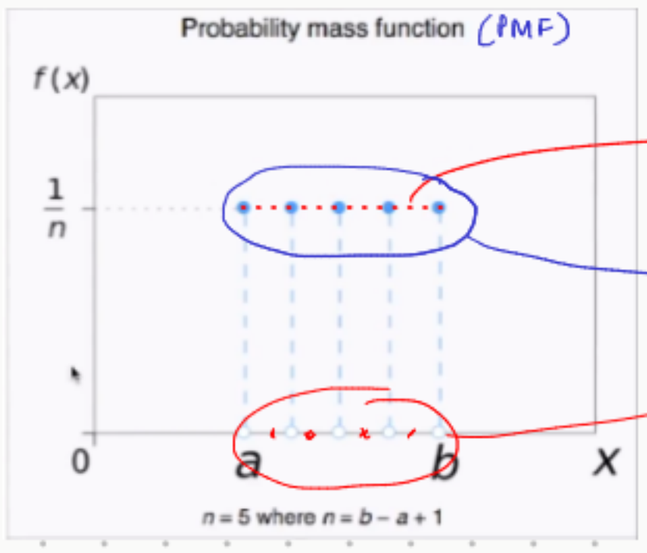
↳ Just like random variables

PDF (Probability Density Functions) are drawn for continuous random variables.
PMF (Probability Mass Functions) are drawn for discrete random variables.

→ Probabity of getting a value in dice is same for all values. This is called equiprobable

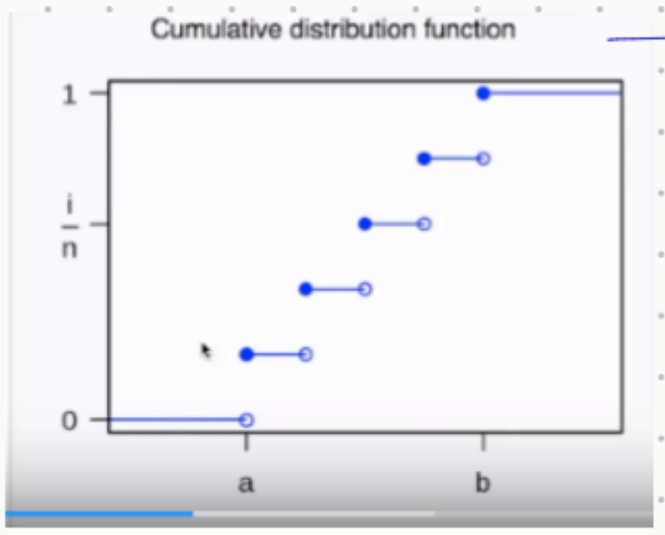
→ CRV has $N(\mu, \sigma)$ as parameters.
DRV has (a, b) as parameters.
↳ both are numbers.
and
$n = b-a+1$. Another formula for calculating $n = \frac{(b-a+k)}{k}$ where k is the common difference.
↳ number of outcomes.

→ In uniform distributions all the variables are equiprobable. ie, $\frac{1}{n}$



This line can't be drawn because these points don't exist in uniform distribution

∴ Only these does exist in PMF



→ CDF

→ This is called a non smooth function

Props of URV :-

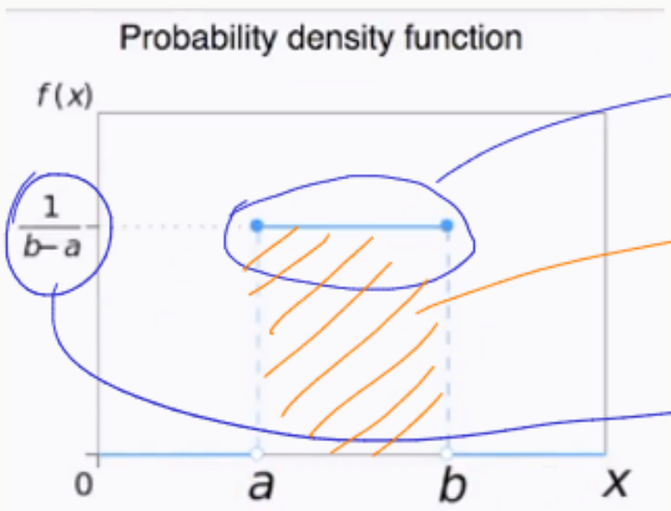
Mean = $\frac{a+b}{2}$

Median = $\frac{a+b}{2}$

Variance = $\frac{(b-a+1)^2-1}{12}$

Skewness = 0

Continuous Random Variable :-

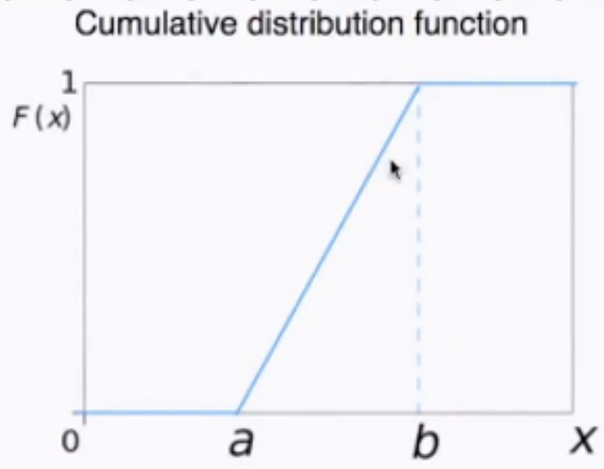


→ Line can exist here because it's continuous.

$\frac{1}{b-a}$ because this area has to be 1. The length is $b-a$. So the height is $\frac{1}{b-a}$.

Using maximum convention

CDF :-



Property of CURV :-

Parms: a, b

$$PDF = \begin{cases} \frac{1}{b-a} & \text{for } x \in [a,b] \\ 0 & \text{for everything else.} \end{cases}$$

Mean = $\frac{a+b}{2}$

Median = $\frac{a+b}{2}$

Variance = $\frac{(b-a)^2}{12}$

PDF → Density of data at the point

PMF → Probability of finding that point there

Random Number generators :-

→ Most random number generators generate uniform random variables unless explicitly specified.

→ python's random.random() picks a number b/w 0 & 1 with uniform probability.

ex :-

→ Picking 30 random values from iris

```
n=150 // Length of iris
m=30  // Sample size
p = m/n    // 0.2
sample_data = []
for i in range(0,n):
    if random.random() <= p :
        sample_data.append(Iris[i,:])
len(sample_data)
```

↳ ≈30

random.random() picks value b/w 1 & n with equal probability.
chances of getting val ≤ 0.2 = 20%.
30 of 150 = 20%.

→ An application of continuous uniform random variables.

Bernoulli & Binomial Distributions:-

→ Discrete Distributions
→ Bernoulli has only two outcomes. Value 1 has probability p & value 0 has probability of 1-p

ex:-

$X \sim$ Bernoulli ($p=0.5$)

$$PMF = \begin{cases} q=(1-p) & \text{for } k=0 \\ p & \text{for } k=1 \end{cases}$$

mean = p
variance = pq

Binomial Random Variable:-

Let X be tossing a coin

$X \sim$ Bernoulli ($p=0.5$)

Let Y = {Number of heads when coin is tossed n times (Let n=30)}

$Y \in \{0, 1, 2, 3, \ldots, 10\}$

$Y \sim$ Binomial (n, p)

n → number of trials
p → probability of value 1

Parameters :-

$$PMF = \binom{n}{k} p^k (1-p)^{n-k} \quad \text{for } P(Y=k)$$

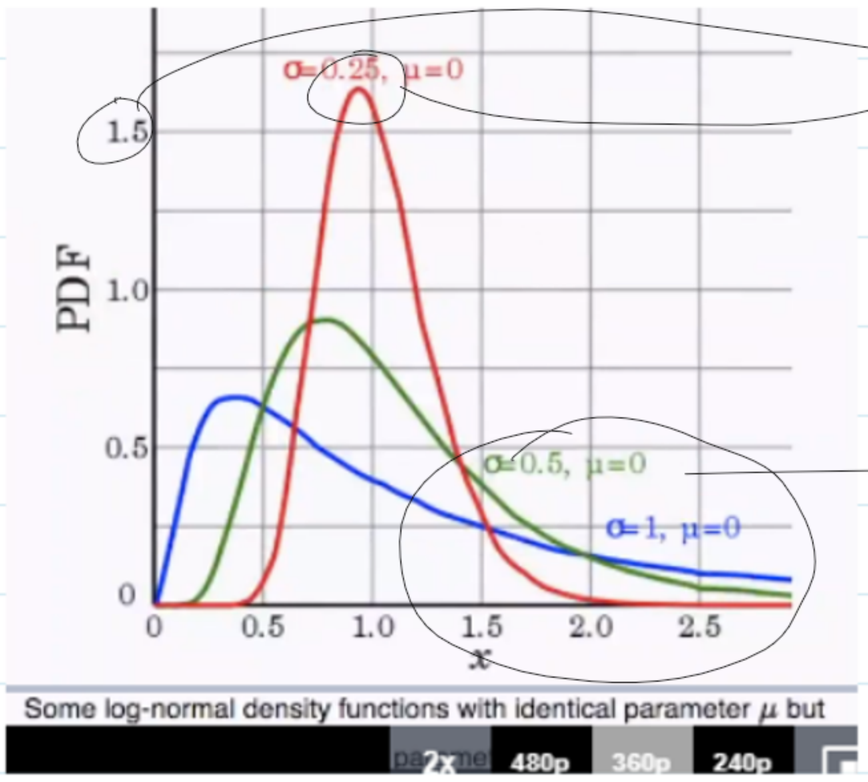
Not used often in Machine Learning.

→ If in Binomial distribution, n=1 ⇒ it's a Bernoulli's distribution.
→ Bernoulli's is unbiased.
→ Haberman's dataset is not Bernoulli Dist as Attributes of Haberman are continuous random variable & Bernoulli, Binomial is performed on discrete random variable having fixed outcomes.

Log Normal Distribution :-

$X \sim$ Log-Normal(μ, σ) if and only if $\log(X)$ is Normally distributed.

↳ Natural log aka ln



Some log-normal density functions with identical parameter μ but

The PDF here is > 1, but its integral, the CDF will be lesser than 1 because the region is narrow
(The PDF can be > 1 but it can't be >1 for bigger ranges)

as σ increases, curve becomes more skewed.

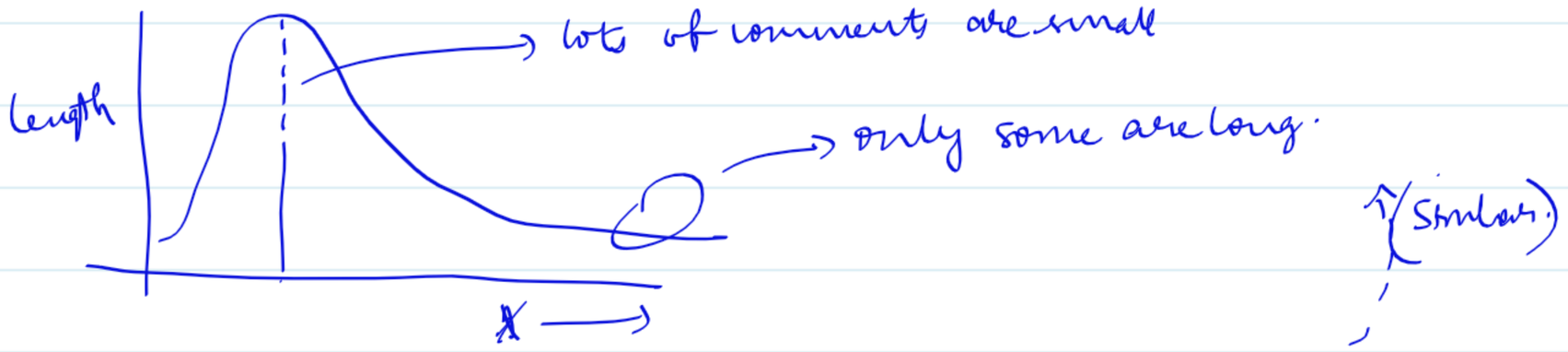
→ For log-normal, the mean is not where the curve peaks.

$$PDF = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{(\ln(x-\mu)^2)}{2\sigma^2}}$$

$$\text{Skewness} = (e^{\sigma^2} + 2)\sqrt{e^{\sigma^2} - 1}$$, Not 0 like seen in Normal Distribution.

Applications :-

→ Lengths of comments in internet forum posts follow log-normal distribution.



(Similar)

→ User's dwell time on online articles follows a log normal distribution.

→ In economics, 97-99% population's income follows log-normal. Distribution of higher income follows Pareto Distribution.

→ If we are given a log normal distribution, we can easily find its normal distribution & apply ML models.

Checking $X \sim$ Log-normal(μ, σ) or not :-

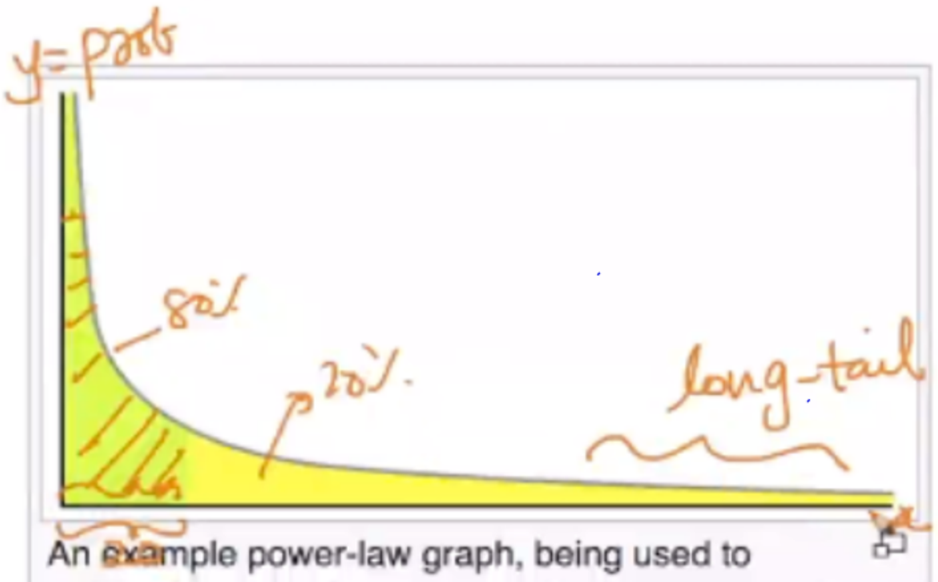
$x_1, x_2, x_3, x_4, \ldots, x_n$

$\ln(x_1)\ \ln(x_2) \ldots \ln(x_n)$

$(y_1)\ (y_2) \ldots (y_n)$ ⟶ QQ-plot ⟶ check if it's gaussian.

→ log normal more common than normal distributions IRL.

Power-Law :-



An example power-law graph, being used to demonstrate ranking of popularity. To the right is the long tail, and to the left are the few that dominate (also known as the 80–20 rule).

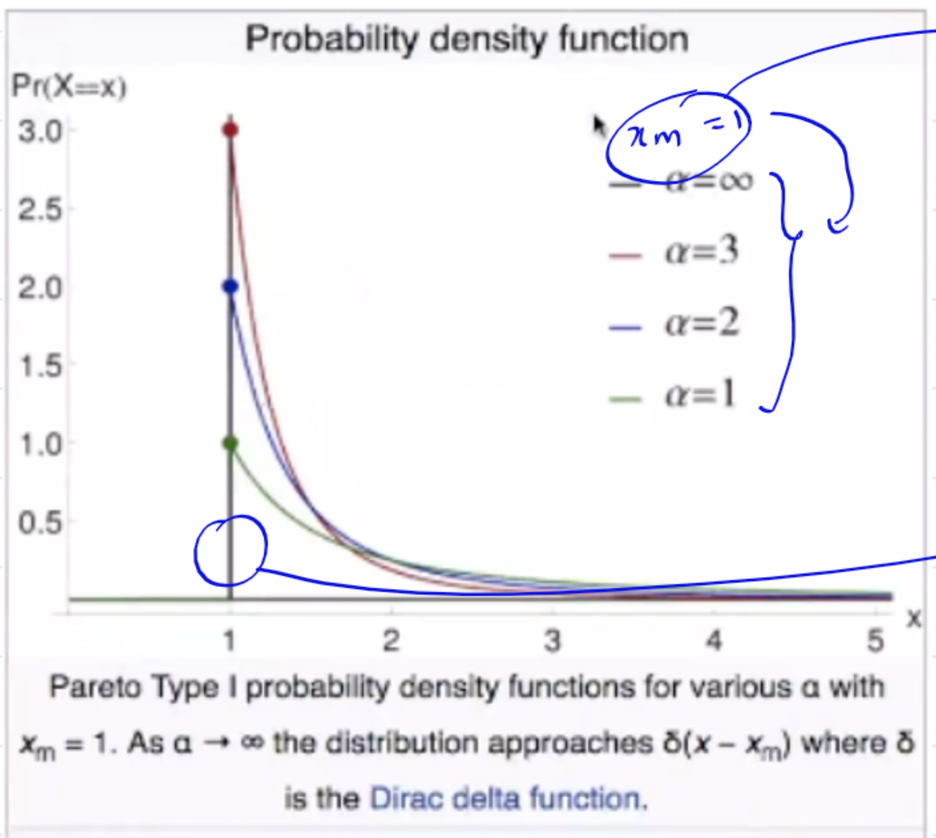
If distribution follows power-law, it's called as a Pareto distribution

Occurs a lot in nature.

Parameter of pareto distribution :-

$x_m > 0$ (scale) $\longrightarrow (\mu)$

$\alpha > 0$ (shape) $\longrightarrow (\sigma)$



Pareto Type I probability density functions for various α with $x_m = 1$. As $\alpha \to \infty$ the distribution approaches $\delta(x - x_m)$ where δ is the Dirac delta function.

peak

as $\alpha \downarrow \rightarrow$ tail size $\uparrow$

$\alpha = \infty$ (Dirac Delta function)

Applications :-

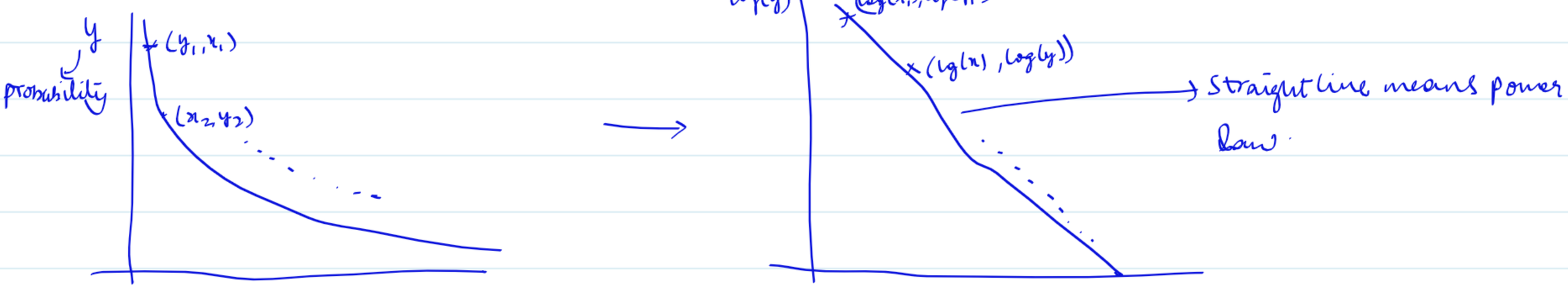
→ Size of human settlements (villages, cities etc.)

→ Filesize distribution of Internet.

→ HDD error rates.

How to check for pareto/powerlaw? :-

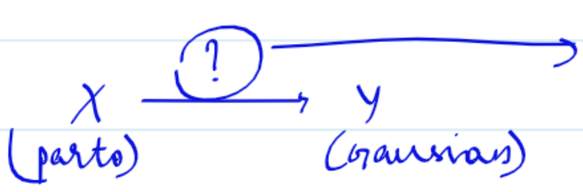
→ We use log-log plots.



→ To check for parto, use Q-Q plot with parto & observations.

Box Cox Transform :-

In ML we sometimes make assumptions that data is Gaussian Normal Distributed.



We use power transform / Box Cox transform for this conversion.

$X \rightarrow$ Pareto $\rightarrow [x_1, x_2, x_3, \ldots, x_n] \rightarrow X$

$Y \rightarrow$ Normal $\rightarrow [y_1, y_2, y_3, \ldots, y_n] \rightarrow Y$

Steps:-

① boxcox(X) $\rightarrow$ returns 'λ' [λ computation is complex and No need to memorize]

② $y_j = \begin{cases} \frac{x_i^{\lambda} - 1}{\lambda} & \text{if } \lambda \neq 0 \\ \log(x_i) & \text{if } \lambda = 0 \end{cases}$ $\forall\ j : 1 \rightarrow n$

$\Rightarrow x_i$ is log-normal

How!

scipy.stats.boxcox(x) $\rightarrow$ returns boxcox power transformed array and λ value.

Note:- Box Cox transform is not guaranteed to work on all pareto/powerlaw distributions. It only works on some of them. We need to apply boxcox & observe QQ plot to be certain.

Applications of Gaussian Distributions :-

→ useful in generating random numbers.

→ There are 100s of distributions.

Why are there so many? :- As well studied distribution gives a theoretical model on how a random variable behaves. There are 100s of distributions.

Most commonly used distributions :-

① Bernoulli
② Uniform Distribution
③ Binomial Distribution
④ Normal Distribution
⑤ Poisson Distribution
⑥ Exponential Distribution

Co-Variance :-

- If there are two datasets that are related, to find the relationship b/w these, we have 3 types of measurements

① covariance ② Pearson correlation coefficient ③ Spearman Rank correlation coefficient

ex:- students height & weight

	h	w
s_1	150	54
s_2	162	72
s_3	135	42
s_4	180	80

is there a relationship b/w height & weight?

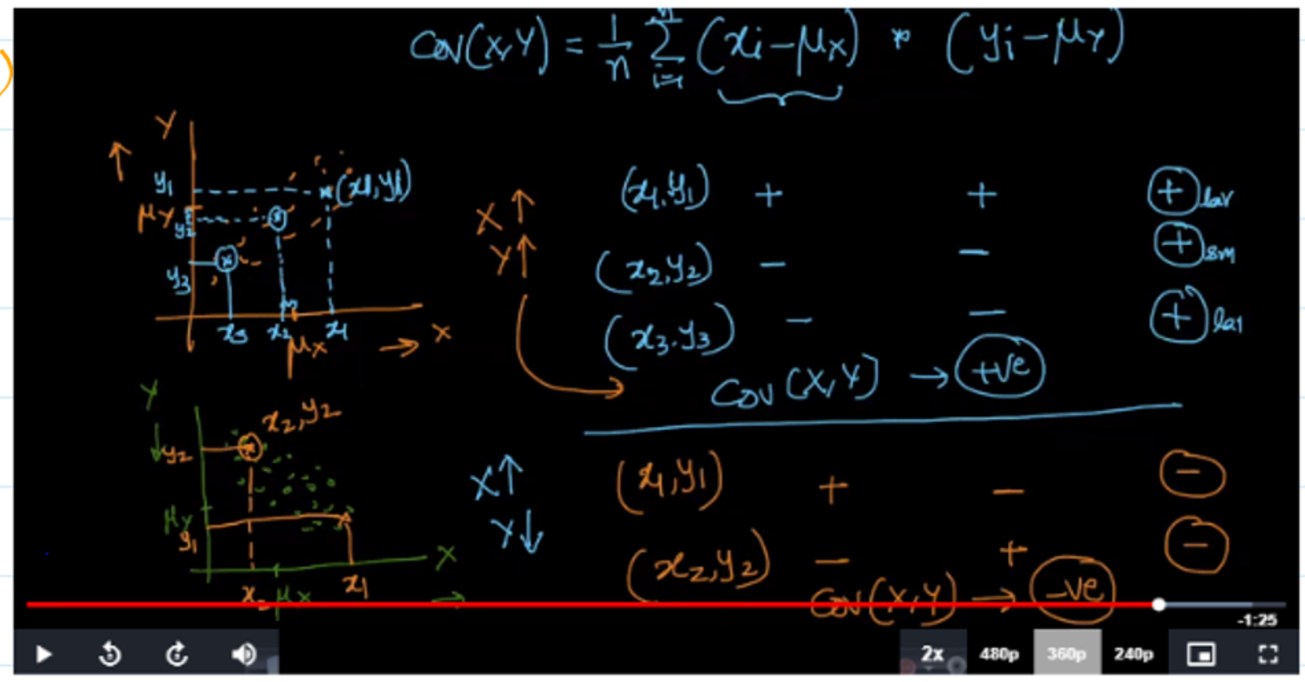
$$cov(x, y) = \frac{1}{n} \sum (x_i - \mu_x) * (y_i - \mu_y)$$

$$\left(\text{similar to } \quad var(x) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu_x) * (y_i - \mu_x) \right)$$

$cov(x, x) = var(x)$

$cov(x, y)$ = +ve if $x \uparrow \Rightarrow y \uparrow$ } (reasoning)

= -ve if $x \uparrow \Rightarrow y \downarrow$ }



Drawbacks :-

cov (height, weight) ≠ cov (height, weight)
(cms) (kg) (ft) (lbs)

same data but diff covariance.

The sign will be the same

Magnitude of covariance is also important as it tells how strongly related two RVs are.

→ Correlation = $\frac{\text{covariance}(x, y)}{\sigma_x \sigma_y}$

→ Extreme outliers drastically affect covariance. So they need to be removed with the noise & other outliers.

If we replace mean in covariance formula with medians it's called robust covariance.

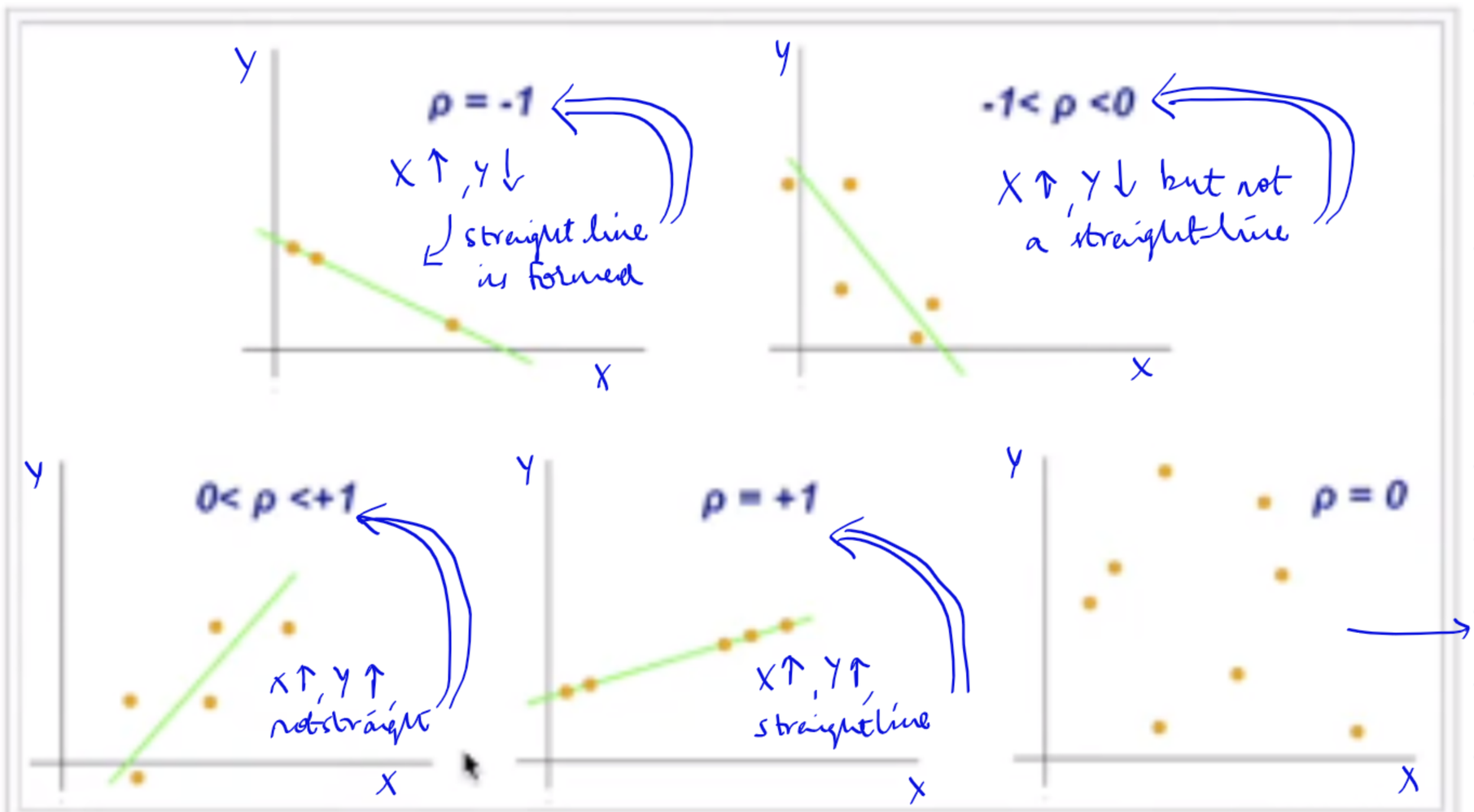
Pearson Correlation Coefficient (PCC) :-

$$\rho_{x,y} = \frac{cov(x, y)}{\sigma_x \sigma_y}$$

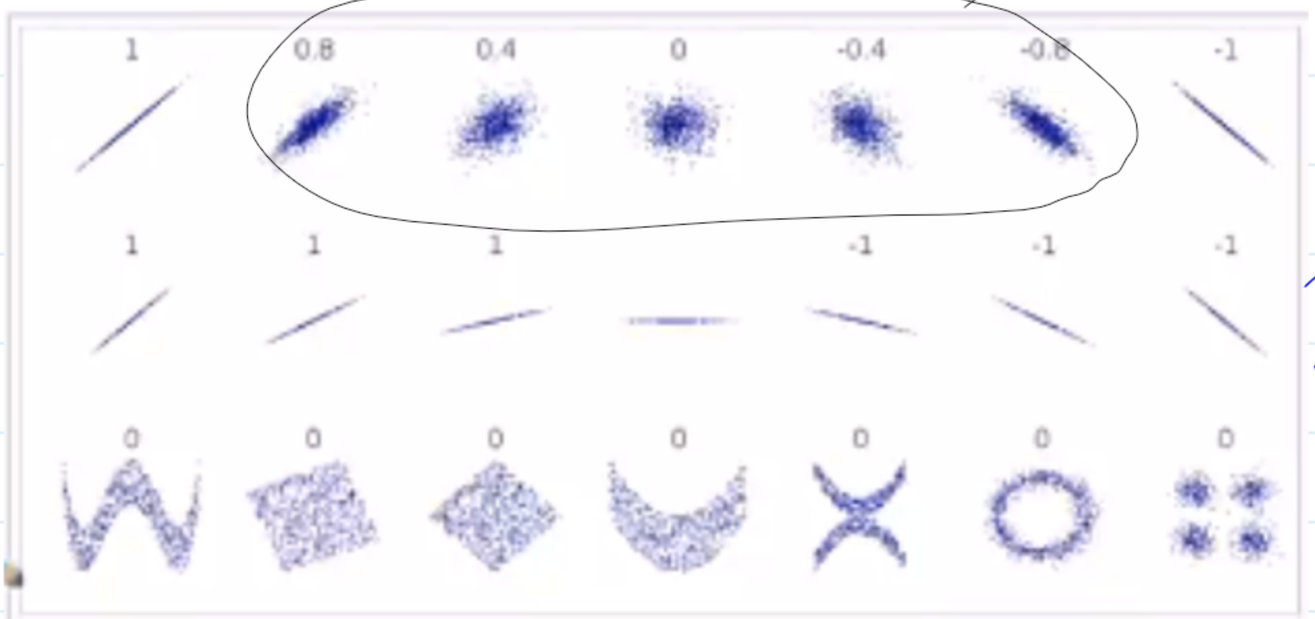
→ Cov does not take variability into consideration. This does.

→ $-1 \leq \rho \leq +1$

→ PCC is heavily biased towards linear dependant variables -



Examples of scatter diagrams with different values of correlation coefficient (ρ)



As it becomes flatter, value decreases.

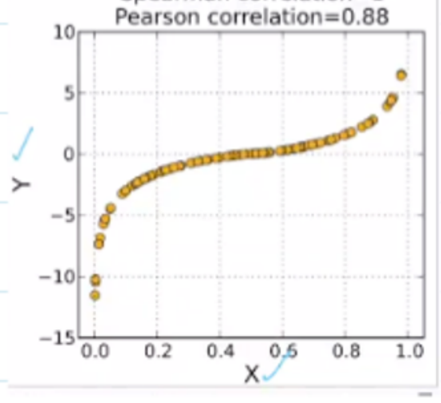
} → PCC does not care about the slope of the line.

} → Does not capture complex / non linear relationships

Monotonically non decreasing → if $x_2 > x_1$ and corresponding $y_2 \geq y_1$

monotonically increasing → if $x_2 > x_1$ and corresponding $y_2 > y_1$

} →



→ Correlation is what defines the change in one item when another item is modified.

Spearmen Correlation Coefficient:-

→ Useful when we have slightly non linear relationships.

ex:-

	X	Y	r_x	r_y
S_1	160	52	4	3
S_2	150	66	2	4
S_3	170	68	5	5
S_4	140	46	1	1
S_5	158	51	3	2

Smallest x

Largest x

During rank calculation, if two vals have same value, then their mean is taken as rank. Fractional rank (or) Same rank depends on the implementation.

r = Spearmen coefficient

$r = P_{r_x, r_y}$ → Ranks are strictly increasing.

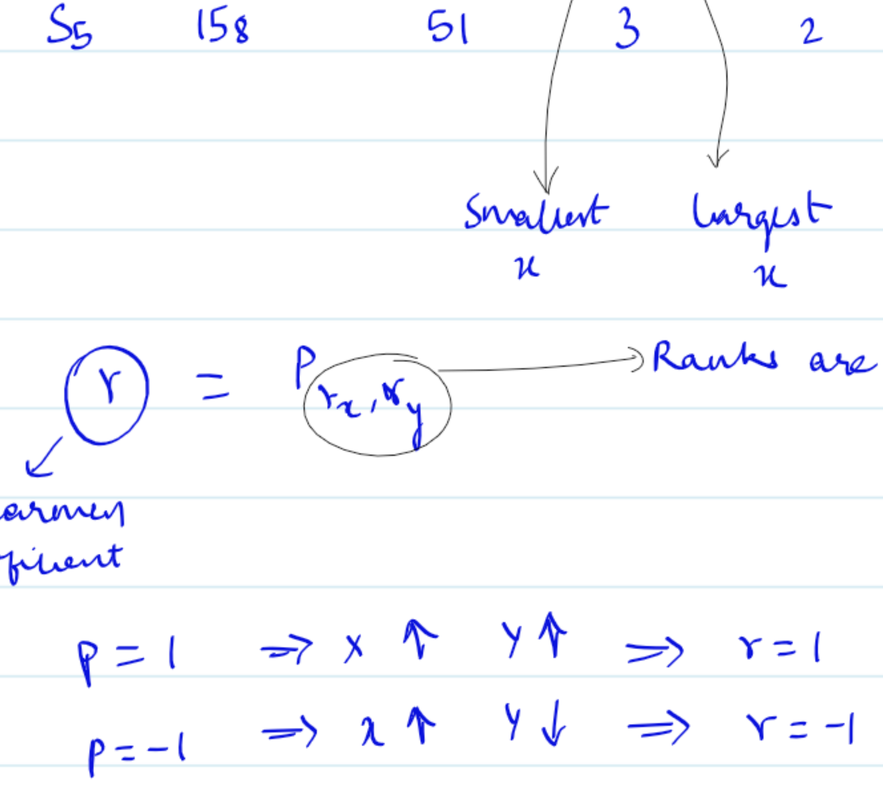
$p = 1 \Rightarrow x \uparrow \; y \uparrow \Rightarrow r = 1$

$p = -1 \Rightarrow x \uparrow \; y \downarrow \Rightarrow r = -1$

} → irrespective of whether they are linear or not. As long as there is monotonicity, it will work.

→ Correlation does not imply causation. For causations there are things like causal models.

ex:-



Applications of Correlations :-

ex:- (Salary & home square footage) → used by real estate brokers.
(# of years of education & income) → not looking for cause-Correlation only.
(Time spent on ecommerce website & money spent in the next 24 hours) → Amazon, Flipkart etc,
(# of unique visitors in one day & $ sales in one day) → ecommerce websites.
(dosage of medicine & blood sugar level of person) → medicine.

→ Ideas from causal inference can be used to determine causations.

Confidence Interval :-

Let X : heights with $x_1, x_2, x_3, \cdots, x_{10}$

Random Sample with size 10

estimate the population mean of $X = \mu$

point estimate ← { $\mu \approx \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$ → Simple Average.

μ → population mean
$\bar{x}$ → Sample mean.

As n increases, $\bar{x} \to \mu$

Point estimate of $\mu = \frac{1}{10} \sum_{i=1}^{10} x_i = 168.5$ cm (Assume)

We can say $\mu \in [162.1, 174.9]$ with 95% probability. → This is better than just saying this

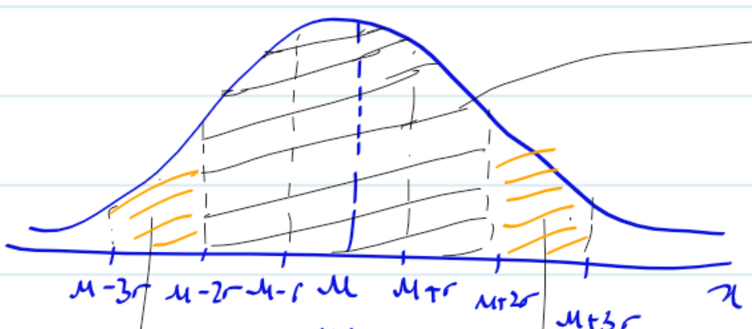
μ → population mean

Interval + Confidence → Confidence Interval.

This statement means that when new samples are taken & their means are calculated, there is a 95% chance that it falls in this interval. Doesn't mean that 95% of the data lies in this interval.

Computing Confidence Interval gives a distribution :- We can calculate CI easily if we know the underlying distribution.

Let $X \sim N(\mu, \sigma)$ with $\mu = 168$, $\sigma = 5$ cm

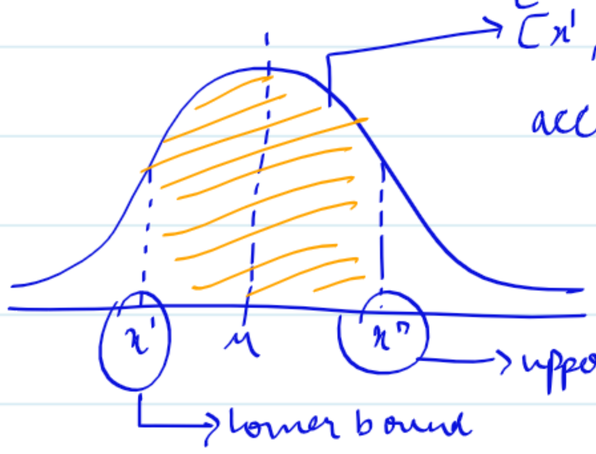


→ $(\mu - 2\sigma, \mu + 2\sigma)$ contains 95% of the data

95% → (Confidence interval)

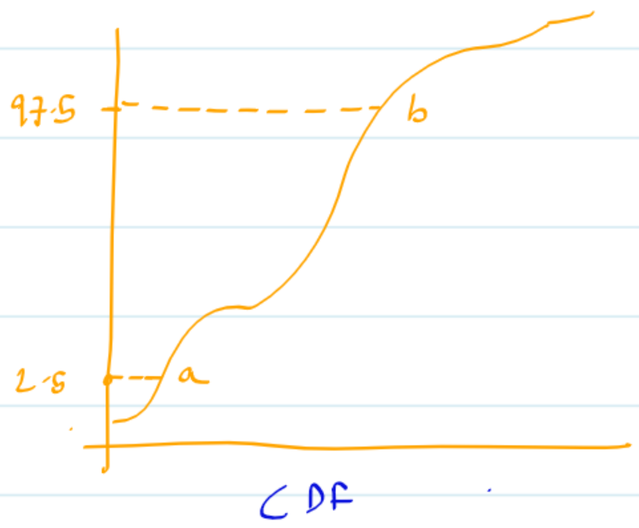
2.5% + 2.5% = 5%

→ Remaining 5% of the data.



[90%]
→ $[x', x'']$ with 90% confidence. We can use normal distribution tables to get the accurate values.

→ CDF can be used to calculate confidence interval.



[a, b] has confidence interval of 95%.

Confidence Interval of Random Variable :-

Let $X \sim (F)(\mu, \sigma)$ unknown dist. $\{x_1, x_2, x_3, \ldots, x_{10}\}$ $n = 10$

What is the 95% CI of μ of this dist?

↳ case ① :- σ is given

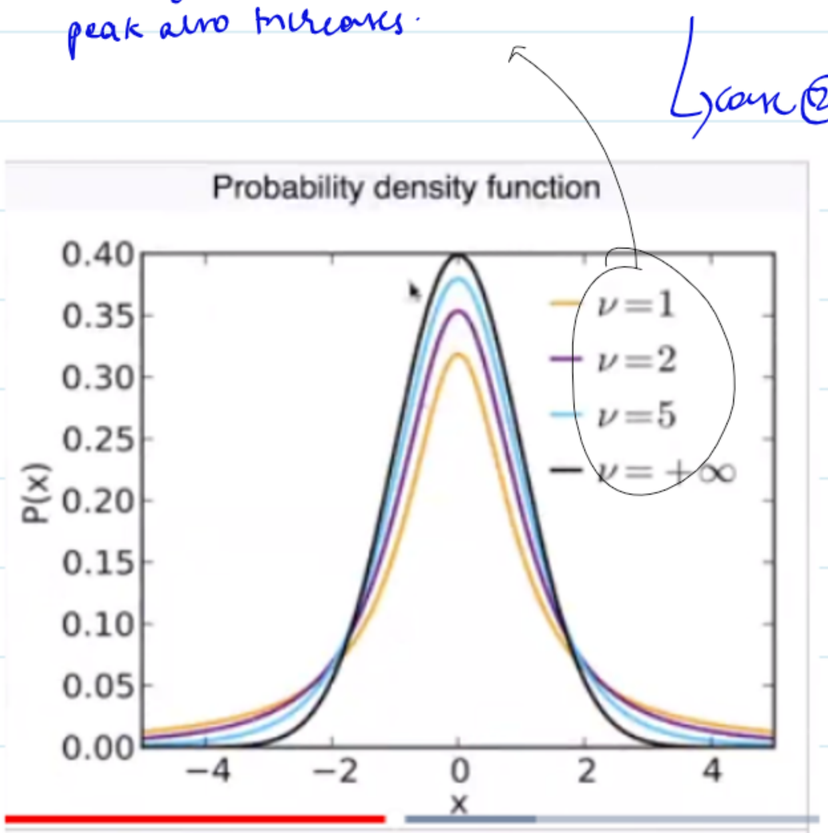
Using CLT, $\bar{x} = \frac{1}{10} \sum_{i=1}^{n} x_i$

$\bar{x} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$ (μ: mean, σ: standard deviation)

$\Rightarrow \mu \in \left[\bar{x} - \frac{2\sigma}{\sqrt{n}} \;,\; \bar{x} + \frac{2\sigma}{\sqrt{n}}\right]$ with 95% of CI.

$(\mu - 2\sigma)$ $(\mu + 2\sigma)$

Looks like a gaussian curve.
As degrees of freedom increases, peak also increases.



t-dist is created to estimate mean of RV when σ is unknown.

↳ case ② :- σ is also unknown. ⇒ CLT can't be applied.

In this case we use t-distribution. It states

$\bar{x} \sim t(n-1)$

↳ degrees of freedom.

→ To estimate stuff like σ / 90th percentile etc; we use Bootstrap based Confidence Intervals.

Confidence Interval Using Bootstrap :-

→ prog based techniques for CI & other statistics

Let $X \sim F(\mu, \sigma)$ task: estimate CI for median of X

X: sample size 10 $\{x_1, x_2, x_3, \ldots, x_{10}\}$

Using only this sample, compute median of X.

Steps:- ① Generate new samples using this sample. ⇒ Some values may repeat. It's called sampling with repetition.

$x_1^{(1)}, x_2^{(1)}, x_3^{(1)}, x_4^{(1)}, \ldots, x_m^{(1)}$ such that $m \leq n$

↑ Random sample of size m generated from S

→ We can use uniform random variables to pick these values ($\cup(1, n)$ - There can be reps)

$S = \{x_1, x_2, x_3, \ldots, x_n\}$

↓ Using sampling with repitition. K samples.

Size $m \leq n$.

$S_1 : x_1^{(1)}, x_2^{(1)}, x_3^{(1)}, \ldots, x_m^{(1)} \longrightarrow m_1$

$S_2 : x_1^{(2)}, x_2^{(2)}, x_3^{(2)}, \ldots, x_m^{(2)} \longrightarrow m_2$

S_3 ⋮

$S_k : x_1^{(k)}, x_2^{(k)}, x_3^{(k)}, \ldots, x_m^{(k)} \longrightarrow m_k$

These are known as Bootstrap samples.

medians. If variances are computed instead, we can calculate variances CI.

Let say we get $m_1, m_2, \ldots, m_{1000}$

↓ sort them in increasing order.

$m_1', m_2', m_3', \ldots, m_{1000}'$

$m_1', \ldots m_{25}' \ldots m_{975}' \ldots, m_{1000}'$

⇓

95% confidence interval = $[m_{25}', m_{975}']$

→ This technique is called non-parametric technique. (No assumptions are made about distribution)

→ Used a Lot in ML.

How? :-

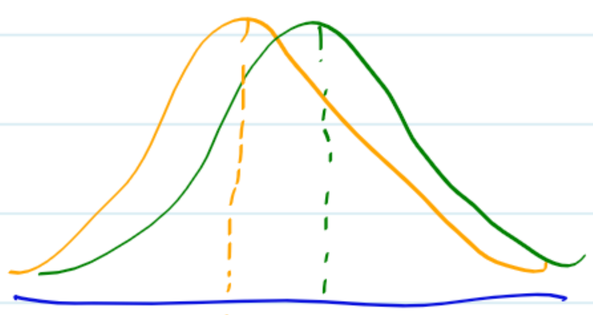
```
from sklearn.utils import resample
⋮
s = resample(x, n_samples = m)
```

→ To generate bootstrap sample.

m → Sample size.

Hypothesis Testing :-

Assume there are two classes cl1, cl2. Both class heights are taken. Histograms are plotted.



Question :- Is there a difference in heights of both classes.

Sol :- ① Choosing a test statistic

↳ $\mu_2 - \mu_1$ → if it's 0, then there's no difference

② Null Hypothesis (H_0) :-

This uses proof by contradiction.

H_0 : Assume there's no diff in μ_1, μ_2.

Alternative Hypothesis (H_1) :- Assume there's diff in μ_1 & μ_2.

We assume H_0 is true & prove H_1 wrong (or) We assume H_1 is true & prove H_0 is wrong.

③ p-value :- It says what is the prob of observing $(\mu_2 - \mu_1)$ if H_0 is true.

⇒ Assume H_0 is true. ⇒ $\mu_1 \approx \mu_2$ ⇒ $cl1 \approx cl2$

if p-value = 0.9 ⇒ probability of observing $u_2 - u_1$ is 0.9 if H_0 is true

Since p-val is high, we accept H_0 is true.

if p-value = 0.05 ⇒ 5% chance that occurs if H_0 is true.

Hypothesis Testing (Additional Video):-

Cl1 → 50 students → u_1 (mean)

Cl2 → 50 students → u_2 (mean)

Step 0:- Δ (or) $X = u_2 - u_1 = 10$ (assume) → This is an observation. Ground truth.

Step 1:- What is the probability of observing a value of $X \geq 10$ cm if there was no difference in class heights. given a hypothesis

$P(X \geq 10 \text{ cm} \mid H_0)$

- $X \geq 10$ cm → Teststatistic
- H_0 → Null hypothesis

if p value is small = 0.01 (or) 1%.

→ < 5% is typically considered small

$\Rightarrow P(X \geq 10 \text{ cm} \mid H_0) = 0.01$ or 1%.

⇓

true. (Since its an observation that's been made already)

↓

Since the probability is small, implies H_0 is less probable ⇒ H_0 may not be true.

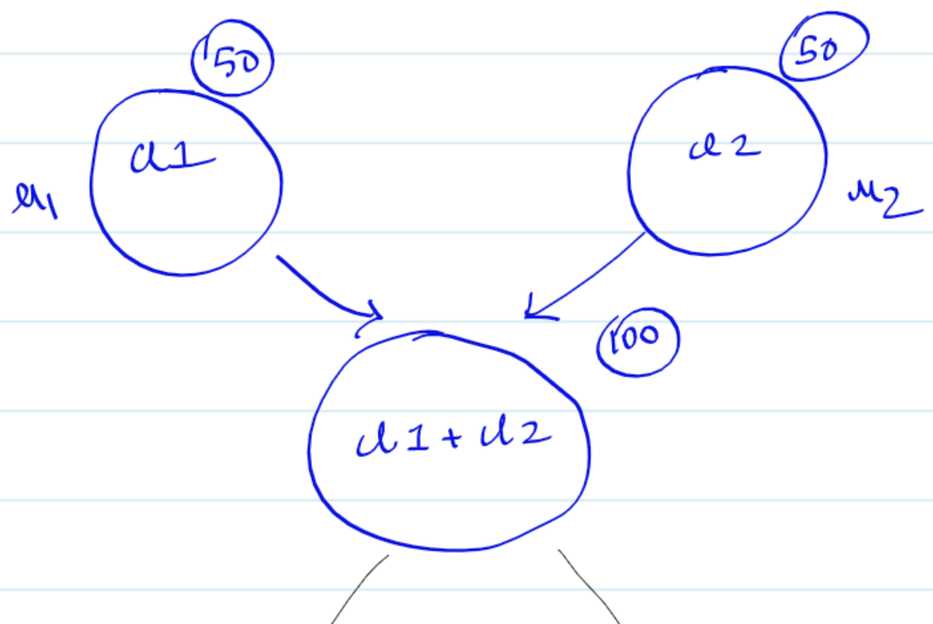
∴ We reject the null hypothesis (H_0)

⇒ The original assumption that there's no difference is incorrect

⇒ There is difference b/w the heights.

How to calculate p-value? We only have sample data from two classes

The way we calculate is, we use resampling & permutationstesting.



$X = u_2 - u_1$

Null hypothesis (H_0):- There is no diff in class heights. We simulate null hypothesis on class heights (combined)

repeated multiple times

$\delta_1, \delta_2, \delta_3, \ldots, \delta_{10,000}$

Simulated X's with H_0 as true.

$\delta = u_2' - u_1'$

↳ simulated diff in class heights with sample size = 50 & with H_0 = true

↳ Since randomly picked.

↳ Sorted: $\delta_1', \delta_2', \delta_3', \ldots, \delta_{10000}'$

≤10: 8000 ; ≥10: 2000

$\Rightarrow P(X \geq 10 \text{ cm} \mid H_0) = 20\% \Rightarrow H_0$ is rejected.

$\delta'_1, \delta'_2, \delta'_3, \ldots\ldots, \delta'_{10000}$

≤ 10 $\quad$ ≥ 10

9900 $\quad$ 100 $\Rightarrow P(X \geq 10cm | H_0) = 1\%$. $\Rightarrow H_0$ is true. Hence accepted.

→ How to pick null hypothesis? We pick H_0 such that it is easy to simulate. Hence we picked that there is no diff b/w the heights.

Hypothesis Testing:-

ex ①:- Given a coin, determine if coin is biased towards heads or not.

$\Rightarrow P(H) > 0.5$

Not biased $\Rightarrow P(H) = 0.5$

Solving using base probability:-

experiment:- Flip a coin 5 times and count number of heads

X → random variable. Also called a test statistic

performing experiment:- H, H, H, H, H $\Rightarrow$ $X = 5$ → observation by experimentation.

$P(X = 5 \mid \text{coin is not biased towards heads}) = P(obs \mid H_0)$

Observation. Assumption.

⇓

Null Hypothesis (H_0):- Coin is not biased towards heads.

$P(X = 5 \mid \overline{H_0}) \Rightarrow \frac{1}{2} * \frac{1}{2} * \frac{1}{2} * \frac{1}{2} * \frac{1}{2} = \frac{1}{2^5} \approx 0.03 = 3\%$.

Five heads in 5 tosses.

→ coin is not biased, $P(H) = \frac{1}{2}$

$P(X = 5 \mid H_0) = 3\%$. $\Rightarrow$ There is 3% chance of getting 5 heads in 5 flips if the coin is not biased

p-value ← [Probability (Observation by experiment | assumption is true) = 3%. → low (<5%)

Since it's low, H_0 may be incorrect.

$\Rightarrow$ Null hypothesis is not true & we reject hypothesis

$\Rightarrow$ We reject the idea that coin is not biased.

Alternative hypothesis (H_1) ← [$\Rightarrow$ Coin IS biased.

Rejecting H_0 = Accepting H_1

Accepting H_1 = Rejecting H_0

→ This expt is reliant on how many times coin has been flipped (Sample Size)

If coin was flipped 3 times, then $P(X \mid H_0) = \frac{1}{2^3} = \frac{1}{8} = (12.5\%)$ → Not low. Significant

$\Rightarrow$ We accept H_0

$\Rightarrow$ Coin is not biased.

3 important things in Hypothesis testing:-

1. Design of the experiment
2. Defining Null hypothesis.
3. Design of test statistic 'X'.

→ Typical p-value in scientific computing = 0.05%. In some cases 0.01 is also used -
↳ when you don't want any outliers at all.

→ In olden days, when there wasn't enough computing power, people used to use distributions & use 68-95-99/ chebyshev to get the pvalue.

K-S Test for Similarity of two distributions:-

Assume $X_1 : [x_1, x_2 \dots x_n]$ (n size sample)

$X_2 : [x_1', x_2', \dots x_m']$ (m size sample)

Question :- Are both X_1 & X_2 coming from the same distributions.

How? :- We plot the CDF of both samples. If they both overlap, that means they belong to the same distributions.

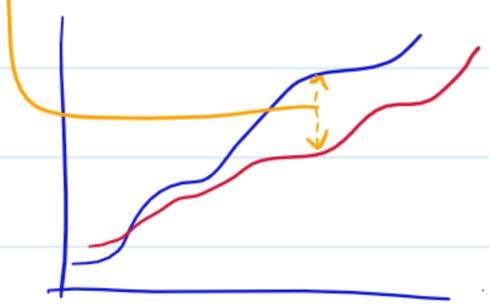
Null hypothesis, H_0 : X_1 & X_2 have the same distribution

Test statistic, difference b/w the CDFs at any given point

=> if CDFs completely overlap, difference = 0

Test statistic, $D_{n,m} = \sup_x |F_{1,n}(x) - F_{2,m}(x)|$

supremum.
The maximal value



→ KS test was used before the concept of resampling & hypothesis testing. So after lot of research, they came up with closed form equations

Null hypothesis is rejected at level α, if $D_{n,m} > c(\alpha)\sqrt{\frac{n+m}{nm}}$ [α = required p-level]

α	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.001
$c(\alpha)$	1.073	1.138	1.224	1.358	1.48	1.628	1.731	1.949

ex :- if $n = 1000$, $m = 5000$, $\alpha = 0.05$, then if $D_{n,m} > 1.358\sqrt{\frac{1000+5000}{1000*5000}}$

=> if $D_{n,m} > 0.047$, we reject H_0 at 0.05 sig level.
↳ Max diff b/w CDFs

In general $C(\alpha) = \sqrt{-\frac{1}{2}\ln\alpha}$

QQ plot is a graphical method whereas as KS test gives us what level of 'x' we are accepting/rejecting.

Implementing KS Test:-

from scipy import stats.

stats.kstest(dataset, 'norm') → type of dist to compare against.

↳ Returns two values → ① $D_{n,m}$
② p-value

→ KS Test is a weak test as it is not designed for testing normality. It is better to use Anderson-Darling test to test for normality. KS Test is good for comparing distributions.

→ As long as p-value > 0.05, we accept the null hypothesis.

Hypothesis testing ②:-

task: city 1 city 2

determine if $\mu(\text{heights}_{\text{city 1}}) = \mu(\text{heights}_{\text{city 2}})$

Population. Not sample.

Getting actual population is expensive. So we use sample mean.

how? :-

city 1 → random ← city 2

$\begin{bmatrix} h_1 \\ h_2 \\ \vdots \\ h_{50} \end{bmatrix}$ $\begin{bmatrix} h_1' \\ h_2' \\ \vdots \\ h_{50}' \end{bmatrix}$

μ_1 → Sample mean ← μ_2

ex:- 162 cm 167 cm

test statistic :- $X = \mu_2 - \mu_1 = 167 - 162 = 5$ cm

Null hypothesis (H_0) :- No diff in population means.

Compute :- $p(x = 5\text{cm} \mid H_0)$ [Probab of observing diff of 5cm in sample mean heights of size 50 b/w C_1 & C_2 if there is no diff in mean heights.]

↳ Difference in sample means with sample size 50 (observation)

case 1 :- $p(x = 5 \mid H_0) = 0.2 = 20\%$.

↳ There is 20% chance of observing diff in means as 5 if H_0 is true

There is no difference in heights.

Significant ⇒ Assumption is true ∴ H_0 is accepted.

case 2 :- $P(x = 5 \mid H_0) = 0.02 = 2\%$. ⇒ low. Assumption is incorrect. Alternative Hypothesis (H_1) is accepted ⇒ Heights are not the same.

Resampling & Permutation test for cities example :-

$x = \mu_1 - \mu_2$ → diff in sample means (Let it be 5 cm)

H_0 = Null hypothesis (No difference in means · For this example)

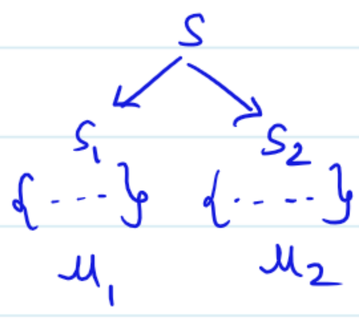
C_1 C_2

$$\begin{bmatrix} h_1 \\ h_2 \\ h_3 \\ \vdots \\ h_{50} \end{bmatrix} \qquad \begin{bmatrix} h_1' \\ h_2' \\ h_3' \\ \vdots \\ h_{50}' \end{bmatrix}$$

Step ① :- $S = \{h_1, h_2, h_3, \ldots, h_{50}, h_1', h_2', h_3', \ldots, h_{50}'\}$

$C_1 \cup C_2$

Step ② :-



→ Randomly selected two sets 50 size each (Resampling)

★ (We are simulating the null hypothesis) ★

$\mu_2 - \mu_1 \longrightarrow 3$ cm (δ')

repeat, $\mu_2 - \mu_1 \longrightarrow -2$ cm (δ^2)

$\mu_2 - \mu_1 \longrightarrow 1$ cm (δ^3)

k times, $\mu_2 - \mu_1 \longrightarrow 6$ cm (δ^K), Let $K = 1000$

Step ③ :- Sort δ_i s

$\delta_1' \le \delta_2' \le \delta_3' \le \delta_4' \ldots\ldots \le \delta_K'$

case ① :- $P(\text{diff} \ge 5\text{cm} \mid H_0)$

$\delta_1' \le \delta_2' \le \delta_3' \le \delta_4' \ldots\ldots \le \delta_{950}' \le \ldots\ldots \delta_{1000}'$

≤ 5 cm — 80% of the data; > 5 cm — 20% of the data

$P(\text{Obs-diff} \mid H_0) = P(x \ge 5\text{cm} \mid H_0) =$ (20%) → p-value

↳ significant (> 5%)

∴ Assumptions must be true · H_0 is accepted ·

case ② :- $\delta_1' \le \delta_2' \le \delta_3' \le \delta_4' \ldots\ldots\ldots \le \delta_{970}' \le \ldots\ldots < \delta_{1000}'$ → Simulated values originated from assuming that null hypothesis is true ·

97% of the data — 3% of the data

$P(x \ge 5\text{cm} \mid H_0) = 3\%$ = p-value is not significant · Assumptions must be incorrect · H_0 is rejected ·

→ if $P(\text{obs} \mid H_0)$ is small, we reject H_0 because the observation is real but we got a small probability implying H_0 is false ·

→ if $P(\text{obs} \mid H_0)$ is very high in this case, it implies error in sample & if we increase sample size, well be fine ·