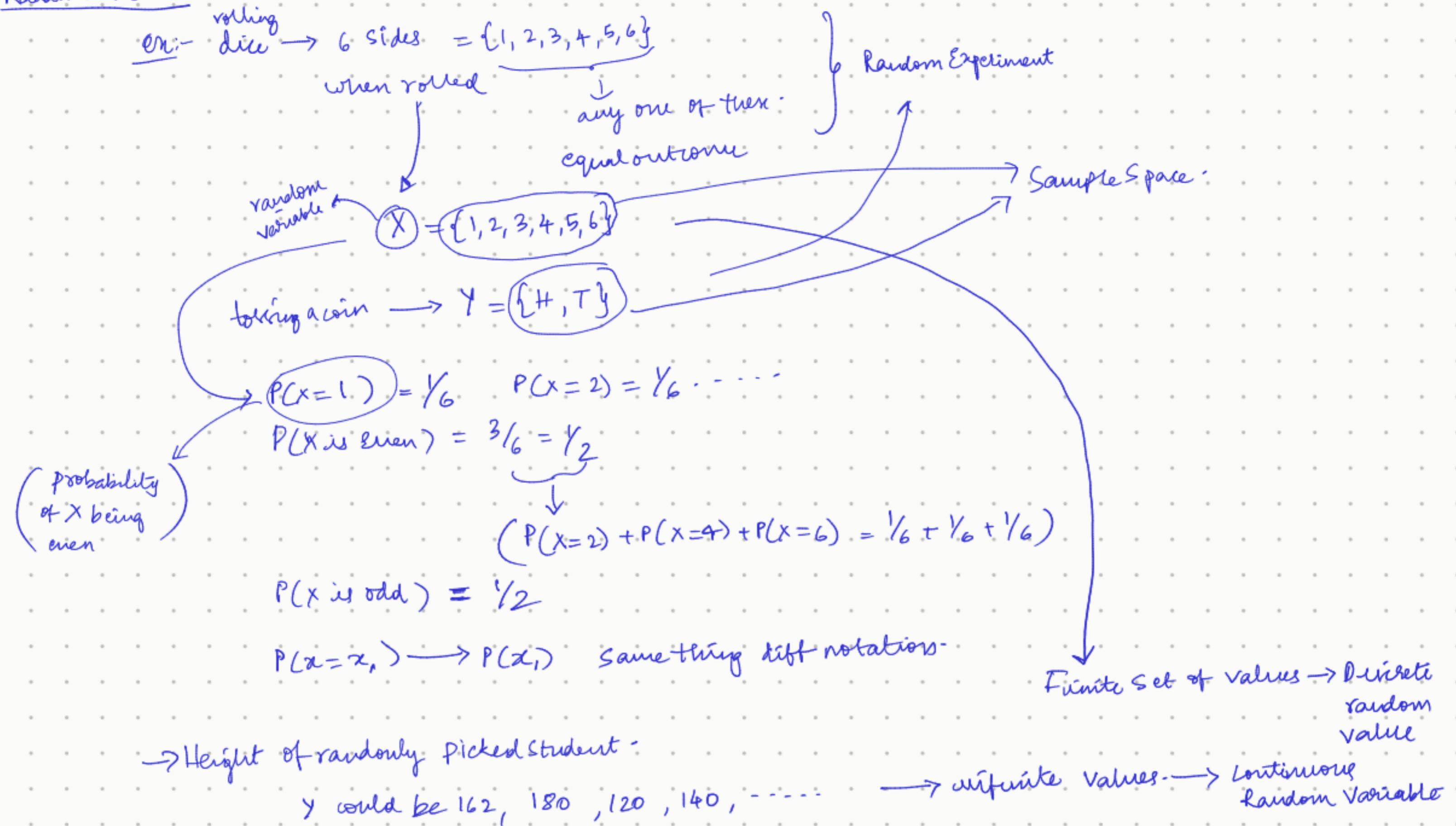


Probability & Statistics :-

Random Variable :-



Outliers :-

y : height of student-

$$\{122.2, 146.4, 132.5, \dots, 121.2, 156.3, 92.7, \dots\}$$

121.2, 156.3, 92.7

would be an outlier

outliers \rightarrow could be human error

(or)

actual height

\rightarrow Outliers can corrupt data

\rightarrow A discrete value is obtained by counting

\rightarrow A continuous value is obtained by measuring.

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

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

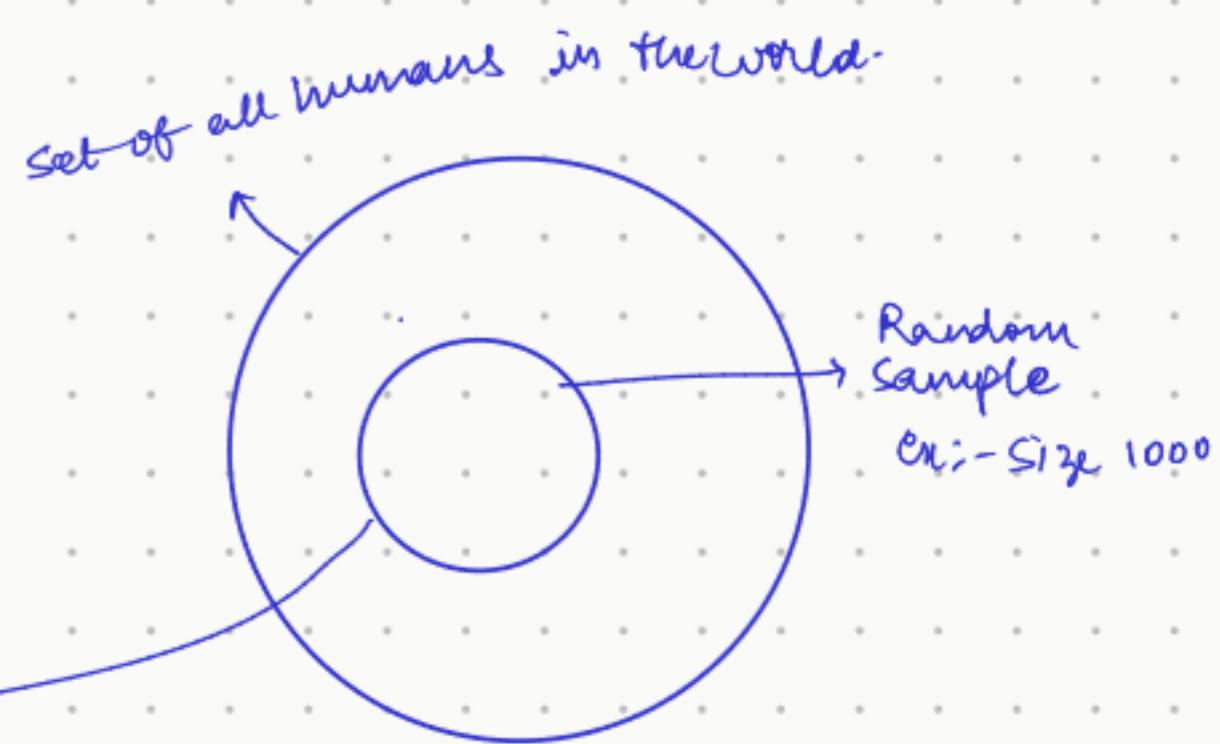
Populations & Sample :-

\rightarrow Estimating the average height of human

$$\bar{x} = \frac{1}{\text{Pop}} \sum_{i=1}^{\text{Pop}} h_i \quad (\text{IMPOSSIBLE})$$

So we estimate:

often represented by $\bar{x} = \frac{1}{1000} \sum_{i=1}^{1000} h_i$



→ As sample size increases

$$\bar{x} = \mu$$

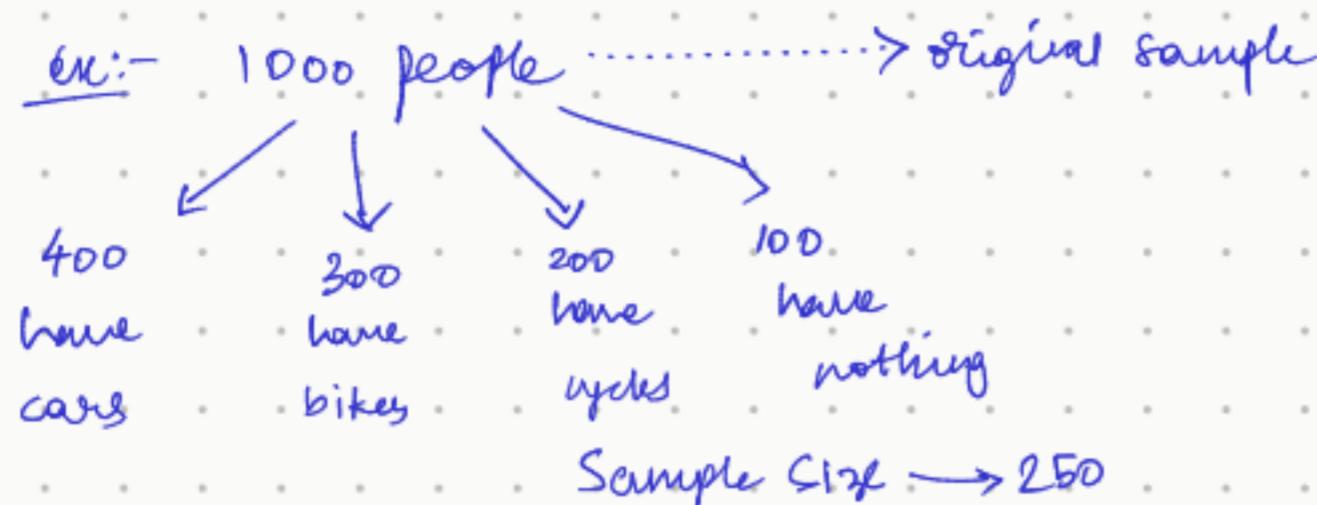
↓
Sample mean

original mean

Sampling is of two types:-

(i) Simple Sampling

(ii) Stratified Sampling → Unbiased Sampling & more accurate results.



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 in proportion to all classes

Gaussian Distribution:- (AKA Normal Distribution)

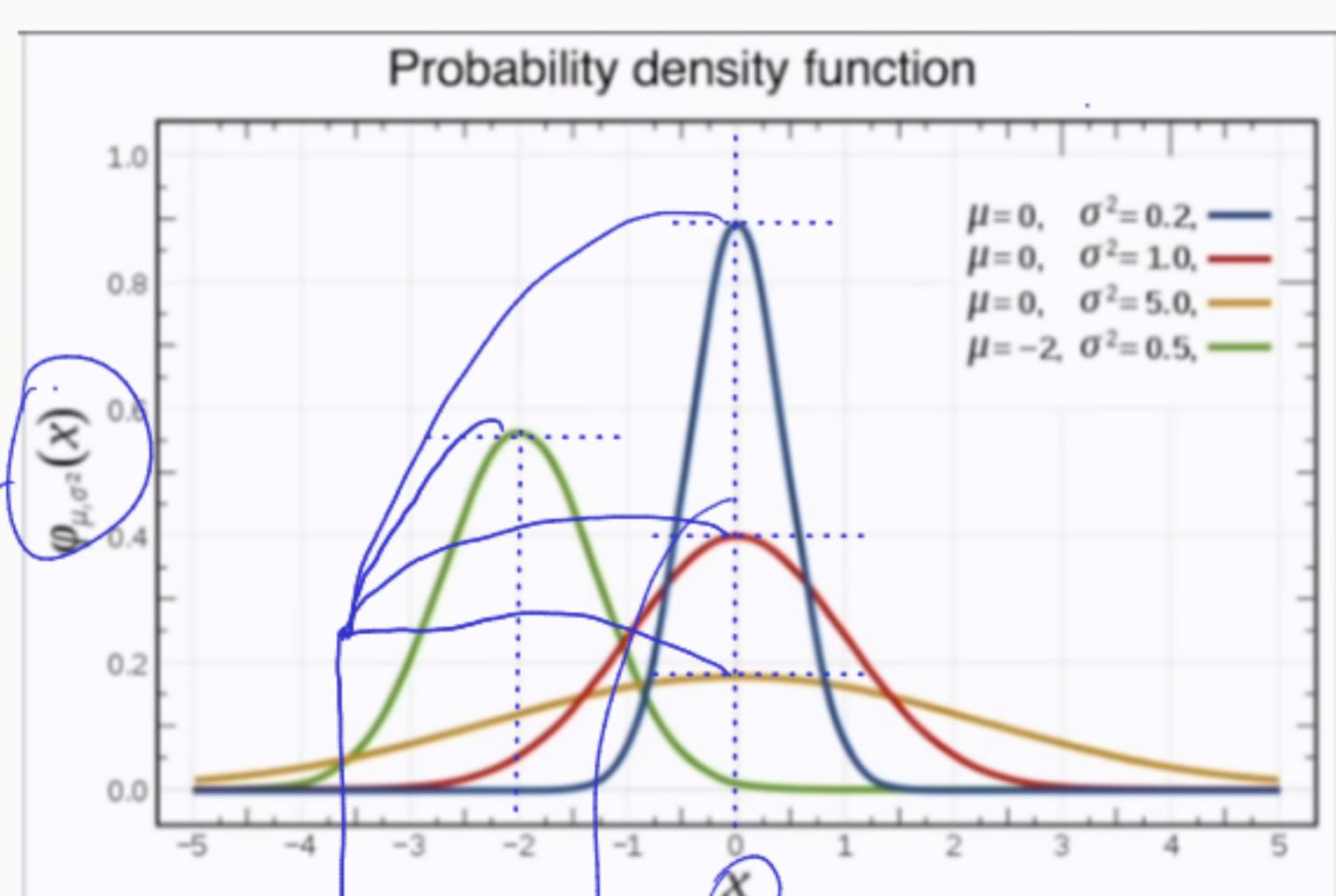
→ If X is a continuous random variable & X has a PDF curve (\mathcal{N}), then we say X has a Gaussian distribution.

Why learn?

- Stuff in nature tends to follow G.D.
- Height, weight of people follow it. (Natural phenomenon)
- They are simple models that summarize R.V.

PDF = Probability Density Functions

Mean, Median & Mode are all same for G.D. There are equal number of measurements above & below the mean



μ = mean
 σ^2 = variance \rightarrow parameters

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

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

Variance is a measure of spread.

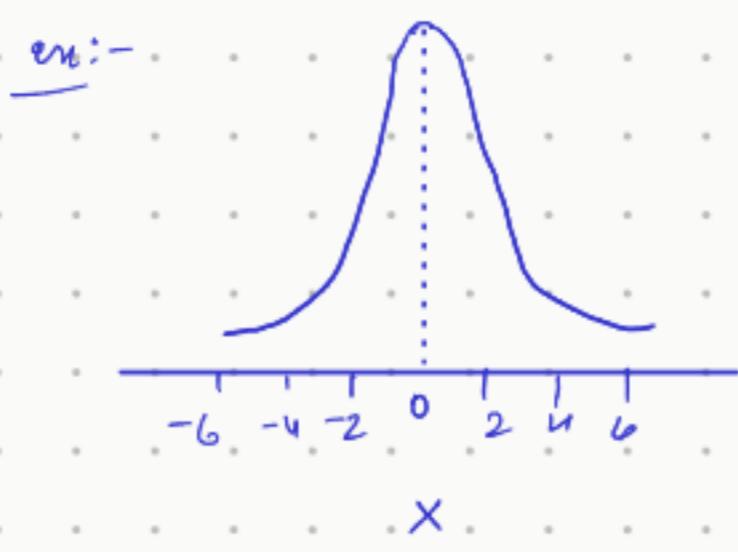
Continuous Random Variable

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

The peak of curve is usually at ' μ '.

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

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



$\rightarrow X \sim N(0, 1)$

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

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

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

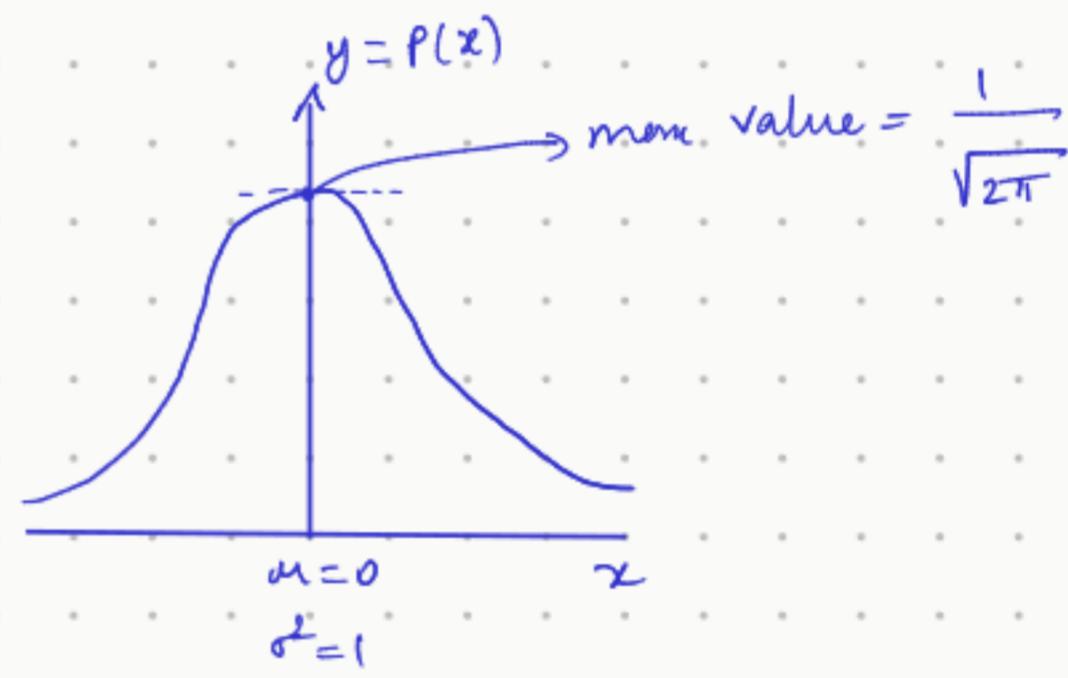
constants

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

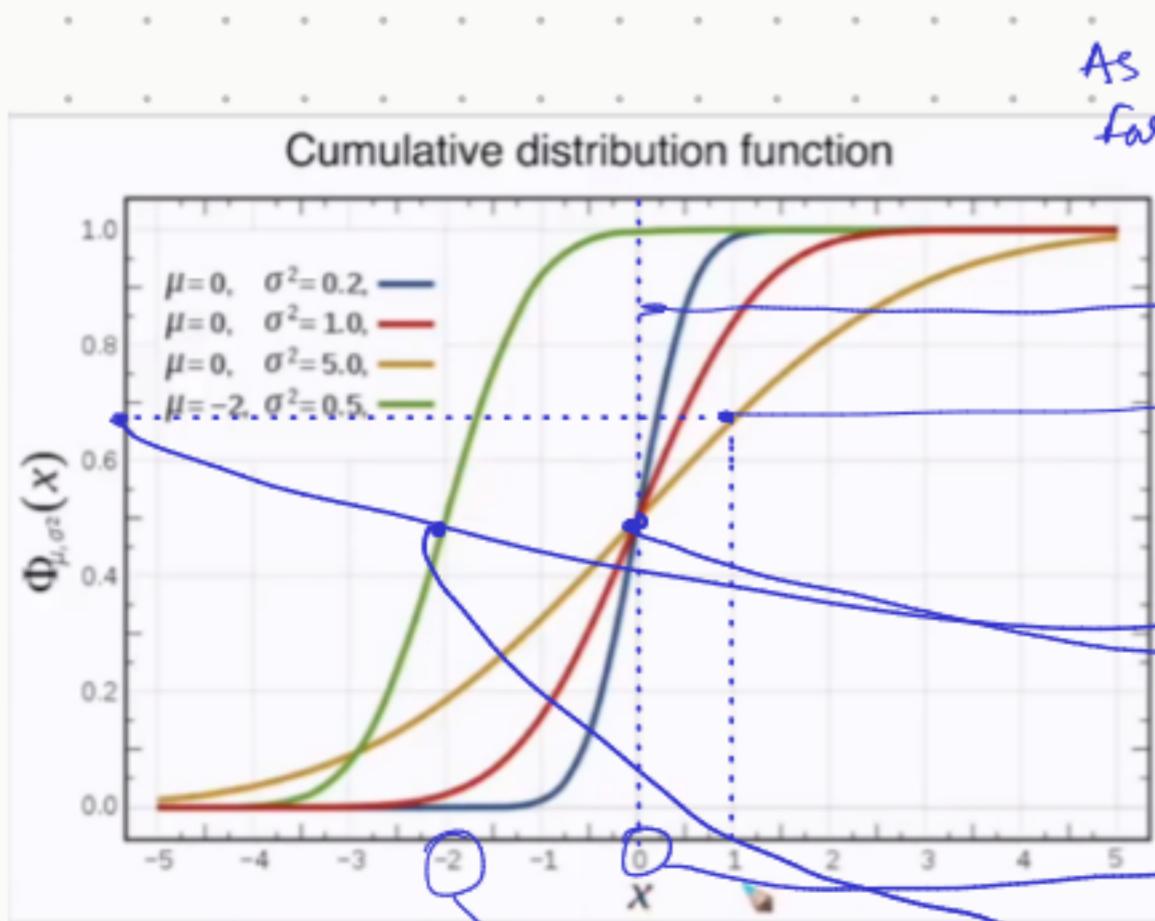
② Graph is symmetric.

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 Function (CDF) of Gaussian Distribution/Normal Distribution:-



As σ^2 increases, CDF goes far from unit line

CDF of a random variable looks like

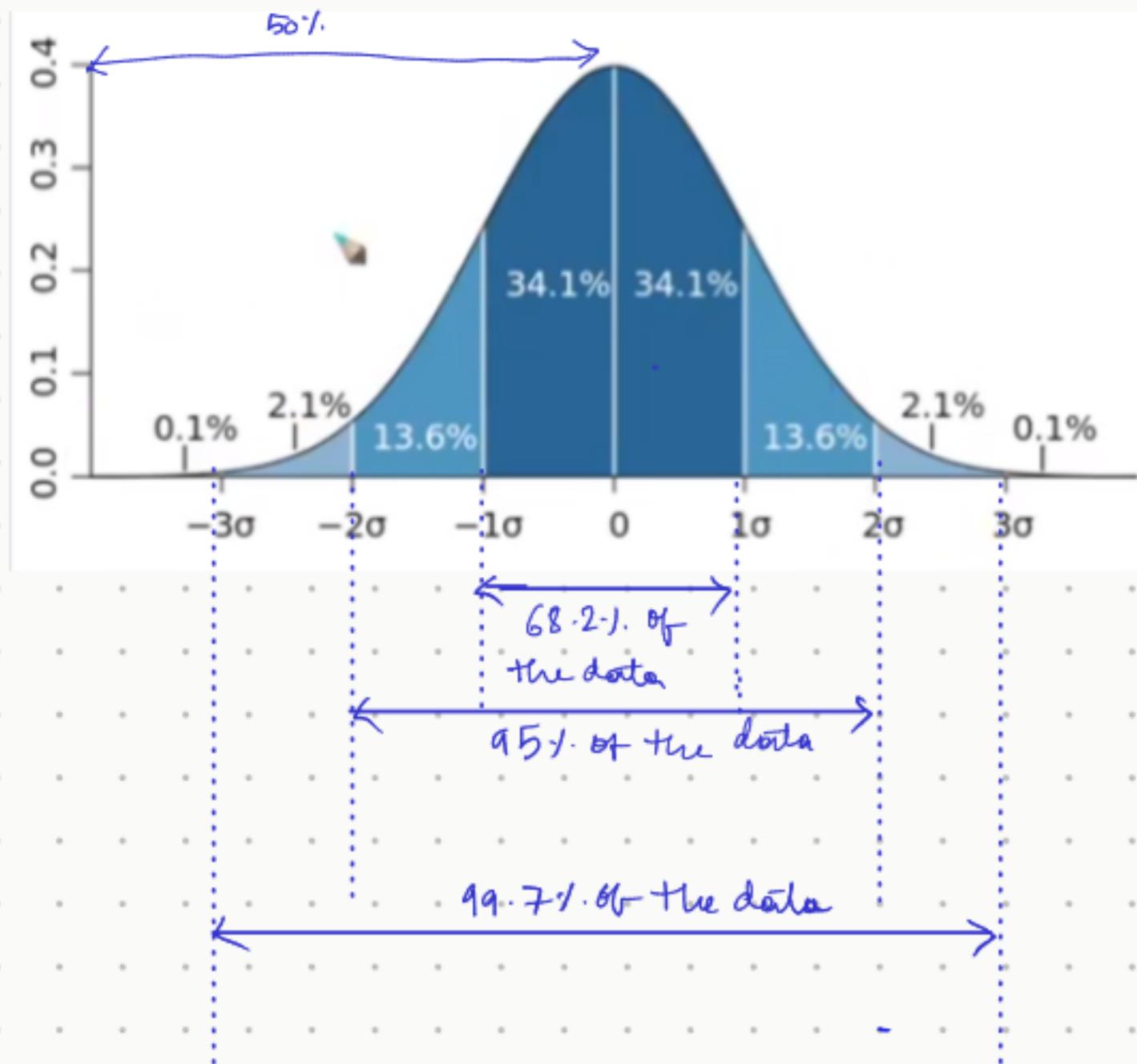
$\rightarrow \mu=0$, center of CDF is at 0.
 $\rightarrow \mu=-2$, center of CDF is at -2.

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

68-95-99.7 rule :-

$$\text{if } \mu=0, \sigma^2=4 \Rightarrow \sigma=2$$

$$X \sim N(0, 4)$$



How is this useful?

Ex:- if human population weight

$$X \sim N(150, 25)$$

$\downarrow \mu$ $\downarrow \sigma$

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

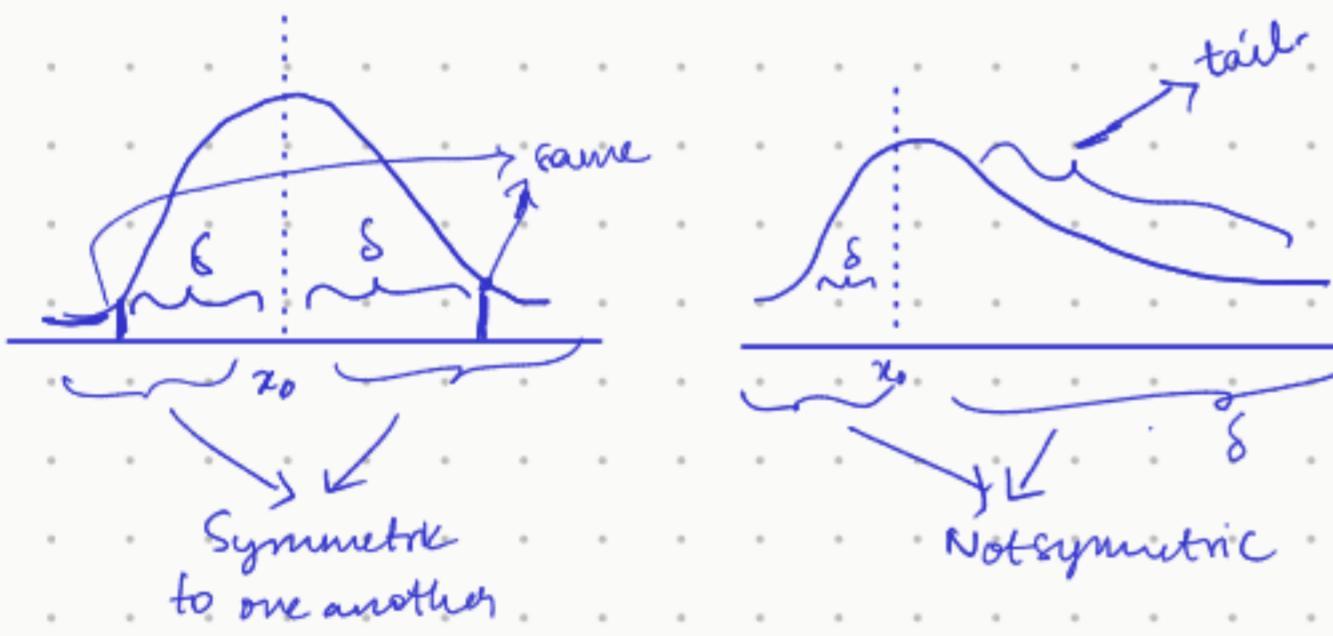
95% of people $(150-50, 150+50)$

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 :-

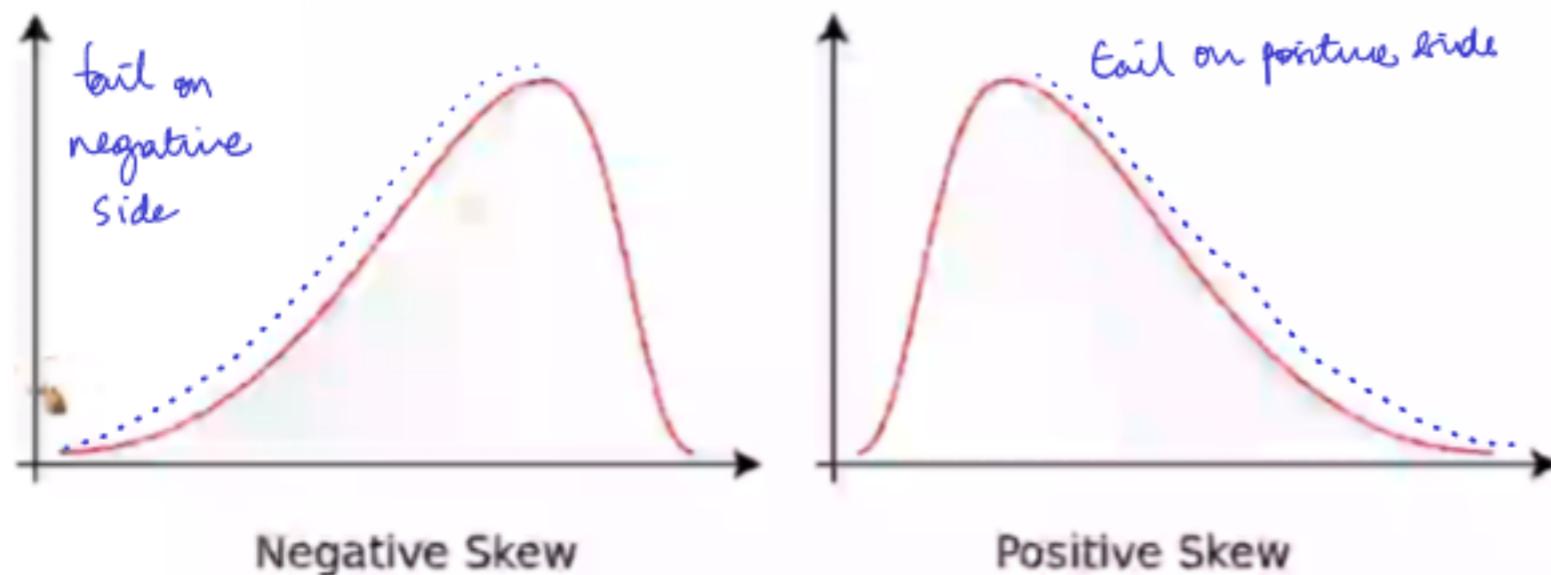
→ They help understand shape of PDF.



→ 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.



$$x = [x_1 \\ x_2 \\ \vdots \\ x_n] \quad \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

$$\text{Skewness} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^3 \quad \text{if } 2 = \text{Variance}$$

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

sample std-deviation

Kurtosis :-

→ Measure of tailedness

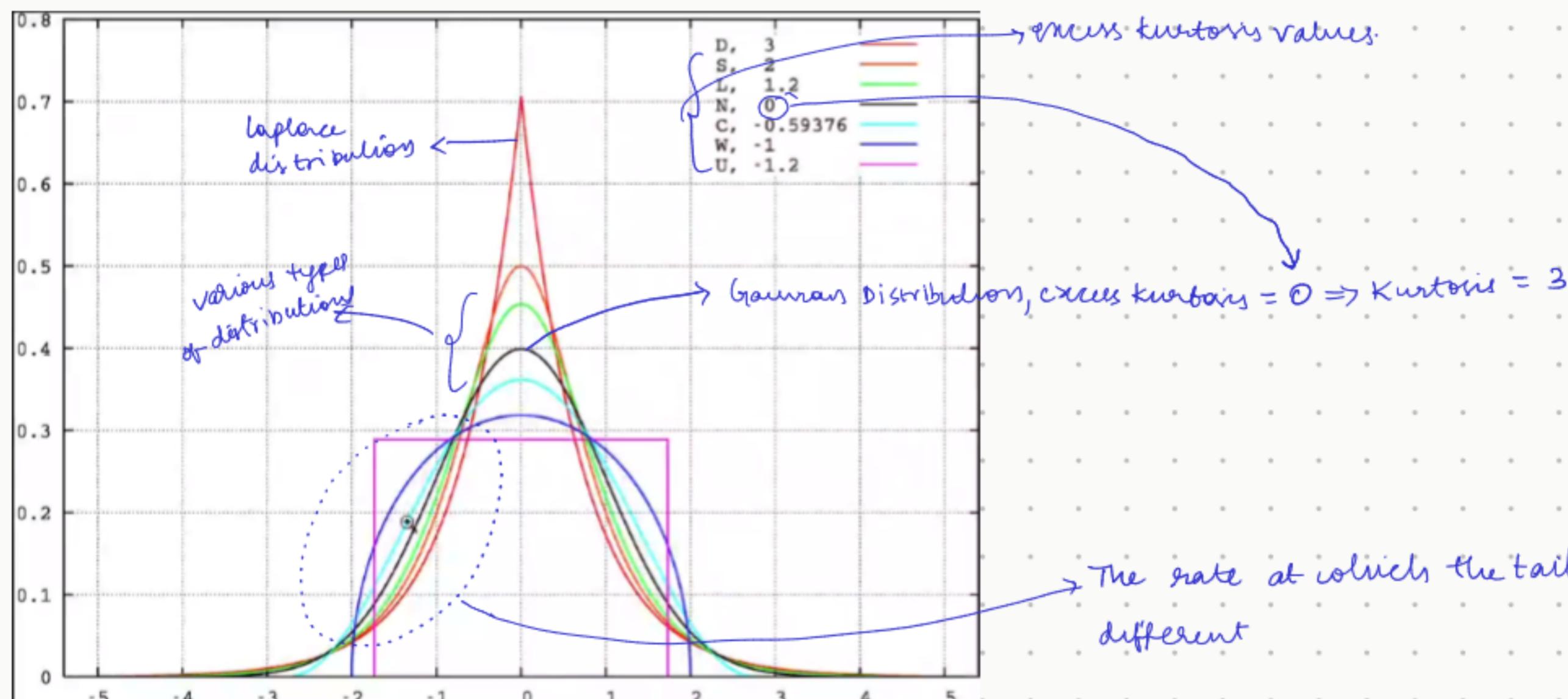
$$\text{excess kurtosis} = \text{kurtosis} - 3$$

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

Variance

Kurtosis of Gaussian Distribution = 3

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



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

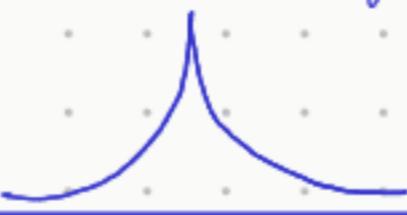
→ Kurtosis is used in analysis of trading, finance etc.)

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

- 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 - It's generally assumed that they follow gaussian distributions for ease of calculations (Pre 2008 financial crisis). As to gaussian distributions chances of losses are less & has good chance of high profits. But now, instead of assuming the data is a gaussian distribution, we look at the actual data & calculate the kurtosis

If kurtosis is $> 3 \rightsquigarrow$



⇒ chances of extreme profits are high & chances of losses are also high.

If kurtosis = 3 \Rightarrow Follow gaussian distribution

If kurtosis $< 3 \Rightarrow$



⇒ 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 $X \sim N(\mu, \sigma^2)$

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

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

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

↑ Standard Normal Variate.

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

$$z = \frac{x - \mu}{\sigma}$$

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

why? :-

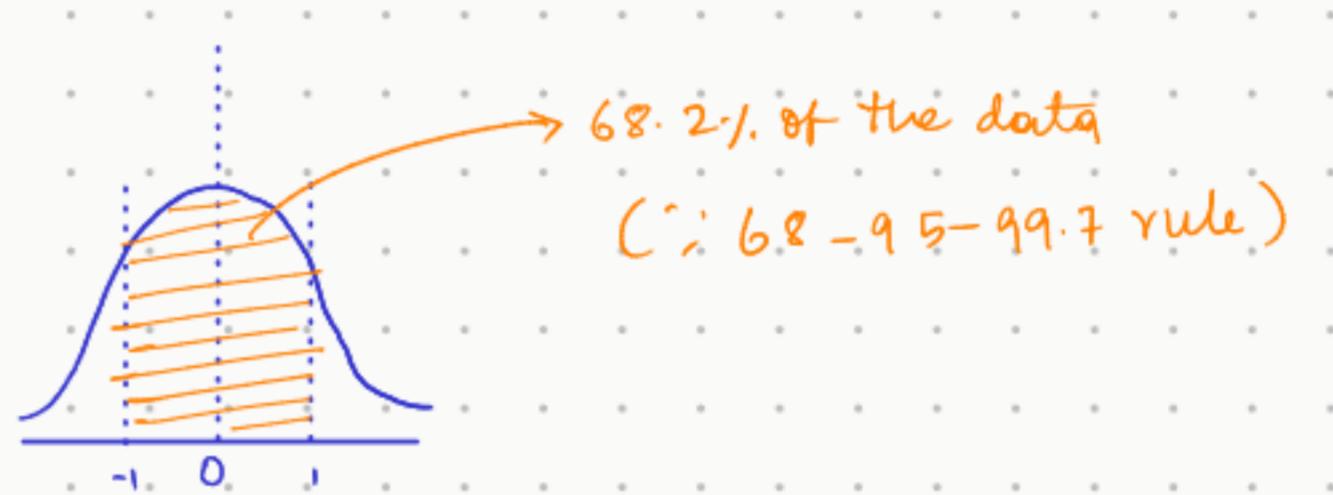
① After standardization, PDF becomes

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

→ We use StandardScalar for standardization & MinMaxScaler for Normalization.

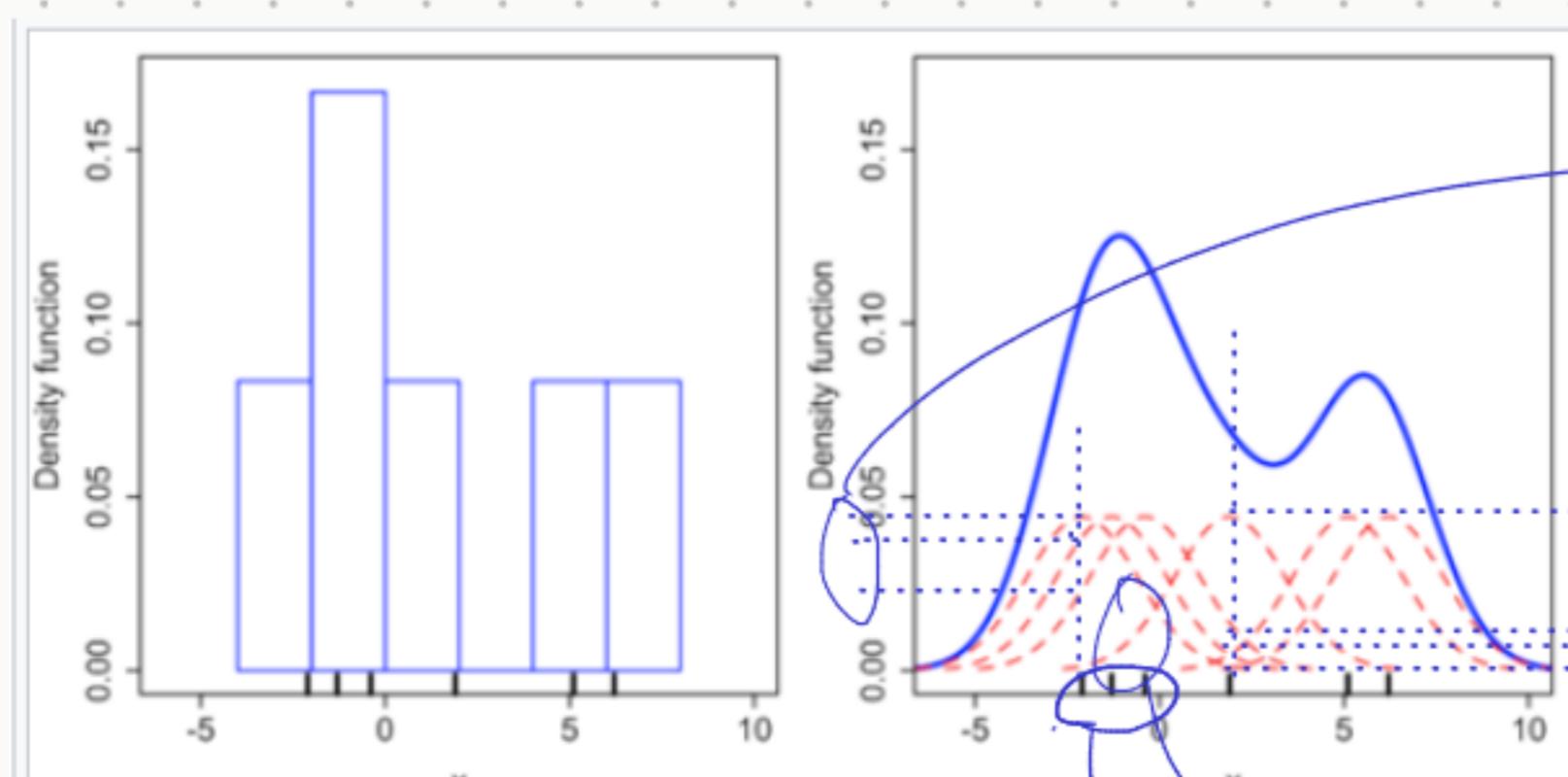
Two techniques of feature scaling -

$$\text{Normalization} \rightarrow x_i' = \frac{x_i - x_{\min}}{(x_{\max} - x_{\min})}$$



Kernel Density Estimation:-

→ Used for smoothing histograms to obtain PDFs

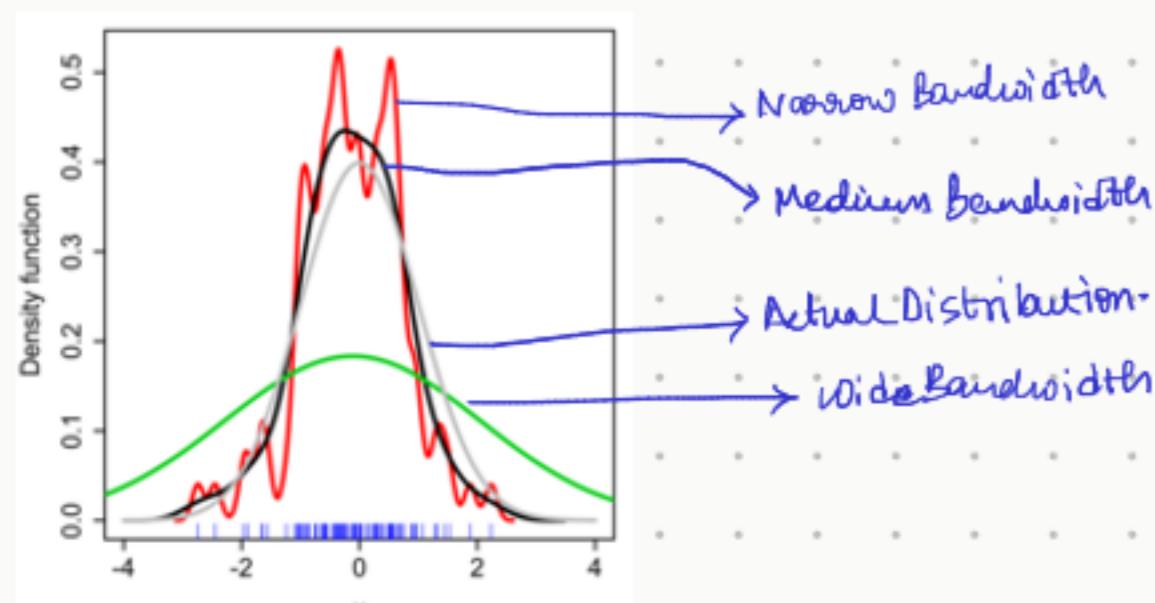


Comparison of the histogram (left) and kernel density estimate (right) constructed using the same data. The six individual kernels are the red dashed curves, the kernel density estimate the blue curves. The data points are the rug plot on the horizontal axis.

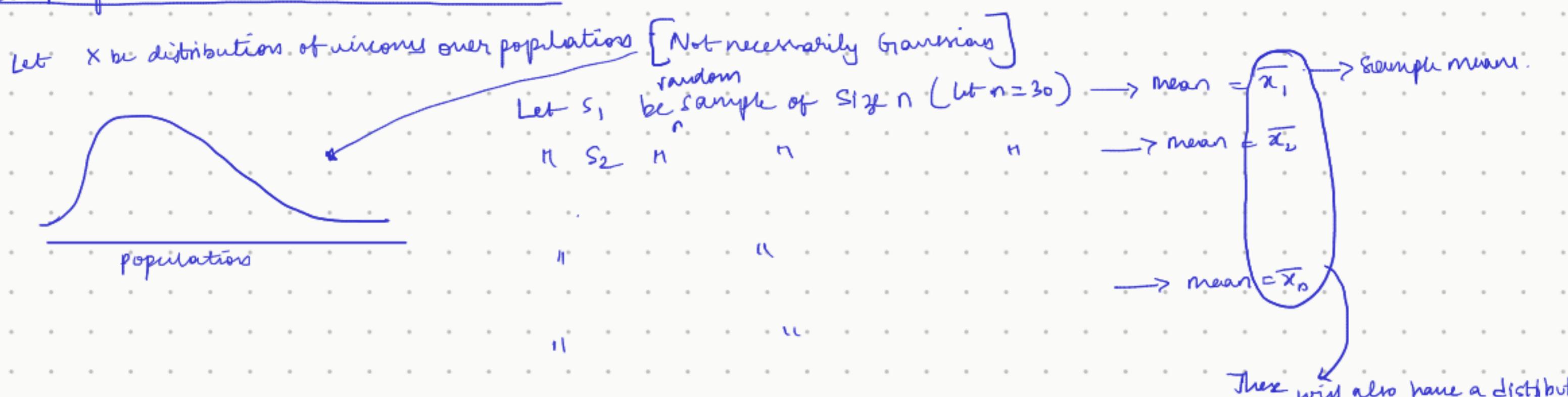
→ Variance of the gaussian kernel is known as bandwidths.

→ Gaussian Kernel -

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



Sampling Distributions & Central Limit Theorem



These will also have a distribution

The distribution \bar{x}_1 = Sampling distribution of sample mean.

Central Limit Theorem :- If original distributions 'X' has finite mean (there can be infinite mean ex:- parabola) & variance σ^2 Samples are created of size 'n' where sample means are $\bar{x}_1, \bar{x}_2, \bar{x}_3, \dots$ whose distribution is called sampling distribution of sampling mean, central limit theorem states that

$$\bar{x}_1 \rightarrow N(\mu, \frac{\sigma^2}{n}) \text{ as } n \rightarrow \infty$$

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

mean is same as original or that of distribution

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

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

Let X be a random variable

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

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

↓
graphical method

more powerful

How? :- ① Sort X & calculate percentiles.

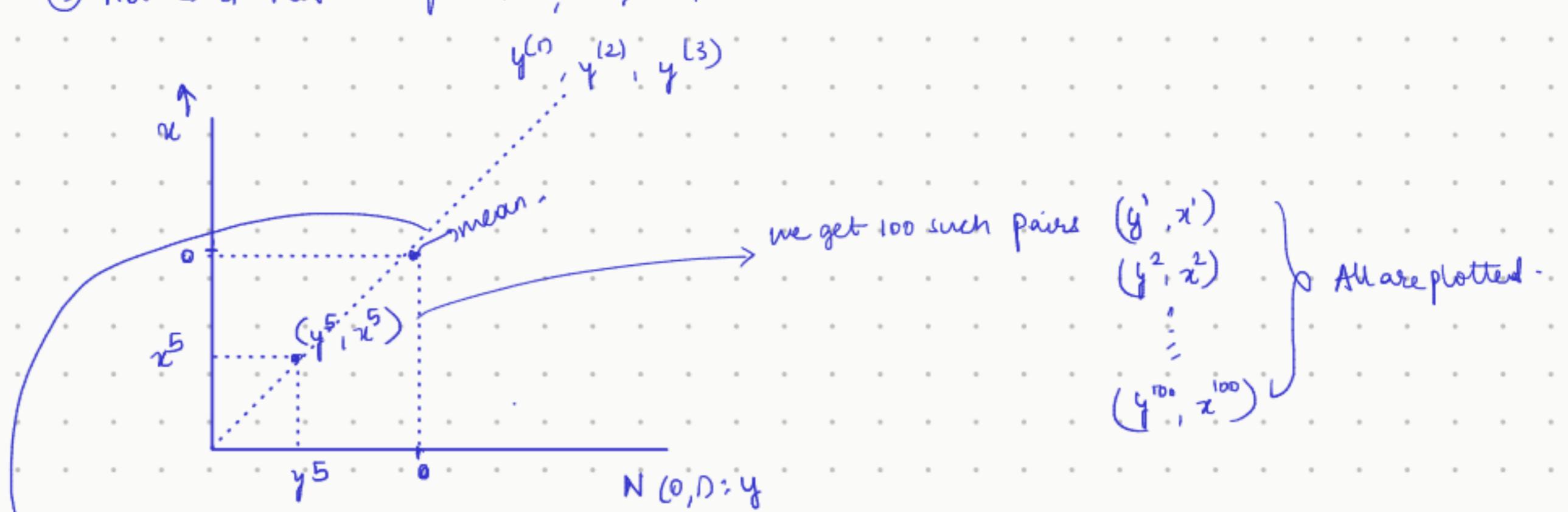
$$\begin{aligned} x'_1, x'_2, x'_3, \dots, x'_{500} \\ x'_1 \rightarrow 1^{st} \text{ percentile} \rightarrow x^{(1)} \\ x'_{10} \rightarrow 10^{th} \text{ percentile} \rightarrow x^{(2)} \\ \vdots \\ x'_{500} \rightarrow 100^{th} \text{ percentile} \rightarrow x^{(100)} \end{aligned}$$

② take $Y \sim N(0,1)$

$$\begin{aligned} Y: y_1, y_2, y_3, \dots, y_{1000} \\ \downarrow \\ y'_1, y'_2, y'_3, \dots, y'_{1000} \\ \downarrow \\ y^{(1)}, y^{(2)}, y^{(3)}, \dots, y^{(100)} \end{aligned}$$

These are called as Theoretical Quantiles

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



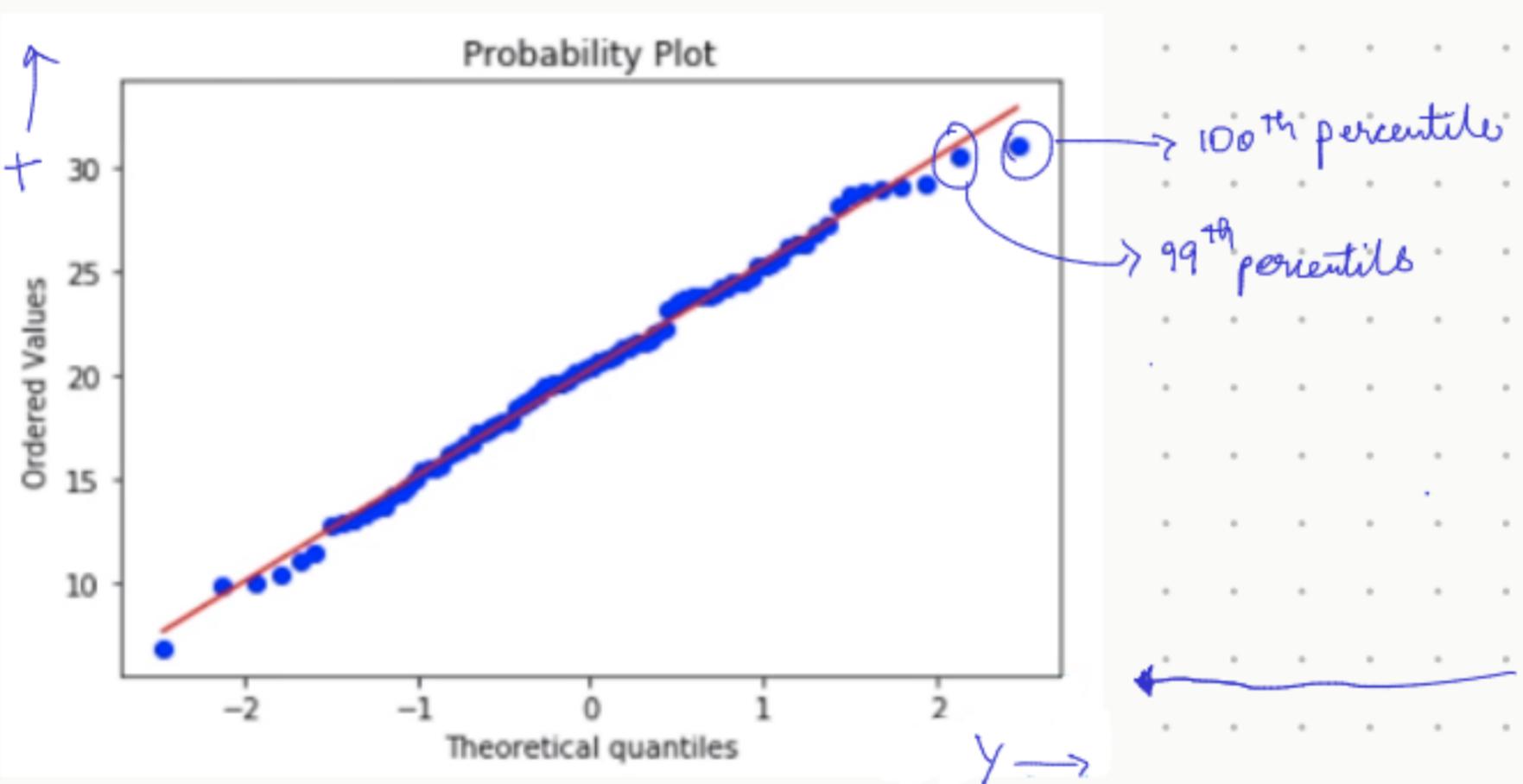
If (y_i, x_i) for $i \geq 100$, lie roughly on a straight line, then y & x have similar distributions.

$\Rightarrow x$ also has gaussian distribution.

→ `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 $x \sim y$ 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)
```

Generating x :-

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

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 deviation from original numbers is allowed in samples.

Chebyshew'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?

\uparrow (finite)
 \uparrow (nonzero & finite)

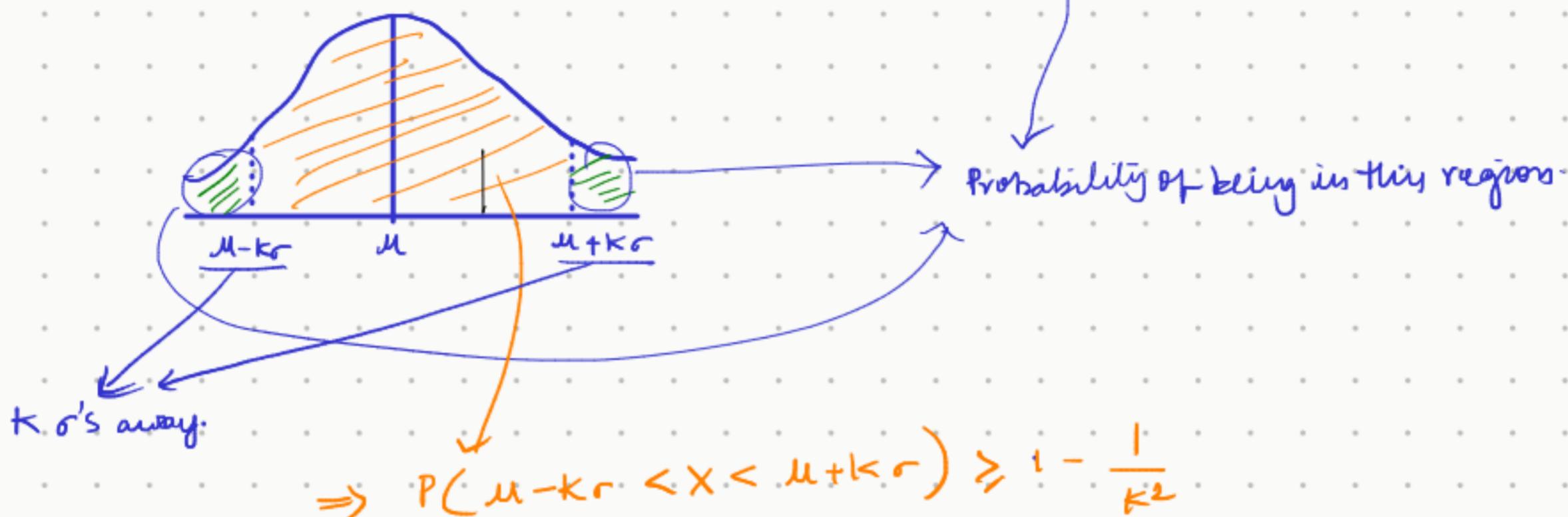
→ Chebyshew's inequality states that

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

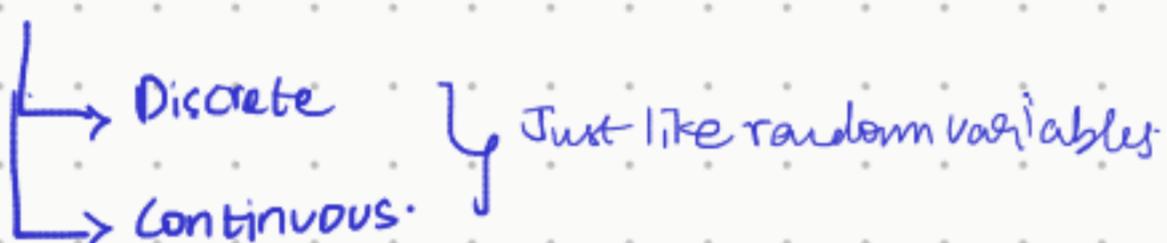


$$\Rightarrow P(\mu - k\sigma < X < \mu + k\sigma) \geq 1 - \frac{1}{k^2}$$

→ From Chebyshew's inequality we can conclude that 75% of any distribution 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 Chebyshew as well.

Uniform Distributions



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

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

→ Probability 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

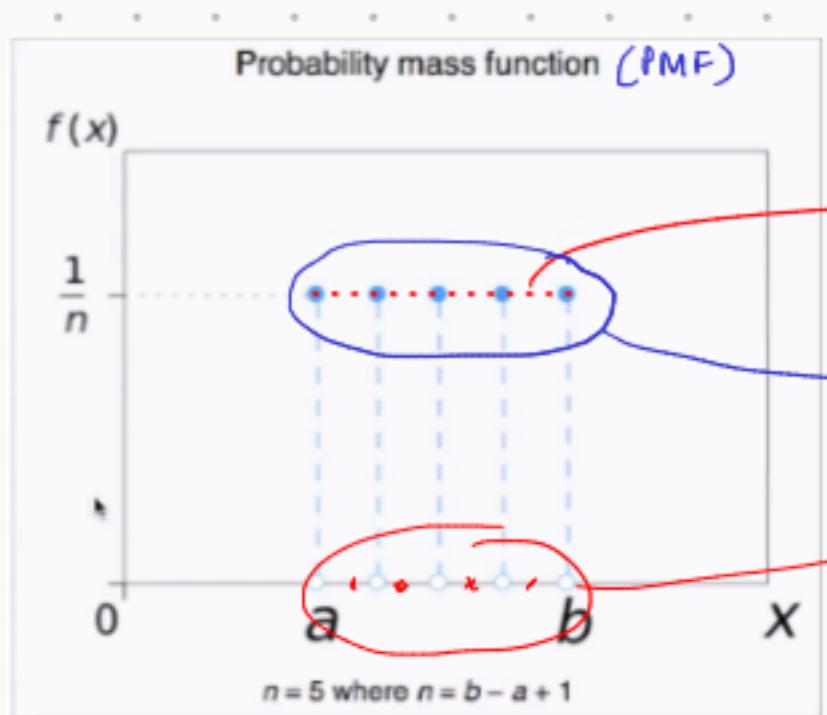
↳ both are numbers.

and

$n = b - a + 1$. Another formula for calculating $n = \frac{(b-a+1)}{k}$ where k is the common difference.

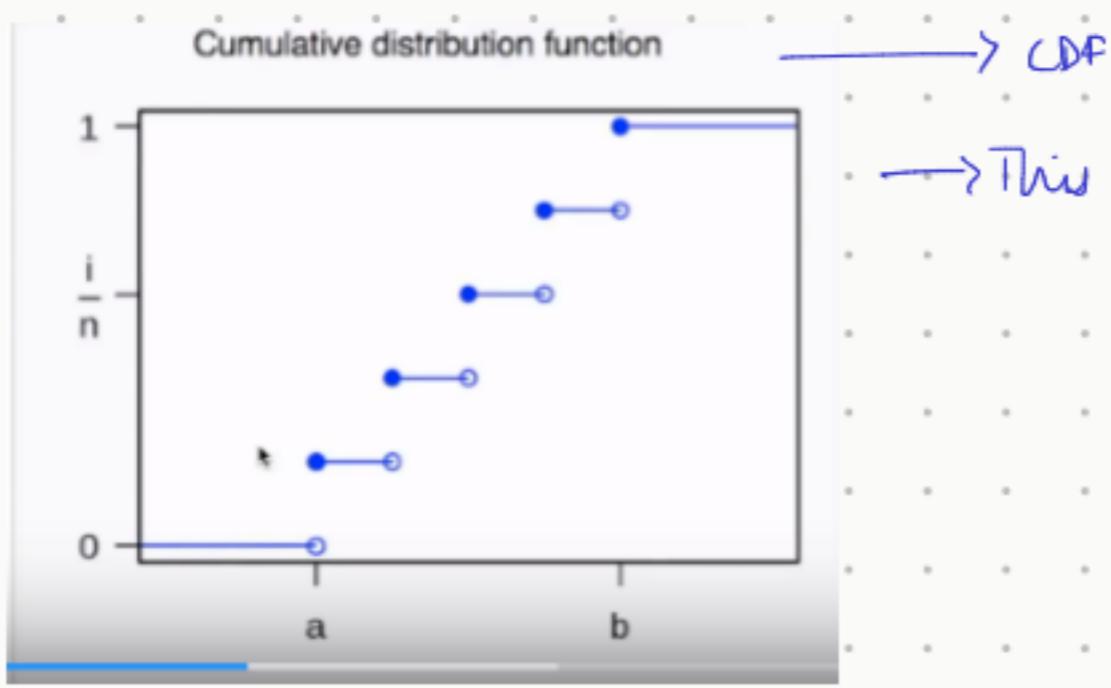
↳ number of outcomes.

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



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

∴ Only these does exist in PMF



→ This is called a non smooth function

Props of URV :-

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

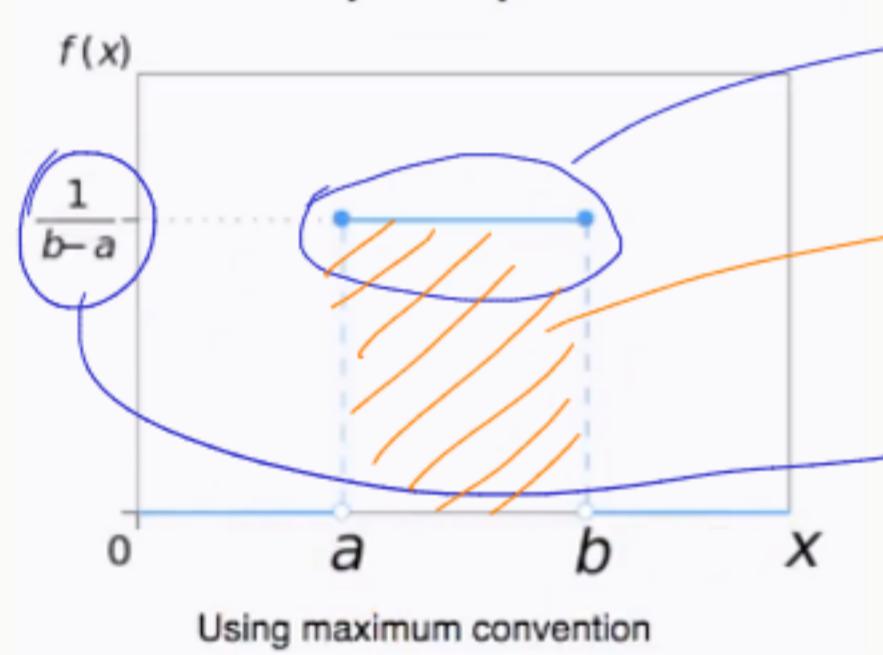
$$\text{Median} = \frac{a+b}{2}$$

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

$$\text{Skewness} = 0$$

Continuous Random Variable :-

Probability density function

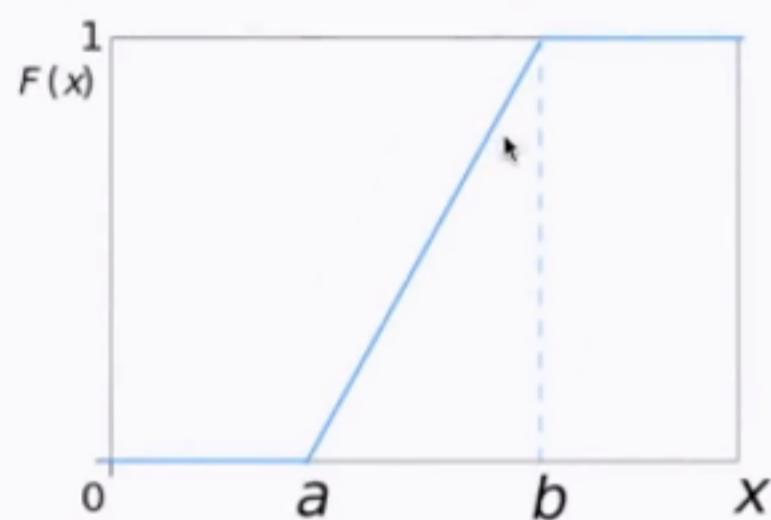


→ Line can exist here because it's continuous.

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

CDF

Cumulative distribution function



Properties of CURV :-

Parms: a, b

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

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

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

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

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.

on:-

→ 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)

```

$\text{random.random}()$ picks value b/w $1 \leq n$ with equal probability.
 chances of getting val $\leq 0.2 = 20\%$.
 $30 \text{ or } 150 = 20\%$.

$\downarrow \approx 30$

→ 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:-

$$X \sim \text{Bernoulli}(p=0.5)$$

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

$$\text{Mean} = p$$

$$\text{Variance} = pq$$

Binomial Random Variable:-

Let X be taking a coin

$$X \sim \text{Bernoulli}(p=0.5)$$

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

$$Y \in \{0, 1, 2, 3, \dots, 30\}$$

$Y \sim \text{Binomial}(n, p)$

probability of value 1
number of trials

Parameters:-

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

Not used often in Machine Learning.

→ If in Binomial distribution, $n=1 \Rightarrow$ it's a Bernoulli distribution.

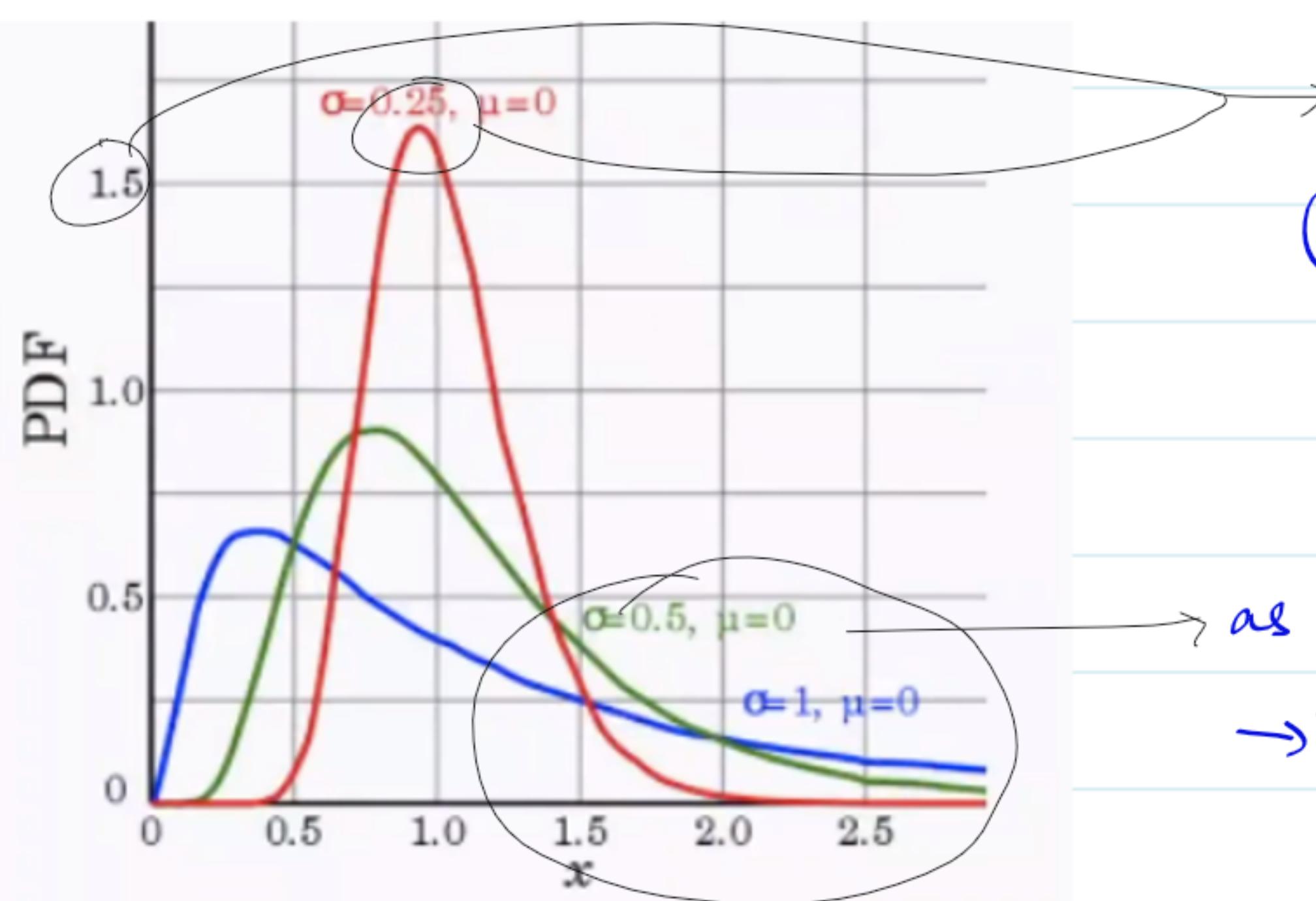
→ Bernoulli is unbiased.

→ 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 :-

$X \sim \text{Log-Normal}(\mu, \sigma)$ if and only if $\log(X)$ is normally distributed.

↳ Natural log aka ln



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

as σ increases, curve becomes more skewed.

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

Some log-normal density functions with identical parameter μ but

paper 480p 360p 240p

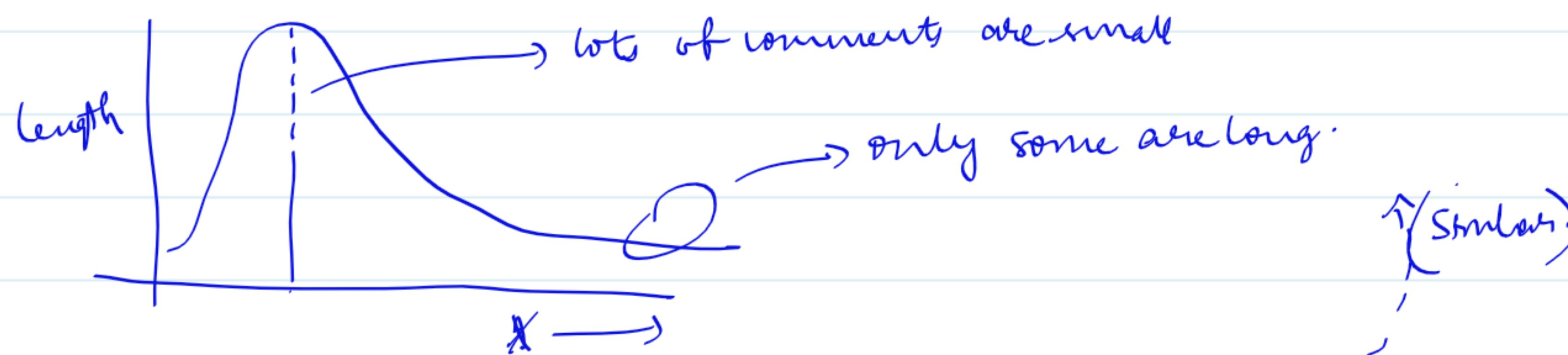
$$\text{PDF} = \frac{1}{\pi \sigma \sqrt{2\pi}} e^{-\frac{(\ln(x-\mu))^2}{2\sigma^2}}$$

$$\text{Skewness} = (e^{\sigma^2} + 2)\sqrt{e^{\sigma^2} - 1}$$

Not 0 like seen in Normal Distribution.

Applications :-

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



→ 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 $X \sim \text{Log-Normal}(\mu, \sigma)$ or not :-

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

↓ ↓ ↓ ↓

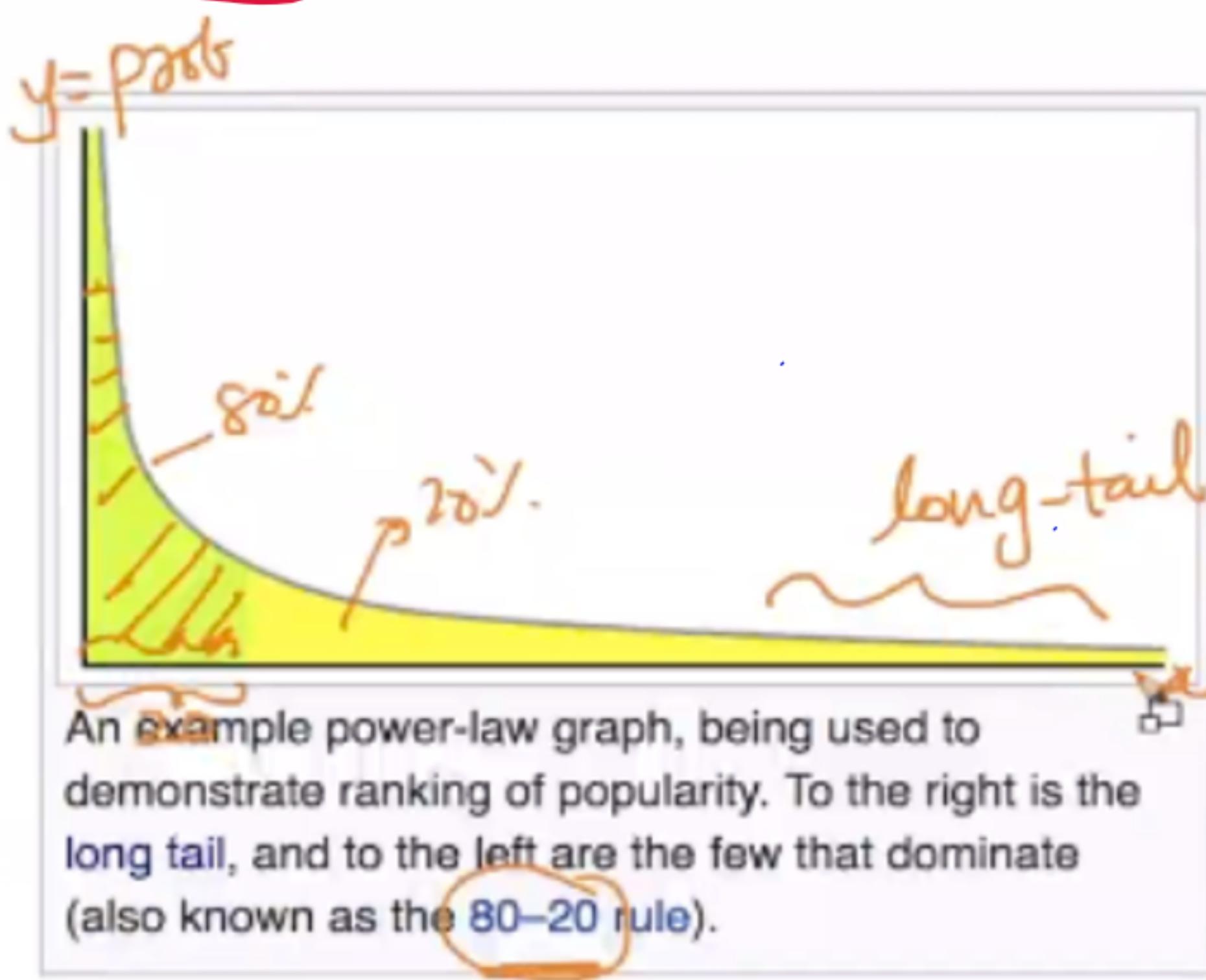
$$\ln(x_1), \ln(x_2), \dots, \ln(x_n)$$

(y_1), (y_2), \dots, (y_n)

→ Q-Q - plot → check if it's gaussian.

→ Log normal more common than normal distributions IRL.

Power-Law :-



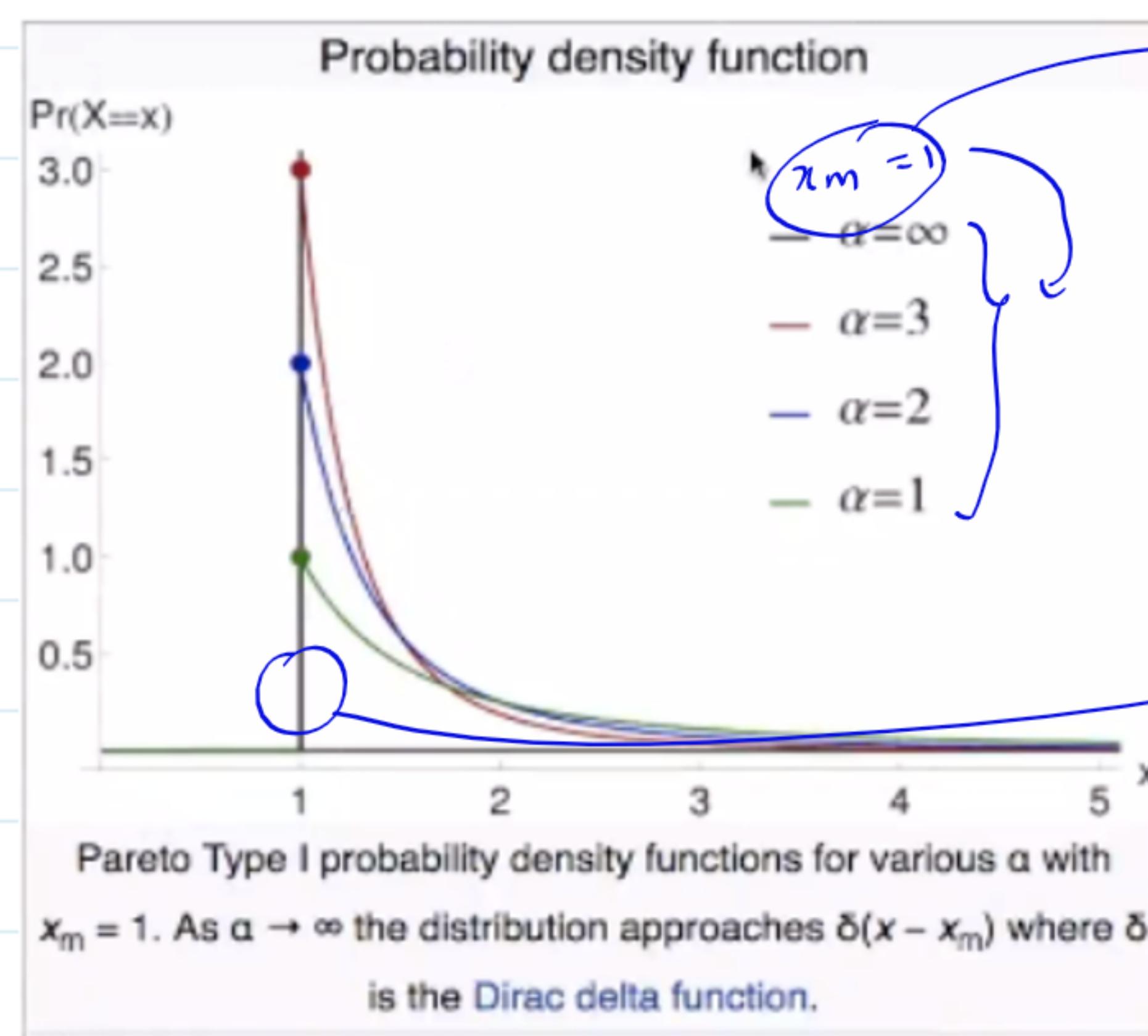
If distribution follows power-law, it's called as a Pareto distribution.

Ocurs a lot in nature.

Parameters of pareto distribution:-

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

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



peak

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

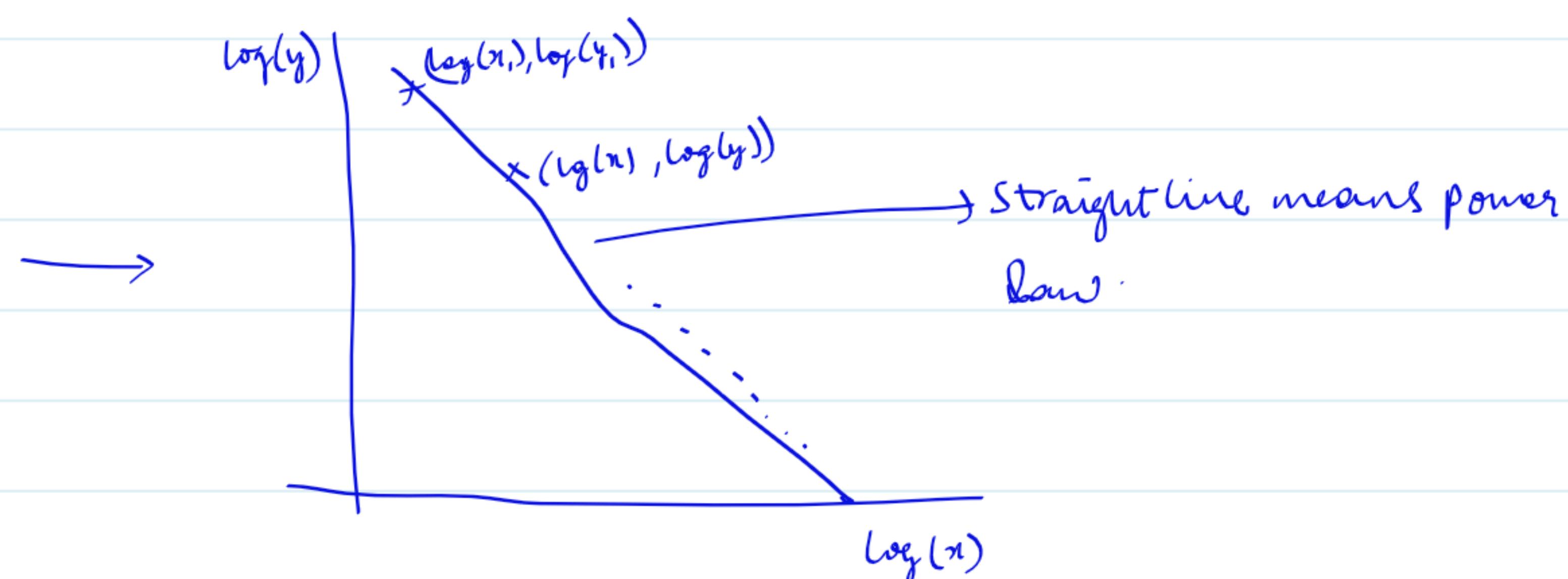
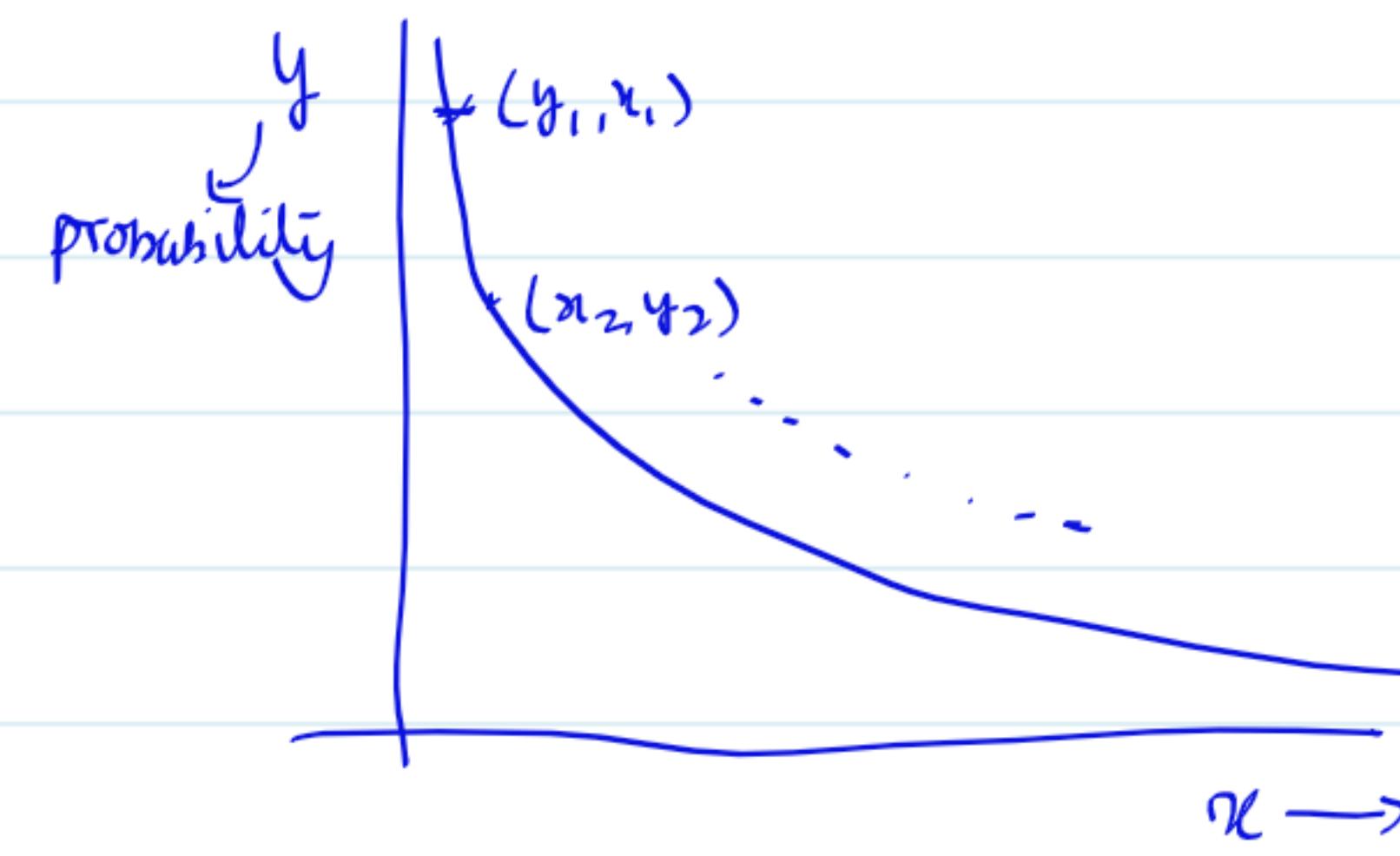
$\alpha = \infty$ (Dirac Delta function)

Applications :-

- size of human settlements (villages, cities etc.)
- Filesize distribution of Internet
- HTTP error rates

How to check for pareto/power law? :-

→ We use log-log plots.

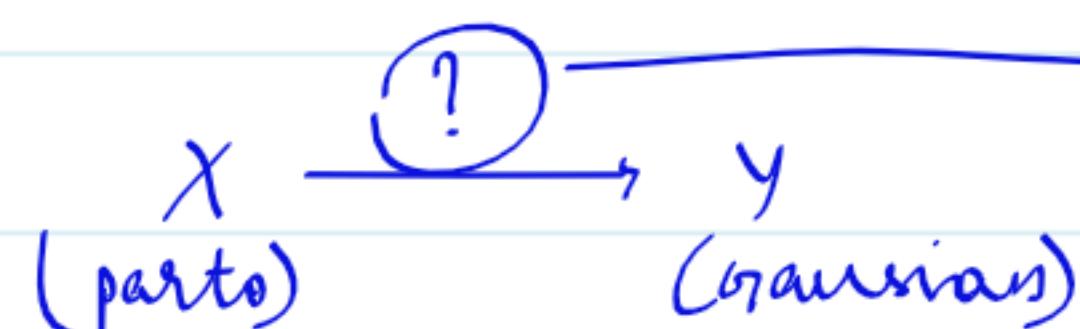


→ To check for pareto, we do log-log plot with power & observation.

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.



$x \rightarrow \text{Pareto} \rightarrow [x_1, x_2, x_3, \dots, x_n] \rightarrow x$

$y \rightarrow \text{Normal} \rightarrow [y_1, y_2, y_3, \dots, y_n] \rightarrow y$

Steps:-

① $\text{boxcox}(x) \rightarrow$ returns ' λ ' [λ computation is complete and no need to memoize]

② $y_i = \begin{cases} \frac{x_i - 1}{\lambda} & \text{if } \lambda \neq 0 \\ \log(x_i) & \text{if } \lambda = 0 \end{cases}$

$\forall i: 1 \rightarrow n$

$\Rightarrow x_i$ is log-normal

How?

`scipy.stats.boxcox(x)` → returns boxcox power transformed array and λ value.

Note:- Boxcox 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? :- Any **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 Distributions

④ Normal Distributions

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?

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

$$\left(\text{similar to } \text{var}(x) = \frac{1}{n} \sum_{i=1}^n (x_i - \mu_x)^2 \right)$$

$$\text{cov}(x, x) = \text{var}(x)$$

$\text{cov}(x, y) = \text{the if } x \uparrow \Rightarrow y \uparrow$ (reasoning)
 $= -\text{ve if } x \uparrow \Rightarrow y \downarrow$

Drawbacks :-

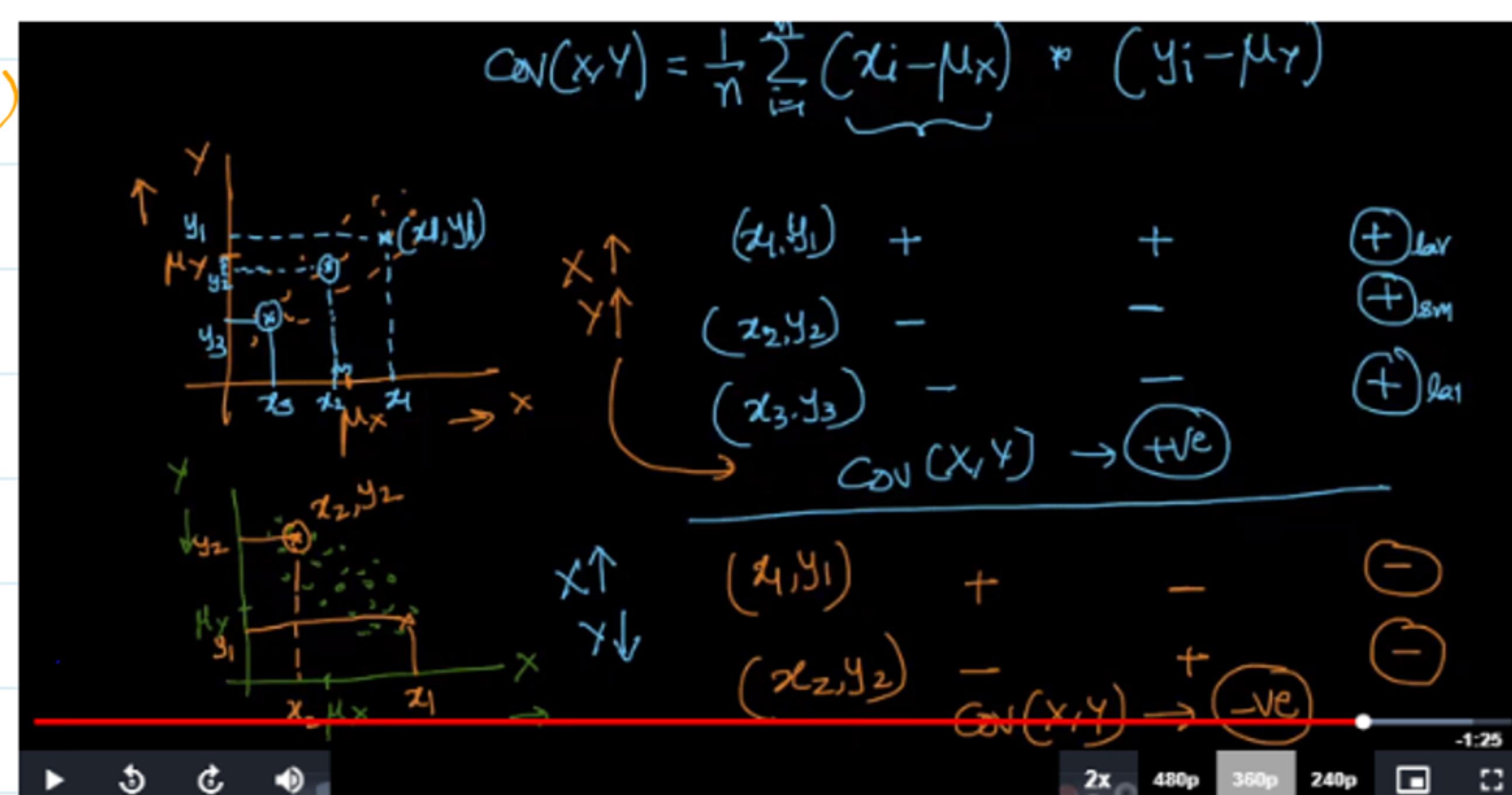
$$\text{cov}(\text{height, weight}) + \text{cov}(\text{height, weight})$$

(cm) (kg) (ft) (lbs)

same data but diff covariances

The sign will be the same

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



$$\rightarrow \text{Correlation} = \frac{\text{Covariance}(x, y)}{\sigma_x \sigma_y}$$

\rightarrow 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.

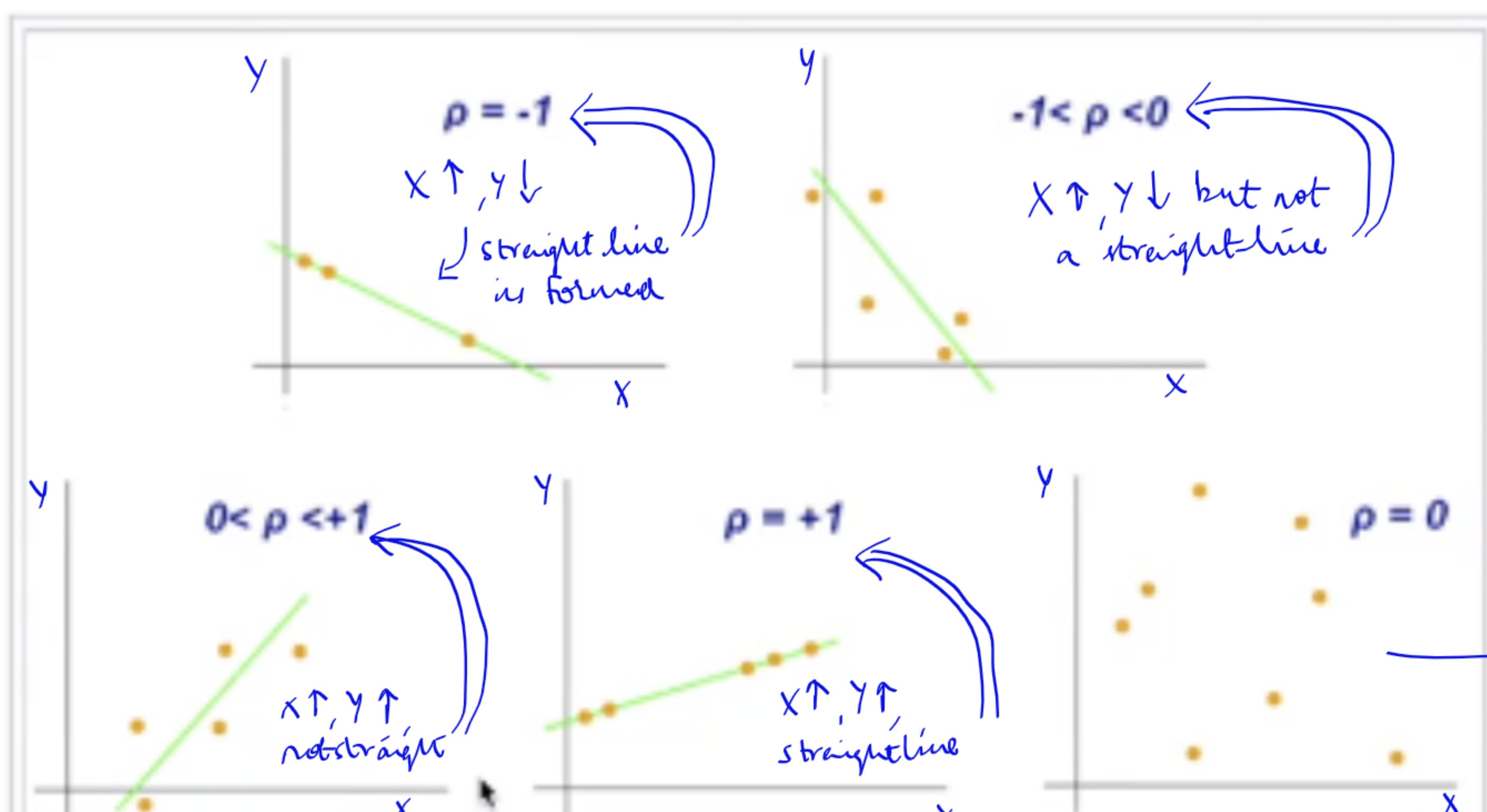
Pearson Correlation Coefficient (PCC) :-

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

\rightarrow Cov does not take variability into considerations. This does.

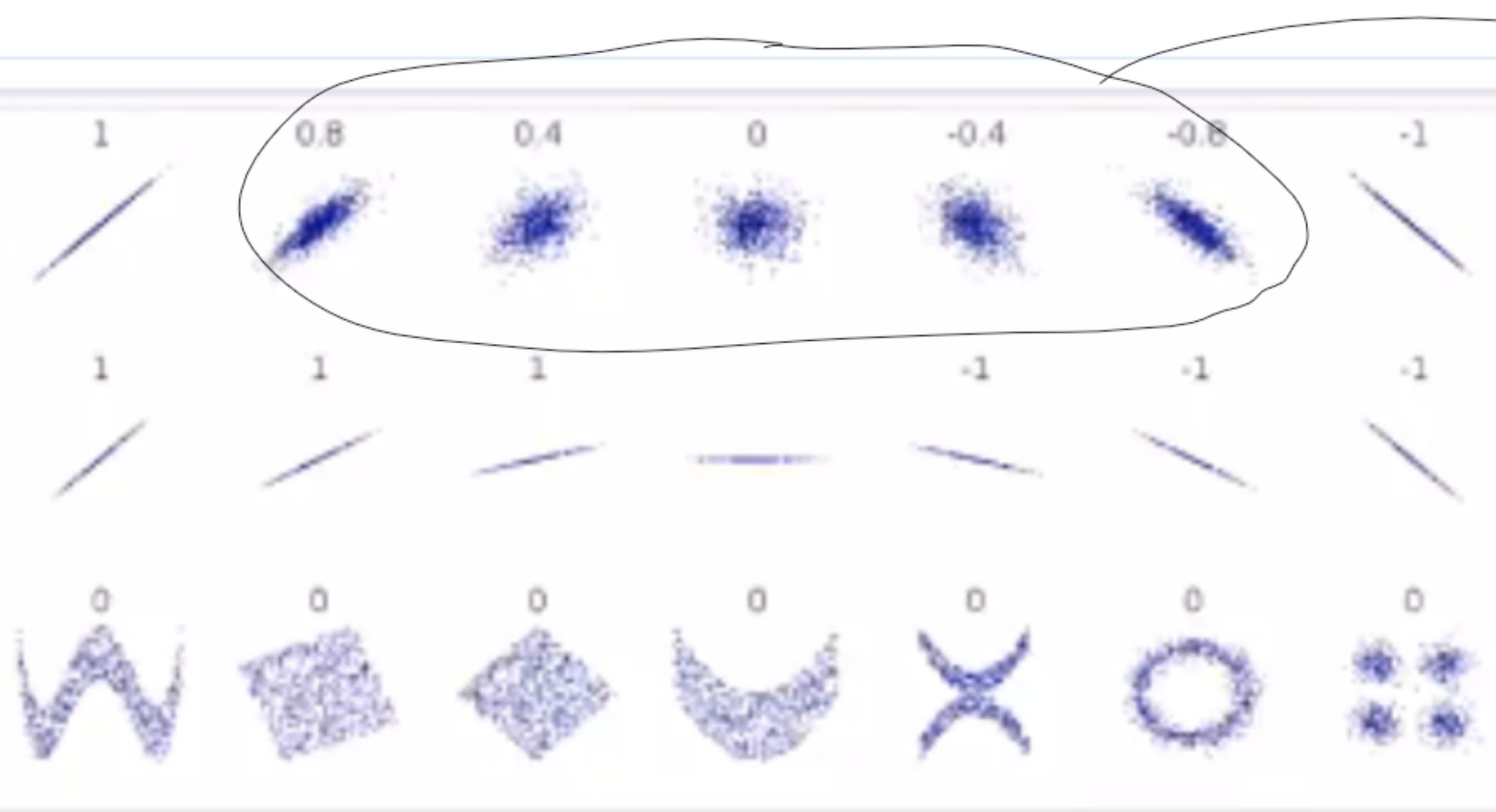
$$-1 \leq p \leq +1$$

\rightarrow PCC is heavily biased towards linearly dependant variables -



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

No relations at all.
 \Rightarrow No line
 $\Rightarrow p=0$.



→ As it becomes flatter, value decreases.

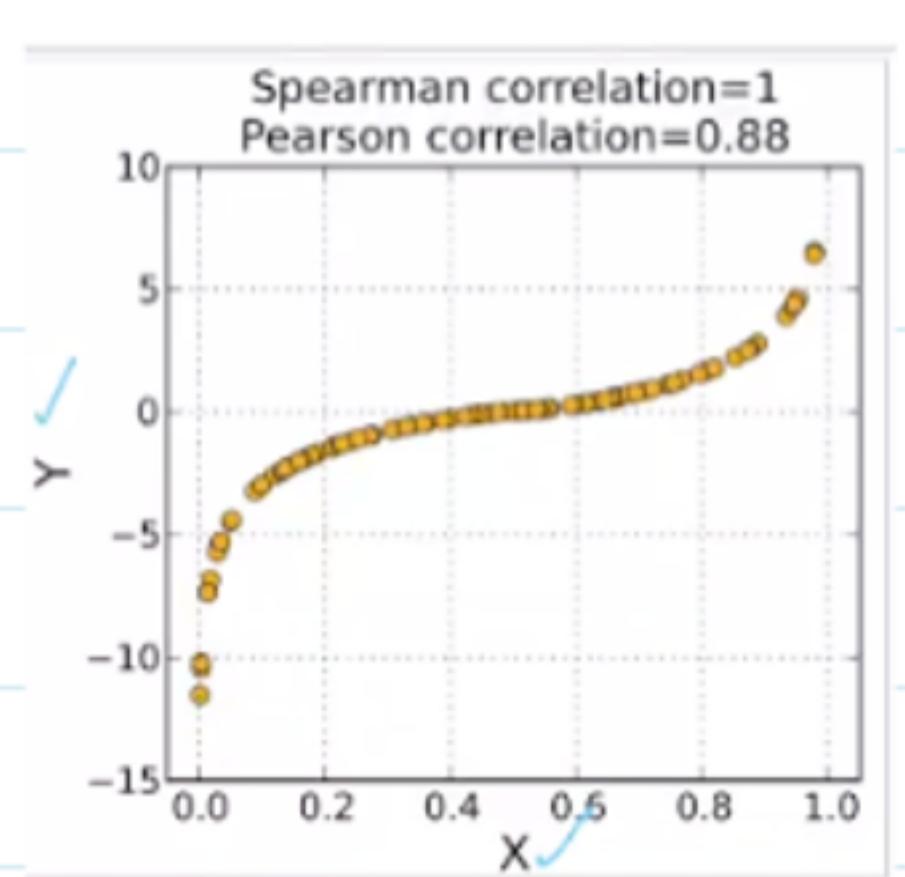
$\downarrow \rightarrow$ PCC does not care about the slope of the line.

$\downarrow \rightarrow$ Does not capture complex/non-linear relationships.

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

monotonically increasing \rightarrow if $x_2 > x_1$ and corresponding $y_2 > y_1$

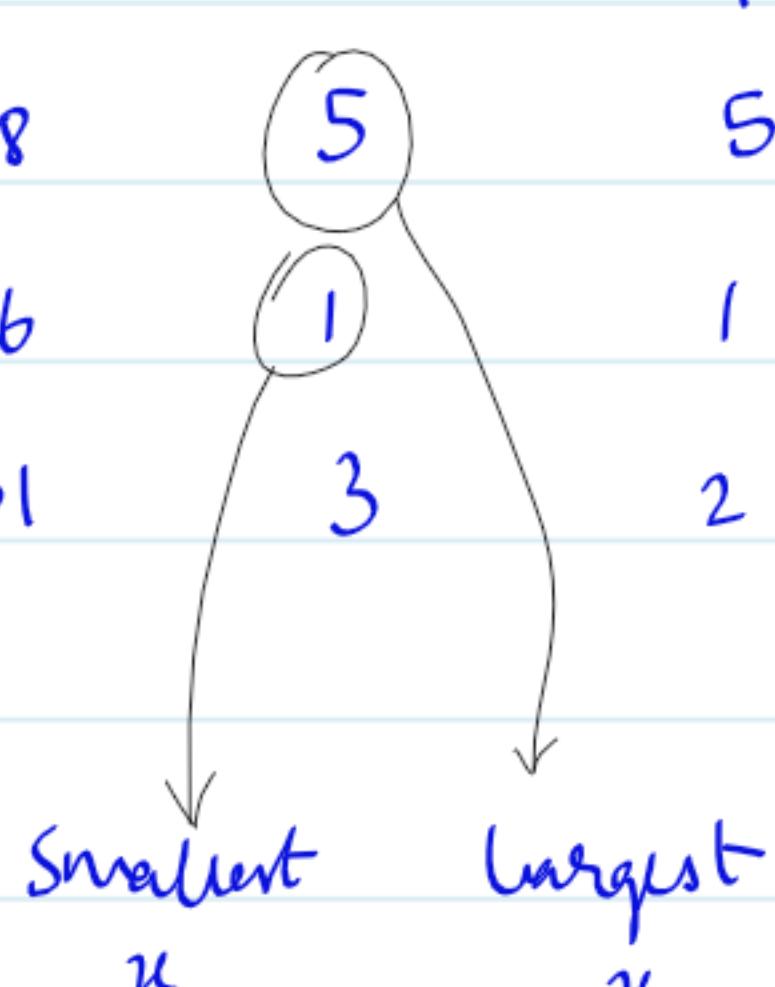
→ Correlation is what defines the change in one item when another item is modified.



Spearman 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



During rank calculation, if two vals have same value, then their mean is taken as rank. Fractional rank (0.5)

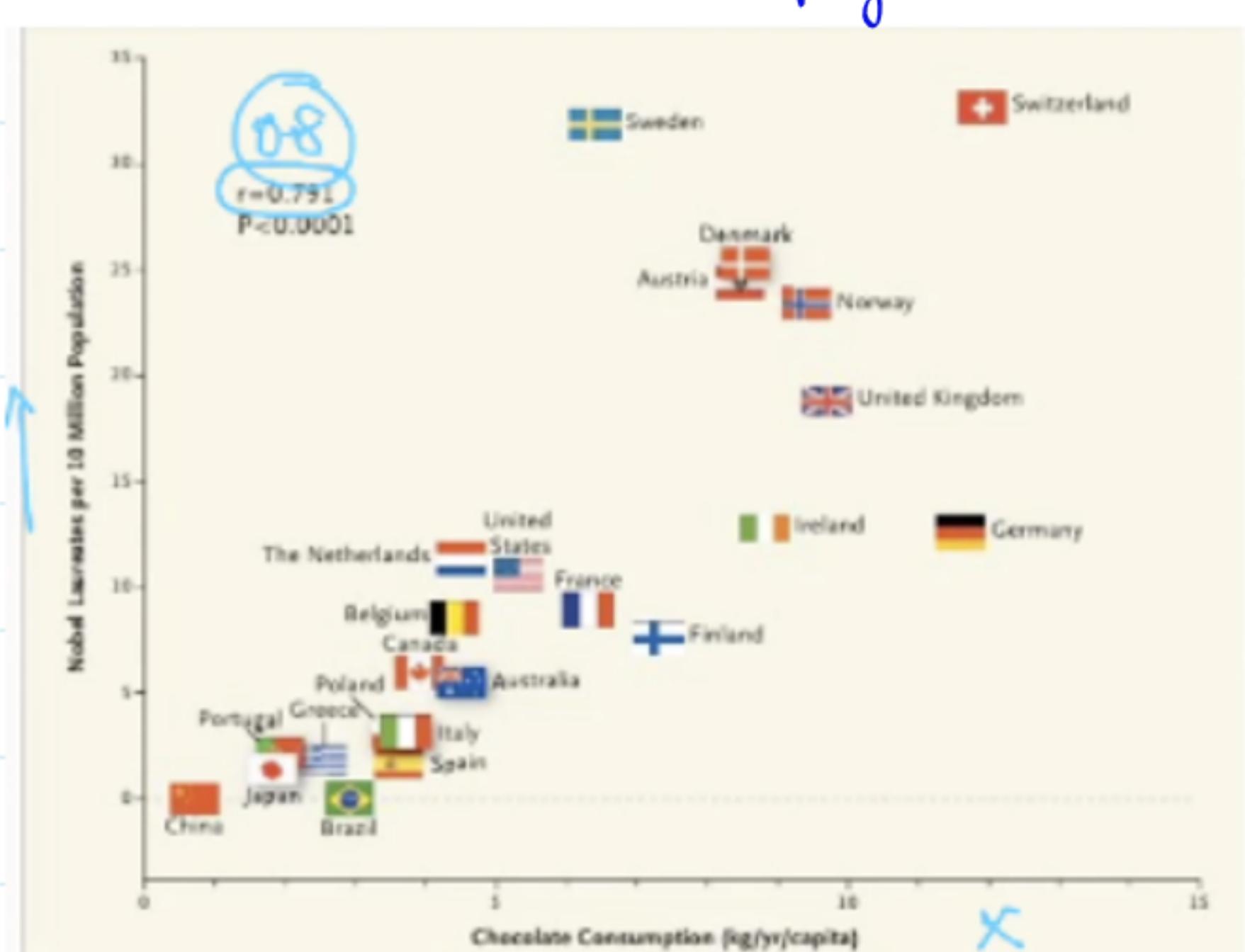
Same rank depends on the implementation.

$$\text{Spearman coefficient} \quad r = P_{r_x, r_y} \quad \rightarrow \text{Ranks are strictly increasing.}$$

$r = 1 \Rightarrow x \uparrow \quad y \uparrow \Rightarrow r = 1 \quad \downarrow \rightarrow$ irrespective of whether they are linear or not.
 $r = -1 \Rightarrow x \uparrow \quad y \downarrow \Rightarrow r = -1 \quad \downarrow \rightarrow$ As long as there is monotonicity, it will work.

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

ex:-



Applications of Correlations :-

ex:- (Salary & home square footage) \rightarrow used by real estate brokers.

(# of years of education & income) \rightarrow not looking for cause-correlation only.

(Time spent on commerce website & money spent in the next 24 hours) \rightarrow Amazon, Flipkart etc.

(# of unique visitors in one day & \$ sales in one day) \rightarrow ecommerce websites.

(dosage of medicine & blood sugar level of person) \rightarrow medicine.

\rightarrow Ideas from causal inference can be used to determine causations.

Confidence Interval:-

Let X : heights with $x_1, x_2, x_3, \dots, x_{10}$
Random Sample with size 10

Point estimate $\{ \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \}$ \rightarrow Simple Average.
Population mean μ \rightarrow Sample mean.
As n increases, $\bar{x} \rightarrow \mu$

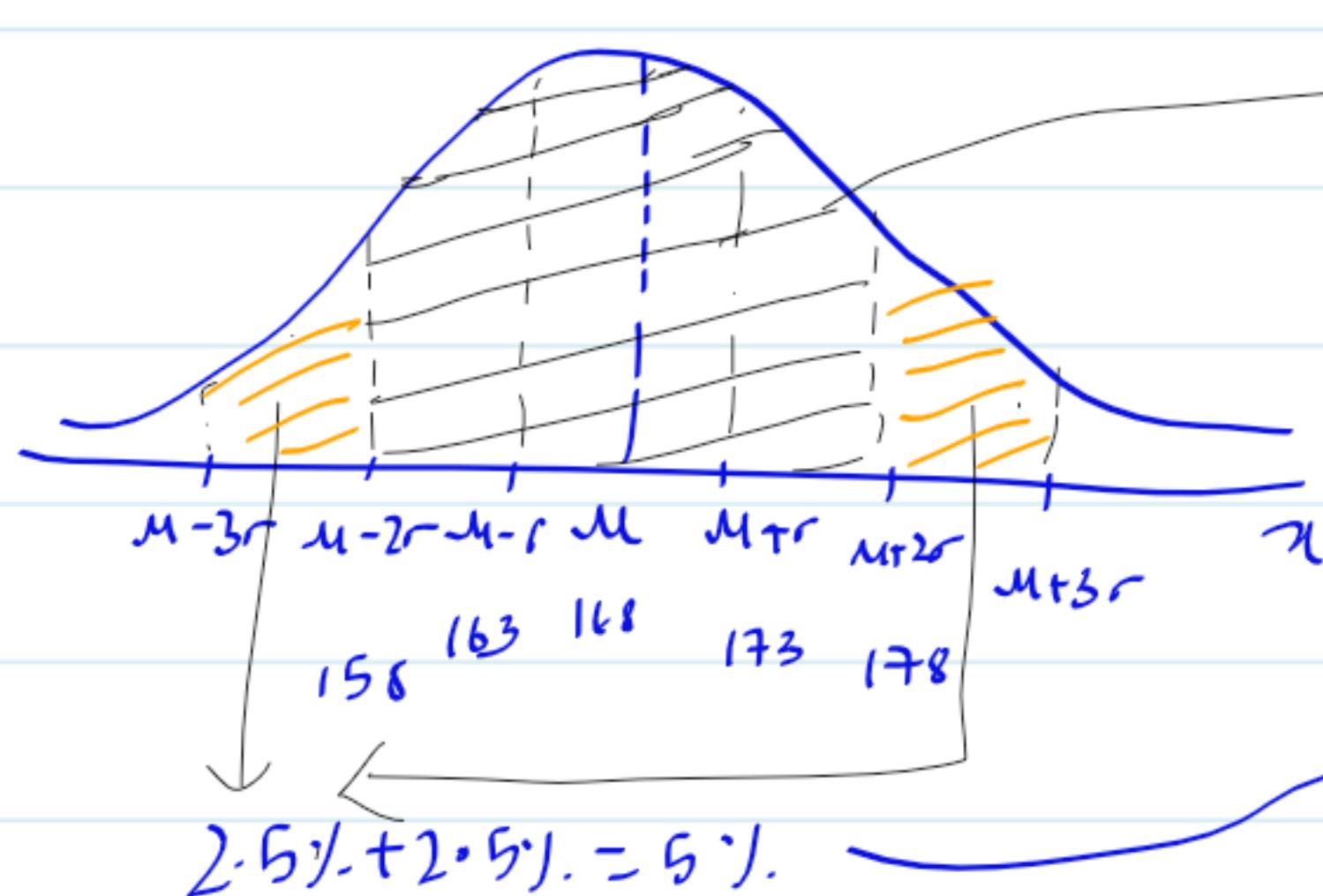
Point estimate of $\mu = \frac{1}{10} \sum_{i=1}^{10} x_i = 168.5$ cm (Assume)

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

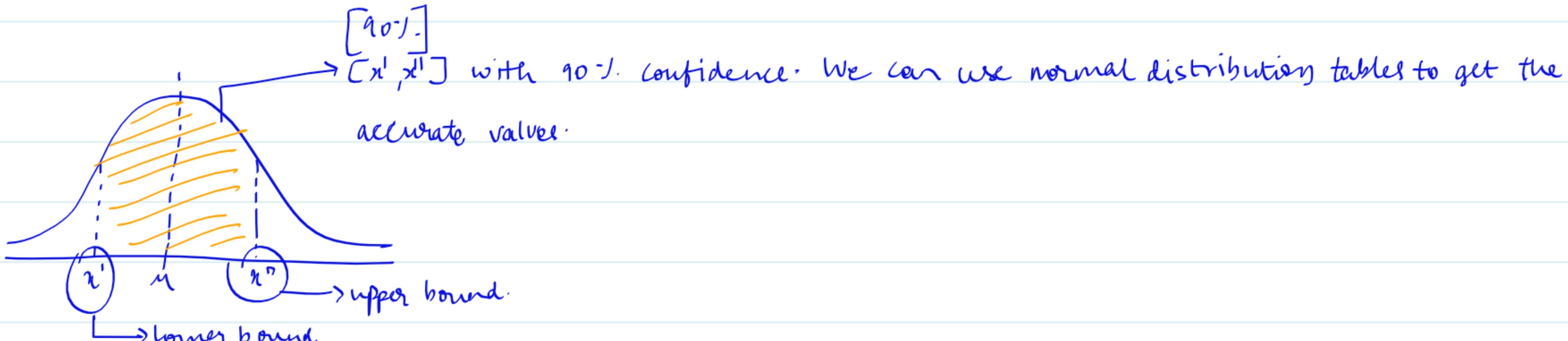
Computing Confidence Interval given a distribution:- We can calculate it easily if we know the underlying distribution.

Let $X \sim N(\mu, \sigma^2)$ with $\mu = 168$, $\sigma = 5$ cm

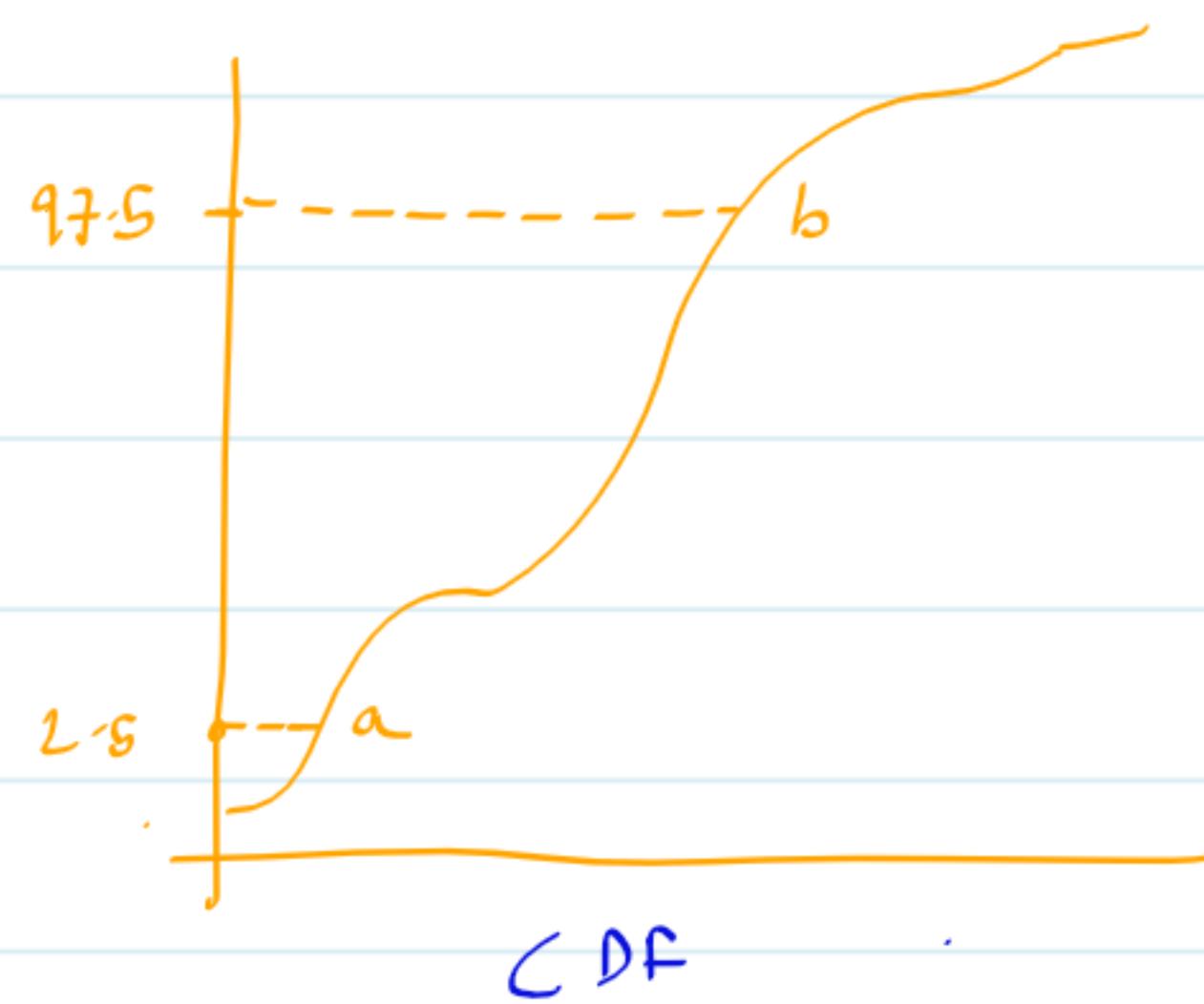


$(\mu - 2\sigma, \mu + 2\sigma)$ contains 95% of the data \rightarrow C (Confidence interval)

Remaining 5% of the data.



→ 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.

$$\{x_1, x_2, x_3, \dots, x_{10}\} \quad n = 10$$

What is the 95% CI of μ of this dist?

↳ case ① :- σ is given

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

$$\bar{x} \sim N\left(\mu, \frac{\sigma^2}{10}\right) \rightarrow \sigma$$

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

$(\bar{x} - 2\sigma) \quad (\bar{x} + 2\sigma)$

Looks like a gaussian curve.

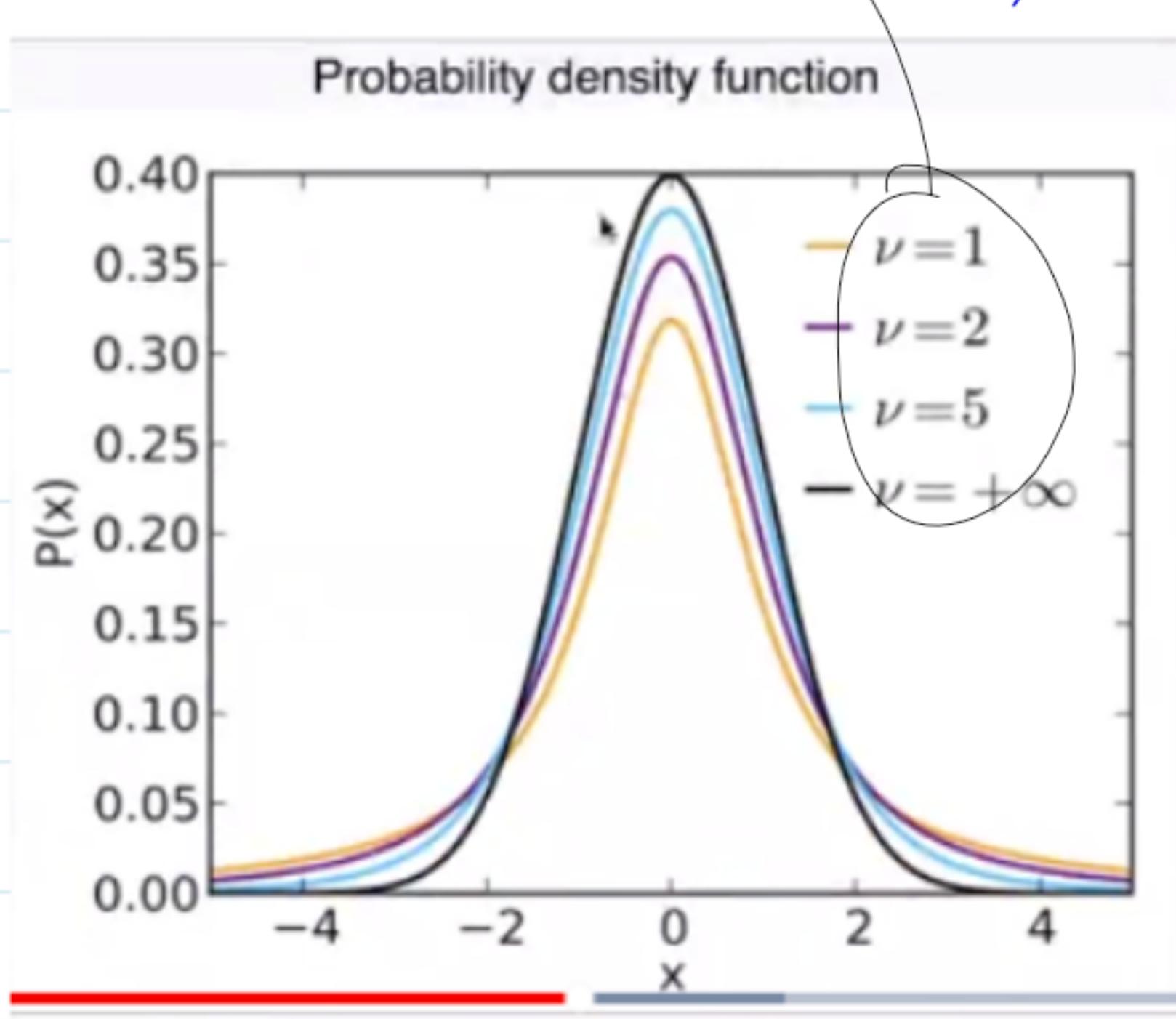
As degrees of freedom increases,
peak also increases.

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

In this case we use t-distribution. It states

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

↳ degrees of freedom.



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

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

Confidence Interval Using Bootstrap :-

→ prob based techniques for CI & other statistics

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

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

Using only this sample, compute medians of X .

= some values may repeat. It's called sampling with repetitions.

steps:- ① Generate new samples using this sample.

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

↑

Random sample of size m generated from S

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

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

↓ Using sampling with repetition.
↓ K samples.

Since $m \leq n$.
These are known as **Bootstrap samples**.

$$\begin{cases} S_1 : x_1^{(1)}, x_2^{(1)}, x_3^{(1)}, \dots, x_m^{(1)} \rightarrow m_1 \\ S_2 : x_1^{(2)}, x_2^{(2)}, x_3^{(2)}, \dots, x_m^{(2)} \rightarrow m_2 \\ S_3 : \vdots \\ \vdots \\ S_K : x_1^{(K)}, x_2^{(K)}, x_3^{(K)}, \dots, x_m^{(K)} \rightarrow m_K \end{cases}$$

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

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

↓ sort them in increasing order.

$$\begin{matrix} m'_1, m'_2, m'_3, \dots, m'_{25} \\ \vdots \\ m'_{25}, \dots, m'_{975}, \dots, m'_{1000} \end{matrix}$$

\downarrow

$$95\% \text{ confidence interval} = [m'_{25}, m'_{975}]$$

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

→ Used a lot in ML.

How? :-

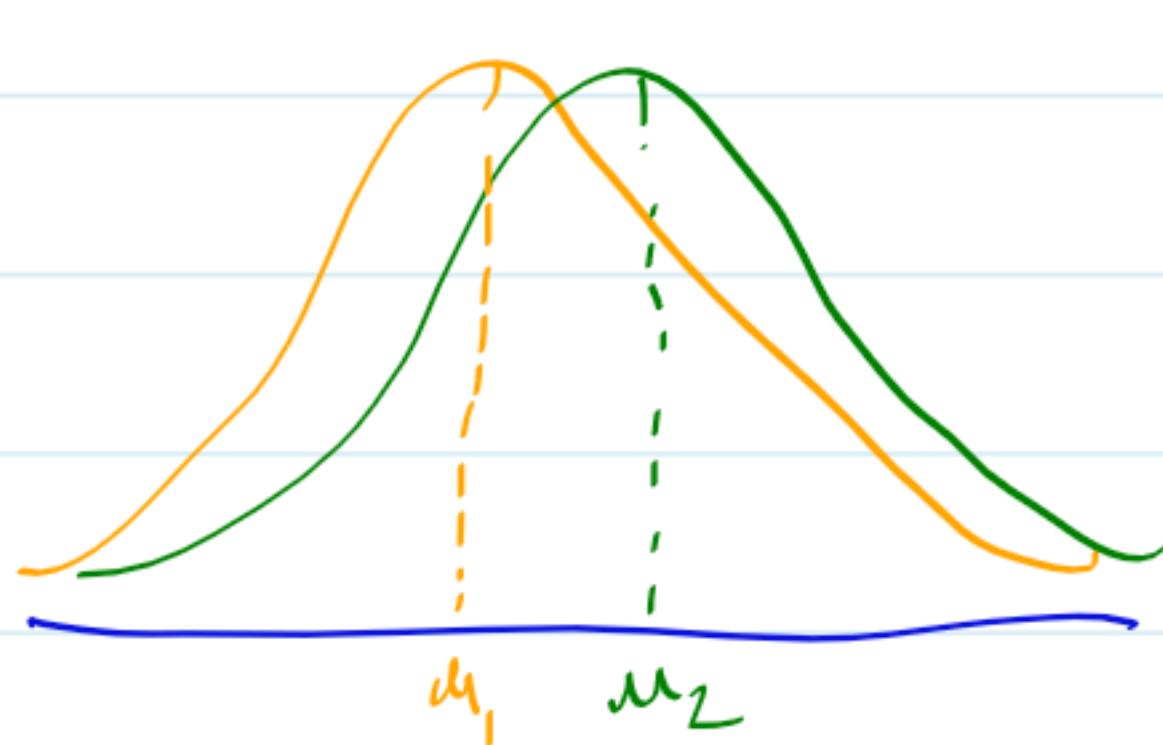
```
from sklearn.utils import resample
```

:

$s = \text{resample}(x, n_samples = \textcircled{m})$ → To generate bootstrap sample.
Sample size.

Hypothesis Testing :-

Assume there are two classes $c1, c2$. Both class heights are taken. Histograms are plotted.



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

Sol :- ① Choosing a test statistic

$\leftarrow \mu_2 - \mu_1 \rightarrow$ 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 $\mu_1 \neq \mu_2$.

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

\Rightarrow Assume H_0 is true. $\Rightarrow \mu_1 \approx \mu_2 \Rightarrow c1 \approx c2$

if p-value = 0.9 \Rightarrow probability of observing $\mu_2 - \mu_1$ is 0.9 if H_0 is true

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

if p-value = 0.05 \Rightarrow 5% chance that 10cm if H_0 is true -

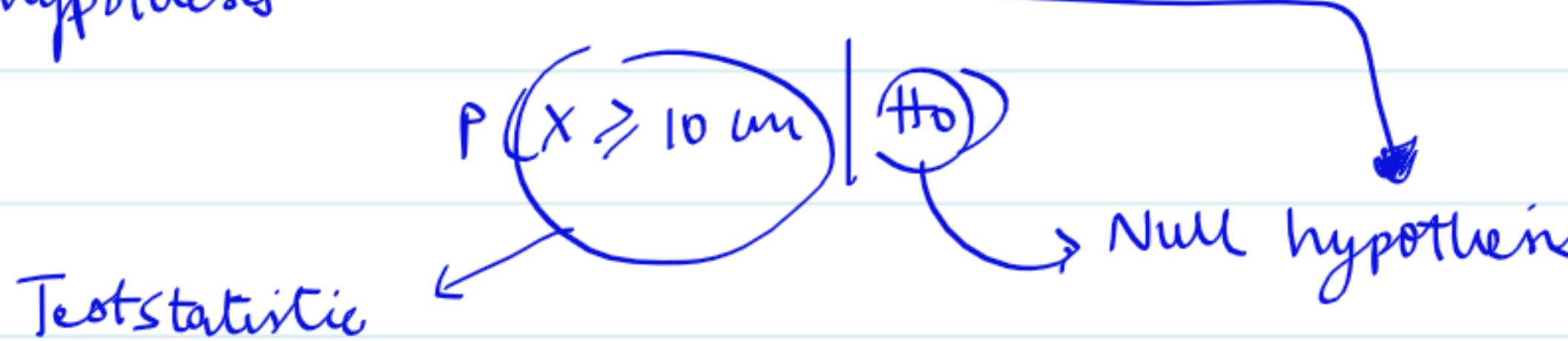
Hypothesis Testing (Additional Video):-

C1 \rightarrow 50 students $\rightarrow \mu_1$ (mean)

C2 \rightarrow 50 students $\rightarrow \mu_2$ (mean)

Step 0: $\Delta \text{ or } X = \mu_2 - \mu_1 = 10$ (assume) \Rightarrow 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



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

$< 5\%$ is typically considered small
 $\Rightarrow P(X \geq 10 \text{ cm} | H_0) = 0.01 \text{ or } 1\%$.

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

Since the probability is small, implies H_0 is less probable $\Rightarrow H_0$ may not be true.

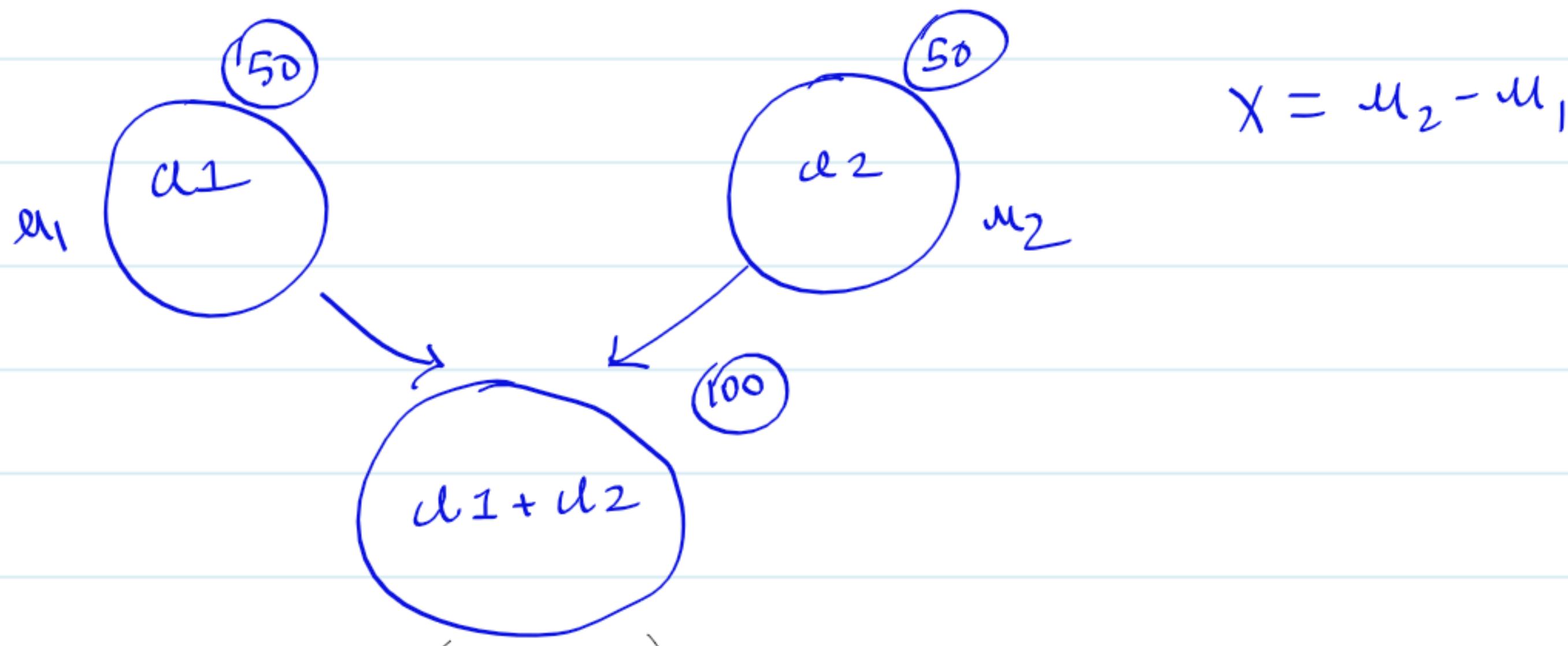
\therefore We reject the null hypothesis (H_0)

\Rightarrow The original assumption that there's no difference is incorrect

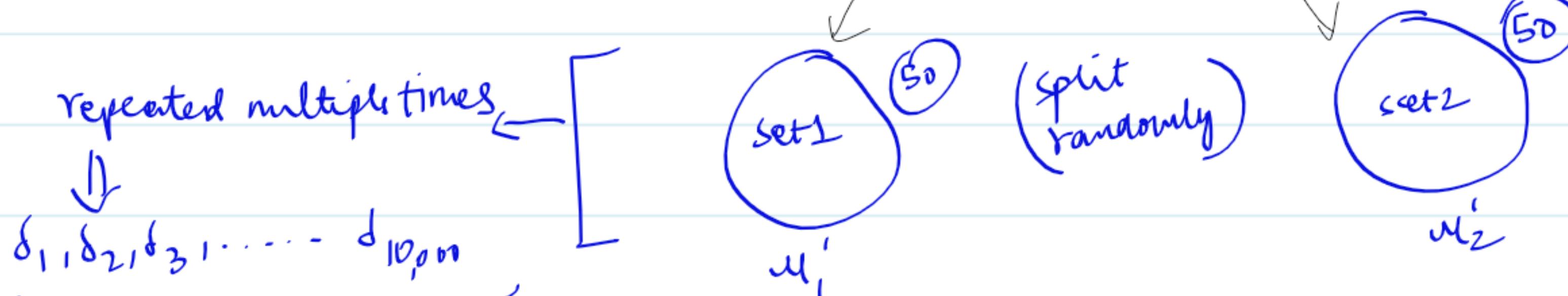
\Rightarrow There is difference b/w the heights -

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

The way we calculate is, we use resampling & permutation testing.



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



Simulated δ 's with H_0 as true.

\Leftrightarrow sorted: $\delta'_1, \delta'_2, \delta'_3, \dots, \delta'_{10000}$

≤ 10
8000

≥ 10
2000

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

\Leftrightarrow Simulated diff in class heights with sample size = 50 * with $H_0 = \text{true}$

\Leftrightarrow Since randomly picked -

$$\delta'_1, \delta'_2, \delta'_3, \dots \dots \dots \quad | \quad \delta'_{10000} \\ \leq 10 \quad \geq 10 \\ 9900 \quad 100 \Rightarrow P(X \geq 10 \text{ cm} | H_0) = 1\% \Rightarrow H_0 \text{ is true. Hence accepted.}$$

→ 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 basic 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.
 $P(X = 5)$ coin is not biased towards heads $= P(\text{obs} | H_0)$
 Observation Assumption.

\downarrow
 $P(X = 5 | H_0) = \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} = \frac{1}{2^5} \approx 0.03 = 3\%$.
 Since heads is 5 times.
 coin is not biased, $P(H) = \frac{1}{2}$

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

$\int_{\text{p-value}}^{\text{Probability (observation by experiment | assumption is true)}} = 3\%$.
 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 $\left[\Rightarrow \text{coin IS biased.}\right]$

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 $P(X | H_0) = \frac{1}{2^3} = \frac{1}{8} = 12.5\%$ → 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 distribution:-

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 distribution?

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

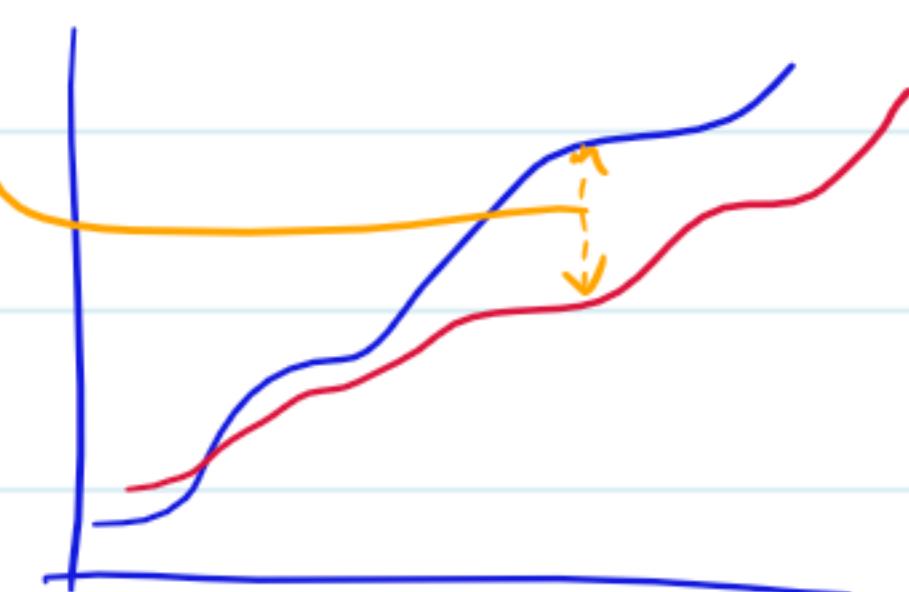
Null hypothesis, $H_0 : X_1$ & X_2 have the same distribution

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 equation

Null hypothesis is rejected at level α , if

$$D_{n,m} > c(\alpha) \sqrt{\frac{n+m}{nm}}$$

$[\alpha = \text{required p-level}]$

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 $D_{n,m} > 1.358 \sqrt{\frac{1000+5000}{1000 \times 5000}}$

⇒ 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(data1, '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) :-

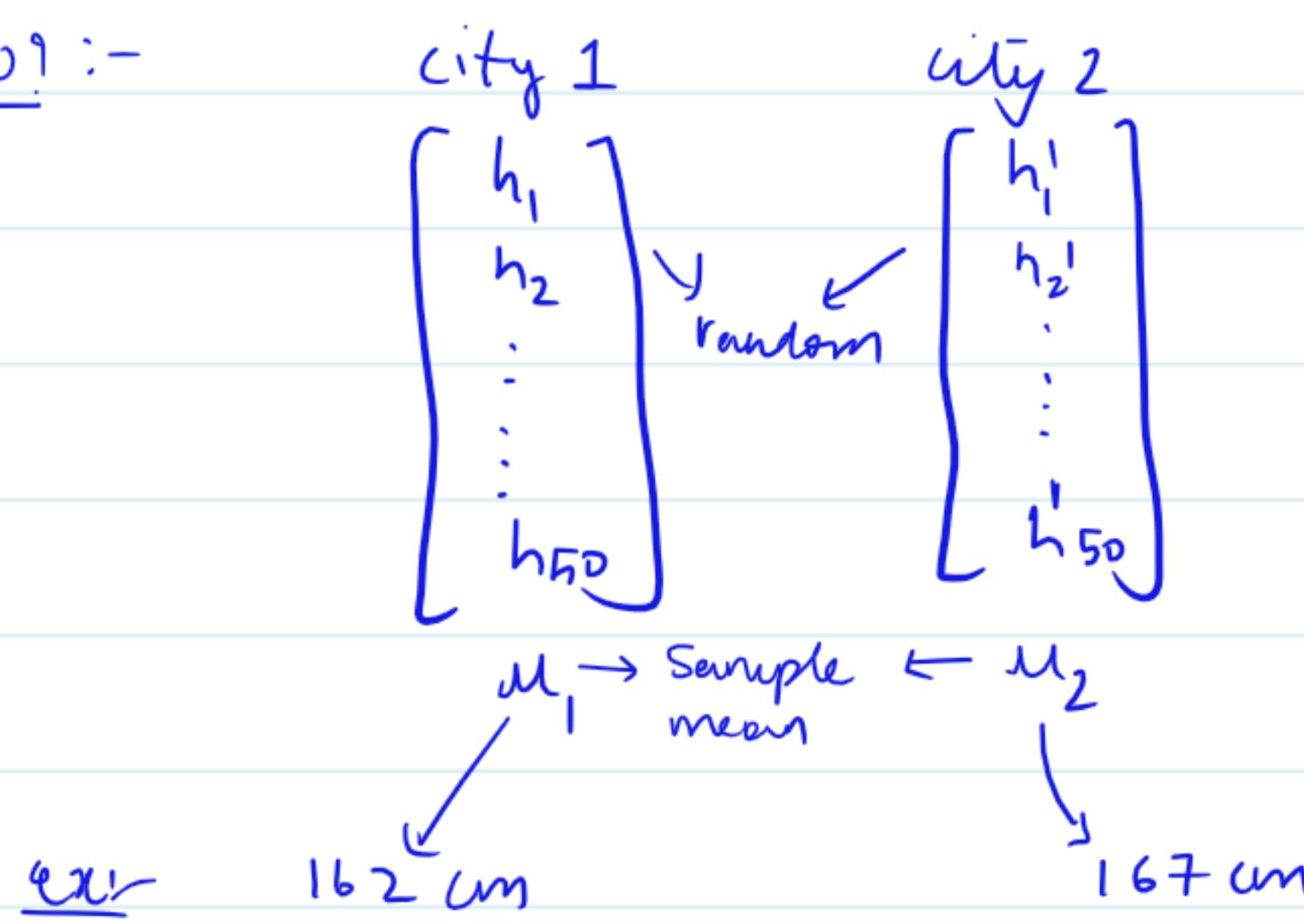
task: city 1 city 2

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 :-



test statistic :- $X = m_2 - m_1 = 167 - 162 = 5 \text{ cm}$

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

compute :- $p(X=5 | H_0)$ [Prob 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 :- $p(X=5 | H_0) = 0.2 = 20\%$

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

There is no difference in heights.

Significant \Rightarrow Assumption is true. $\therefore H_0$ is accepted.

case 2 :- $p(X=5 | H_0) = 0.02 = 2\%$ \Rightarrow low. Assumption is incorrect. Alternative hypothesis (H_1) is accepted \Rightarrow Heights are not the same.

Resampling & Permutation test for cities example

$\bar{x} = \bar{m}_1 - \bar{m}_2 \rightarrow$ diff in sample means (Let it be 5cm)

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

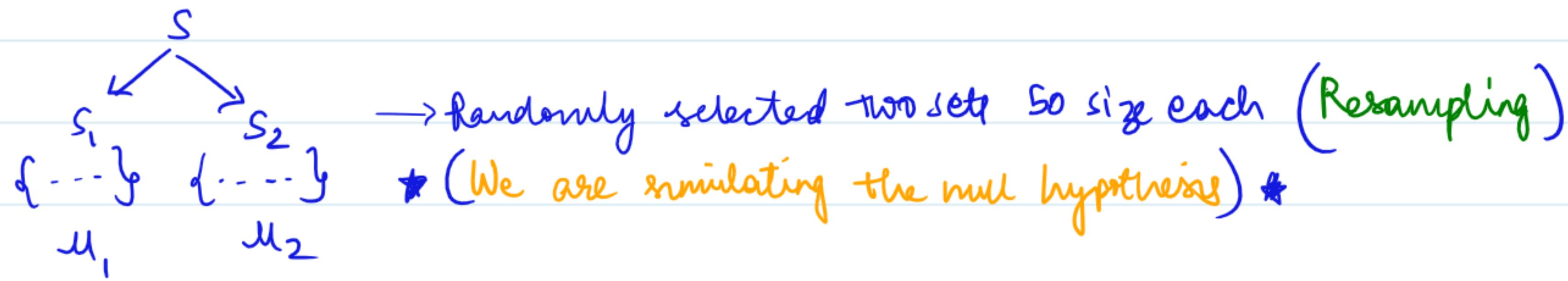
$$C_1 \quad C_2$$

$$\begin{bmatrix} h_1 \\ h_2 \\ h_3 \\ \vdots \\ h_{50} \end{bmatrix} \quad \begin{bmatrix} h'_1 \\ h'_2 \\ h'_3 \\ \vdots \\ h'_{50} \end{bmatrix}$$

Step 1 :- $S = \{h_1, h_2, h_3, \dots, h_{50}, h'_1, h'_2, h'_3, \dots, h'_{50}\}$

$C_1 \cup C_2$

Step 2 :-



$$m_2 - m_1 \rightarrow 3 \text{ cm } (\delta^1)$$

$$\text{repeat, } m_2 - m_1 \rightarrow -2 \text{ cm } (\delta^2)$$

$$\text{"}, m_2 - m_1 \rightarrow 1 \text{ cm } (\delta^3)$$

⋮ ⋮ ⋮ ⋮

$$k \text{ times, } m_2 - m_1 \rightarrow 6 \text{ cm } (\delta^k), \text{ Let } k=1000$$

Step 3 :- Sort $\delta_{1,2}$

$$\delta'_1 \leq \delta'_2 \leq \delta'_3 \leq \delta'_4 \dots \leq \delta'_k$$

Case 1 :- $P(\text{diff} \geq 5 \text{ cm} | H_0)$

$$\underbrace{\delta'_1 \leq \delta'_2 \leq \delta'_3 \leq \delta'_4 \dots \leq \delta'_{100}}_{\leq 5 \text{ cm}} \leq \dots \leq \delta'_{1000} \underbrace{\dots}_{> 5 \text{ cm}}$$

20% of the data

$$P(\text{obs-diff} | H_0) = P(x \geq 5 \text{ cm} | H_0) = 20\%$$

↳ significant ($> 5\%$)

∴ Assumption must be true · H_0 is accepted.

Case 2 :- $\delta'_1 \leq \delta'_2 \leq \delta'_3 \leq \delta'_4 \dots \leq \delta'_{970} \leq \dots \leq \delta'_{1000}$ → Simulated values originated from
97% of the data 3% of the data assuming that null hypothesis is true.

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

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

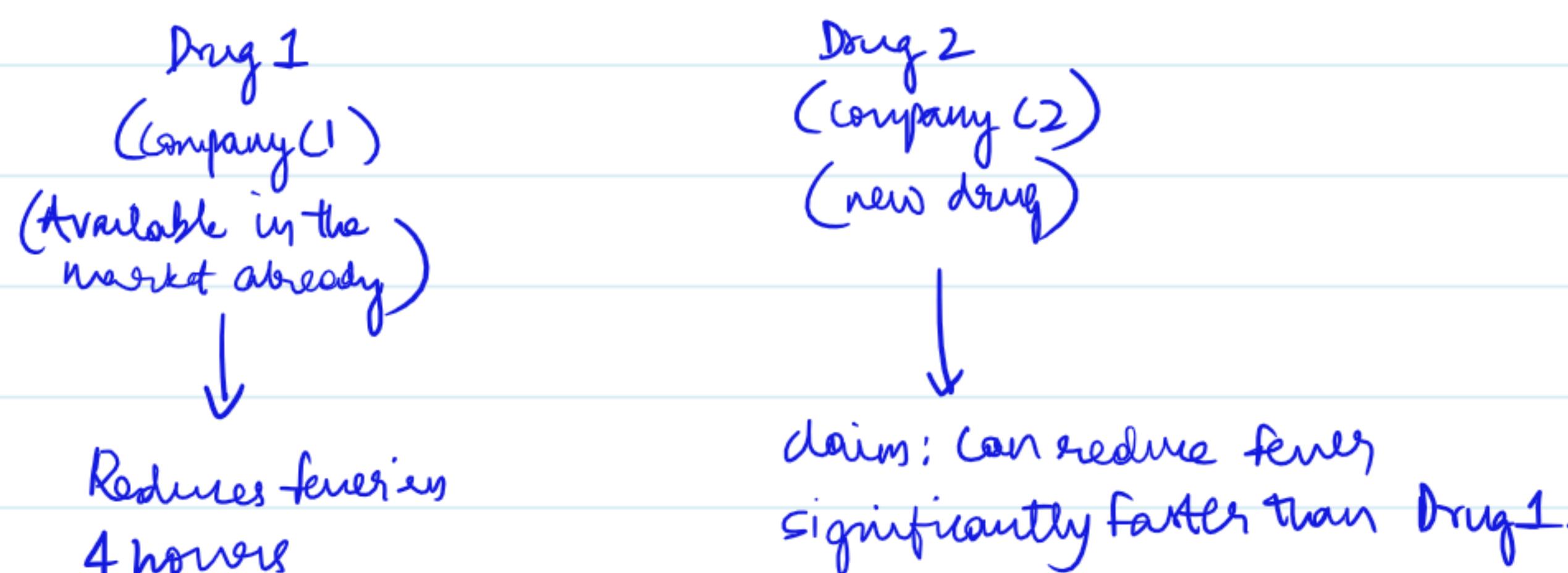
→ if $P(\text{obs} | H_0)$ is very high in this case, it implies error in sample & if we increase sample size, well be fine ·

How to use hypothesis testing :-

KS testing → used hypothesis testing to determine if two RV have the same dist or not.

Applications of hypothesis testing:

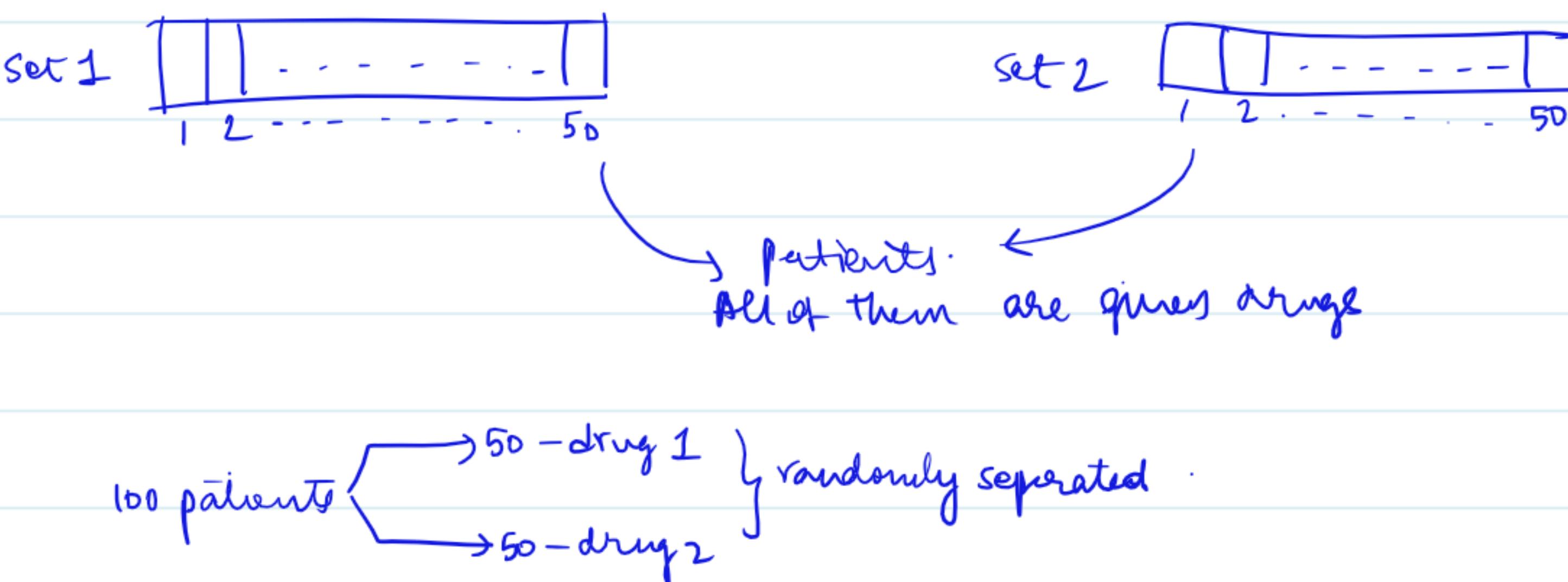
ex:- Designing Medicine



Experiment → To determine if claims is true or not.

Drug 1

Drug 2



Each patient's fever recovery time after taking drug is tracked.

$$\mu_1 = 4 \text{ hrs}$$

$$\mu_2 = 2 \text{ hrs}$$

Instead of concluding drug 2 is better from this example, we perform hypothesis testing.

① $H_0: D_1 \text{ & } D_2 \text{ take same time to reduce fever} \rightarrow \text{Null hypothesis}$

$H_1: D_1 \text{ & } D_2 \text{ take different times} \rightarrow \text{Alternate hypothesis}$

② Test Statistic :- $X = \mu_2 - \mu_1 = 4 - 2 = 2 \text{ hours}$

compute $P(X \geq 2 | H_0)$.

Let $P(X \geq 2 | H_0)$ is small $\rightarrow < 5\%$. \Rightarrow The probability of observing $X \geq 2$ is very small ($< 5\%$)

\Rightarrow Observation & H_0 do not agree with each other.

$\Rightarrow H_0$ is incorrect since observation is true & real

\Rightarrow We reject H_0 & accept H_1 .

\rightarrow Resampling & permutation test are used to find $P(X \geq 2 | H_0)$

③ Significance level (α) :- Generally significance level = 5%.

If p-Value $< \alpha$, we reject H_0 . Else we accept H_0 .

We take smaller α if we want to be very very sure (ex:- Medicine, $\alpha = 1\%$).
Commerce, $\alpha = 3\%$.

Proportional Sampling :-

Let d be an array

d_1	d_2	d_3	d_4	d_5
2.0	6.0	1.2	5.8	20.0

1 2 3 4 5

$n = 5$

task :- pick an element from the n elements such that Probability of picking an element is proportional to d_i

Normally, each has the same probability.

But according to proportional sampling, d_5 should have the highest probability.

d_3 should have the least probability.

How?

Step 1a :- $s = \sum_{i=1}^n d_i = 35$ (in this case) [compute the sum]

Step 1b :- $d'_i = d_i/s$ [normalizing values using the sum]

Now all values lie b/w 0 & 1 and sum of all values (d'_i) = 1

Step 1c :- Computing Cumulative Normalized Sum

$d'_1 = 0.0571$	$\tilde{d}_1 = d'_1 = 0.0571$
$d'_2 = 0.171428$	$\tilde{d}_2 = \tilde{d}_1 + d'_2 = 0.228528$
$d'_3 = 0.0343$	$\tilde{d}_3 = \tilde{d}_2 + d'_3 = 0.262828$
$d'_4 = 0.1657$	$\tilde{d}_4 = \tilde{d}_3 + d'_4 = 0.428528$
$d'_5 = 0.5714$	$\tilde{d}_5 = \tilde{d}_4 + d'_5 = 1.00$



Cumulative Normalized Value

Step 2 :- Sample one value from uniform random variable b/w 0 & 1

$r = \text{numpy.random.uniform}(0.0, 1.0, 1)$

Range Number of values.

Step 3 :- Proportional Sampling :

```

if  $r \leq \tilde{d}_1$  return 1
else if  $r \leq \tilde{d}_2$  return 2
else if  $r \leq \tilde{d}_3$  return 3
else if  $r \leq \tilde{d}_4$  return 4
else if  $r \leq \tilde{d}_5$  return 5

```