Probability & statistics

- fundamental area
- histograms, PDF, COF, mean, var, std-dev
  ↳ prob & stat

→ random variable :-

eg :- dice : six-sides → {1, 2, 3, 4, 5, 6}
experiment → roll a fair dice ↑ equally likely

r.v → X = { }
↳ it can be anything {1, 2, ... 6}

→ tossing coin :-

r.v → Y = {H, T}
↳ equally likely

→ alway read eqn and code in english.

→ dice

* p(x is even) = prob. that x → r.v is even = $\frac{1}{2}$

Types of random variable :-

- Discrete random value :-

  → eg :- dice → X = {1, 2, ... 6}

  ↳ a r.v which have set of outcomes

  ↳ a r.v can take one value in a finite set of values is called d.r.v.

- Continous r.v :-

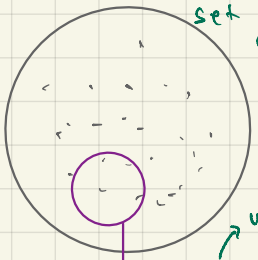
  eg :- height of randomly picked student → 120cm ... 190cm
  ↳ here the set is not finite it's continous so, c.r.v.

Outlier :-

mean : ⟶ Outlier

Variance :- MAD

Population & sample :-

→ 

set of all ppl in the world

→ estimate the average/mean height of a human

we have taken random sample.

→ sample (sample of size 1000)

mean of a population ← $\mu = \frac{1}{7B} \cdot \sum_{i=1}^{7B} h_i$ → we can't find that.

→ a "sample" is a subset of population.

mean of a sample ← $\bar{h} = \mu_{est.} = \frac{1}{1000} \times \sum_{i=1}^{1000} h_i$

↳ heights in my sample.

→ As sample size inc^{n} $\bar{x} = x$ our sample mean will converge towards population mean.

Gaussian or Normal distribution and its PDF

PDF of a Gaussian disb r.v

→ disb. are very simple models for real world.

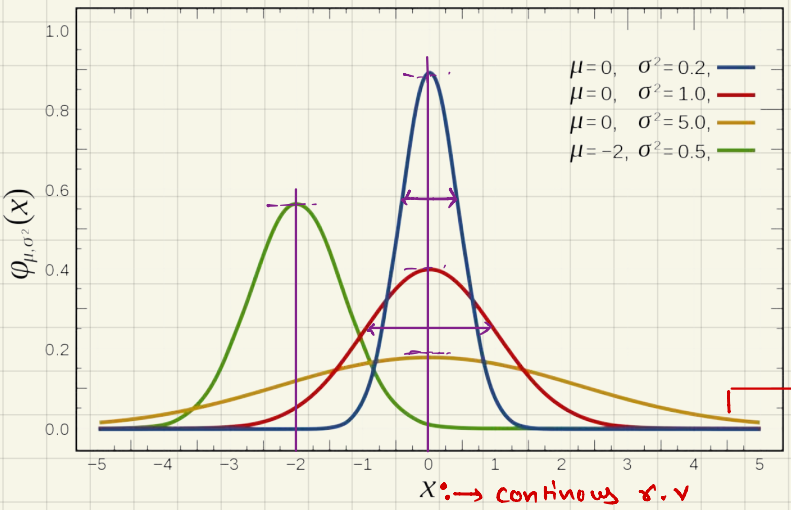
→ If you know the PDF you can plot CDF.

Let's assume :-

X : continous r.v
and
X PDF is look like "bell shaped"
then we will say 'X' has a distribution which is a gaussian disb^n.

→ why should we learn about this disb^n ?

→ because lot of things in nature follow this disb^n. (not everything)
for eg:- P.L (petal-length) follow gaussian distribution.



μ :→ mean
σ^2 → variance → spread / space
↳ parameters of disb^n

If σ^2 changes → spreadness changes
If μ changes → peak of hill changes

→ PDF of a continous r.v

→ If you tell me my 'X' (continous r.v) follows Gaussian disb^n and if you give me it's mean (μ) and variance (σ^2) nothing else, just these two, then I can tell you what it's PDF is. (without having one value/observation)

→ And if I know PDF, I can easily plot CDF.

If PL is G.D
then μ, σ^2

parameters of a Gaussian disb: μ, σ^2

Normal / gaussian

$X \sim N(\mu, \sigma^2)$ → X follows gaussian disb^n

follows · mean · variance

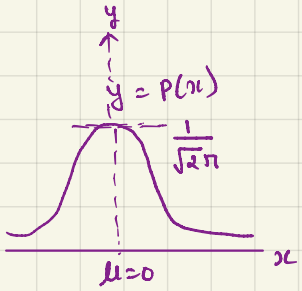
→ probability of finding r.v(x) is this

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

Let assume,

$\mu = 0$ & $\sigma^2 = 1$

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



$$y = e^{-x^2}$$

if $x \uparrow$ then $y \downarrow$ → the rate of reducing 'y' is 'square' of 'exponent'

if $x \downarrow$, then $y \downarrow$

Conclusions

- As x moves away from μ; $y \downarrow$
- symmetric along μ.
- x moves away from μ, 'y' reduces exp.$(-x^2)$

→ $y = \exp(-x)^2$

$x = 0$ $y = 1$

$x = 1$ $y = \exp(-1)$ $= 0.3678$

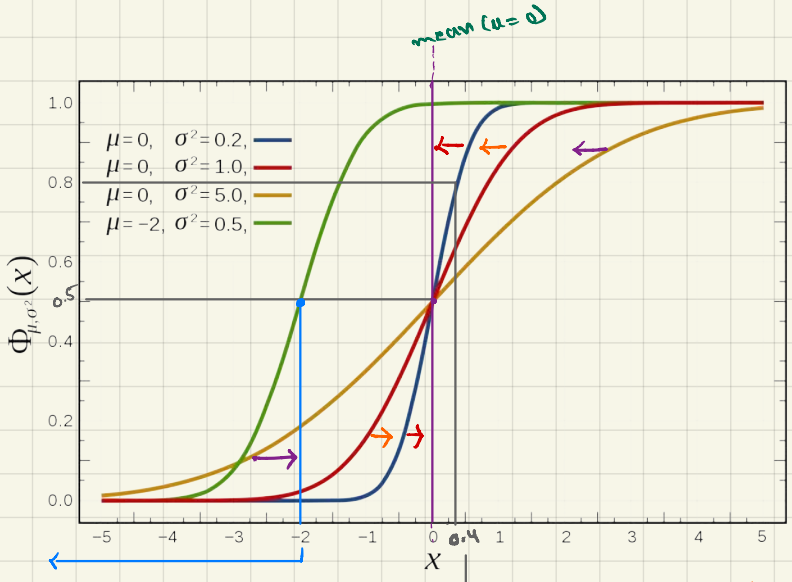
→ 2x

$x = 2$ $y = \exp(-4)$ $= 0.018$ → 20x → $\frac{0.018}{0.36}$

→ 1.5x

$x = 3$ $y = \exp(-9)$ $= 0.000123$ → 100x ↓

CDF



→ as σ increases our graph go more away from mean.

→ Because our PDF (shown above) is symmetric our CDF is also symmetric. so, 50% our pt. lies one side of mean & 50% of pt. lies on other side.

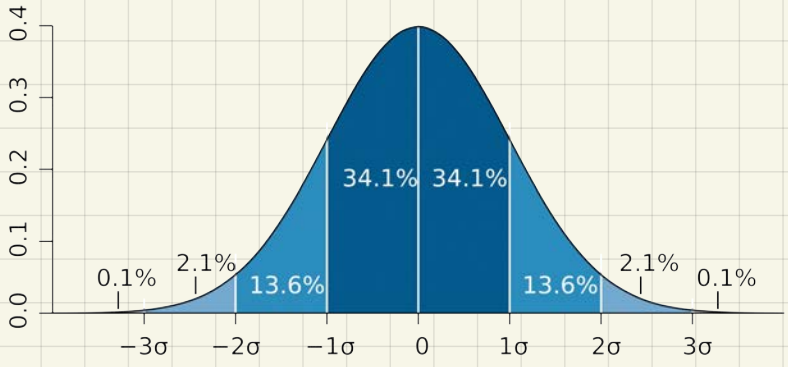
for green, half of my pt. lie before "−2" & after "−2".

(y-axis of CDF tells that) → (means prob. of x having less than equal to 0.4 is 80%)

↳ This tells me that $P(x \leq 0.4)$ is 80%.

Std. deviation (σ)

→ If we know our x → follows Gaussian dist^n and the value of mean & std. deviation is known to us then what can we predict?



A plot of normal distribution (or bell-shaped curve) where each band has a width of 1 standard deviation – See also: 68–95–99.7 rule.

→ If we know $x \sim N(\mu, \sigma^2)$

$\mu = 0$

$\sigma^2 = 4 \dashrightarrow \sigma = \pm 2$

obs. → (mean, std. devi)

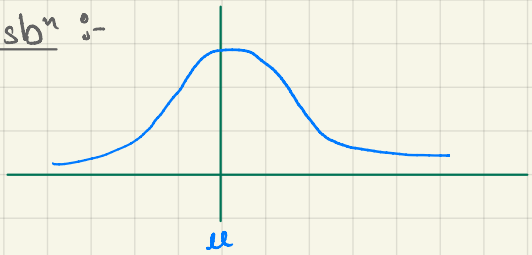
b/w 0 & 2 our 34.1% pt. lies.

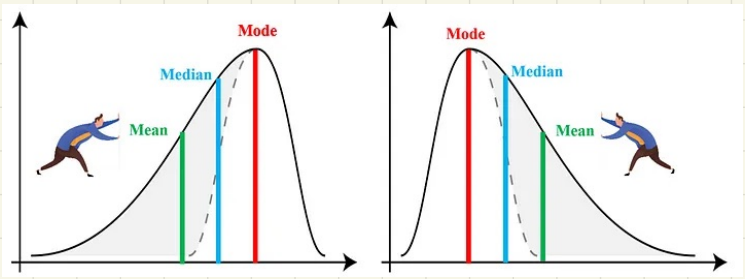
b/w −2 & 2 our 68.2% pt. lies.

b/w −4 & 4 our 95% pt. lies

Symmetric, skew, and kurtosis distribution

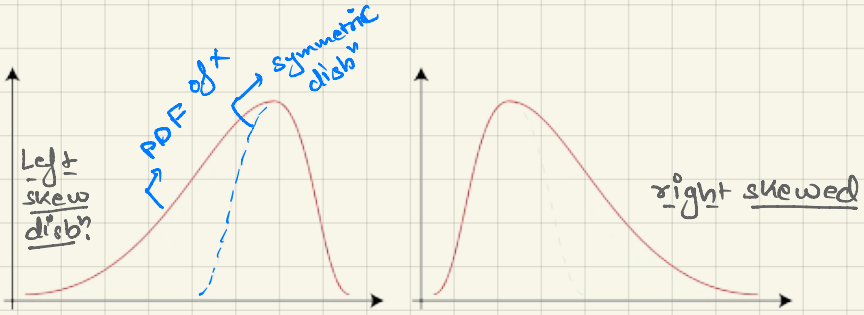
Symmetric disbⁿ :-





μ :→ mean → location or middle
σ^2 :→ spreadness

Skew disbⁿ :- It tells you how dissimilar is the disbⁿ from symmetric disbⁿ.



Negative Skew Positive Skew

There are four mathematical measures of relative skewness.

(a) Karl Pearson's Coefficient of Skewness

$$SK = \frac{\text{Mean} - \text{Mode}}{\text{S.D.}}$$

If mode is ill-defined, then we take

$$SK = \frac{3(\text{Mean} - \text{Median})}{\text{S.D.}}$$

(b) Bowley's Coefficient of Skewness

This is based on quartiles and median and is defined as

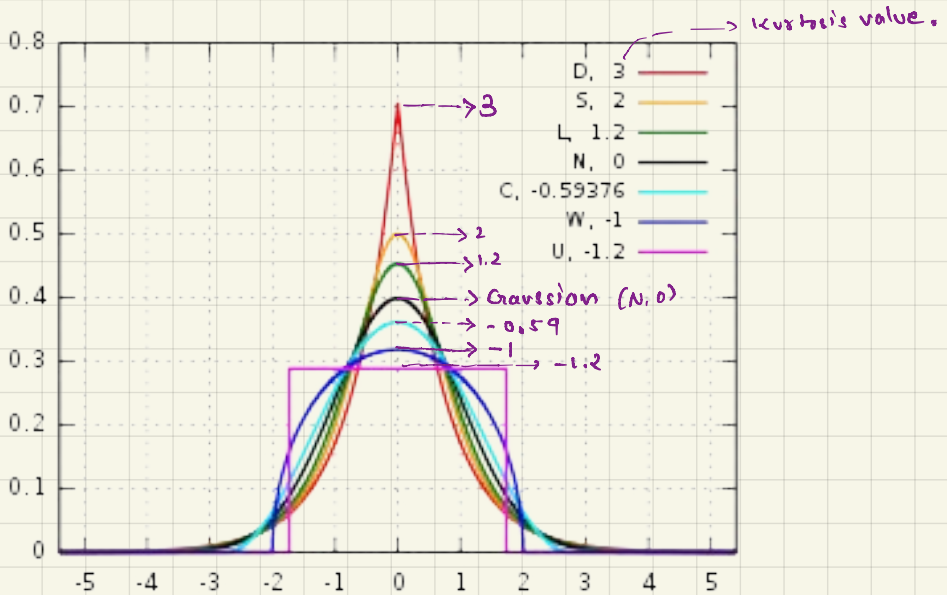
$$SK = \frac{Q_3 + Q_1 - 2\,\text{Median}}{Q_3 - Q_1}$$

This formula is useful when the mode is ill-defined or the distribution has open end classes or unequal class-intervals.

If this value is between:

1. -0.5 and 0.5, the distribution of the value is almost symmetrical
2. -1 and -0.5, the data is negatively skewed, and if it is between 0.5 to 1, the data is positively skewed. The skewness is moderate.
3. If the skewness is lower than -1 (negatively skewed) or greater than 1 (positively skewed), the data is highly skewed.

Kurtosis :- How peaked your distribution is. (it measures that)



Standard normal variate (z) and standardization :-

→ If my r.v (x) follows normal distⁿ and I want to standardized it because after standardization I can say that my 68% data lie b/w → -1 & 1

95% data lie b/w → -2 & 2

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

↳ How to read that → X (r.v) follows N.D & have mean (μ) & variance (σ)

x (feature) having dim. 'n', follows ND & ..... or

$X = [x_1, x_2, \dots\dots\dots x_{50}]$ (assume)

all this are observation of this r.v.

Standardization → method in which I can transform any mean & variance into a std. data.

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

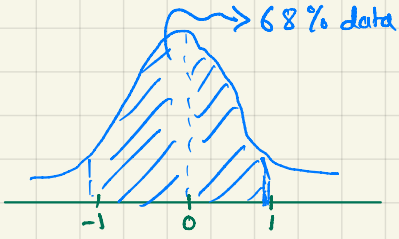
$$X' \sim N(0,1)$$

↳ standard Normal variate

$\Rightarrow X \sim N(\mu, \sigma^2)$ —standardization→ $X' \sim N(0,1)$ (n-dim.)

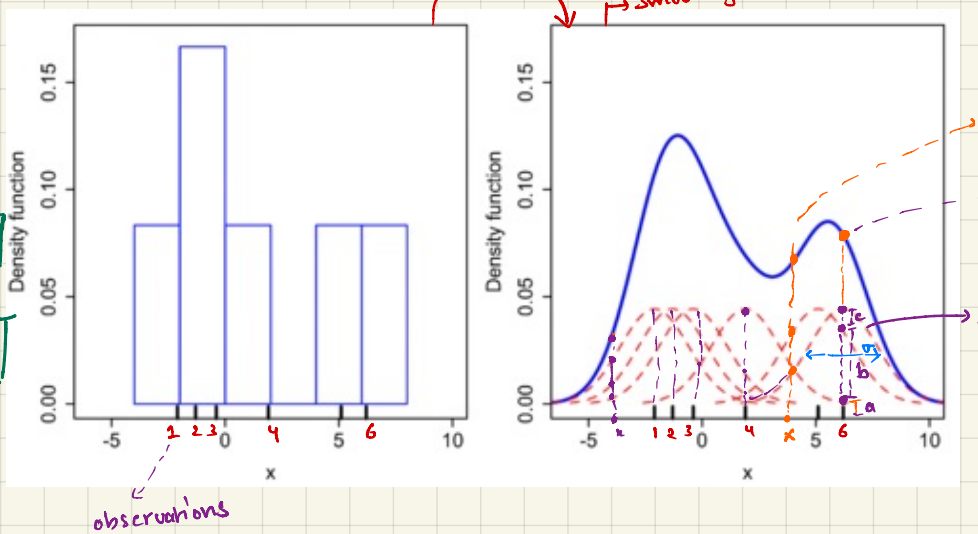
or

$$z \sim N(0,1) = \frac{X - \mu}{\sigma}$$



Kernel density estimation :- a way of converting histogram into PDF.

we can do smoothing using Kernel → smooth form of histogram.



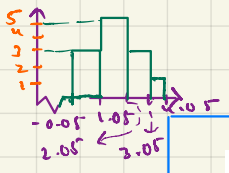
→ I want to calculate height of PDF at 'x'? → Just by adding the Kernels.

→ I calculated this value by adding a + b + c.

→ I have plotted a bell-shaped Kernel or gaussian Kernel here, with center or mean at 6th point.

→ We have taken 'mean' in consideration what about Variance? → The variance here is also called bandwidth of Kernel

Bandwidth



class limit → 0-1, 1-2, 2-3, 3-4

class boundaries → -0.05 - 1.05, 1.05 - 2.05, 2.05 - 3.05, 3.05 - 4.05

eg:- assume

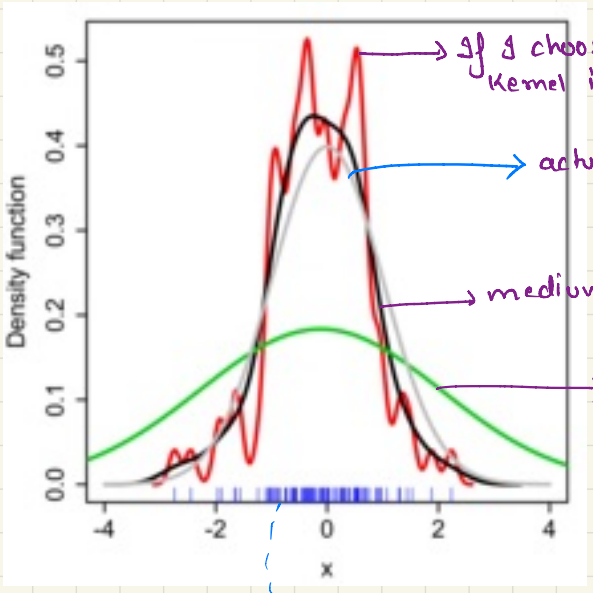
pt: {0, 0.5, 1, 1.2, 1.3, 1.4, 1.8, 2, 2.1, 2.8, 3, 4}

→ histogram having bins "4"

since, our data points are integer. we can do to set the class boundary subtract or add by 0.05

Class boundaries	frequency
-0.05 - 1.05	3
1.05 - 2.05	5
2.05 - 3.05	3
3.05 - 4.05	1

→ 4 bins



→ If I choose my b.w 'small' then my kernel is very narrow or zaggered. (histogram with lots of bin).

→ actual distⁿ.

→ medium b/w.

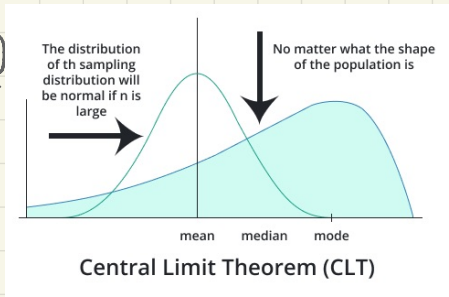
→ very large 'b.w'.

→ observations.

→ To find the write b.w what does 'seaborn' do is?

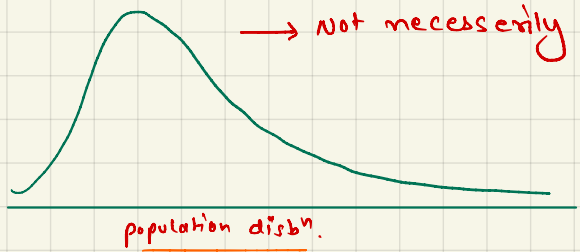
↳ They choose a very small 'b.w' initially & see how much zaggered is our PDF is and then they incⁿ the 'b.w' till the time the line become smooth enough.

Sampling distⁿ and central limit theorem (CLT)



Central Limit Theorem (CLT)

X : (distⁿ of incomes)



→ not necessarily Gaussian.

Sample mean

→ I take 'random sample' of size $n(=30)$ Let → $S_1 \rightarrow \bar{x}_1$

→ again I take 'random sample' of size $(n = 30) \rightarrow S_2 \rightarrow \bar{x}_2$

I took independently. It doesn't depend on one another

→ ——— → $S_m \rightarrow \bar{x}_m$

→ $\bar{x}_1, \bar{x}_2, \bar{x}_3$ ——— $\bar{x}_m$: → m sample means

→ $\bar{x}_i$ ~ disb^n

↳ disb^n of sample means

→ disb of $\bar{x}_i$ = Sampling disb^n of sample-mean

CLT (only for finite mean & variance)

If X: (population disb^n): have finite μ & σ^2

infinite mean & varr
↳ (pareto disb^n.)

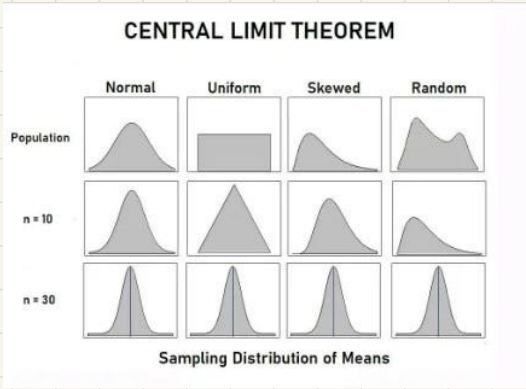
samples ← S_1, S_2, S_3 ---------- S_m (same, same)
of size 'n'.

sample means ← $\bar{x}_1$, $\bar{x}_2$, $\bar{x}_3$ ---------- $\bar{x}_m$ → disb^n of $\bar{x}_i$ = sampling disb^n of sample mean.

CLT:-

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

gaussian disb^n



⇒ eg:- X: μ, σ^2 (income)

any disb^n need not be Gaussian.

n: Let n = 30 → S_1, S_2, ---------- S_m

Let, m = 1000

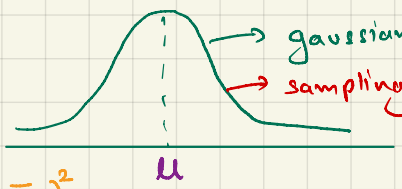
1000 × 30 = 30K (people income)
m n

$\bar{x}_1$, $\bar{x}_2$ ---------- $\bar{x}_m$ (m = 1000)

mean income in each sample

$$\rightarrow \bar{x}_\mu = \frac{\bar{x}_1 + \bar{x}_2 + \cdots + \bar{x}_m}{m}$$

Plot



$\mu \approx$ mean of $\bar{x}_i$'s $= \bar{x}_\mu$

$$\frac{\sigma^2}{n} \approx \text{variance of } \bar{x}_i\text{'s} \approx \frac{\sum_{i=1}^{m} (\bar{x}_i - \bar{x}_\mu)^2}{m}$$

↳ rule of thumb → $n \geq 30$ → not necessarily $n \to \infty$

Note:- By looking at '30 k' data points / observations. I am able to estimate the whole population → mean, variance

→ To understand the population mean or variance (μ, σ^2) of any disb^n (not necessary gaussian) I just need to know they are finite or well defined.

Quantile - Quantile (Q-Q) plot :-

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

Q) is 'X' Gaussian disb?
- → QQ Plot (graphical method)
- → statistical testing (KS, AD)

Steps for Q-Q plot

→ sort x_i's & compute percentiles

$x_1, x_2 \ldots\ldots x_{500}$

↓ sort (asc.)

$x_1', x_2', \ldots\ldots x_{500}'$

↓ Percentiles

① ② ③ ⋯ (100) → 100th percentile

$x_5', x_{10}', x_{15}', \ldots\ldots x_{500}'$

↓

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

1st - percentile value of x_i's

→ std. Normal / gaussian disbn

→ $Y \sim N(0,1)$ ⟶ (create this)

$y_1, y_2, \ldots\ldots y_{1000}$ → 1000 obsc. from $N(0,1)$

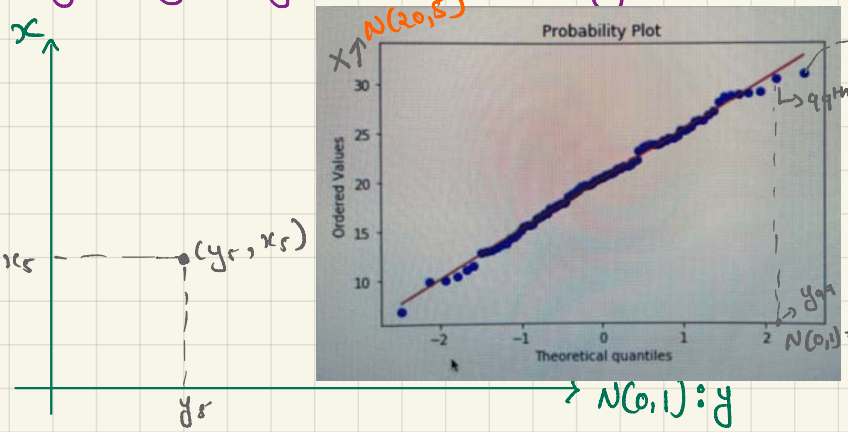
↓ sort (asc.)

$y_1', y_2', y_3' \ldots\ldots y_{1000}'$

↓ percentiles

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

→



→ $(y^{(i)}, x^{(i)})$ ∀ i : 1 → 100

↳ q_i^{th} percentile

if $(y^{(i)}, x^{(i)})$ i : 1 → 100 lie on a st. line then X & y have a similar disbn.

⇒ Q-Q plot ↳ $x \sim N(\mu, \sigma^2)$?

→ Theoritical quantiles

Q-Q Plot

```
#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)

# 0 to 100th percentiles of std-normal
for i in range(0,101):
    print(i, np.percentile(std_normal,i))
```

(mean) → (σ) (std. dev.)

→ generate random no. from a normal/gaussian distb having mean → 0 & std. dev. → 1 & generate '1000' observation/sample. and put all value in std_normal.

```
0 -2.9941560608546496
1 -2.331155489997273
2 -2.102462010297869
3 -1.9030070976001443
4 -1.7878574029822945
5 -1.6383831947251253
6 -1.5935117657666127
7 -1.5315554277910335
8 -1.4362998197560297
9 -1.394737614240833
10 -1.3132161092517054
...
94 1.4626263448353378
95 1.5608572685665627
96 1.6966233988361676
97 1.9147750543381148
98 2.0157626321951363
99 2.246560915109006
100 2.7918505373969498
```

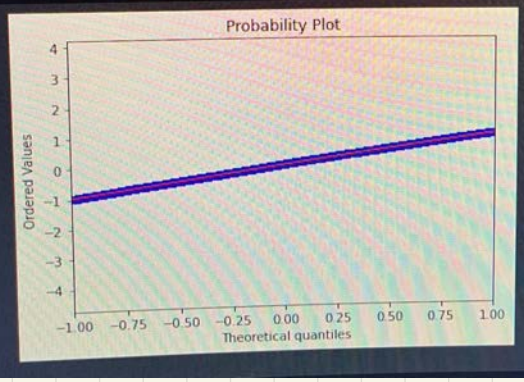
→ my theoretical value goes from (-3 to 3) approximately

```
[ ] import matplotlib.pyplot as plt
```

```
# generate 100 sanples from N(20,5)
measurements = np.random.normal(loc = 0, scale = 1, size=10000)
#try size=1000

plt.xlim(-1,1)

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

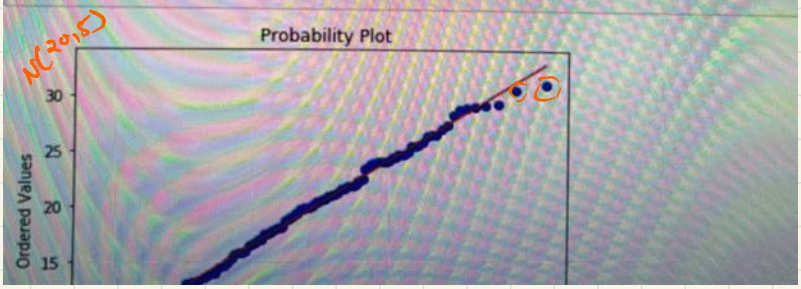
Q-Q plot → use this library for plotting.

→ it says I want to compare it with 'norm' or std_normal variate.



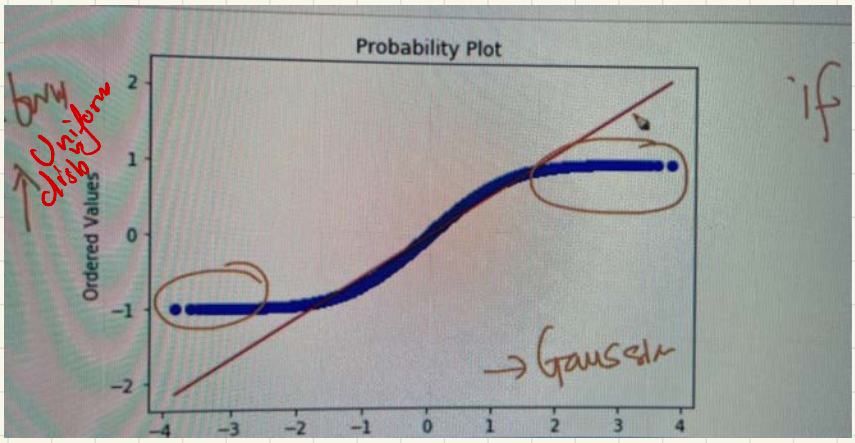

```
# generate 100 sanples from N(20,5)
measurements = np.random.normal(loc = 20, scale = 5, size=100)
#try size=1000

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

→ if we inc^n size → 1000
↳ we will see more & more pt. lie on st. line.



→ if we dec^n size → 50
↳ we will see more & more deviation from line.

∴ If no. of observation/sample is small, it's hard to interpret Q-Q plot.



→ suppose if we generate data from "uniform" disb^n

→ if we inc^n size here from 100 → 1000
↳ we will see more deviation

→ if we dec^n size here 100 → 50
↳ we will see our deviation is less. [Just opp. the above concept]

Q-Q plot

→ is $x \sim N(\mu, \sigma^2)$ → is x comming from Normal disb^n

→ is x, y : r.v { does x & y have the same disb^n.

→ we can answer both by using Q-Q plot.

How/where to use distributions?

→ r.v, pdf, cdf, Gaussian → 68-95-99.7 rule

probability → data analysis → answering questions about data

Q) company → xyz

Task :- order t-shirts for all employees → 100K

size → S, M, L, XL

a) How many 'XL' t-shirts should you order?

→ ① collect data for all 100K employees.

→ ② { typically ppl. height ≥ 180cm → (XL) t-shirt

[160cm, 180cm] → (L) t-shirt

domain knowledge.

→ here I assume my sample (μ & σ) are approx ≃ population (μ, σ) (100K)

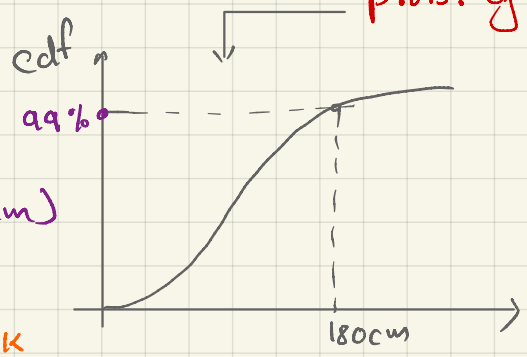
→ collect heights of (500) random employees

mean, μ , std. dev, σ

→ somebody told me heights $\sim N(\mu, \sigma)$ (doctor told me)

↳ prob. of (height $\geq$ 180cm) that I can calculate

↳ How? → I know, μ & σ and I know it follow normal disb^n then I can plot CDF & PDF.



So, prob. (height $\geq$ 180cm) is 1% → so,

100K x 1% = 1K

↳ I will order '1K' of 'xL' t-shirt for employee.

assumptions → i) you can predict size of t-shirt from heights.

ii) height is a Gaussian disb^n & I am finding μ, σ for sample.

→ Gaussian disb.^{n} → Theoritical model

↓

observed in many natural phenomena.

Q) Salaries(S) → of all ppl. (100K)

↳ If $S \sim N(\mu, \sigma)$

using CDF I can answer that.

↳ how many employees make a salary $\geq$ 100K $

↳ " " " " " " $50K $\leq S \leq$ $70K'

→ How do you know your S (salary) → Gaussian?

↳ Q-Q test (or other technique)

⇒ If salaries are not gaussian disb^n we can't use CDF.

↳ to check that we can use 'Q-Q' plot.

→ Gaussian disb^n → give us theoretical model.

Chebyshev's Inequality → help us to find the outliers for a feature.

X : Gaussian - distⁿ → 68 - 95 - 99% rule (1σ, 2σ, 3σ)

X ← heights of students

↳ If we know ①, ② & ③ $X \sim N(\mu, \sigma)$ (① ② ③)

$\Rightarrow prob(\mu - 2\sigma \le X \le \mu + 2\sigma) = 95\%$

assume, $\mu = 150$, $\sigma = 10$

↳ 95% of students heights lie in the [130, 170]

Q) what if I don't know the distⁿ.

But I know (μ, σ) → we can also calculate this by using "CLT".

↓ finite ↳ 'non-zero' and 'finite'

→ Can I say : x % of data lies within $\mu - 2\sigma$ & $\mu + 2\sigma$

y % of " " " $\mu - 1.5\sigma$ & $\mu + 1.5\sigma$

Can I find x % & y % ?

why?

Salaries of individuals → don't know the distⁿ

μ, σ ---→ (I know)

↳ $10K

↳ $40K

(2σ) Q) what %age of individual have salary in the range of [20K, 60K]?

(3σ) Q) " " " " " " " " " " [10K, 70K]?

↳ Here, comes the concept → Chebyshev's inequality:

→ X : r.v finite mean = μ

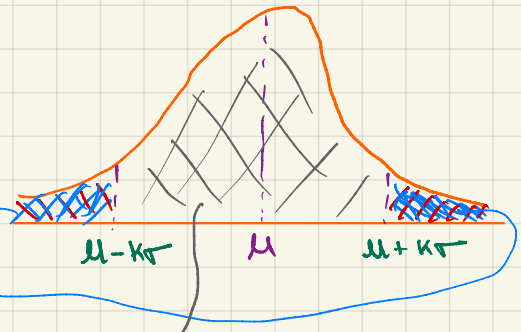
non-zero & finite std-dev = σ

↓ don't know the distⁿ.

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

$\rightarrow$ $x \geq \mu + K\sigma$

$x \leq \mu - K\sigma$



$\rightarrow$ Let, assume our distⁿ look like that

$$P(x \geq \mu + K\sigma \text{ or } x \leq \mu - K\sigma) \leq \frac{1}{K^2}$$

OR

$$P(\mu - K\sigma < x < \mu + K\sigma) > 1 - \frac{1}{K^2}$$

Let's go to Quesⁿ $\rightarrow$ b/w salary $\longrightarrow$ [20K , 60K]

$\downarrow$

$20K = \mu - 2\sigma$

$40K = \mu$

$60K = \mu + 2\sigma$

from chebyshev's inequality

$\hookrightarrow$ $P(\mu - 2\sigma < x < \mu + 2\sigma) > 1 - \frac{1}{K^2}$

here $K = 2$

$$P(\$20K < x < \$60K) > 1 - \frac{1}{2^2}$$

$$P(\$20K < x < \$60K) > 0.75$$

$\rightarrow$ it means, atleast 75 % ppl. salaries lie in this region.

$\rightarrow$ In this we can't say accurately like 95% ppl salaries is lie in that region. & can only give "range". (in gaussian distb^n)

a) $\rightarrow$ $[10K \leq x \leq 70K] > 1 - \frac{1}{9}$ $\approx$ 90% $\rightarrow$ greater than 90% ppl. salaries lie in that region.

$\rightarrow$ link to understand more.

https://www.learndatasci.com/glossary/chebyshevs-inequality/

Uniform distribution

→ here uniform means, the value of 'x' w.r.t 'y' is const.

- discrete
- continous

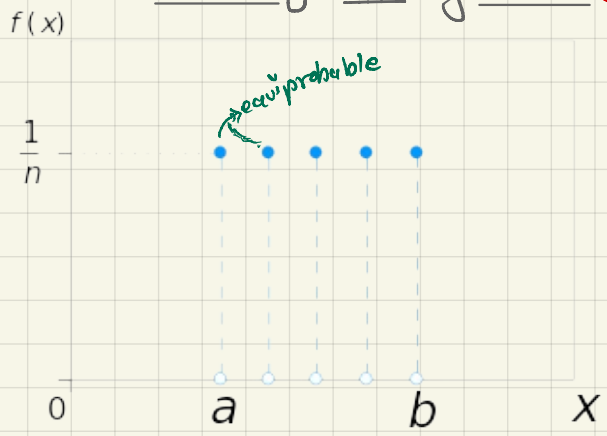
earlier we learend r.v ↔ discrete / continous

Discrete uniform dist^n

PDF → continous r.v

Pmf → discrete r.v

Probability mass function (Pmf)



notation :- $U(a, b)$

parameter :- $a \in$ integer, $b \in$ integer, $b \geq a$

# of outcomes we have ← $n = b - a + 1$

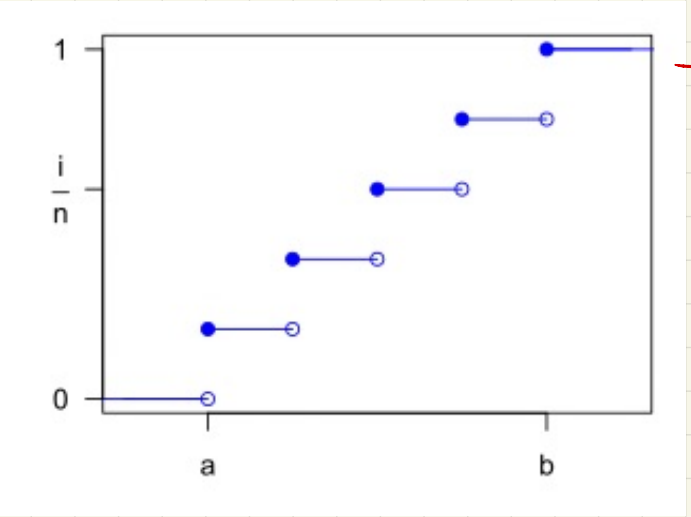
PDF :- $\frac{1}{n}$

eg:- In a dice → $a = 1$, $b = 6$, $n = 6 - 1 + 1 = 6$ (total outcomes) → probability of every outcome $= \frac{1}{n} = \frac{1}{6}$

=> if you know the parameters you know all the information about the distribution :- $N(\mu, \sigma)$ → Gaussian

$U(a, b)$ → uniform

CDF



→ non-smooth funct