Introduction to Stats:-

1. Basics to Advance → Descriptive stats
                     → Inferential stats

What is Statistics?

Statistics is the science of collecting, organizing and analyzing data. { Better Decision making }

Data: Facts (or) pieces of info that can be measured.

Descriptive Stats:

It consists of organizing and summarizing data

↳ eg: mean, median, mode etc...

Inferential stats:

Technique where in we use the data that we have measured to form conclusions.

eg: Are the marks of the students of this classroom similar to the marks of the Maths classroom in the college...

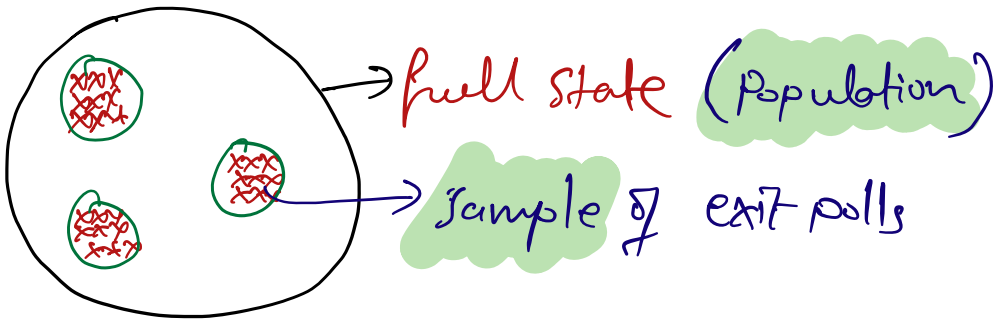
eg ①:

Classroom – Sample

college – population

eg ②:

election



population → N

sample → n

note:

we selected the samples randomly, to make it more systematic we use various sampling techniques

① Simple Random Sampling:

Every member of the population (N) has an equal chance of being selected as a sample (n)

② Stratified Sampling:

where the population (N) is split into non-overlapping groups

eg:

(1) Gender → Male
Female

(2) Age Group (0-10) (10-20) (20-40) (40-100)

(3) Profession Doctors, Eng, Actors

(3) Systematic Sampling:

from population (N) → we keep every (n^{th}) individual

eg: Mall → survey (Covid)

↳ 8^{th} person → survey

(4) Convenience Sampling:

Sample is taken from a group of people easy to contact (or) reach.

It is also called Grab Sampling (or) availability Sampling.

eg: Data Science.

↳ we consider a certain domain

(8)

Variable Measurement Scales

(1) Nominal data {"Categorical data"}
(2) ordinal → order of data matters, Value does not
(3) Interval → order matters, values also matter,
(4) Ratio natural zero is not present

Frequency Distribution:

eg: Sample data – Rose, Lilly, sunflower, Rose, Lilly, sunflower, Rose, Lilly, Lilly

flower	frequency
Rose	3
Lilly	4
Sunflower	2

→ from this we can plot diff graphs.

Cummulative frequency:

Rose – 3 ⇒ Lilly – 7 ⇒ sunflw – 9

note:

→ if Variable is discrete we draw "Barchart"

→ if Variable is Continuous we draw "Histogram"

pdf is smoothening of histogram

Agedai

1. Measure of Central Tendency
2. measure of dispersion
3. Gaussian Dist
4. Z-score
5. Standard Normal Dist

1. Arithmetic Mean for population & Sample :

Mean (Avg)

population (N)

Sample (n)

eg: $X = \{1, 1, 2, 2, 3, 3, 4, 5, 5, 6\}$

$$\mu = \sum_{i=1}^{N} \frac{x_i}{N}$$

$$= \frac{32}{10} = 3.2$$

$$\bar{x} = \sum_{i=1}^{n} \frac{x_i}{n}$$

Central Measure Tendency:

(1) Mean (2) Median (3) Mode

(1) ⇒ Refers to the measure used to determine the centre of the distribution of data.

(2) ⇒ (1) Sort the numbers
(2) the central element } Good for outliers

(3) ⇒ Most frequent element } bad for outliers.

Measure of dispersion: → Spread

(1) Variance

- to understand how two dispersions are diff

population Variance

$$\sigma^2 = \sum_{i=1}^{N} \frac{(x_i - \mu)^2}{N}$$

Sample Variance

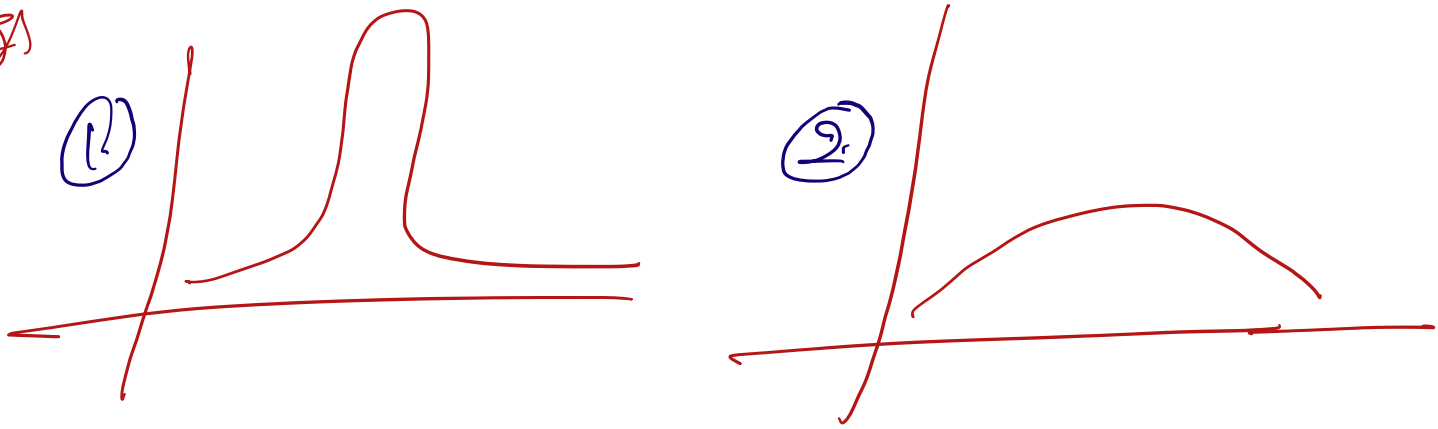
$$s^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1}$$

eg:

x	μ	$x-\mu$	$(x-\mu)^2$
1	2.83	−1.83	3.34
2	2.83	−0.83	0.6889
2	2.83	+0.83	0.6889
3	2.83	0.17	0.03
4	2.83	1.17	1.37
5	2.83	2.17	4.71
			10.84

$\mu = 2.83$

eg:



Variance is more in (2), cuz the distribution is more, the space b/w the points from avg should be more...

Standard deviation:

$$\sigma = \sqrt{\text{Variance}}$$

Percentile :

percentage: (eg) 1, 2, 3, 4, 5

% of nums that are odd?

$3/6 \times 100 = 50\%$

percentile: A percentile is a value below which a certain percentage of observation lie

Five Number Summary: [To remove outlier]

1. Minimum
2. First Quartile (Q_1)
3. median
4. Third Quartile (Q_3)
5. Maximum

$$\frac{\text{no. of value less than that number}}{\text{total samples}} \times 100$$

IQR = Inter Quartile Range

Lower fence = $Q_1 - 1.5\,(IQR)$

Upper fence = $Q_3 + 1.5\,(IQR)$

$$IQR = Q_3 - Q_1$$

ex1

{1, 2, 2, 2, 3, 3, 4, 5, 5, 5, 6, 6, 6, 6, 7, 8, 8, 9, 27}

$Q_1 = 25\%\ (n+1) \Rightarrow \frac{25}{100}(19+1) \Rightarrow 5^{th}$ index

(total elements + 1)

$\Rightarrow Q_1 = 3$

$Q_3 - Q_1$

IQR = 7 − 3 = 4

By 75% → $Q_3 = 7$

This formula is derived from percentile.

Lower fence = 3 − 1.5(4) = 3 − 6 = −3

higher = 7 + 6 = 13

range = [−3 ⟷ 13]

⇒ min = 1

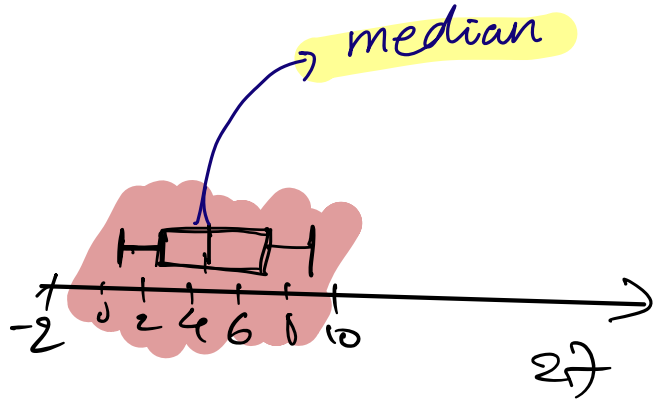
$Q_1 = 3$

median = 5

$Q_3 = 7$

max = 9

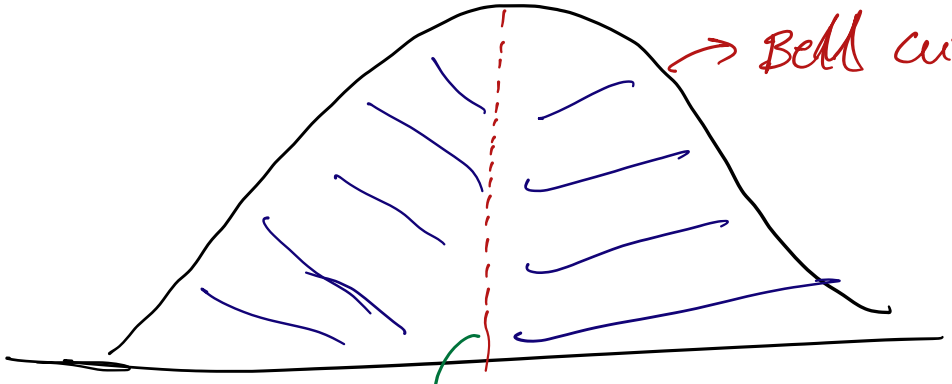
boxplot



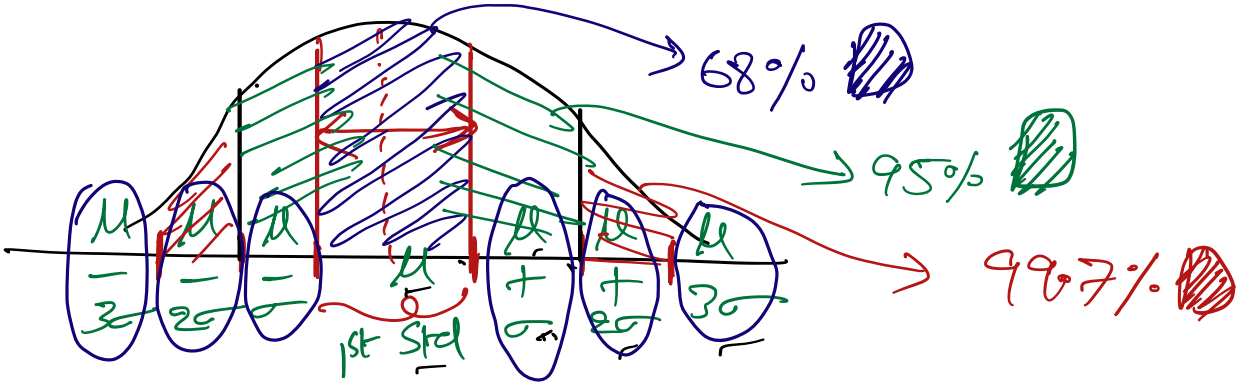
Distributions:

1. Normal Distr / Gaussian Distr
2. Standard Normal Distr
3. z score
4. Log normal Distr
5. Bernoulis Distr
6. Binomial Distr

Gaussian / Normal Distribution:



→ Normal distribution (or) Gaussian Distribution



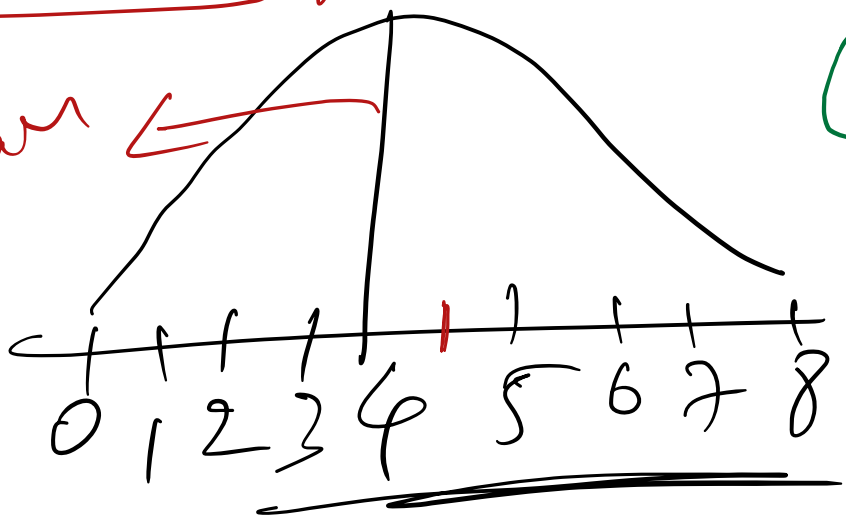
Empirical formula: 1st 2nd 3rd 68 – 95 – 99.7% Rule

% of distribution

eg:

wt, Iris dataset etc.....

mean $\mu = 4$ SD $\sigma = 1$



let,

4.5 → where it will fall in terms of SD

1. +0.5 sd

2. 4.75 ⇒ +0.75 sd

to find this we use

z-score, tells us how much SD is the point far from the mean

Z score $= \frac{x_i - \mu}{\sigma}$

x_i → data value

μ → mean

σ → S.D

$= \frac{4.75 - 4}{1} = 0.75$ SD

→ Coz, this is +ve number then it is on right

→ if it is -ve then it will be on left

⇒ After applying z-score then the values will change....

$z(1) = \frac{1-4}{1} = -3$,

$z(2) = \frac{2-4}{1} = -2$,

$z(3) = \frac{3-4}{1} = -1$,

$\{-3, -2, -1, 0, 1, 2, 3\}$

they look like Standardized values...

They are called as **Standard Normal Dist**

⇒ $[\mu = 0, \sigma = 1]$

eg:

Age	Salary	weight
24	40K	70
25	80K	80
26	60K	55
27	70K	45
(yrs)	(Rs)	(kg)

If we want to make $\mu = 0$ and $\sigma = 1$

unit

We call this as Standardization

Zscore formula is applied internally

Normalization:

↳ If we want to change all the values in b/w range of 0 and 1 then we use normalization

↳ we use MinMax Scaler for this process.

↳ (0 to 1)

MinMax Scaler:

→ Transform features by scaling each feature to a given range.

→ This estimator scales and translates each feature individually such that it is in the given range on the training set.

eg 0 - 1

1.

$$X\_std = \frac{(X - X.min(axis=0))}{(X.max(axis=0) - X.min(axis=0))}$$

2.

$$X\_scaled = X\_std * (max - min) + min$$

where min, max = feature_range.

note:

this transformation is always used as an alternative to

mean = 0,
Variance = 1

ex: ODI series

2021 series Avg 2021 = 250
S.D = 10
V.K Avg Score = ~~260~~ 240

2020 Series Avg 2020 = 260
S.D = 12
V.K Avg score = 245

Compared to both the series in which year VK score was better?

Sol. to find out this value, we find the z-score

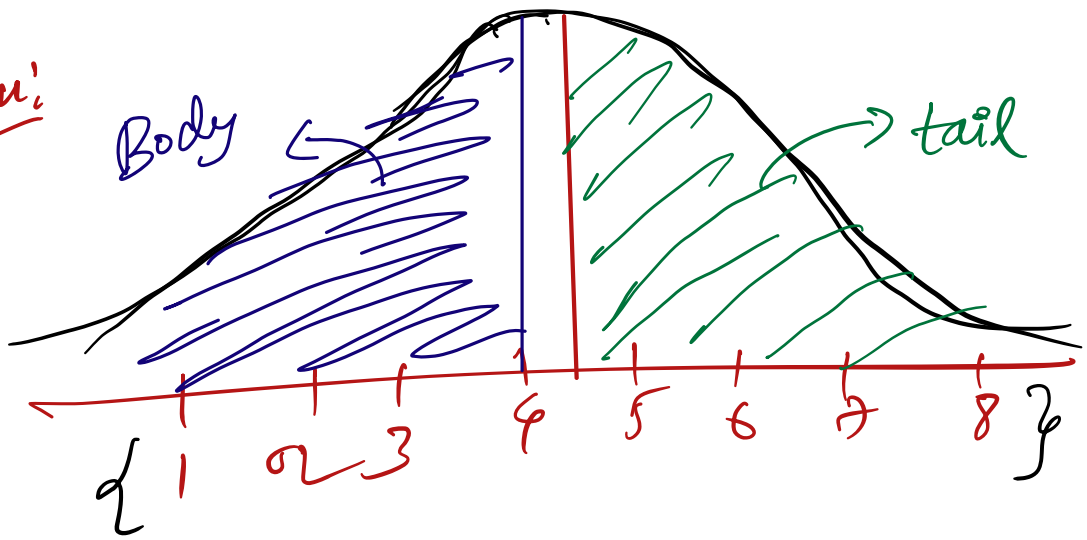
$$\text{Z-Score} = \frac{x_i - \mu}{\sigma}$$

$$\Rightarrow \boxed{2021} = \frac{260 - 250}{10} = 1$$

$$\Rightarrow \boxed{2020} \quad \frac{245 - 260}{12} = -1.25$$

↳ (Best)

Qu:



what percentage of scores fall above 4.25?

Note:

Z-score finds the area of the body curve

$$z = \frac{x_i - \mu}{\sigma} = \frac{4.25 - 4}{1} = 0.25$$

Note:

In online we have values for all kinds of z-score and then you can find the area required.

0.4013 from table

$\Rightarrow$ 40%

Agenda:

1. Probability
2. P & C
3. Confidence Interval
4. P value
5. Hypothesis testing

probability:

→ What is probability
→ Addition Rule

↳ Mutual Exclusive

↳ Non-Mutual Exclusive

→ Multiplication Rule

↳ Independent ⇒ (Conditional prob)

↳ dependent

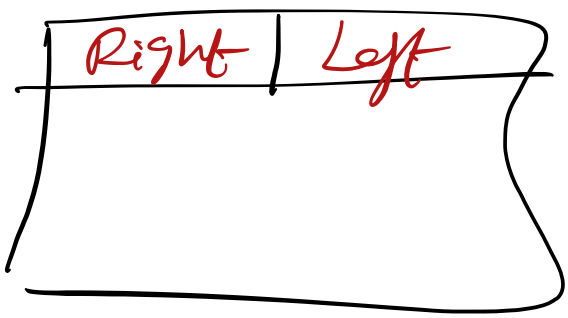
↳ (Naive Bayes)

Per & Comb:

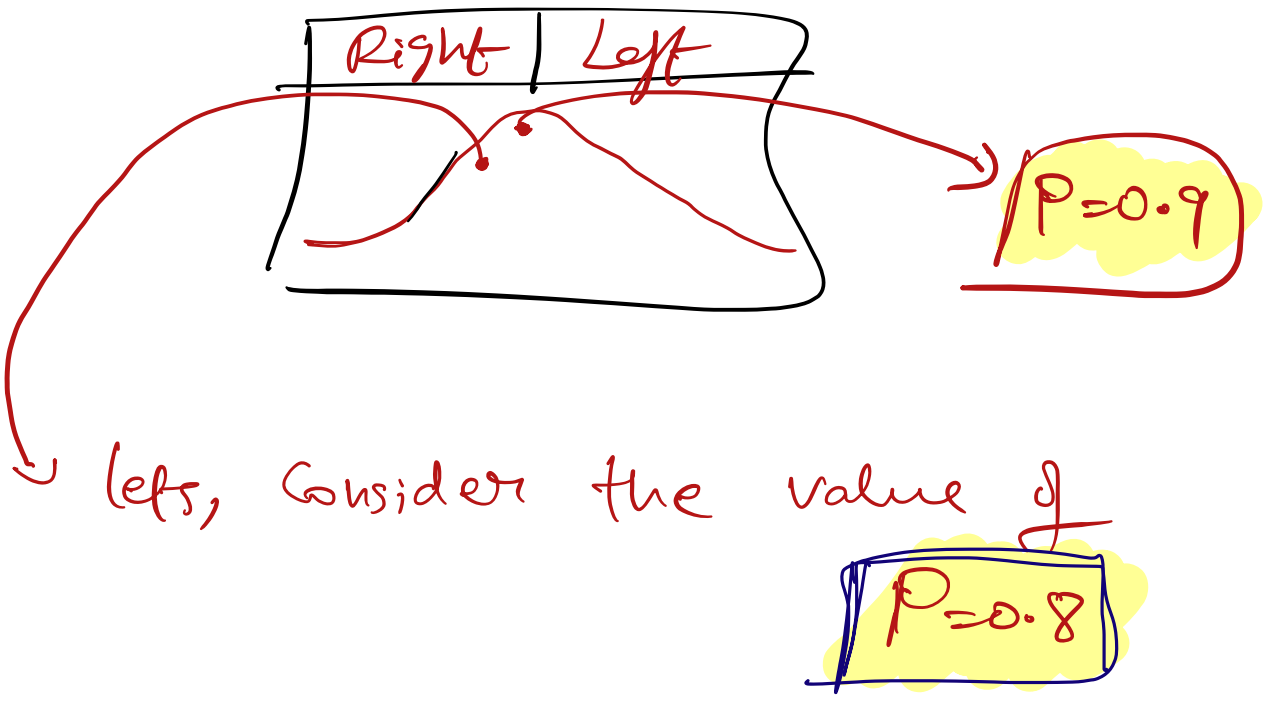
^nP_r, ^nC_r, circular arrangement, repetations, etc----

P value:

Lets Consider a laptop mouse pad.



1. we will use right click or left click, when we want to select something.
2. Else most of the time we just play our finger on the bottom area.
3. So if we plot the graph for the area we touch mostly it will look smtg like,



⇒ Every 100 times we touch the pad, there can be chance of 80 times we toucch this specific region.

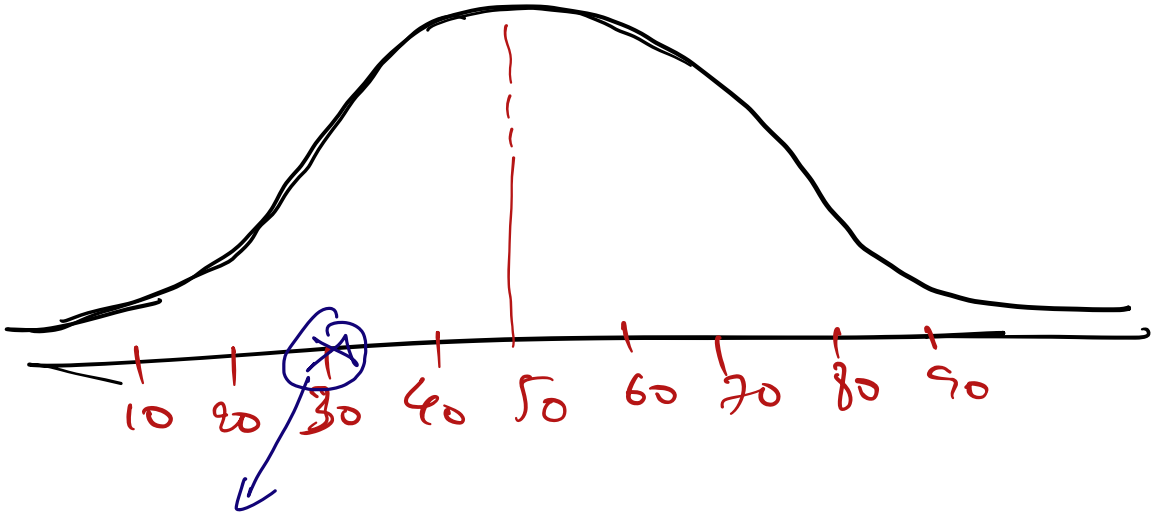
Hypothesis Testing:

1. Null Hypothesis: Coin is fair
2. Alternate hypothesis: Coin is Unfair
3. Experiment
4. Reject (or) Accept the null hypothesis

ex: lets take a coin example, and we tossed it 100 times

$$P(H) = 0.5 \qquad P(T) = 0.5$$

if we are drawing a SD graph.



If we say we got 30 times head, is that fair coin (or) an unfair coin?

Note:

⇒ our value should be as close as possible to the mean value (50)

⇒ for this, to determine, till how far is the points accepted from the mean, it is calculated by using.

Significance Value

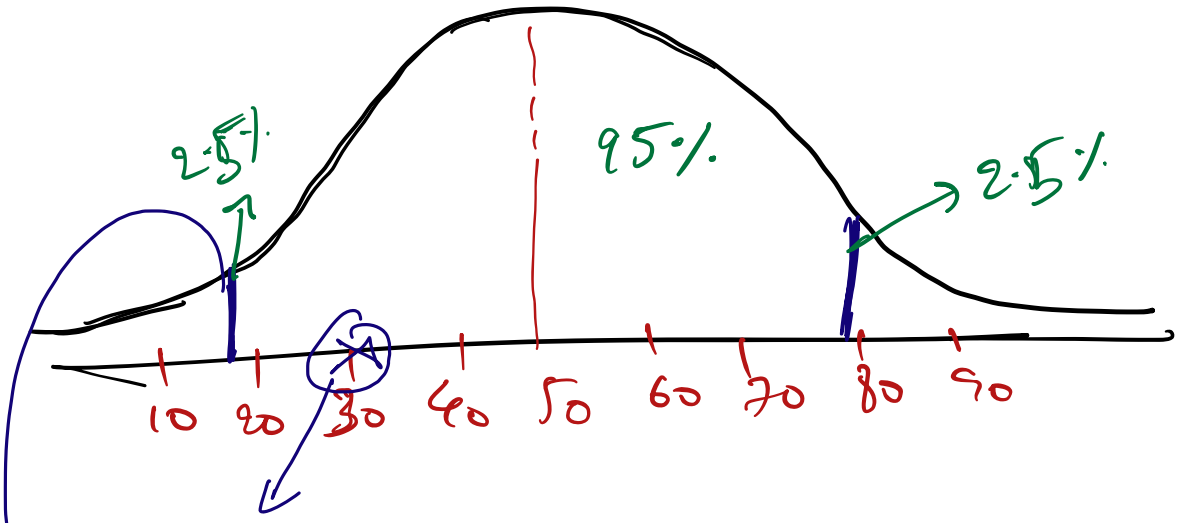
↳ rep by (α)

bt, $\alpha = 0.05 \Rightarrow 5\%$ {Domain Expert}

$100 - 5 = 95\%$ → Confidence interval

↓

on graph



→ Using two tail test we found out these are our 2.5% from front and back

note:

⇒ If the expirement fall in the 95% Confidence interval we call it a fair coin

⇒ else it is a unfair coin

Eg: we got 10 heads out of 100 trails

It is outside of CI

⇒ it is unfair

⊗ Null hypothesis ⇒ (H_0)

⊗ Alternate hypothesis ⇒ (H_1)

Reality Check:

Null hypothesis is True (or)
Null hypothesis is False

Decision:

H_0 is true (or) H_0 is false

Outcome 1 :-

we reject the null hypothesis, when in reality it is false

↳ Yes ↳ Good

O/p 2:

we reject the null hypothesis, when in reality it is true

↳ No

Type 1 Error

Op 3:

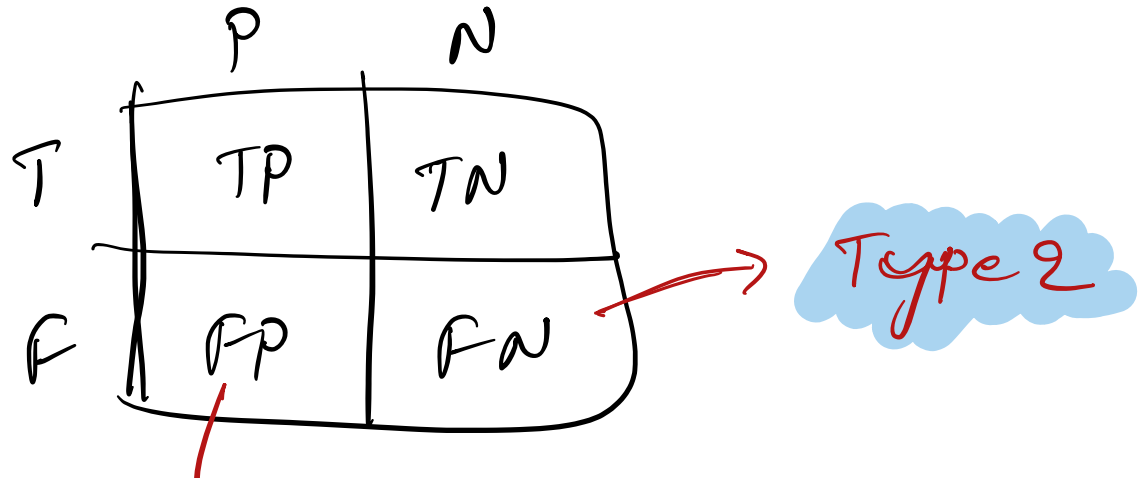
we Accept the null hypothesis, when in reality it is false

→ No → Type 2 Error

Op 4:

we Accept the null hypothesis, when in reality it is true

→ Yes → Good



Type 1

1 tail and 2 tailed test:

eg: Colleges in Karnataka have an 85% placement ratio. A new College was recently opened and it was found that a sample of 150 students had a placement rate of 88% with a S.D (4%) Does this college has a diff placement rate compared to other Colleges?

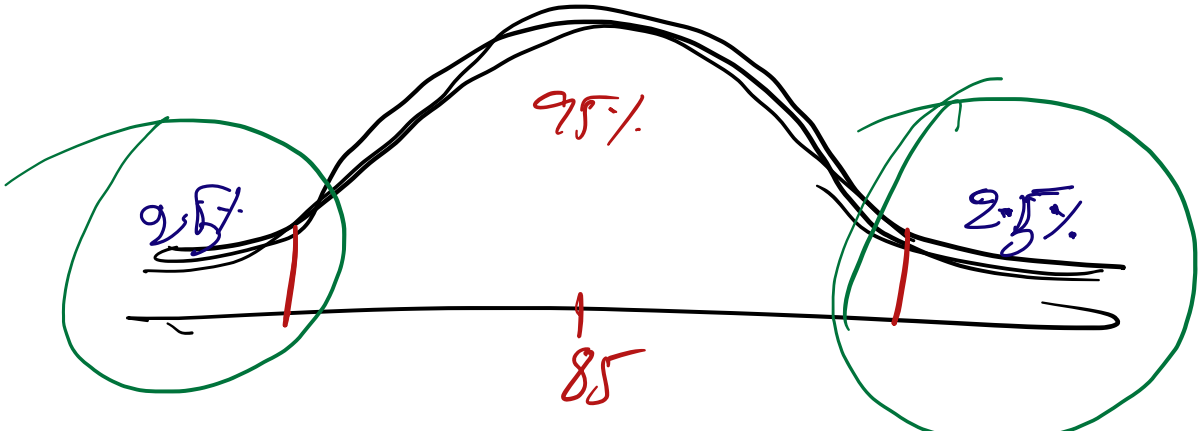
→ 85%.

Sol:

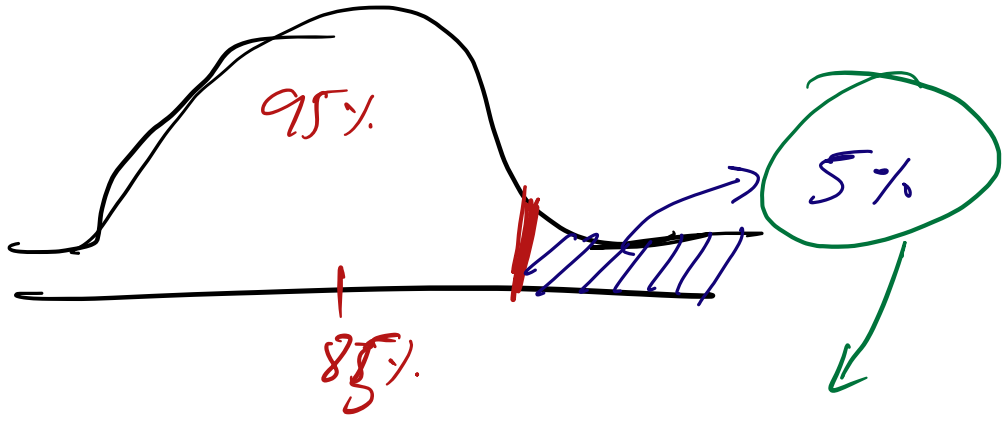
let, $\alpha = 0.05$ (Sig. Val)

then C.I ⇒ 95%

→ two tailed



Q Does the college have a placement rate greater than 85% ?

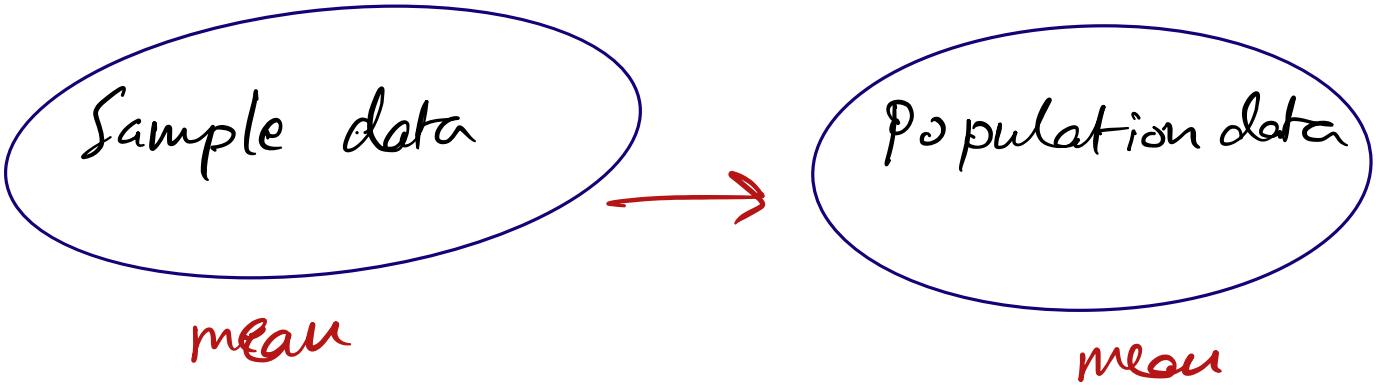


one tailed test

Confidence Interval:

point Estimate: The value of any statistic that estimates the value of a parameter.

Inferential Stats:



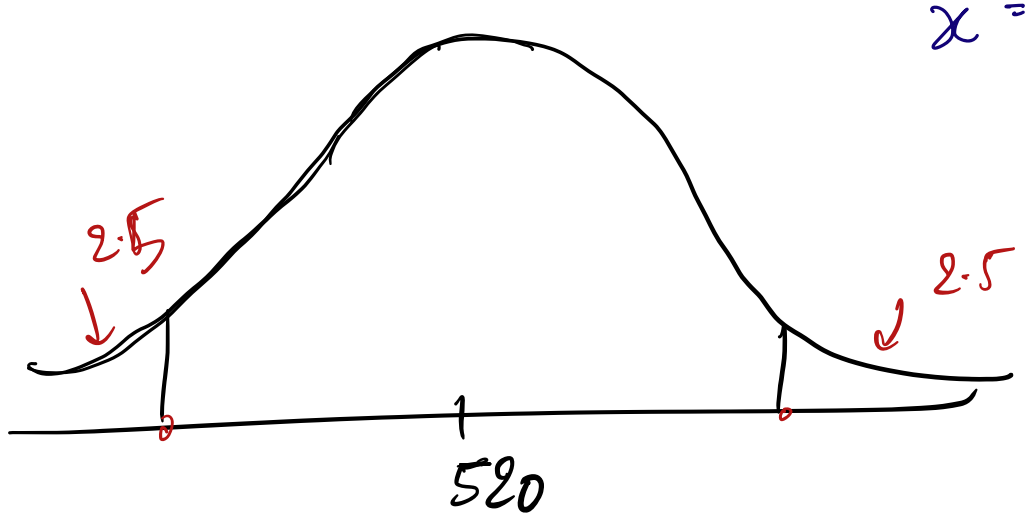
⇒ with the help of sample data, we will estimate population data

$$CI = \text{Point Estimate} \pm \text{Margin of Error}$$

ex: On the quant test of CAT Exam, the standard deviation is known to be 100. A sample of 25 test takers has a mean of 520 score. Construct a 95% CI about the mean?

Sol: $\sigma = 100$, $n = 25$, $\alpha = 0.05$,

$\bar{x} = 520$



⇒ Point Estimate ± margin of error

$n \geq 30$ ⇒ generally

$$\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$ → Standard Error

Upper bound = $\bar{x} + z_{\frac{0.05}{2}} \frac{100}{\sqrt{25}}$

Lower bound = $\bar{x} - z_{\frac{0.05}{2}} \frac{100}{\sqrt{25}}$

note:

$z_{\frac{0.05}{2}} = z_{0.025}$ → you can find this from online table of z-score.

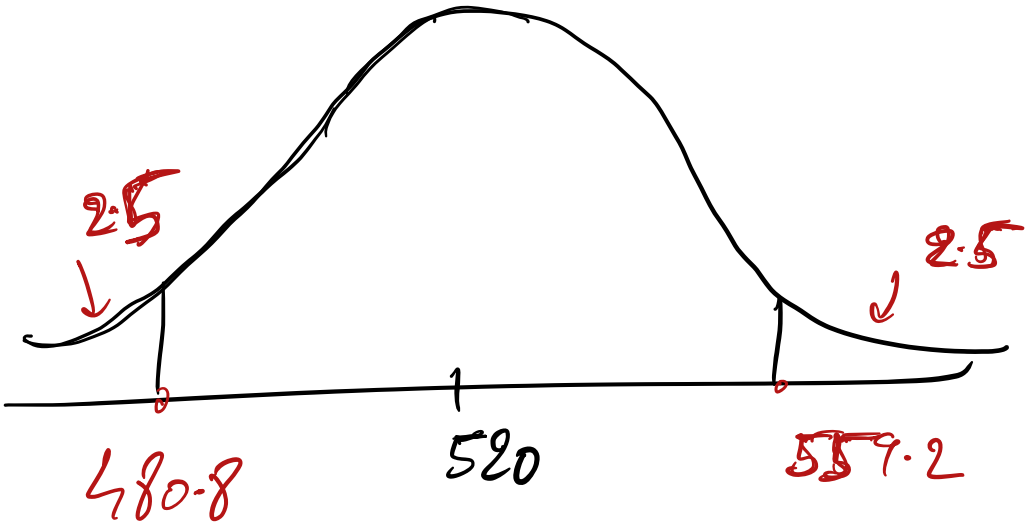
$\alpha = 0.025$

$1 - \alpha = 0.975$

↳ from table it lies at 1.96

Upper = $520 + 1.96(20) = 559.2$

Lower = $520 - 1.96(20) = 480.8$



Q. On the quant test of CAT exam, a sample of 25 test takers has a mean of 520 with a SD of 80. Construct 95% CI about the mean.

Sol. Here, population std is not given

↳ hence, we use t-test

= point estimate ± margin of error

$$= \bar{x} \pm t_{\frac{\alpha}{2}} \left(\frac{s}{\sqrt{n}}\right)$$ → Standard Error

Upper bound = $\bar{x} + t_{\frac{0.05}{2}} \left(\frac{s}{\sqrt{n}}\right)$

Degree of freedom $\Rightarrow$ $n-1 \Rightarrow 25-1 = 24$

$= 520 + 2.064\ (80/5)$

$= 553.024$

Lower bound = $\bar{x} - t_{\frac{0.05}{2}} \left(\frac{s}{\sqrt{n}}\right)$

$= 520 - 2.064 \left(\frac{80}{5}\right)$

$= 486.97$

$[486.97 \longleftrightarrow 553.024]$

One Sample Z-test:

(1) In the population, the average IQ is 100 with a SD of 15. Researches wants to test a new medication to see if there is a positive (or) negative effect on intelligence, (or) no effect at All. A sample of 30 participants who have taken the medication has a mean of 140. Did the medication effect the intelligence?

Sol. (1) Define Null hypothesis

$H_0 \Rightarrow \mu = 100$

$\alpha = 0.05$

C.I = 95%

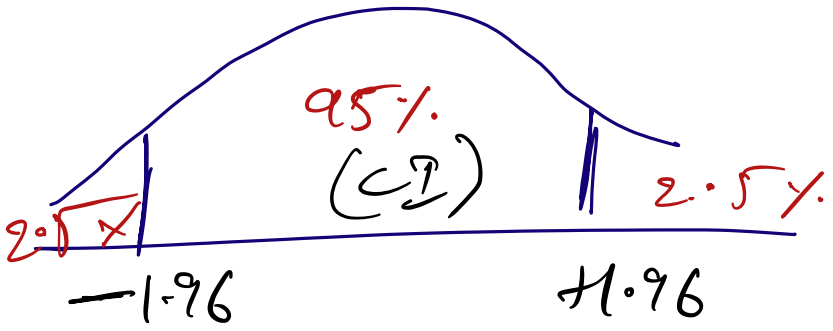
(2) Alternate hypothesis

$H_1 \Rightarrow \mu \neq 100$

(3) State Alpha $\alpha = 0.05$

(4) State Decision Rule

(2 tail test)



(5.) Calculate z test:

$$z = \frac{\bar{x} - \mu}{\left(\frac{\sigma}{\sqrt{n}}\right)}$$

note:

(1) for sample value we take standard error

(2) for population value we take only (σ) not standard error.

$$\frac{140 - 100}{\frac{15}{\sqrt{30}}} = 14.60$$

(6.) State our Decision

↳ 14.60 > 1.96

If z is < than -1.96 (or) > 1.96, reject the null hypothesis.

(a) did the medication improve the intelligence (or) decrease.

↳ [Improved]

One Sample T-test:

- Z-test ⇒ population std
- T-test ⇒ unknown " "

(a) population the average IQ = 100

$n = 30$ $\bar{x} = 140$ $s = 20$

Did the medication affect the int?

$\alpha = 0.05$

Sol.

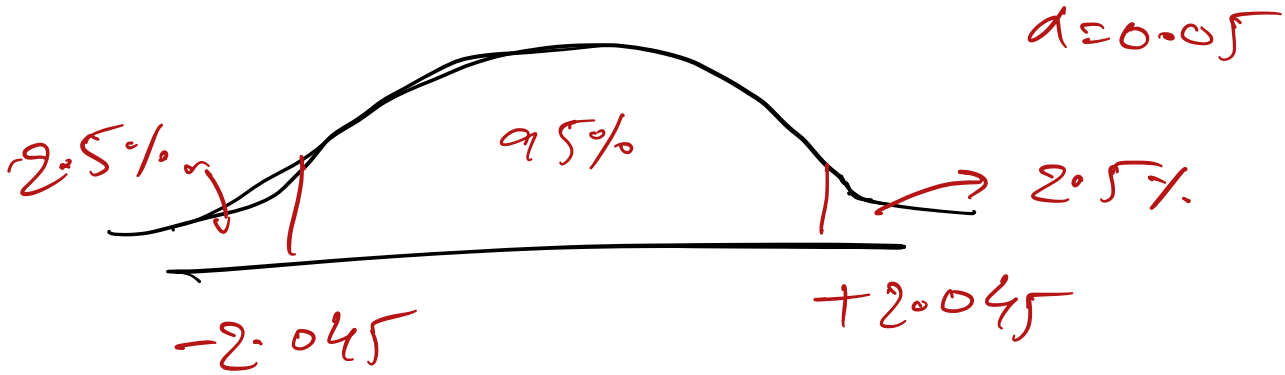
(1) $H_0 \Rightarrow \mu = 100$

(2) $H_1 \Rightarrow \mu \neq 100$

(3) Calculate the degree of freedom

$$n-1 = 30-1 = 29$$

(4) State Decision rule



(5) T test

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

$\bar{x} = 140, \mu = 100, s = 20, n = 30$

$= 10.96$

$t = 10.96 > 2.045$

⇒ Reject null hypothesis

P ≤ Sig value

→ [increased the intelligence]

Agenda:

1. Chi Square
2. Covariance
3. pearson Correlation Coefficient
4. Spearman Rank Correlation
5. f test (ANOVA)

Chi Square:

→ chi square test claims about population proportions

→ It is a non-parametric test that is performed on categorical variables can be either nominal or ordinal

data.

(8.) In the 2000 Indian Census, the age of the individual in a small town were found to be the following?

Less than 18	18-35	>35
20%	30%	50%

In 2010, age of $n = 500$ individuals were sampled. Below are the results

<18	18-35	>35
121	288	91

Using $\alpha = 0.05$, would you conclude the population distribution of ages has changed in the last 10 yrs?

Sol: note: when we have data in terms of proportion then we need to

implement non-parametric test.

<18	18-35	>35	
121	288	91	(obs)
100	150	250	(Exp)

(1) H_0 = The data meets the distribution of 2000 Census

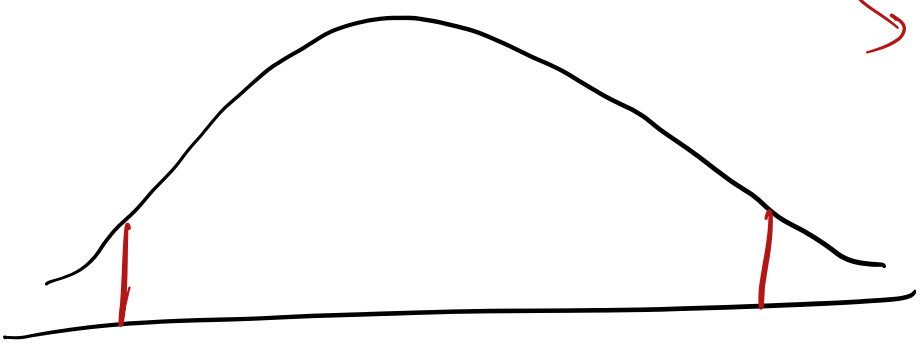
H_1 = The data does not meet the " " "

(2) $\alpha = 0.05$ (95% CI)

(3) Degree of freedom $\Rightarrow n-1$

$\Rightarrow 3-1=2$

(4) Check in the Chi Square table to find Decision Boundary



χ^2

If χ^2 is greater than 5.99 then reject the H_0

(5.) Calculate the test statistics

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

$$= \frac{(121-100)^2}{100} + \frac{(288-150)^2}{150} + \frac{(91-250)^2}{250}$$

$$= 232.94$$

$$\chi^2 = 232.94 > 5.99$$

{Reject the Null hypothesis}

note:

If p-value < Significance value

then we Reject the null hypothesis

Covariance:

Relation b/w values

$$Cov(x, y) = \sum_{i=1}^{n} \frac{(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

= +ve (or) -ve (or) 0

note:

This value can be anty in terms of magnitude.

To get a constraint for this we use

pearson correlation

it restricts all the values b/w

(-1 to +1)

The more it is towards +1 ⇒ +ve correlation

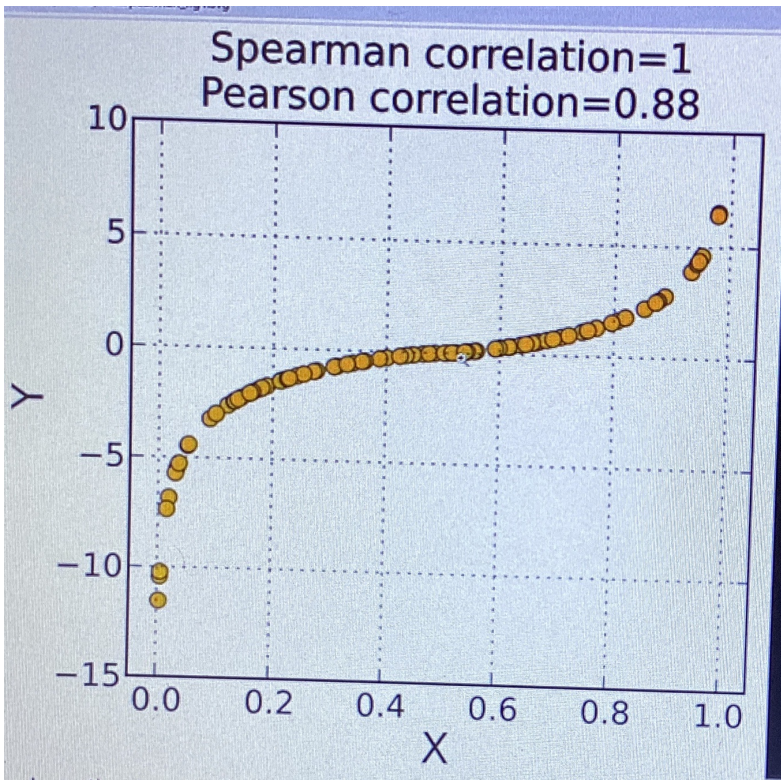
The more " " -1 ⇒ -ve corln

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

$\{-1$ to $+1\}$ range

→ It captures linear properties well.

Spearman's rank correlation:



⇒ along with linear, non-linear

properties can also be captured well.

$$\rho_{(x,y)} = \frac{Cov(R(x), R(y))}{R_{\sigma_x} \; R_{\sigma_y}}$$

ex)

x (Ht)	y (wt)	R(x)	R(y)
170	75	2	2
160	62	3	3
150	60	4	4
145	55	5	5
180	85	1	1