previous part were based on discriptive statistics and now we are moving to infrential statistics

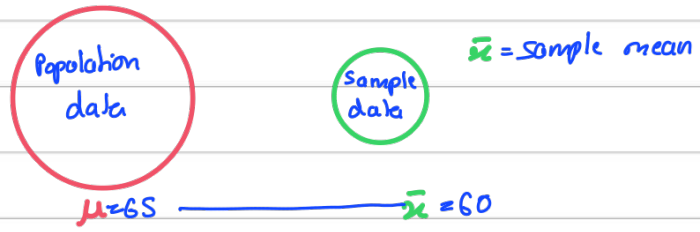
Infrential Statistics: In this type of statistics we make diff infrences/conclusions from a sample data wrt finding any population parameter.

Ty {Estimate}: It is a specified observed numerical value used to estimate an unknown population parameter.

Types of Estimates:-

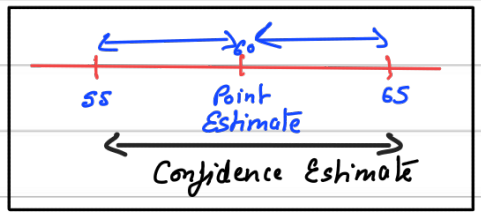
① **Point Estimate:** single numerical value used to estimate an unknown population parameter.

Ex:- Sample mean is a point estimate of a population data.



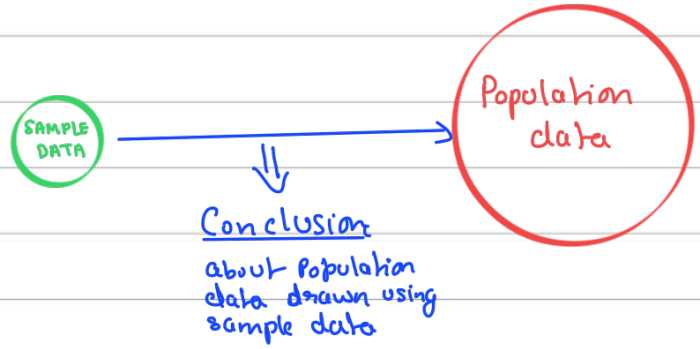
② **Interval Estimate:** Range of values used to estimate the unknown population parameter.

when we have a pt estimate along with have a range which we get by adding & substracting margin of error and this is c/a "Confidence Interval"



an this kind of interval is know as an **Interval Estimation** made for a sample mean

III {Hypothesis Testing & Mechanism}



these conclusion or inferences are drawn using "Hypothesis Testing"

Hypothesis Testing Mechanism:- # suppose person committed crime

Step ① Null Hypothesis (H_0) :

- it contains the assumption you are begining with

acc^n to H_0, court say "Incent until proven guilty"

Step ② Alternate Hypothesis (H_1)

- it is nothing but opposite of null hypothesis

acc^n to H_1, he is guilty

Step ③ Experiments : Statistical Analysis

at this stage various (DNA/Fringerprint test are conducted)

Step ④ Accept or Reject the Null hypothesis

Q) A College at district A states its average pass %age of students are 85%. A new college opened in the district and it was found that a sample of student 100 have a pass %age of 90% with a std deviation of 4%. Does this college have a diff. passed %age.

Sol^n → 1st step defining null hypothesis (H_0):

$H_0 = \mu$ (avg pass %age) = 85%

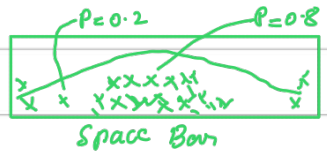
2nd step defining alternate Hypothesis (H_1):-

$H_1 = \mu \neq 85\%$

III) P-VALUE

It is a number, calculated from a stastical test, that describes how likely you are to have found a particular set of observations if the null hypothesis were true. P values are used in hypothesis testing to help decide weather to reject the null hypothesis.

Simple Example: below is given a space bar where the x represents the place where the space bar has been touched :-



at region where P=0.2, it means at that region space bar has been touched 20 times out of 100 touches

region where P=0.8, that means keyboard's been touched 80 times out of 100.

Q) Do Hypothesis testing on a coin toss to determine weather the coin is fair or not

P(H) = 0.5 , P(T) = 0.5 ⇒ fair coin

P(H) = 0.6 , P(T) = 0.4 ⇒ somewhat fair

P(H) = 0.7 , P(T) = 0.3 ⇒ Toss biased towards Heads

Soln — ① H_1 : Coin is fair

② H_0 : Not a fair coin

③ Experiment : tossing for 100 times

IV Hypothesis Testing and Statistical Analysis

There are various types of Statistical Analysis.

① Z-test } - deal with avg. data } < Z-table, t table
② t-test }

③ CHI-Square test } - deals with Categorical data

④ ANNOVA test } - deals with variance.

1) Z-test :-

Requirements for using Z-test :-

① must know population std

② Sample size (n) ≥ 30

Q1) Avg. height of all the residents in a city is 168cm with a $\sigma = 3.9$. A doctor believes the mean to be different. He measured the height of 36 individuals and found the avg. height to be 169.5 cm

a) State null & alternate hypothesis

b) At a 95% confidence level, is there enough evidence to reject the null hypothesis.

Sol^n - population mean → 168 cm with std. (σ) = 3.9

Sample size (n) = 36 & sample mean ($\bar{x}$) = 169.5 cm

As,

population mean is given & $n > 30$ ∴ we will be using Z-test

S1) a) Null Hypothesis (H_0) - $\mu = 168$ cm

(H_1) = $\mu \neq 168$ cm — $\boxed{\text{2 tailed test}}$

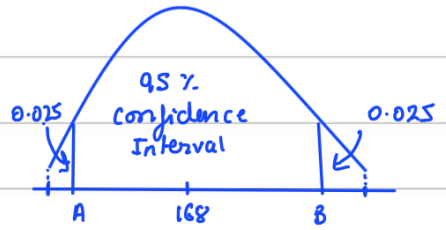
coz μ can be > 168 or $\mu < 168$ cm

S2) b) given confidence interval : 95%

∴, $\boxed{\text{Significance value : } 1 - CI}$ $= 1 - \frac{95}{100} = 1 - 0.95 = \underline{\underline{0.05}}$

Now,

S3) Based of C.I we will draw Decision Boundry



finding Z-score for B

$1-0.025 = 0.9750$ (+ve)

referring z-table we get

$\boxed{Z(B) = +1.96}$

finding z-score for A

"

"

$\boxed{Z(A) = -1.96}$

5V
$\boxed{0.025}$

Now

If new z-score is < -1.96 or $> +1.96$ then we must reject the Null Hypothesis (H_0):

S4) Applying Z-test

2.31

0.1044

0.01044

- for population data z-score $= \bar{x} - \mu / \sigma$
- for sample data z-score $(Z_d) = \dfrac{\bar{x} - \mu}{\sigma/\sqrt{n}}$

$$= \frac{169.5 - 168}{3.9/\sqrt{36}}$$

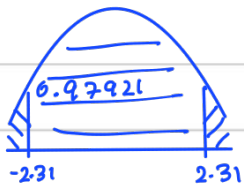
$$\boxed{Z_d = 2.31}$$

S5) Conclusion: based on above Z_d score we can conclude that as $Z_d\ (2.31) > 1.96$ ∴, we can reject the Null Hypothesis

$\Rightarrow \mu \neq 168$ and since $Z_d = 2.31 = +ve$ ∴ avg height $\neq 168$ seems to be ring i.e > 168.

Now,

New decision boundry is :-



using z-table area left of $-2.31 = \underline{.1044}$

∴ area on right of $2.31 = .1044$

new P value $= .1044 + .1044 = \underline{0.02088}$

as P value (new) < 0.05 ∴ we reject Null Hypothesis

Finding the confidence interval:

↳ $1 - 0.02088 = 0.97921 \approx 97.921\ \%$

Q1) A factory Manufactures bulbs with a avg warranty of 5 yrs with std of 0.50. A worker believes that the bulb will malfunction in less than 5 yrs. He tests a sample of 40 bulbs and find the avg. time to be 4.8 yrs

a) State Null & Alternate Hypothesis

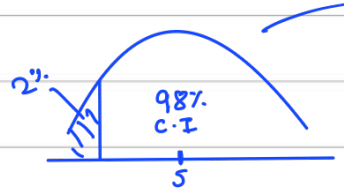
b) at 2% significance level, is there enough evidance to support the idea that warranty should be revised?

Soln. → S-1) – $H_0 \Rightarrow \mu = 5$ yrs

S2) → $H_1 \Rightarrow \mu < 5$ yrs # checking only <5yr & not ≠5 ∴ it is a $\boxed{\text{1 tailed Test}}$

S3) Decision Boundary

only 2% on left side cuz its a 1 tailed test



S4) Z-test:-

$$Z_d = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{4.8 - 5}{0.50/\sqrt{40}} = -2.53$$

for $Z_d = -2.53$ area under the curve is: 0.00570 P-value

So;

Now as P-value (0.00570) < Significance Value (0.02)

∴ reject the Null Hypothesis

In Z-Statistics when we perform any analysis using Z-score we require population standard deviation (σ)

But,

What happen if population std(σ) is not given? How can we analyse the sample data then?

In such cases we use <u>Student's t distribution</u>

wrt this distribution

$$\boxed{t = \frac{\bar{x} - \mu}{s/\sqrt{n}} \quad , \; s \rightarrow \text{sample standard deviation}}$$

t-table is used for Student's t distribution analysis which has degree of freedom (df)

$\boxed{df = n-1}$, n = no. of sample data

Why df = n-1

Suppose 3 chairs are empty inside a room & 3 ppl looking for chair, 1st person to enter the room will have 3 choices to sit, 2nd person will have 2 choices & no choice will be available of 3rd person and he has to sit on a chair which is left empty

So, with 3 ppl only 2 choices are available

$\therefore$ DF = n-1

2) t-stats : t-test :- 1 sample t-test

Q1) In a population the avg IQ is 100. A team of researchers want to test a new medication to see if it has either a +ve or -ve effect on intelligence, or no effect at all. A sample of 30 participants who have taken the medication has a mean of 140 with a standard deviation of 20. Did the medication affect intelligence?

Sol^n → $\mu = 100$

Sample, $n = 30$, $\bar{x} = 140$, $s = 20$ – sample std

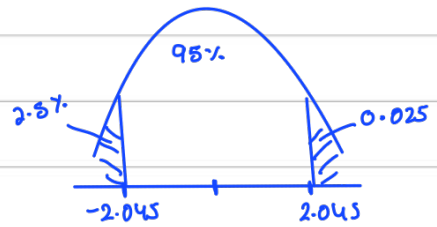
☆ Let's initialize our Confidence Interval (C.I) = 95% and Significance value (α) = 0.05 ≈ 5%

S1) $H_0 \rightarrow \mu = 100$

S2) $H_1 \rightarrow \mu \neq 100$ (∴, 2 tailed test)

S3) Computing Degree of Freedom for t-test: Dof = 30-1 = 29

S4) Decision Rule:-



Using t-table at Dof=29 & α(two tail) = 0.05

$\boxed{\text{t-value} = 2.045}$

☆ if → (t test < -2.045 & t test > 2.045) → reject Null Hypothesis.

S5) t-test:-

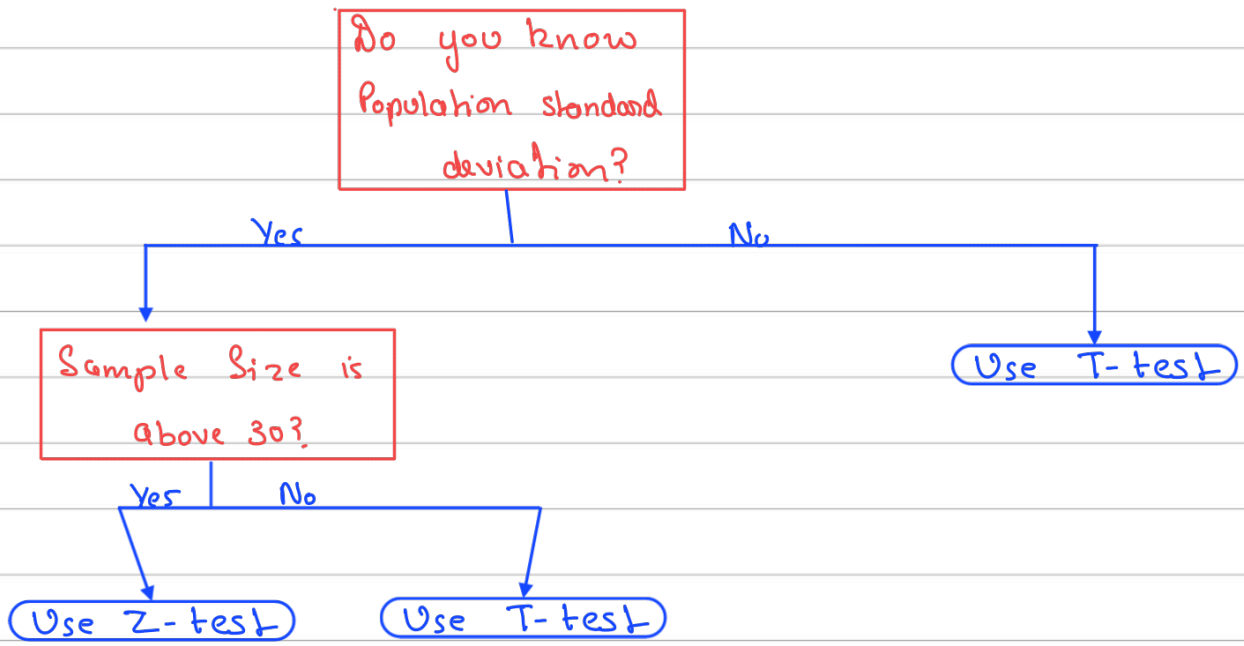
$$t = \frac{\bar{x} - \mu}{s\sqrt{n}} = \frac{140 - 100}{20\sqrt{30}} = \underline{\underline{10.96}}$$

Since,

t-test = 10.96 > 2.045 ∴, Rejecting the null Hypothesis

S6) Conclusion:- → Medication used has affected the intellegence
→ also as t-test value = 10.96 > 2.045 & +ve ∴ intellegence has increased.

Q4 When to use T-test Vs Z-test



11/03/23

Type-1 and Type-2 Error

Reality: Null Hypothesis is True or Null Hypothesis is False
Decision: Null Hypothesis is True or Null Hypothesis is False

Outcome 1:- We reject the Null Hypothesis when in reality it is false → Good Scenario

Outcome 2:- We reject the Null Hypothesis when in reality it is True → Type 1 Error

Outcome 3:- We retain the Null Hypothesis when in reality it is false → Type 2 Error

Outcome 4:- We retain the Null Hypothesis when in reality it is True → Good Scenario

{Confidence Interval & Margin of Error}

Till now we had point estimate i.e sample mean($\bar{x}$) and we were trying to make some assumptions about the population mean μ.

Q1 What is the problem with using point estimate?

Sol^n → Suppose now pt. estimate $\bar{x}$ is used to estimate the population mean (μ)

$$\boxed{\bar{x}} \longrightarrow \boxed{\mu}$$

if $\bar{x}=2.5$ $\qquad$ $\mu=3$

But depending on diff. samples the $\bar{x}$ value can be greater than or less tha or equals to $\mu=3$.

So; in order to avoid this,

We define a range of Confidence interval which is used to correctly define the population μ.

Confidence Interval:-

$$\boxed{\begin{aligned} C.I &= \text{Point Estimate} \pm \text{Margin of Error} \\ &= \bar{x} \pm \underbrace{\boxed{Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \text{ (for z-test)}}}_{\text{Margin Error (for z-test)}} \\ &= \bar{x} + \underbrace{\boxed{t_{\alpha/2}\frac{\sigma}{\sqrt{n}} \text{ (for t-test)}}}_{\text{margin error (for t test)}} \end{aligned}}$$

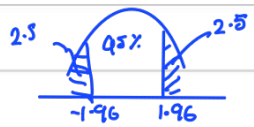
, α → Significance value

Q1 In the verbal section of CAT exam, the std is known to be 100. A sample of 30 test takers has a mean of 520. Construct 95% CI about the mean?

Sol^n → we know $CI = \bar{x} + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$

we know for 95% confidence Index $\alpha = 5\% \approx \underline{0.05}$

∴,



$Z_{\alpha/2} = 1.96$ & -1.96

Lower CI $= 520 - (1.96)\frac{100}{\sqrt{25}} = 480.8$

Higher CI $= 520 + (1.96)\frac{100}{\sqrt{25}} = 559.2$

So, with 95% assurance we can say that our final mean CAT score will be lying b/w 480.8 and 559.2.

Bayes Statistics (Bayes Theorem)

Bayes statistics is an approach to data analysis and parameter estimation based on bayes' theorem.

Baye's Theorem:- # completely based on probabilities

Probability —— Independent Events / Dependent Events

Independent Event

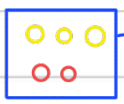
ex: Rolling a dice
$\{1, 2, 3, 4, 5, 6\}$
$P(1) = \frac{1}{6}$, $P(2) = \frac{1}{6}$ .....

ex: Tossing a coin
$P(H) = 0.5$ & $P(T) = 0.5$

Probabilities does not dependent of other outcomes.

Dependent Events:

ex: A Box of marbel with 3 yellow marbles and 2 red marbles



Red → $P(R) = \frac{2}{5}$ $\xrightarrow[\text{this event}]{\text{after}}$ $P(Y) = \frac{3}{4}$

$\therefore$

$$\boxed{P(R \,\&\, Y) = P(R) \times \boxed{P(Y/R)}} \quad \text{—(1)}$$

→ Conditional Probability

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

$= 6/20$

Deriving Baye's Theorem:-

from above eq ① we know $P(A \& B) = P(A) \times \boxed{P(B/A)}$

↳ Probability of B after event A already happened

as, $P(A \& B) = P(B \& A)$

$\Rightarrow P(A) \times P(B/A) = P(B) \times P(A/B)$

5★

$$\Rightarrow \boxed{P(B/A) = \frac{P(B) \times P(A/B)}{P(A)}}$$

Baye's Theorem

also $$\boxed{P(A/B) = \frac{P(A) \times P(B/A)}{P(B)}}$$

where, $A, B \rightarrow$ Events

$P(A/B) \rightarrow$ Probability of A given B is true

$P(B/A) \rightarrow$ Probability of B given A is true

$P(A)$ & $P(B) \rightarrow$ Independent probabilities of A & B.

Qy Use Baye's Theorem to predict the output price of the house from the bellow data set

Data set:-

Size of House x_1	No of Rooms x_2	Location x_3	Price y

features — outcome

Sol^n → using Baye's theorem to predict price:-

$$\boxed{P(y/x_1 x_2 x_3) = \frac{P(y) \times P(x_1 x_2 x_3 / y)}{P(x_1 x_2 x_3)}}$$

12/03/23

CHI SQUARE TEST

- The "Chi Square Tesh for Goodness of fit" test claims about population proportions.
- It is a non-parametric test that is performed on the categorical data (Nominal or Ordinal)

Ex:→ There is a population of male who likes different colour bikes. and there is a theory that 1/3 ppl like yellow, 1/3 like Red & 1/3 like orange.

	Theory	Sample
Yellow bike	1/3	22
Red bike	1/3	17
Orange bike	1/3	59

Theory Categorical distribution → observed categorical distribution

Now,
using Goodness of fit Chi^2 test we will confirm that weather the sample supports the theory or not

Goodness of fit test:-

Qy In a science class of 75 students, 11 are left handed. Does this class fit the theory that 12% of the ppl are left handed.

Soln→

	O(observed)	E(expected: 11% are lefty)
left handed	11	9 (11% of 75)
Right handed	64	66
	75	75

Now Chi^2 test is performed to make sure that weather this observed value is Goodness of fit for the expected values.

In Chi^2 test we compare 2 categorical variable i.e Theoritical /expected and Observed Categorical distribution (Sample).

Performing Chi^2 Test:-

Q1) In 2010 census of the city, the weight of the individual in a smal city were found to be the following:

<50kg	50-75	>75
20%	30%	50%

In 2020, weight of n=500 individuals were sampled. Below are the results

<50kg	50-75	>75
140	160	200

Using $\alpha = 0.05$, would you conclude the population differences of weights changed in the last 10 years?

Sol^n

Expected data (2010)

<50kg	50-75	>75
20%	30%	50%

Sample / Observed data (2020)

<50kg	50-75	>75
140	160	200

Expected data (2020)

<50kg	50-75	>75
20% of 500 = 100	30% of 500 = 150	50% of 500 = 250

S1) Null Hypothesis (H_0): The data meets the expectation

S2) Alternate Hypothesis (H_1): The data does not meet the expectation

S3) as $\alpha = 0.05$ ∴, Confidence Interval (CI) = 95%

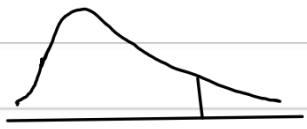
S4) Degree of Freedom:-

$$df = K-1 = 3-1 = 2$$

S5) Decision Boundary:

☆ In case of Chi^2 test we don't have a symmetrical distribution.

distribution looks like:



•) using CHI^2 table find critical value for $\alpha = 0.05$

$\therefore$ CV = 5.991 (using χ table)

If χ^2 is greater than 5.991, Reject H_0

If χ^2 is less than 5.991, fail to reject H_0

Step Calculating χ^2 test statistics :-

Expected data (2020)

<50kg	50-75	>75
20% of 500 = 100	30% of 500 = 150	50% of 500 = 250

Sample / Observed data (2020)

<50kg	50-75	>75
140	160	200

$$\boxed{\chi^2 = \sum \frac{(O-E)^2}{E}}$$

$$\Rightarrow \chi^2 = \frac{(140-100)^2}{100} + \frac{(160-150)^2}{150} + \frac{(200-250)^2}{250}$$

$$= \frac{1600}{100} + \frac{100}{150} + \frac{2500}{250} =$$

$$= 26 + \frac{10}{15}$$

$$= \frac{390+10}{15} = \frac{400}{15} = \frac{80}{3} = 26.66$$

as $26.66 > 5.91$ $\therefore$ we reject H_0 (null hypothesis)

Step Conclusion:- The weights of 2020 population are different than those expected in the 2010 population.

CHI SQUARE TEST for PYTHON