→ Conclusion or Inferences about population data

Date
Page

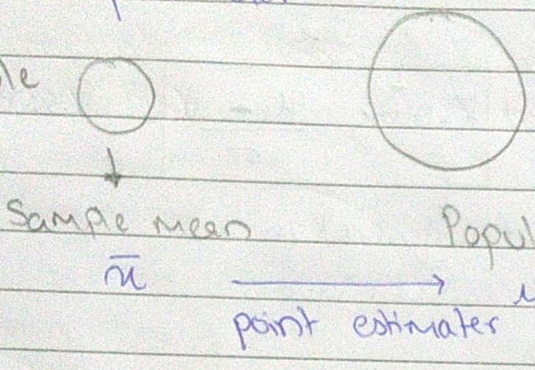
INFERENTIAL STATS

* Estimate

It is an observed numerical value used to estimate an unknown population parameter.

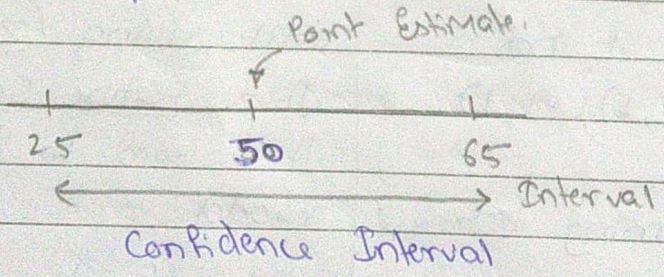
1) Point Estimate

Single numerical value used to estimate the unknown population parameter.

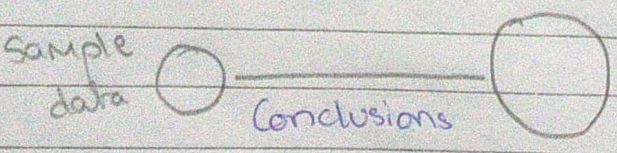


2) Interval Estimate

Range of values used to estimate the unknown population parameter



* Hypothesis & Hypothesis Testing Mechanism



↳ Made of hypothesis testing

→ what is given in question always goes to null hypothesis

Date
Page

1) Null hypothesis (H_0)
- Assumption you are beginning with
- Ex: The person is not guilty

2) Alternate hypothesis (H_1)
- opposite of null hypothesis
- Ex: The person is guilty

→ z test, t test, Anova, chi-square

3) Evidences
→ performing experiments ⇒ Statistical Analysis
- Ex: (finger test, DNA, etc.. test)

4) we fail to reject the null hypothesis / Reject the null hypothesis

* P value

P value is a number, calculated from a statistical test, that describes how likely you are to have found a practical set of observations if null hypothesis were true.

- P values are used in hypothesis testing to help decide whether to reject the null hypothesis

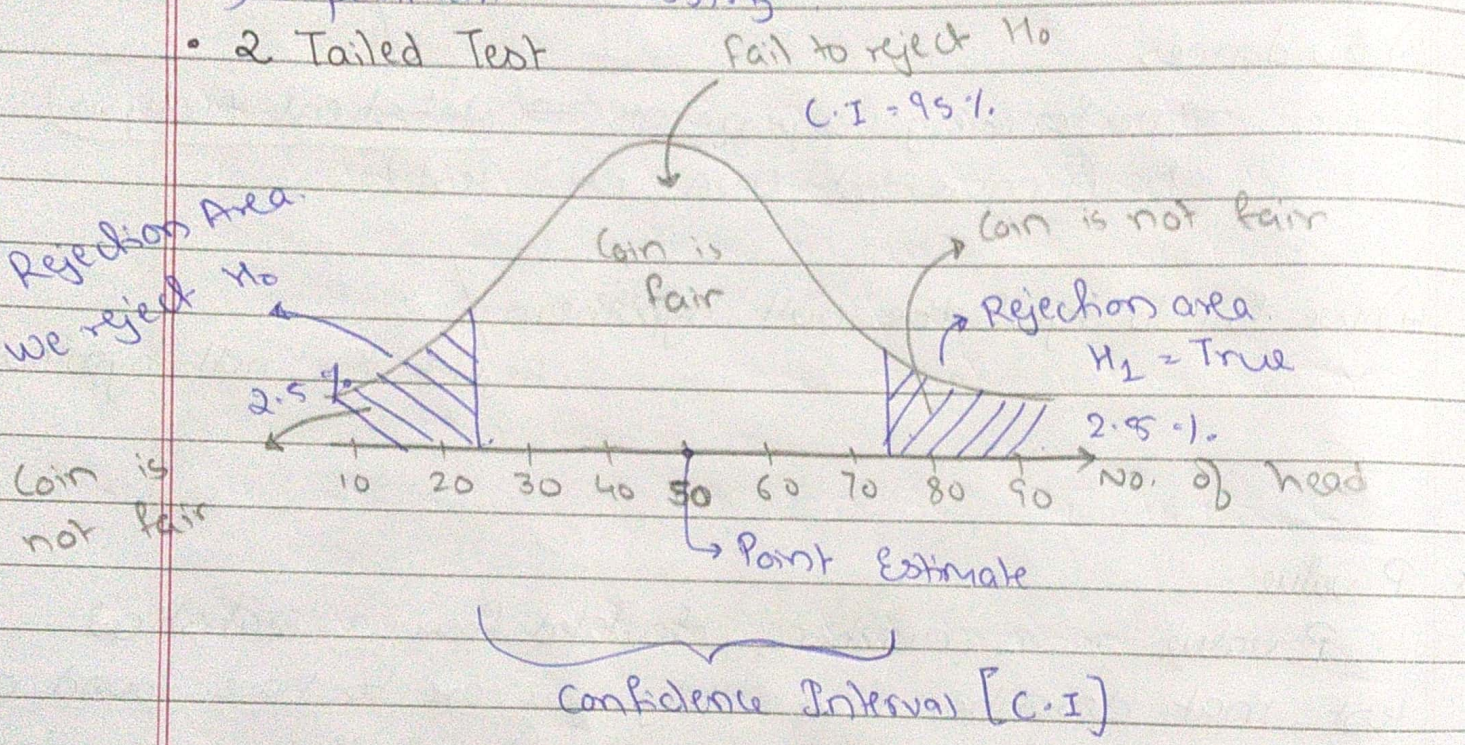
→ Ex: Coin is fair or not

Hypothesis Testing:

1) null hypothesis H_0 = Coin is fair
2) Alternate hypothesis H_1 = Coin is not fair

3) Experiment → Tossing

• 2 Tailed Test



• Significance value (α)

$$\alpha = 1 - C.I = 0.05$$

P value < Significance

⇓

we reject null hypothesis [Coin is not fair]

else

we fail to reject null hypothesis [Coin is fair]

★ Confidence Interval & Margin of Error

We construct a confidence interval to help estimate what is actual value of unknown population Mean

Point Estimate ± Margin of Error

$$\bar{x} \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

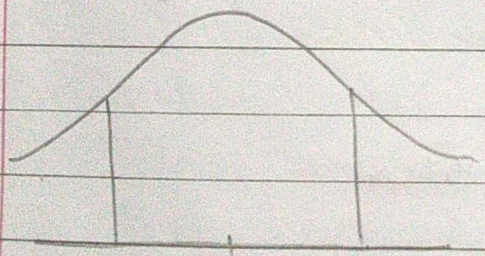
(Z → Z table; $\frac{\sigma}{\sqrt{n}}$ → standard error; α → Significance value)

Q. In verbal section of CAT Exam, standard deviation is known to be 100. A sample of 25 test takers has a Mean of 520. Construct a 95% C.I about the Mean?

Ans. $\sigma = 100$, $n = 25$, $\bar{x} = 520$, $C.I = 0.95$, $\alpha = 0.05$

$$Z_{0.05/2} = Z_{0.025}$$

$$\bar{x} + Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$



$\therefore$ Lower C.I $= 520 - (1.96) * \frac{100}{\sqrt{25}} = 480.8$

480.8

559.2

$\therefore$ Higher C.I $= 520 + (1.96) * \frac{100}{\sqrt{25}} = 559.2$

$\therefore$ I am 95% confident/sure that Mean CAT score lied between 480.8 and 559.2

→ Hypothesis Testing & Statistical Analysis

1) Z test } → Average
2) t test }
3) chi Square → Categorical
4) Annova → variance

• Z test :

q The average heights of all residents in a city is 168 cm with $\sigma = 3.9$. A doctor believes the mean to be different. He measured the height of 36 individuals & found the avg height to be 169.5 cm
1) State null & alternate hypothesis
2) At a 95% C.I, is there enough evidence to reject null hypothesis

Ans. $\mu = 168$ cm, $\sigma = 3.9$, $n = 36$, $\bar{x} = 169.5$ cm

* whenever, population standard deviation is given used Z-test

a) null hypothesis H_0 $\mu = 168$ cm
b) Alternate hypothesis H_1 $\mu \neq 168$ cm

c) C.I = 0.95 $\alpha = 1 - 0.95 = 0.05$

d) Statistical Analysis

$$Z\text{-test} = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

$\sigma/\sqrt{n}$ → Standard Error

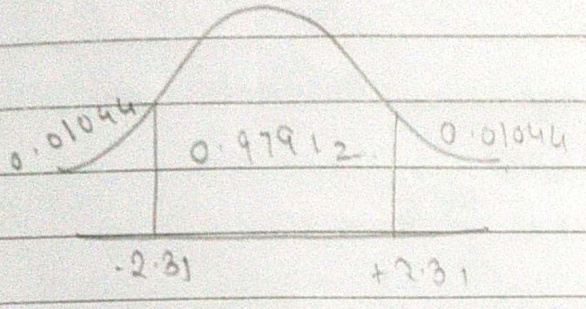
$$= \frac{169.5 - 168}{3.9/\sqrt{36}} = 2.31$$

Conclusion:

If Z-test value is less than -1.96 or greater than +1.96 we reject the null hypothesis

2.31 > 1.96 ⟹ reject the null hypothesis

- P-value



P-value = 0.01044 + 0.01044
= 0.02088

If P-value < significance value
0.02088 < 0.05
∴ Reject the null hypothesis.

- T test → when we have limited sample size (ie $n < 30$)
→ when we don't know population standard deviation

Q. In the population the average IQ is 100. A team of researchers want to test a new medication to see if it has either 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 with a standard deviation of 20. Did the medication affect intelligence? CI=95

Ans Given:

$\mu = 100$, $n = 30$, $\bar{x} = 140$, $s = 20$, $C.I = 0.95$, $\alpha = 0.05$

① null hypothesis H_0 : $\mu = 100$
② Alternate hypothesis H_1 : $\mu \neq 100$ {2 tail test}

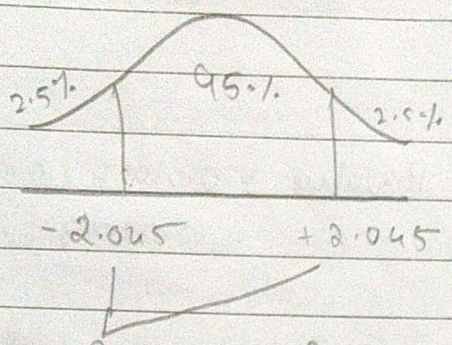
• 1 tail test : sample mean = population mean / true

Date
Page

③ $\alpha = 0.05$ $CI = 0.95$

④ Degree of freedom [dof] $= n - 1 = 30 - 1 = 29$

⑤ Decision boundary



values from t-table,
2 tailed test

Conclusion : if t-test is less than -2.045 & greater than 2.045, reject the null hypothesis

⑥ t test statistics

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

⑦ Conclusion

$10.96 > 2.045$ ∴ we reject null hypothesis

Here we don't know population standard deviation that's why we are doing t-Test

★ When to use T-test Vs Z-test

Do you know the population Std σ?

- Yes → Is the sample size above 30?
  - Yes → Use Z test
  - No → Use T test
- No → Use t-test

★ DISPERSION

Spread of the Data.

① Range : Upper Limit − lower Limit

10, 50, 9, 8, 25, 60

Sort → 8, 9, 10, 25, 50, 60 → upper

lower

60 − 8 = 52

② Percentile [Quantile, Decile]

↳ 5 no. of summary → which describe's the data

IQR = $q_3 - q_1$

= 75% − 25%

= 50%

11, 12, 13, 16, 16, 17, 18, 21, 22

50

Median value

$q_1 - (IQR * 1.5)$ ..... 25% | 25% | 25% | 25% ..... $q_3 + (IQR * 1.5)$

q_1 q_2 q_3 q_4

q_1

75

Chi Square Test

The chi square test for goodness of fit claims about population proportions [categorical variables]

- It is a non parametric test [i.e will not use mean, median or mode] that is performed on categorical data [nominal, ordinal]

Eg: Population of Male who likes different color of bikes

	Theory	Sample
Yellow Bike	1/3	22
Orange Bike	1/3	17
Red Bike	1/3	59 → observed categorical Distribution [count]

↓

Theory categorical distribution [percentage]

talks more about proportions of distribution.

Q. In a student class of 100 students, 30 are right handed. Does this class fit the theory 12% of people are right handed

	Observed	Expected	
Right handed	30	12	} Theory
Left handed	70	88	

↓

Observed

→ Chi-Square : For Goodness of Fit

Q In 2010 census of the city, the weight of individuals in a small city were found to be following

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

→ Expected

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

<50	50-75	>75
140	160	200

→ Observed

Confidence Interval = 95%

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

Ans.

Expected

<50	50-75	>75
$\frac{20 \times 500}{100}$	$\frac{30 \times 500}{100}$	$\frac{50 \times 500}{100}$
100	150	250

① Null hypothesis : H_0 : The data meets the expectations

② Alternate hypothesis : H_1 : The data does not meet the expecta

③ $\alpha = 0.05$, CI = 95%

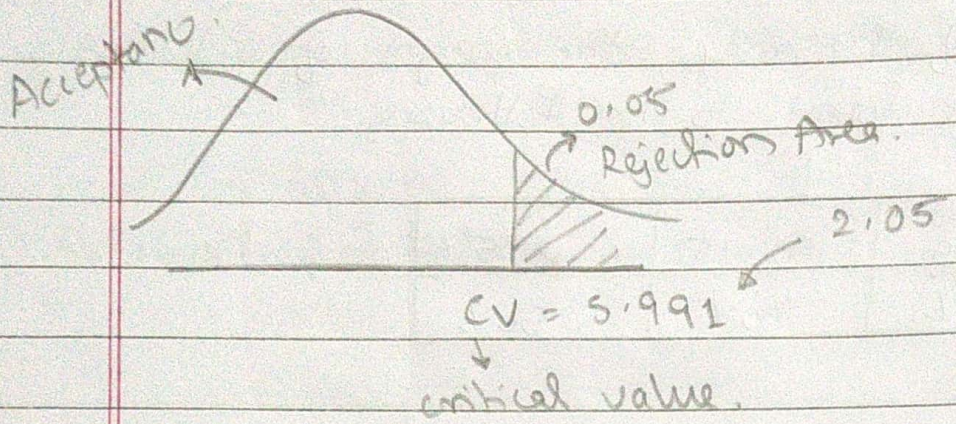
④ Degree of freedom

$df = k - 1 = 3 - 1 = 2$

→ chi-square table is always 1-tail table

⑤ Decision Boundary → CHI SQUARE Table

- Right ~~set~~ Skewed



If χ^2 is greater than 5.991, Reject H_0
else
we fail to reject H_0.

⑥ Calculate Chi-Square Test Statistics

$$\chi^2 = \sum_{i=1}^{k} \frac{(O-E)^2}{E}$$

O: Observed
E: Expected

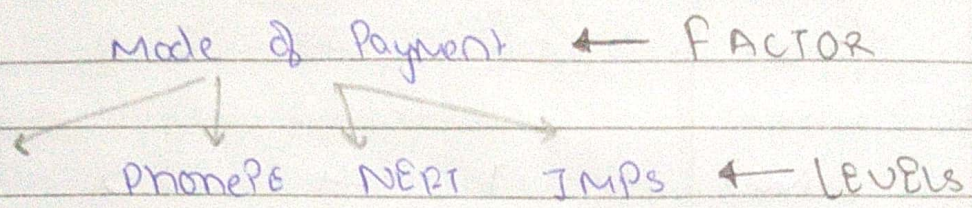
$$= \frac{(140-100)^2}{100} + \frac{(160-150)^2}{150} + \frac{(200-250)^2}{250}$$

$$= 16 + 0.66 + 10$$

$$\boxed{\chi^2 = 26.66}$$

$26.66 > 5.99$, Reject H_0

- ANOVA [Analysis of Variance]
  It is a statistical method used to compare the mean of 2 or more groups

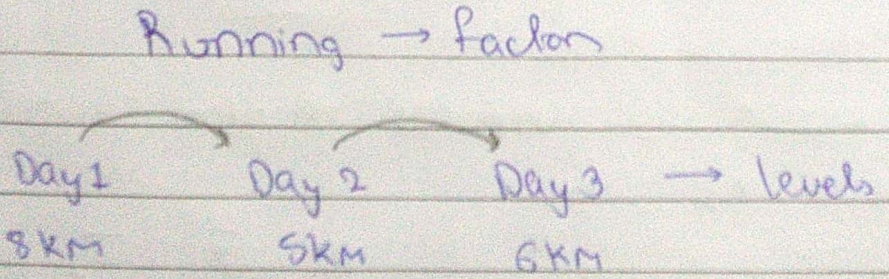


→ Types of Anova

1) One way Anova : One factor with atleast 2 levels, these levels are independent

Eg: Factor is brand of soda, & you collect data on Coke, Pepsi, Sprite to find out if there is difference in price per 100 ml

2) Repeated Measures Anova : One factor with atleast 2 levels, levels are dependent



3) Factorial Anova : Two or more factors [each of which with atleast 2 levels], levels can be either dependent or independent

Running ← factor

independent		Day1	Day2	Day3 ← dependent
Gender ↑ factor	Male	8	5	6
	female	7	4	3

→ Hypothesis Testing In Anova

- Null hypothesis $H_0 : \mu_1 = \mu_2 = \mu_3 = \cdots = \mu_K$
- Alternate hypothesis H_1 : Atleast one of mean is not equal

→ F Test Statistics

$$F = \frac{\text{Variation between Samples}}{\text{variation within Samples}}$$

variance between sample

	x_1	x_2	x_3
variance within sample	1	6	5
	2	7	6
	4	3	3
	5	2	2
	3	1	4

$H_0 = \bar{X}_1 = \bar{X}_2 = \bar{X}_3$

H_1 : Atleast one sample Mean is not equal.

$\bar{X}_1 = 3 \quad \bar{X}_2 = 19/5 \quad \bar{X}_3 = 4$

Q Doctors want to test a new medication which reduces headache. They split the participant into 3 condition [15Mg, 30Mg, 45 Mg]. Later on doctor ask the patient to rate the headache between [1-10]. Are there any differences between the 3 conditions using alpha = 0.05?

	15Mg	30Mg	45 Mg
N: Total number of samples.	9	7	4
	8	6	3
	7	6	2
a: No. of categories	8	7	3
	8	8	4
n: number of sample in 1 column	9	7	3
	8	6	2

① Null hypothesis H_0 : $\mu_{15} = \mu_{30} = \mu_{45}$

② Alternate hypothesis H_1 : Atleast one Mean is not equal

③ State significance value

$\alpha = 0.05 \Rightarrow C.I = 0.95$

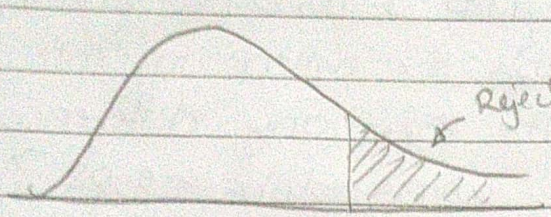
④ Calculate degree of freedom [df]

$N = 21$, $a = 3$, $n = 7$

df between = $a - 1 = 3 - 1 = 2$ $(2, 18) \Rightarrow$ F test table

df within = $N - a = 21 - 3 = 18$

Decision Boundary $\alpha = 0.05$



3.5546 ← (2, 18) : F test table

⑤ Calculate F-test Statistics

	SS [Sum of Squares]	df	MS [Mean Square]	F
Between	98.67	2	49.34	
within	10.29	18	0.54	86.56
Total	108.96	20		

① $SS_{between} = \frac{\Sigma(\Sigma a_i)^2}{n} - \frac{T^2}{N}$

98.67 / 2

10.29 / 18

15 Mg = 9+8+7+8+8+9+8 = 57

30 Mg = 7+6+6+7+8+7+6 = 47

45 Mg = 4+3+2+3+4+3+2 = 21

$\Rightarrow T^2$

$\Sigma y^2 = 9^2+8^2+7^2+8^2+8^2+9^2+8^2+7^2+6^2+6^2+7^2+\ldots$

$= 853$

$= \frac{57^2+47^2+21^2}{7} - \frac{[57+47+21]}{21}$

$= 98.67$

② $SS_{within} = \Sigma y^2 - \frac{\Sigma(\Sigma a_i)^2}{n}$

$= 853 - \left[\frac{57^2+47^2+21^2}{7}\right]$

$= 10.29$

$$F = \frac{\text{variation between samples}}{\text{variation within samples}} = \frac{\text{MS between}}{\text{MS within}}$$

$$= \frac{49.34}{0.54} = 86.56$$

$86.56 > 3.556$, we reject null hypothesis.