Unit - 3

Sampling :-

- Fundamental concept in statistics that involves selecting subset of individuals / items from a large population.
- It is used to draw conclusion about entire population.

Population - The entire group of individuals, items or data points under consideration

Sample - A subset of population for analysis.

Types of Sampling

- Probability Sampling

A probability sample means that every member of the population has a chance of being selected.

- used in quantative research

Four types of probability sampling -

1. Simple random sampling
2. Systematic sampling
3. Stratified sampling
4. Cluster sampling

- Non Probability Sampling

Individuals are selected on non-random criteria & not every individual has a chance of being selected.

- easier & cheaper to access
- high risk of sample biasness
- used in exploratary & qualitative reasearch

Four types of Non-probability Sampling –

1) Convenience Sampling
2) Purposive Sampling
3) Snowball Sampling
4) Quota Sampling

Probability Sampling techniques –

1. Simple random Sampling :- Every member of the population has equal chance
   - sample frame includes whole population.
   - use tools like random number generator

   eg There are 1000 employees & selecting 100 out of these 1000.

2. Systematic Sampling :- similar to simple random sampling
   - selection is done by choosing individual at regular interval.

   eg Out of 100, 10 students are choosen such as (6^{th}, 16^{th}, 26^{th}, ... 86^{th}, 96^{th}) student is selected in sample.

3. Stratified Sampling :- Dividing the population into subpopulation based on relevant criteria.
   - Then sample is created from subgroup (sub population) using systematic sampling.

   eg Out of 1000, 800-female & 200male then sample of 100 is generated 80female & 20male. (Strata based on gender)

4 Cluster Sampling :- Dividing the population into subgroups, but each subgroup must have similar characteristics

- Instead of randomly selecting individuals, we select entire subgroups as sample.

eg Company has offices in 10 cities, you make sample by selecting 3 offices (3 cities). (here every office is a cluster).

* Multistage sampling - If the cluster is too much larger, then sample individuals from within each clusters, this is known as Multistage sampling.

Non-Probability Sampling —

1 Convenience Sampling :- Sample simply includes the individuals who happen to be most accessible to the researcher.

- Easy & inexpensive way to gather initial data. (there is no way to tell if sample represent population.

eg College surveys like asking nearby students about certain things regarding college.

2. Voluntary

2. Purposive Sampling - Also known as judgemental sampling, involves the researcher using their expertise to select a sample that is most useful to purpose of the research.

- Also used in qualitative research, where the researcher wants to gain detailed knowledge about a specific phenomenon. (population is small & specific).

eg Taking opinion of 4th year students regarding placement

3 Snowball sampling - Used when population is hard to access, Recruiting participants via other participants. Since people increase one after another like "snowball", your access increases.

- Sampling biasness happens in this.

eg: Gathering data of homeless people from one homeless person, thus the sample increases.

4 Quota Sampling - Relies on the non-random selection of a predetermined number or proportion of units. This is called quota.

- First divide population into mutually exclusive subgroups (strata) & then sample units until you reach your quota

eg population - 1000 divide strata - vegetarian, non-veg. & vegans recruit 200 for each then compare..

Camlin Page
Date / /

Difference b/w Parameter & Statistic

Parameter	Statistic
- Numerical value that describe characteristics of population	- Numerical value that describe characteristics of sample.
- Represented by Geek letters mean = μ S.d. = σ size = N	- Represented by Roman letters Mean = $\bar{x}$ Sd = s size = n
- Parameters are fixed & unchangable, usually known	- Statistics can be variable depending on different samples
- Calculated using data from entire population	- Computed for different samples acc. to data.
- Used to make inference about a population	- Used to make inference about population based on a sample.

Test of Significance

- Statistical method used to determine whether an observed effect (result) is likely to be genuine or occurred by chance.
- These test help to draw meaningful conclusion from data.
- Null Hypothesis (H_0): Hypothesis is correct but rejected
- Alternate Hypothesis (H_1): Hypothesis is wrong but accepted.

- **Type 1 error : (α) Rejecting H_0 while H_0 is true**

Null Hypo :- Hypothesis which is tested for possible rejection under the assumption that it is true, is called type 1 error / Null hypothesis.

- **Type 2 error (β): Accepting H_0 while H_0 is false**

Alternate :- Hypothesis which is tested for possible acceptance under the assumption that it is false, is called ~~type 2 error~~ / Alternate hypothesis.

Type 1 error (α): The error of rejecting H_0 when H_0 is true is called type 1 error.

- It gives false positive conclusion.
- It is denoted by alpha (α).

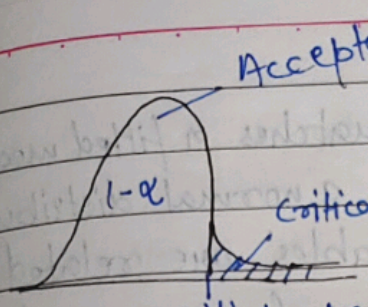
$$\alpha = P(\text{Rejecting } H_0 / H_0 \text{ is true})$$

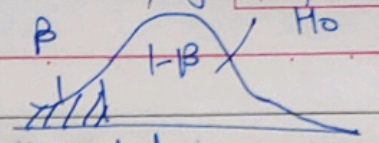
Type 2 error (β): The error of accepting H_0 when H_0 is false is called type 2 error.

- It gives false negative conclusion
- It is denoted by beta (β)

$$\beta = P(\text{Accepting } H_0 / H_0 \text{ is false})$$

- Critical region : If the statistic value falls in critical region, the null hypothesis is rejected.
- Critical value : Boundary value that speraates critical region from acceptance region.

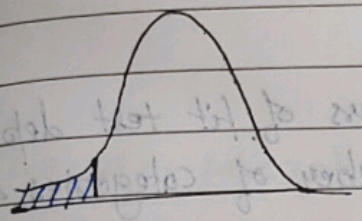




Area = α

Right tailed

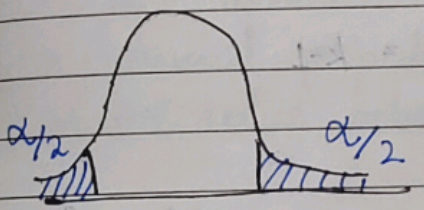
H_0:



Area = α

Acceptance = $1-\alpha$

Left tailed



Area = $\frac{\alpha}{2} + \frac{\alpha}{2} = \alpha$

Two tailed

$H_0: \mu = \bar{x}$

$H_1: \mu \neq x$

- Significance level : The probability of rejecting the null hypothesis when it is true. (Standard val)
  - standard value is 0.05 or 5%.
- P-value : The observed probability of rejecting the null hypothesis when it is true. (Experimental val)
  - If P-value < Significant value ∴ Null hypothesis rejected
  - ~~If Reverse~~

Goodness of fit test

A goodness of fit test is a statistical method used to determine if an observed frequency distribution fits a theoretical model or expected distribution.

- Sample data follows a specific probability distribution.

Camlin Page
Date / /

- Measures how well observed data matches a fitted model.
- Determine if a sample follows a normal distribution.
- If categorical variables are related
- Random samples are from same distribution

- Degrees of Freedom: The goodness of fit test depends on the number of categories & constraints in the problem.

degrees of freedom (df) = k-1

Chi Square test (Pearson's chi sq. test) χ^2

- Non parameteric test
- Pearson's chi-sq. test is a statistical test for categorical data. It is used to determine whether your data are significantly different from what you expected.

There are two types of pearson's chi-sq. test
1) Chi-sq. goodness of fit test
2) Chi-sq. Test of Independence

1 Chi-sq Goodness of fit test
- It is use to test whether the frequency distribution of a categorical variable is different from your expectation.

2 Chi-sq Test of Independence
- It use to test whether two categorical variables are related to each other.

Camlin Page
Date / /

$\chi^2 = \sum \left\{ \frac{(O_i - E_i)^2}{E_i} \right\}$ $\quad df = n-1$

where O_i = observed frequency
E_i = Expected frequency
n = number of objects

T-test $(n<30)$

- To test whether the mean of a sample drawn from a normal population derivates significantly from a stated value when variance of population is unknown

One Sample	Two Sample
$t = \frac{\lvert \bar{x} - \mu \rvert}{s/\sqrt{n}}$	$t = \frac{\lvert \bar{x}_1 - \bar{x}_2 \rvert}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$
$df = n-1$	$df = n_1 + n_2 - 2$
$\bar{x}$ = mean of Sample	$\bar{x}_1, \bar{x}_2$ = mean of sam$_1$, sam$_2$
μ = mean of Population	s_1, s_2 = sd of sam$_1$, sam$_2$
s = sd. of sample	n_1, n_2 = sam$_1$, sam$_2$ size
n = sample size	

Z-test $(n>30)$

- To test whether there is significance difference b/w sample mean & population mean, given that variance of population is known.

$$z = \frac{\lvert \bar{x} - \mu \rvert}{SEM}$$

$\bar{x}$ = sample mean
μ = population mean
n = sample size. SEM = St. error of mean

$$SEM = \frac{\sigma}{\sqrt{n}}$$

σ = Sd. population

=> Two Sample Z-test

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

$\bar{x}_1, \bar{x}_2$ = Sample means

n_1, n_2 = Sample size

Significance level with standard value for Z-test

	Signi level	Confi level	2-tailed	1-tailed
0.1	10%	90%	1.65	1.28
0.05	5%	95%	1.96	1.64
0.01	1%	99%	2.58	2.33
0.001	0.1%	99.9%	3.29	3.10

*** (5 marks)

Qw. Write down the procedure of Hypothesis testing?

Ans

Steps :-

1) Setting up Hypothesis — • Null Hypothesis
   • Alternate Hypothesis
2) Set up significance level — (5%, 10%, 15%, ...)
3) Testing the hypothesis —
   - Z-test (n>30)
   - t-test (n<30) (or we have two means $\bar{x}_1, \bar{x}_2$)
   - χ^2-test
   - F-test
   - ANOVA
4) Doing Computation (comparing with Standard value)
5) Making Decision (Acceptance / Rejecting Null)

Que. The mean life time of 81 glucometers produced by a company is found to be 58 months with a s.d. of 10 months. Company claims that mean life time of glucometers is 60 months. Test the hypothesis at 5% level of significance whether the 58 months life time is accepted or not.

Sol^n given $n = 81$ $\mu = 60$ $\bar{x} = 58$
$\sigma = 10$ $sf = 5\%$

H_0: There is no significant difference b/w means $\bar{x} = \mu$
H_1: There is significant difference b/w means $\bar{x} \neq \mu$.

$$\therefore Z = \frac{|\bar{x} - \mu|}{\frac{\sigma}{\sqrt{n}}} = \frac{|58 - 60|}{10/\sqrt{81}}$$

$$Z = \frac{2}{10} \times 9 = 1.8$$

The standard value for two tailed test at 5% level of significance is 1.96

since, $Z = 1.8 < 1.96$

thus H_0 is accepted, means there is no significant difference b/w sample & population mean.

Qn A teacher claims that the mean score of student in his class is greater than 82 with a sd. of 20. If a sample of 81 students was selected at random with a mean score of 90 then check if there is enough evidence to support the claim at a significant level of 0.05.

Sol^n given - $n = 81$ $\mu \geq 82$ $sf = 0.05$
$\sigma = 20$ $\bar{x} = 90$

$\alpha = 5\%$

H_0: Mean score is greater than 82 $\quad \mu > 82$

H_1: Mean score is less than 82 $\quad \mu < 82$

$$z = \frac{|\bar{x} - \mu|}{\frac{\sigma}{\sqrt{n}}}$$

$$z = \frac{|90 - 82|}{\frac{20}{\sqrt{81}}}$$

$$z = \frac{8 \times 9}{20} = 3.60$$

for one-tailed at 0.05 significant level
the standard value is 1.64

since, $z = 3.6 > 1.64$

thus H_0 is rejected

Qu. Is there a significant difference in test score b/w 25 student who recieved inperson instruction & 25 persons who recieved online instruction. The mean test score for inperson group is 80 & for online group is 75 & the s.d. for inperson group is 5 & for online group is 7.

Soln given, $\bar{x}_1 = 80 \quad s_1 = 5 \quad n_1 = 25$

$\bar{x}_2 = 75 \quad s_2 = 7 \quad n_2 = 25$

H_0: There is no significant diff. b/w sample means $x_1 = x_2$

H_1: There is significant diff. b/w sample means.

$$\therefore t = \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

$$t = \frac{80-75}{\sqrt{\frac{25}{25} + \frac{49}{25}}}$$

$$t = \frac{5}{\sqrt{74}} \times 5$$

$$t = \frac{25}{\sqrt{74}} = 2.906 = 2.91$$

since, $df = n_1 + n_2 - 2 = 25 + 25 - 2 = 48$

$sf = 0.05$ standard value $= 2.01$

$\therefore t = 2.91 > 2.01$

H_0 is rejected, that means there is significant diff. b/w means of two given samples.

A normal population has a mean of 6.8 & sd 1.5 a sample of 400 members gave a mean of 6.75, is there any significant difference.

given $\mu = 6.8$ $\bar{x} = 6.75$ $sf = 0.05$

$\sigma = 1.5$ $n = 400$ $H_0: \mu = \bar{x}$ $H_1: \mu \neq \bar{x}$

$$z = \frac{|\bar{x} - \mu|}{\frac{\sigma}{\sqrt{n}}} = \frac{|6.75 - 6.8|}{\frac{1.5}{\sqrt{400}}} = \frac{0.05}{\frac{1.5}{3}} \times \frac{200}{100} = 0.66$$

for two tailed at $sf = 0.05$ standard value $= 1.96$

$z = 0.67 < 1.96$

H_0 is accepted there is no significant diff.

Camlin Page
Date / /

Que. The following table gives the number of accidents that took place in an industry during various days of week. Test if accidents are uniformly distributed over the week.

Days	No. of Accidents (observed)	Expected frequency	$(O-E)^2$	$\frac{(O-E)^2}{E}$
Monday	15	12	9	0.75
Tuesday	11	12	1	0.083
Wednesday	9	12	9	0.75
Thursday	12	12	0	0
Friday	10	12	4	0.33
Saturday	14	12	4	0.33
Sunday	13	12	1	0.083
	84			2.326

Soln

$E = \frac{84}{7} = 12$

H_0: Uniformly distributed
H_1: Non uniformly distributed.

$$\chi^2 = \Sigma \left\{ \frac{(O_i - E_i)^2}{E} \right\}$$

$\chi^2 = 2.326$

$df = N - 1 = 7 - 1 = 6$
Standard Value = 12.59
Since $\chi^2 = 2.326 < 12.59$ ∴ Null hypothesis accepted

Que.

Days	Mon	Tue	Wed	Thus	Fri	Sat
Accidents	14	18	12	11	15	14
Expected	14	14	14	14	14	14
$(O_i - E_i)^2$	0	16	4	9	1	0
$\frac{(O_i - E_i)^2}{E}$	0	1.14	0.28	0.64	0.07	0

$$\chi^2 = \Sigma\left(\frac{(O_i - E_i)^2}{E}\right) = 2.13 < 11.07$$

$df = 6-1 = 5$ standard value = 11.07

$\therefore$ H_0 is accepted.

The life time of Electric bulbs for a random sample of 10 from a large consignment gave the following data can we expect the hypothesis that the avg life of bulb is 4000 hrs.

Sol

Items	Life time (x_i in 1000)	$x - \bar{x}$	
i_1	4.2	-0.2	0.04
i_2	4.6	0.2	0.04
i_3	3.9	0.5	0.25
i_4	4.1	-0.3	0.09
i_5	5.2	0.8	0.64
i_6	3.8	-0.6	0.36
i_7	3.9	-0.5	0.25
i_8	4.3	-0.1	0.01
i_9	4.4	0	0
i_{10}	5.6	1.2	1.44
	44		3.13

Mean $= \frac{44}{10}$

Mean $= 4.4$

$sd = \sqrt{\frac{\Sigma(x_i - \bar{x})^2}{n-1}}$

$sd = \sqrt{\frac{3.13}{9}}$

$sd = 0.58$

Sol H_0: There is no diff. b/w given & observed value $\mu = 4000$

H_1: There is diff.

$\bar{x} = 4.4 \quad s = 0.58 \quad n = 10 \quad df = n-1 = 10-1 = 9$

$$t = \frac{|4.4 - 4|}{\frac{0.58}{\sqrt{10}}} = \frac{0.4 \times \sqrt{10}}{0.58} = 2.18 < 2.262$$

Thus null hypothesis is accepted.

ANOVA = Analysis of variance.

- when we have more than 2 samples

Que. To assess the significance of possible variation in a certain test b/w the schools of a city a common test was given to a number of student taken at random from the 12th class of the three schools concerned. Make the analysis of variance from the below given results.

Sol

A	B	C
2	3	4
4	5	6
6	7	8

①

Null H_0: There is no significance difference $\mu_1 = \mu_2 = \mu_3$

Alternative H_1: There is significance difference

② Calculate the mean of each sample

A	B	C
2	3	4
4	5	6
6	7	8
Total 12	15	18
Mean 4	5	6

$\bar{x}_A = \frac{12}{3} = 4$

$\bar{x}_B = \frac{15}{3} = 5$

$\bar{x}_C = \frac{18}{3} = 6$

Camlin Page
Date / /

Calculate the ground average of means

$$\bar{X} = \frac{\bar{x}_A + \bar{x}_B + \bar{x}_C}{3}$$

$$\bar{X} = \frac{4+5+6}{3}$$

$$\bar{X} = 5$$

Calculation of SSC (Sum of squares b/w samples)

$\bar{X}_A - \bar{X}$	$(\bar{X}_A - \bar{X})^2$	$(\bar{x}_B - \bar{X})$	$(x_B - \bar{X})^2$	$(\bar{x}_c - \bar{X})$	$(\bar{x}_c - \bar{X})^2$
4-5=-1	1	5-5=0	0	6-5=1	1
4-5=-1	1	5-5=0	0	6-5=1	1
4-5=-1	1	5-5=0	0	6-5=1	1
	3		0		3

$\Sigma(\bar{X} - \bar{\bar{X}}) = 3+0+3=6$ (SSC)

↓ (grand avg)

avg

Calculation of SSE (Sum of squares within samples)

$A - \bar{X}_A$	$(A - \bar{X}_A)^2$	$(B - \bar{X}_B)$	$(B - \bar{X}_B)^2$	$(C - \bar{X}_C)$	$(C - \bar{X}_C)^2$
2-4=-2	4	3-5=-2	4	4-6=-2	4
4-4=0	0	5-5=0	0	6-6=0	0
6-4=2	4	7-5=2	4	8-6=2	4
	8		8		8

$\Sigma(A - \bar{X}_A)^2$

$\Sigma(x - \bar{X})^2 = 8+8+8=24$

Sq. of Variations	Sum of Squares	Degree of freedom	Mean Sum of Squares	$F = \frac{MSC}{MSE}$
1 B/w Samples	SSC = 6	$V_1 = c-1$ $= 3-1 = 2$	$MSC = \frac{SSC}{c-1} = 3$	$f = 3/4$ $= 0.75$
2 within Samples	SSE = 24	$V_2 = n-c$ $= 9-3 = 6$	$MSE = \frac{SSE}{n-c} = 4$	

n = no. of samples = 9 c = no. of columns = 3 (A, B, C)

since 0.75 < 5.14 thus H_0 accepted

Anova table

look sample no. (n-c) ↓

(c-1) ← look column no.

(n-c)(c-1) (n-c) standard value.

Que. Calculate the Anova coefficient for following data

Plant	Number	Avg Span	s^2
Hab	5	12	2
Masi	5	16	1
Rose	5	20	4

Solu given:

Plant	No.	Avg	sd
Hab	5	12 ($\bar{X}_A$)	2
Masi	5	16 ($\bar{X}_B$)	1
Rose	5	20 ($\bar{X}_C$)	4

$$\bar{\bar{X}} = \frac{\bar{X}_A + \bar{X}_B + \bar{X}_C}{3} = \frac{12+16+20}{3} = \frac{48}{3} = 16$$

$(\bar{X}_A - \bar{\bar{X}}) = (12-16) = -4 \quad (\bar{X}_A - \bar{\bar{X}})^2 = 16 \quad \Sigma(\bar{X}_A - \bar{\bar{X}}) = 80$

$(\bar{X}_B - \bar{\bar{X}}) = (16-16) = 0 \quad (\bar{X}_B - \bar{\bar{X}})^2 = 0 \quad \Sigma(\bar{X}_B - \bar{\bar{X}}) = 0$

$(\bar{X}_C - \bar{\bar{X}}) = (20-16) = 4 \quad (\bar{X}_C - \bar{\bar{X}})^2 = 16 \quad \Sigma(\bar{X}_C - \bar{\bar{X}}) = 80$

since $n_1 = 5 = n_2 = n_3 \quad \therefore \Sigma(\bar{X}_A - \bar{\bar{X}}) = (\bar{X}_A - \bar{\bar{X}}) \times n_1$

$\therefore SSC = \Sigma(\bar{X} - \bar{\bar{X}}) = 80 + 0 + 80 = 160$

since $s = \sqrt{\frac{\Sigma(x-\bar{x})^2}{n-1}}$

$\therefore s^2 = \frac{\Sigma(x-\bar{x})^2}{n-1}$

for SSE we need $\Sigma(x-\bar{x})$

$\therefore s^2(n-1) = \Sigma(x-\bar{x})^2$

$\Sigma(x_A - \bar{X}_A) = s_1^2(n_1-1) = 4 \times (5-1) = 16$

$\Sigma(x_B - \bar{X}_B) = s_2^2(n_2-1) = 1 \times (5-1) = 4$

$\Sigma(x_C - \bar{X}_C) = s_3^2(n_3-1) = 16 \times (5-1) = 64$

$\therefore SSE = \Sigma(x-\bar{x}) = 16 + 4 + 64 = 84$

Sq. of variations	Sum of Squares	Degree of Freedom	Mean Sum of Sq.	$F = \frac{MSC}{MSE}$
B/w Samples	SSC = 160	$v_1 = c-1$ $= 3-1 = 2$	$MSC = \frac{SSC}{v_1} = \frac{160}{2}$ $MSC = 80$	$F = \frac{80}{7}$
Within Sample	SSE = 84	$v_2 = n-c$ $= 15-3 = 12$	$MSE = \frac{SSE}{v_2} = \frac{84}{12}$ $MSE = 7$	$F = 11.42$

→ Garden centre wala numerical.
→ If the median of a distribution of given below is 28.5 the find x and y.

C.I	0-10	10-20	20-30	30-40	40-50	50-60
Frequency	5	x	20	15	y	5

Total Frequ = 60

→ probability marks (***) in notes

Unit 3 formulas from ppt shared by many.
Numerical → Normal distribution from ppt page no - 21
- sampling and its types. - 10 marks.
- sampling - cluster Vs stratified. (or area)
- prob sampling Vs non-prob sampling
- procedure for hypothesis test
- z-test Vs t-test.
- ANOVA - 10 mark.
No. of accidents - chi-square - 10 mark. (goodness of fit)
- population Vs mean
- Imp of statistics - 2 marks
- what is p-value, critical region - 2m and similar definitions - 3 mark
- teacher claim question.
- ~~Abou~~
- In-person and online wala question
- 4000 hours wala - 10 marks
**** - explain types of z-test. 1D, 2D, 1 sa
- Poisson - 10 mark.