

## Statistics-II Summary - Ankit Pangani

- o **Population:** A set of similar items or events that we take for the study of statistics. It is a collection of data from which a statistical sample is drawn for study.
- o **Parameter:** A value (piece of data) that tells you something about whole population. It is more reliable than sample.
- o **Statistic:** A piece of data that describes only the sample of a population.
- o **Sampling:** Process of selecting a sample from a population for study and research.
- o **Sampling distribution:** Probability distribution of the statistic of the repeated samples taken from a specific population.
- o **Sampling distribution of mean:** Distribution of the <sup>sample</sup> means of the repeated samples taken from a specific population.
 

$\text{mean } E(\bar{x}) = \mu$ $\text{Variance } V(\bar{x}) = \frac{\sigma^2}{n} \left( \frac{n-n}{n-1} \right)$ $\text{Standard Error (S.E)} = \sqrt{V(\bar{x})}$	with replacement (infinite population) $\text{mean } E(\bar{x}) = \mu$ $\text{Variance } V(\bar{x}) = \frac{\sigma^2}{n}$ $\text{Standard Error (S.E)} = \sqrt{V(\bar{x})}$
---	--
- o **Properties of Sampling dist. of mean**
  - It is unbiased estimate of population mean ( $\mu$ )
  - Variance depends on sample size ( $n$ ) i.e.  $V = \frac{\sigma^2}{n}$
- o **Sampling distribution of proportion:**

(proportion): No. of elements with a given characteristic divided by total no. of elements in that group

$$P = \frac{x}{n}, \quad P = \frac{X}{N}, \quad Q = 1 - P$$

where,  $P = \text{popn proportion of success}$   
 $q = 1 - P \quad p = \text{sample propn of success}$

Distribution of sample proportions of the repeated samples taken from a specific population is called sampling distribution of proportion.

→ To study any qualitative study, we need to estimate population proportion.

<b>Infinite popn:</b> $E(p) = P, \quad \text{Var}(p) = \frac{PQ}{n}$	<b>Finite popn:</b> $E(p) = P$ $\text{Var}(p) = \frac{PQ}{n} \left( \frac{N-n}{N-1} \right)$
--	--

Central Limit Theorem:  
States that, as the sample size gets large enough, the sampling distribution of mean is approximately normally distributed.

Inferential statistics: Descriptive statistics

Descriptive statistics describes data (ex: a chart or graph)  
and Inferential statistics allows you to make predictions from that data.

Estimation: Process of estimating the unknown population parameters like population mean, variance from known sample statistics like sample mean, variance.

Estimator: The sample statistic which is used to estimate the unknown population parameter ex: sample mean.

Estimate: The particular value taken by the estimator.

Types of estimation:  
 \* Point estimation: A sample statistic (numerical value) is used to provide an estimate of popn parameter.  
 \* Interval estimation: Probable range is specified within which the value of parameter might be expected to lie.

Properties of good estimator (criteria): SE  
 S → sufficiency      U → consistency  
 E → Efficiency      U → unbiasedness.

Confidence Interval (CI) estimation of population mean ( $\mu$ )

Large sample ( $n > 30$ )  
 $C.I = \bar{x} \pm z_{\alpha} \cdot \frac{\sigma}{\sqrt{n}}$

$C.I = \bar{x} \pm z_{\alpha} \cdot \frac{\sigma}{\sqrt{n}} \cdot \sqrt{\frac{n-1}{N-1}}$   
 If  $N$  given

$\left  \begin{array}{l} \text{large sample } (n > 30) \\ C.I = \bar{x} \pm t_{\alpha, n-1} \frac{s}{\sqrt{n-1}}, \text{ if } s \text{ given} \\ C.I = \bar{x} \pm t_{\alpha, n-1} \frac{s}{\sqrt{n}}, \text{ if } s \text{ calculated} \\ C.I = \bar{x} \pm t_{\alpha, n-1} \frac{s}{\sqrt{n-1}} \cdot \sqrt{\frac{n-1}{N-1}}, \text{ if } N \text{ given.} \end{array} \right.$
---

Confidence Interval estimation of population proportion ( $\pi$ )

Large sample ( $n > 30$ )

$$C.I = p \pm z_{\alpha} \cdot \sqrt{\frac{pq}{n}}$$

$$C.I = p \pm z_{\alpha} \cdot \sqrt{\frac{pq}{n}}$$

If  $N$  is given.

Small sample ( $n < 30$ )

$$C.I = p \pm t_{\alpha, n-1} \sqrt{\frac{pq}{n}}$$

$$C.I = p \pm t_{\alpha, n-1} \sqrt{\frac{pq}{n}}$$

Note:  $t_{\alpha, n-1}$  = critical value of  $t$

$$C.I = p \pm t_{\alpha, n-1} \sqrt{\frac{pq}{n}}$$

$$\text{Sample size } n = \frac{z_{\alpha}^2 \sigma^2}{e^2} \quad (\text{mean}) \quad | \quad n = \frac{z_{\alpha}^2 p q}{e^2} \quad (\text{proportion})$$

Note: Confidence level increases, sample size increases  
\* Sampling error decreases, Sample size decreases  
\* Sample size increases, width of CI decreases

Hypothesis: A tentative theory or supposition of something.

- Hypothesis: A tentative theory or supposition of something.
- Statistical hypothesis: (parametric): A statistical hypothesis is a tentative statement or supposition about the estimated value of one or more population parameters.
- Non-parametric hypothesis: Statistical hypothesis about attributes.
- Note: If hypothesis completely defines the population, then it is simple hypothesis otherwise composite hypothesis.
- Types of hypothesis:
  - \* Null hypothesis: Supposition about the population parameter. It is the hypothesis of difference between parameters which is tested for possible rejection under the assumption that it is true. Set up as:  $H_0: \mu = \mu_0$
  - \* Alternate hypothesis: Complementary to null hypothesis. It is the hypothesis of difference between parameters.  
Set up as:
    - $H_1: \mu \neq \mu_0$  (two tailed)
    - $H_1: \mu > \mu_0$  (one tailed right)
    - $H_1: \mu < \mu_0$  (one tailed left)
- Errors in hypothesis testing on the basis of accepting or rejecting the hypothesis. Two types
  - \* Type-I error: Error of rejecting null hypothesis when it is true. Its probability is  $\alpha$  (level of significance)
  - \* Type-II error: Error of accepting null hypothesis when it is false. Its probability is  $\beta$ .
- Power of test ( $1-\beta$ ): Probability of rejecting null hypothesis when it is actually false. i.e. probability of making correct decision.
- Test-Statistic: Statistic based upon approximate probability distribution. Used to decide whether to accept or reject null hypothesis.  
 $(n \geq 30)$ 
  - \* Z-test: used for large samples in case of test of significance of mean, difference between means, test of proportion & difference between proportions
  - \* t-test: used for small samples ( $n \leq 30$ ) in case of " " "

\* Level of significance: ( $\alpha$ )

Two errors (type I & type II) are inversely related. Both can't be minimized at the same time. usually we fix Type I and minimize Type-II error. Maximum size of type I error prepared in hypothesis testing is L.O.S. It should be chosen on the basis of power of test ( $1-\beta$ ) & if P.O.t is too low,  $\alpha$  should be high.

\* Critical region: (rejection region) & acceptance region

The set of all possible values of statistic is divided into two regions, one leading to rejection & other leading to acceptance of null hypothesis. The division is based on L.O.S & alternate hyp. So, rejection which leads to the rejection of  $H_0$  = rejection region  
" " " acceptance of  $H_0$  = acceptance region

\* Critical value: The value of statistic which separates critical region and acceptance region.

\* Degree of freedom: No. of independent variates i.e. no. of independent choices that make up statistic is called df.  
If sample size =  $n$  and restriction = 1 Then,  $df = n - 1$

\* One-tailed test: A test of statistical hypothesis in which the alternative hypothesis looks for a definite increase or decrease in parameters.  $H_1: \mu > \mu_0$  or  $H_1: \mu < \mu_0$

\* Two-tailed test: A test of statistical hypothesis in which the alternative hypothesis looks for a definite change.  $\mu \neq \mu_0$

\* Steps used in hypothesis testing:

Step-1: State null hypothesis and alternate hypothesis

Step-2: Select the appropriate test statistic (Z-test or t-test)  
& calculated required

Step-3: State the level of significance. Value for test

Step-4: State degree of freedom

Step-5: Find the critical value (tabulated value) from corresponding tables

Step-6: Make decision by comparing the calculated value and tabulated value

Step-7: Make conclusion (Either accept or reject null hypothesis)

o Non-parametric test:	Used when samples are drawn from non normal population. It is used when certain assumptions about population are in weak state. These are mostly based upon ranked data	
Normal dist. test	corresponding non parametric test	Purpose of the test
→ T-test for independent samples	→ Mann Whitney U	compare two independent samples
→ Paired-T-test	→ Wilcoxon matched pair signed rank	→ compare dependant samples
→ One way ANOVA	→ Kruskal Wallis	→ compare three or more group
→ Two way ANOVA	→ Friedman test	→ compare two groups classified by two factors.

### o Advantages of non-parametric test:

- Less time consuming, simple to understand
- Requires no assumption about the population from which sample is selected, doesn't require complicated sampling theory.

### o Disadvantages:

- cannot be used to estimate population parameters
- Less reliable & less powerful than parametric test.
- Lots of information are discarded or unused & lots of tables are needed for tests.

### o Uses :

- When data size is small, weak scaled & show highly skewed nature.

### o Measurement scale:

Used to quantify the variables.

\* Nominal scale: Assigning number or symbol to events such that numbers have no numerical meaning. ex: Jersey no.

\* Ordinal scale: Numbers assigned have ranking to data. ex: Grading

\* Interval scale: Interval betn any two objects are exactly known. Ratio of any two intervals is independent of unit of measurements & zero point. ex: temperature recorded

\* Ratio scale: Includes a true zero point as its origin. ex: heights measured

### o Assumption of Non-parametric tests:

i) sample observations are independent

ii) Variables under study is continuous

iii) Sample p.d.f is continuous

iv) Lower order moments exists. (mean & variance)

## Tests

- \* One Sample test
  - 1) Run test: Non-parametric test used to determine the randomness of the selected sample
  - 2) Binomial Test: Non-parametric test used for data present in either nominal or ordinal scale. It is used to test whether the binomial population has two distinct groups of two equal numbers of outcomes or not. we use p-test.
  - 3) Kolmogorov Smirnov test:  
Non parametric test used to test the goodness of fit. It is alternative to chisquare test for goodness of fit when size is small.
- \* Two Independent Sample test
  - 1) Median Test:  
Non-parametric test used to test the significant difference between two independent distributions.
  - 2) Mann-Whitney U test  
Non-parametric test used to determine whether two independent samples have been drawn from the population with same distribution.
  - 3) Kolmogorov Smirnov Two sample test  
Non-parametric test used to determine whether two independent samples have drawn from the same population.
  - 4) Chisquare for goodness of fit  
used to test the significance difference between observed frequencies ( $O_i$ ) and Expected frequencies ( $E_i$ )
  - 5) Chisquare for independence of attributes.  
characteristics which are capable of being measured qualitatively but not quantitatively are attributes. This test is used to find any association between the attributes.
- \* Paired-Sample Tests.
  - 1) Wilcoxon Matched pair signed rank test  
Used to compare two populations for which observations are paired. Based on magnitude & direction of difference b/w observations

## 2) Kruskal Wallis H test

One-way ANOVA test (single factor analysis) used to test the significant difference between average of three or more independent populations.

## 3) Friedman F test

Two-way ANOVA test (double factor analysis) used to test the significant difference of two or more independent population.

## 4) Cochran Q test

Used for more than two related samples (dependent samples). It is used to test significant difference in frequencies or proportion of three or more related samples.

### o Partial correlation:

Relationship between any two variables keeping all the other remaining variables involved constant.

$$\text{Ex: } r_{12.3} = \frac{r_{12} - r_{13} \cdot r_{23}}{\sqrt{1 - r_{13}^2} \cdot \sqrt{1 - r_{23}^2}} \quad \text{where } r_{12} = \frac{n \sum x_1 x_2 - \bar{x}_1 \cdot \bar{x}_2}{\sqrt{n \sum x_1^2 - (\bar{x}_1)^2} \cdot \sqrt{n \sum x_2^2 - (\bar{x}_2)^2}}$$

### o Coefficient of partial determination ( $r_{x-y-z}$ )<sup>2</sup>

Square of partial correlation coefficient. Used to measure variation in one variable as explained by other variable keeping the next variable constant.

$$\text{Ex: } r_{12.3} = 0.8 \text{ then } r_{12.3}^2 = 0.64 = 64\%$$

It means 64% of total variation on  $x_1$  is explained by  $x_2$  when  $x_3$  is constant.

### o Multiple correlation:

Relationship among three or more variables simultaneously. Here, relationship of a variable with two or more variables is studied at the same time. Ex:  $R_{1.2.3} = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2 \cdot r_{12} \cdot r_{13} \cdot r_{23}}{1 - r_{23}^2}}$

### o Coefficient of multiple determination ( $R_{x-y-z}$ )<sup>2</sup>

Square of multiple correlation coefficient. Used to measure variation in one variable as explained by two remaining variables.

$$\text{Ex: } R_{1.2.3} = 0.7 \text{ then } R_{1.2.3}^2 = 0.49 = 49\%$$

It means 49% of total variation on  $x_1$  is explained by  $x_2$  and  $x_3$ .

## o Multiple Linear regression:

Linear functional relationship of one dependant variable(response) with two or more independent(explanatory) variables.

$$\text{Ex: } Y = b_0 + b_1 x_1 + b_2 x_2 + \epsilon$$

where  $y$  = dependant variable

$x_1, x_2$ : independent variables

$\epsilon$  = error.

$b_0$  = average value of  $y$  when  $x_1$  and  $x_2$  are zero

$b_1$  = regression coefficient of  $y$  on  $x_1$  keeping  $x_2$  constant.

Measures the change in  $y$  per unit change in  $x_1$  keeping  $x_2$  constant

$b_2$  = regression coefficient of  $y$  on  $x_2$ . Measures change in  $y$  per unit change in  $x_2$  keeping  $x_1$  constant

$\epsilon$  = random error.

## o Measures of variation:-

Total variation is divided into explained variation (SSR) and unexplained variation (SSE).

$$\text{i.e. } TSS = SSR + SSE$$

$$TSS = \sum (Y - \bar{Y})^2 = \sum y^2 - n \bar{Y}^2$$

$$SSE = \sum (Y - \hat{Y})^2 = \sum y^2 - b_0 \sum y - b_1 \sum x_1 - b_2 \sum x_2$$

ANOVA TABLE:

SV	DF	SS	MSS	F Ratio	F crit	Where
Regression	$k$	SSR	$MSP = \frac{SSR}{k}$	$F = \frac{MSP}{MSE}$		$k = \text{no. of independent vars}$
Error/residual	$n-k-1$	SSE	$MSE = \frac{SSE}{n-k-1}$			$n = \text{no. of obs}$
Total	$n-1$	TSS				

o Standard Error of estimate: It is the square root of variance computed from the sample data. It measures the average variation or scatterness of observed data point around regression line. It is used to measure reliability of regression eqn. Regression line having more SSE is less reliable.

$$Se = \sqrt{\frac{SSE}{n-k-1}} \quad \text{If } se=0 \text{ (no variance)} \\ Se \neq 0 \text{ more reliable.}$$

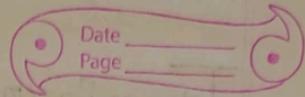
o Coefficient of determination ( $R^2$ ): Measures the proportion of variation in dependent variable explained by set of indep. variables.

$$R^2 = \frac{SSR}{TSS}, 0 \leq R^2 \leq 1. \text{ Higher value of } R^2, \text{ more reliable the fitted eqn is.}$$

o Adjusted  $R^2$ : Used for comparing two or more regression models that predict same dependant var. with different no. of indep. var.

- Test of Significance of Regression coefficient:  
Used to test significance of individual regression coefficients.  
i.e. It is used to test whether there is a significant linear relationship between dependent and independent variable.  
 $t$  test is used.
- Test of significance of overall regression coefficient.  
Used to test whether there is a significant linear relationship between dependant variable and set of independent variables.
- Stochastic process:  
A family of random variable  $\{x(t) | t \in T\}$  defined on a given probability space, indexed by the parameter  $t$ , where  $t$  varies over an index set  $T$ .
- The values assumed by random variable  $x(t)$  are called states and the set of all possible values forms the state space of the process. State space is denoted by  $I$ .
- If state space is discrete, then it is called discrete state process also known as chain.
- If state space is continuous, " " " continuous state process.
- If index set is discrete, then it is called discrete parameter process
- If index set is continuous, " " " continuous parameter process
- Markov process:  
Stochastic process such that the probability distribution of its future development depends only on present state and not how process arrived in that state (i.e. past states).
- Markov chain: It is a stochastic model, in which the probability of each event depends only on the state attained in previous event.
- Transitional probability: The probability of moving from one state to another state or remain in the same state in a single period of time. The probability of moving from one state to another depends upon the probability of preceding states, so it is also called conditional probability.
- Transitional Prob. Matrix: Matrix obtained by using one step state probabilities of various steps.

## Unit-5



### Experimental design (design of experiments)

Design of experiment is the branch of Statistics that deals with the design and analysis of experiments. In any experiment, the main aim is to increase the precision and minimize the error. Hence validity of experiment depends on how it is designed.

Design of experiment is the design of any task that helps to determine the relationship between the causes and effects of various experimental treatments.

### Objectives of experimental design

- i) To study the effects of various treatments.
- ii) To study the mathematical relationship between various treatments and their effects.
- iii) To estimate error and control error.
- iv) Proper interpretation of results.

### Steps in design of experiments:

- i) The problem which has to be solved by experiment should be stated in a clear & significant way.
- ii) The hypothesis in accordance with the experiment should be formulated in a clear way.
- iii) The experimental technique should be divided in a systematic way.
- iv) Result that come from design should be examined carefully.
- v) The conclusion is drawn after experiment carefully.

## Terminology in experimental design.

- \* Experiment : It is means of getting an answer to a question that the experimenter has in his mind. It is basically the systematic procedure carried out under controlled conditions in order to discover an unknown effect, to test or establish a hypothesis or to illustrate a known effect.
  - Absolute exp: determining absolute values
  - Some characteristics: like finding correlation coeff.
  - Comparative exp: comparing different types of fertilizers, cultivation methods.
- \* Treatment : It is the factor that affects the output of any experiment. These are the inputs whose outcomes are estimated & compared. Ex:- Corn field is divided into 4, each part is 'treated' with different fertilizers to see which produces more corn.
- \* Experimental unit : The smallest division of the experimental material, in which treatments are applied & effects of treatment are measured.  
Ex: among patients admitted in a hospital, a patient is E.U.
- \* Yields (effects) (response variables)  
The outcomes of the experiment are yields.  
In agriculture, experimental production of crops on using different fertilizers are yields.

\* Blocks: The experimental field is divided into relatively homogeneous subgroups or strata which is homogeneous or uniform among themselves than the field, as a whole called as block.

### \* Experimental error

In experimental design, the total variation is divided into two: ~~explained~~ variation & unexplained variation i.e. residual/error. The unexplained random part of variation is termed as experimental error.

It is a technical term and doesn't mean mistake. Occurs due to: Inherent variability in experimental unit, Small errors associated with measurements, Lack of representative of sample.

→ It can be estimated by replication and can be controlled by local control.

### \* Precision

It is the amount of information or sensitivity of an experiment. Given by reciprocal of variance of mean. If an experiment is replicated r times, precision of experiment:  $\text{Precision} = \frac{1}{\sigma^2}$ ,  $\sigma^2$  is variance.

The precision increases as the replicates increase or variance decreases.

## \* Efficiency of design

Consider any two designs  $D_1 \times D_2$  with replication,  $V_1$  and  $V_2$ . & variance  $\sigma_1^2, \sigma_2^2$ , then ratio of precision of design  $D_1 \times D_2$  is called the relative efficiency of design  $D_1$  w.r.t.  $D_2$ .

Given by,  $E = \frac{V_1}{\sigma_1^2} / \frac{V_2}{\sigma_2^2}$  if  $E=1$ , both  $D_1 \times D_2$  are equal  
 $E < 1$ ,  $D_1$  is more efficient  
 $E > 1$ ,  $D_2$  is less efficient

## → Basic Principles of Experimental Design

According to R.A. Fisher, a good experimental design must possess following three principles

- i) Replication: It is the repetition of treatments under investigation. A treatment is repeated a large number of times in order to obtain a reliable result that is possible from single observation.  
→ It works along with randomization to provide an estimate of treatment effect.
- Along with local control, to minimize the error.
- Most effective way to increase precision is to increase k.

- ii) Randomization: It is the process of allocating treatments to various plots (exp. units) in a random manner to ensure that each treatment will have equal chance of being assigned to an experimental unit. It removes biasness & is essential for valid estimate of exp. error.

- iii) Local Control: It means the control of all factors except the ones about which we are investigating. It is another method to reduce or control the unexplained variation. In L.C., we divide fields into blocks & each block is divided into parts equal to no. of treatments.

## CRD Design (completely randomized design)

It is the simplest of all designs which is based upon two principles of design namely Replication and randomization. The total variation is divided into two components i.e. Treatment & Error. It is used in case of homogeneous experimental materials.

### Assumptions (same for all)

- i) All the observations are independent
- ii) All the observations are drawn from population having constant variance.
- iii) All the treatments should be homogeneous as far as possible.
- iv) Various treatment & environmental effects should be additive in nature.

### Advantages

- i) It is easy to layout
- ii) It is simple to statistical analysis due to one way classification
- iii) If some observations are missing, the analysis still remains simple.
- iv) It allows max. no. of d.f. for MSS

### Disadvantages

- i) Principle of local control is not used.
- ii) Suitable only for small treatments, replication & homogeneous experimental material only
- iii) In green house, laboratory

## Randomised Block design (RBD)

When the experimental material is not homogenous, then RBD is better than CRD.

It is based upon all the principles of a good experimental design, namely - - - -

The total variation is divided into three - - -

The treatment are allocated in random manner but it shouldn't be repeated in either row or column

### Advantages

- i) RBD provides better result than CRD
- ii) There is no restriction in no. of treatment & Replication, but at least two Replication is necessary
- iii) If some observations are missing, analysis still remains simple.

## Latin Square design (LSD)

When the experimental material is not homogeneous, then LSD is better than RBD. In RBD ~~totally~~, local control is used according to two way grouping ie rows & columns. Hence, it is used when two sources of errors are to be controlled simultaneously.

The no. of treatments, replications are equal. The treatments are associated in such a way that, each of the treatment occurs once & only once.

### Advantages

- i) Due to two way grouping in Local control, more variation than CRD & RBD.
- ii) The statistical analysis remains simple, if some observations are missing.

### Disadvantages

- i) The assumption that observations are independent is not always true.
- ii) It is suitable only for treatments 5 to 10.
- iii) It is not easy in field layout.

