# Statistics-II Summary - Ankit Pangeni

- **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.
- **Parameter:** A value (piece of data) that tells you something about whole population. It is more reliable than sample.
- **Statistic:** A piece of data that describes only the sample of a population.
- **Sampling:** Process of selecting a sample from a population for study and research.
- **Sampling distribution:** Probability distribution of the statistic of the repeated samples taken from a specific population.
- **Sampling distribution of mean:** Distribution of the sample means of the repeated samples taken from a specific population.

| without replacement (finite population) | with replacement (infinite) |
|---|---|
| mean $(E(\bar{x})) = \mu$ | mean $(E(\bar{x})) = \mu$ |
| Variance $V(\bar{x}) = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)$ | Variance $(V(\bar{x})) = \frac{\sigma^2}{n}$ |
| Standard error $(S.E) = \sqrt{V(\bar{x})}$ | Standard error $(S.E) = \sqrt{V(\bar{x})}$ |

- **Properties of sampling dist. of mean**
  - → It is unbiased estimate of population mean $(\mu)$
  - → Variance depends on sample ~~mea~~ size $(n)$ i.e. $V = \frac{\sigma^2}{n}$
- **Sampling distribution of proportion:**
  (proportion): No. of elements with a given characteristics divided by total no. of elements in that group

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

$$q = 1 - p$$

where, $P$ = popⁿ proportion of success.
$p$ = 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.

| Infinite popⁿ: | Finite popⁿ: |
|---|---|
| $E(p) = P$, $Var(p) = \frac{PQ}{n}$ | $E(p) = P$, $Var(p) = \frac{PQ}{n}\left(\frac{N-n}{N-1}\right)$ |

- 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 pop$^n$ parameter.
  * Interval estimation: Probable range is specified within which the value of parameter might be expected to lie.

- Properties of good estimator (criterias): SECU
  S → sufficiency  C → consistency
  E → Efficiency  U → unbiasedness.

- Confidence Interval (CI) estimation of population mean(μ)

  large sample ($n \geq 30$)

  $C.I = \bar{x} \pm z_\alpha \cdot \frac{\sigma}{\sqrt{n}}$

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

  small sample ($n < 30$)

  $C.I = \bar{x} \pm t_{\alpha, n-1} \frac{s}{\sqrt{n-1}}$ If s = given (biased)

  $CI = \bar{x} \pm t_{\alpha, n-1} \frac{s}{\sqrt{n}}$ If s = calculated (unbiased)

  $CI = \bar{x} \pm t_{\alpha, n-1} \frac{s}{\sqrt{n-1}} \cdot \sqrt{\frac{N-n}{N}}$ If N = given.

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

  Large sample ($n \geq 30$)

  $CI = p \pm z_\alpha \cdot \sqrt{\frac{PQ}{n}}$

  $CI = p \pm z_\alpha \cdot \sqrt{\frac{PQ}{n}} \cdot \sqrt{\frac{N-n}{N-1}}$ If N is given.

  $CI = p \pm z_\alpha \cdot \sqrt{\frac{PQ}{n}}$

  small sample ($n < 30$)

  $CI = p \pm t_{\alpha, n-1} \sqrt{\frac{PQ}{n}}$

  $CI = p \pm t_{\alpha, n-1} \sqrt{\frac{PQ}{n}}$

  note $z_\alpha$ = critical value of z

  $CI = p \pm t_{\alpha, n-1} \sqrt{\frac{PQ}{n-1}} \cdot \sqrt{\frac{N-n}{N}}$

- Sample size $n = \frac{z_\alpha^2 \sigma^2}{e^2}$ (mean) | $n = \frac{z_\alpha^2 PQ}{e^2}$ (proportion)

Note: (Confidence level increases, sample size increases
* Samping error decreases, Sample size decreases
* Sample size increases, width of C.I decreases

- 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 parameter.
- 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 no 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 type
  * 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. ie. prbability of making correct decision.

- Test-Statistic: Statistic based upon approximate probability distribution. Used to decide whether to accept or reject null hypothesis.
  * Z-test: Used for large samples $(n \geq 30)$ in case of test of significance of mean, difference between means, test of proportion & difference between proportions
  * t-test: used for small samples $(n < 30)$ 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 1·0·5. 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 acception of null hypothesis. The divison is based on 1·0·5 & 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 ie. no. of independent choices that make up statistics 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 parameter. , $H_1: \mu > \mu_0$ or $H: \mu < \mu_0$

* 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 statistics (z-test or t-test) & calculated required value for test
Step-3: State the level of significance.
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)

- 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. |

- 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.

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

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

- 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 interval is independent of unit of measurements & zero point. Ex: temperature recorded
  - * Ratio scale: Includes a true zero point as its origin. Ex: heights measured

- Assumption of Non-parametric test:
  - 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 number 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 use 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 betn 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 independend population

4) Cochran Q test
used for more than two related samples (dependant samples) It is used to test significant difference in frequencies or proportions of three or more related samples.

- Partial correlation:
Relationship between any two variables keeping all the other remaining variables involved constant.
Ex: $r_{12.3} = \frac{r_{12} - r_{13} \cdot r_{23}}{\sqrt{1 - r_{13}^2} \cdot \sqrt{1 - r_{23}^2}}$ where $r_{12} = \frac{n\Sigma x_1 x_2 - \Sigma x_1 \cdot \Sigma x_2}{\sqrt{n\Sigma x_1^2 - (\Sigma x_1)^2} \cdot \sqrt{n\Sigma x_2^2 - (\Sigma x_2)^2}}$

- Coefficient of partial determination $(r_{xy.z})^2$
Square of partial correlation coefficient. Used to measure variation in one variable as explained by other variable keeping the next variable constant.
Ex: $r_{12.3} = 0.8$ 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.

- 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: $R_{1.23} = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2 \cdot r_{12} \cdot r_{13} \cdot r_{23}}{1 - r_{23}^2}}$

- Coefficient of multiple determination $(R_{x.yz})^2$
Square of multiple correlation coefficient. Used to measure variation in one variable as explained by two remaining variables
Ex: $R_{1.23} = 0.7$ then $R_{1.23}^2 = 0.49 = 49\%$
It means 49% of total variation on $x_1$ is explained by $x_2$ and $x_3$.

- Multiple Linear regression:
  Linear functional relationship of one dependant variable (response) with two or more independent (explanatory) variables.
  Ex: $Y = b_0 + b_1x_1 + b_2x_2 + \varepsilon$
  where, $Y$ = dependant variable
  $x_1, x_2$ = independent variables
  $\varepsilon$ = 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
  $\varepsilon$ = random Error.

- Measures of variation:
  Total variation is divided into explained variation (SSR) and unexplained variation (SSE).
  ie: $TSS = SSR + SSE$
  $TSS = \Sigma(Y-\bar{Y})^2 = \Sigma Y^2 - n\bar{Y}^2$
  $SSE = \Sigma(Y-\hat{Y})^2 = \Sigma Y^2 - b_0\Sigma Y - b_1\Sigma YX_1 - b_2\Sigma YX_2$

ANOVA TABLE:

| SV | DF | SS | MSS | F cal | F<1 |
|---|---|---|---|---|---|
| Regression | k | SSR | $MSR = \frac{SSR}{k}$ | $F = \frac{MSR}{MSE}$ | |
| Error/residual | n-k-1 | SSE | $MSE = \frac{SSE}{n-k-1}$ | | |
| Total | n-1 | TSS | | | |

Then,
k = no. of independent var
n = no. of obs

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

- Coefficient of determination ($R^2$): Measures the proportion of variation in dependent variable explained by set of indep. variables.
  $R^2 = \frac{SSR}{TSS}$ $0 \le R^2 \le 1$. Higher value of $R^2$, more reliable the fitted eqn is.

- Adjusted $R^2$: Used for compairing two or more regression model 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.
  ie. It is used to test whether there is a significant linear relationship between dependent and independant variable.
  t test is used.
- Test of significance of overall regression coefficient
  Used to test whether there is a significant linear ~~line~~ relationship between dependant variable and set of independant 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 value 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 (ie. past states)
- Markov chain: It is a stochastic model, which is a sequence of events 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 preceeding states, so it is also called conditional probability
- Transitional Prob. Matrix: Matrix obtained by using one step state probabilities of various states.

# Unit-5

## Experimental design (design of experiment)

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

## Terminology in experimental design.

* **Experiment** : It is means of getting an answer to a question that the experimenter has in its 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 of some characteristics like finding correlation coef.
  - → 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 ~~unit~~ divison of the experimental material, in which treatments are applied & effects of treatments are measured. Ex: among patients admitted in a hospital, a patient is E.U.

* **Yields (effects)** (response variable)
  The outcomes of the experiments are yields. In agriculture, experiment: 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 ie. 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: Inherient 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 contr.

* 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: $\frac{r}{\sigma^2}$, $\sigma^2$ is variance

The precision increases as the replication increases or variance decreases.

## * Efficiency of design

Consider any two designs $D_1$ & $D_2$ with replications $r_1$ and $r_2$ & variance $\sigma_1^2$ & $\sigma_2^2$, then ratio of precision of design $D_1$ & $D_2$ is called the relative efficiency of design $D_1$ w.r. to $D_2$.

Given by, $E = \frac{V_1}{\sigma_1^2} \Big/ \frac{V_2}{\sigma_2^2}$

if $E=1$, both $D_1$ & $D_2$ are equ
$E<1$, $D_1$ is more effic
$E>1$, $D_1$ is les effic

## → Basic Principles of Experimental Desn

According to R.A. Fisher, a good experimental design must posses following three princip

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 observat
→ It works along with randomizatn to provide an estimate of treatment effec
→ Along with local contro, to minimize the err
→ Most effective way to increase precision is to increa

ii) **Randomization:-** It is the process of allocating treatm to various plots (exp. units) in a random manner to ensure that each treatment will have equal chanc of being assigned to an experimental unit. It remove biasedness & is essential for valid estimate of exp. erro

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

Then treatm ents are randomly assig ned to these parts of bl

Himal
Date
Page

# CRD Design (completely randomized design)

It is the simplest of all designs which is based upon two principles of design ~~mainly~~ namely replication and randomization. The total variation is divided into two component ie. 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 | Disadvantages |
|---|---|
| i) It is easy to layout | i) Principle of local control is not used. |
| ii) It is simple to statistical analysis due to one way classification | ii) Suitable only for small treatments, replications & homogeneous exp material only |
| iii) If some observations are missing, the analysis still remains simple. | Uses |
| iv) It allows max. no. of d.f for MSS | i) 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. Name ---
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 ~~total~~, local control is used according to two way grouping ie rows x columns. Hence, it is used when two sources of errors are to be controlled simultaneously.

The no. of treat, repli. are equal. ~~Then~~ Treatment is allocated in such a way that, each of the treatment occurs once & only once.

## Advantages

i) Due to two way group in Local control, more variation than CRD & RBD.
ii) The statistical analysis remains sample, if some observations are miss.

## Disadvantage

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