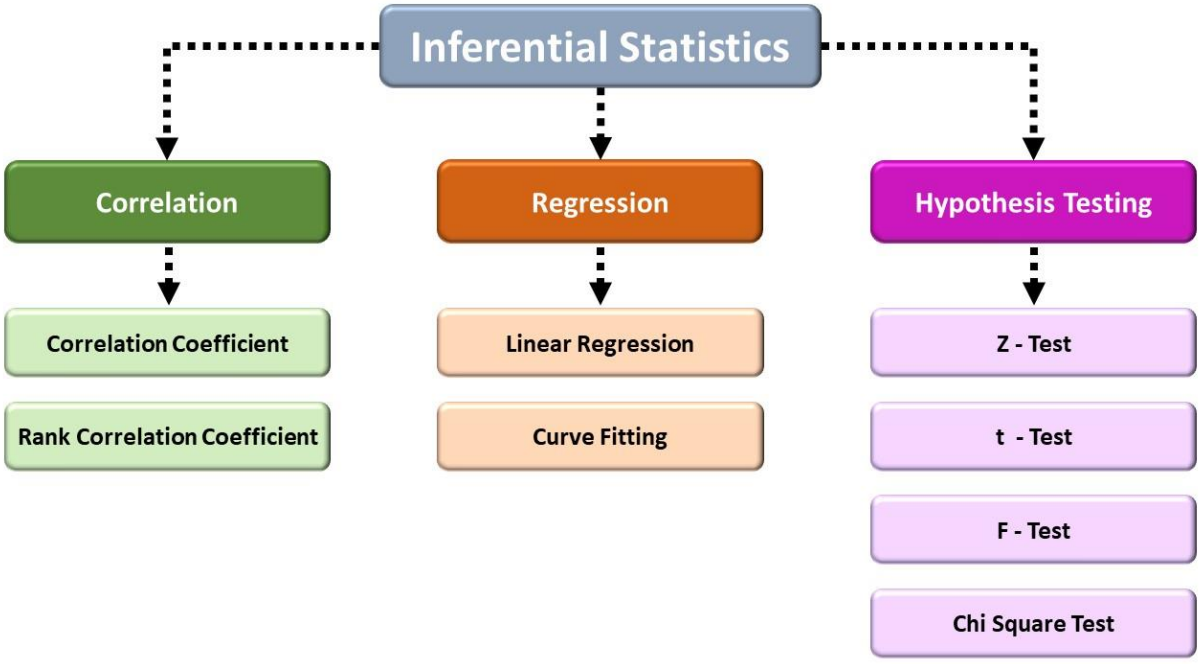
Index

Unit – 4 ⇝ Inferential Statistics – I

Introduction

→ In quantitative research, after collecting data, the first step of statistical analysis is to describe characteristics of the responses, such as the average of one variable (e.g., age), or the relation between two variables (e.g., age and creativity).

→ After summarizing the data, we use **Inferential Statistics**, which help us to decide whether your data confirms or refutes your statement/hypothesis or it can be generalizable to a larger population or not.



Unit – 4.1 ⇝ Correlation and Regression

Introduction

→ Correlation and regression are the most commonly used techniques for investigating the relationship between two quantitative variables.

→ Correlation refers to the relationship of two or more variables. Regression establishes a functional relationship between the variables.

→ The coefficient of correlation is a relative measure whereas the regression coefficient is an absolute figure.

Correlation

→ Two variables are known as **correlated** if a change in one variable affects a change in the other variable. Such a data connecting two variables is called bivariate data.

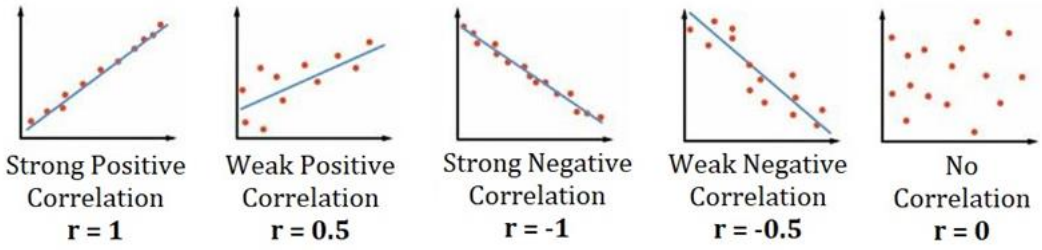
→ For Example:

- Relationship between heights and weights.
- Relationship between price and demand of commodity.
- Relationship between age of husband and age of wife.

→ When two variables are correlated with each other, it is important to know the amount or extent of correlation between them.

→ The numerical measure of correlation of degree of relationship existing between two variables is known as the coefficient of correlation.

→ It is denoted by **r** and it is always lying between -1 and 1.



→ The value of r is ± 0.9 or ± 0.8 etc. shows high degree of relationship between the variables while ± 0.2 or ± 0.1 etc. shows low degree of correlation.

Types of Correlation

→ Correlation is classified into four types.

- Positive Correlation
- Negative Correlation
- Linear Correlation
- Nonlinear Correlation

Positive Correlation

→ If both the variables vary in same direction, then such correlation is known as **positive correlation**.

→ In other words, if the value of one variable increases, the value of other variables also increases, or, if the value of one variable decreases, the value of other variables also decreases.

→ For Example:

- The correlation between heights and weights of group of persons is a positive correlation.

Height(cm)	150	152	155	160	162	165
Weight(kg)	60	62	64	65	67	69

Negative Correlation

→ If both the variables vary in opposite direction, then such correlation is known as **negative correlation**.

→ In other words, if the value of one variable increases, the value of other variables decreases, or, if the value of one variable decreases, the value of other variables increases.

→ For Example:

- The correlation between the price and demand of a commodity is a negative correlation.

Price (₹ per unit)	10	8	6	5	4	1
Demand(units)	100	200	300	400	500	600

Linear Correlation

→ If the ratio of change between two variables is constant, then such correlation is known as **linear correlation**.

→ If such variables are plotted on a graph paper, a straight line is obtained.

→ For Example:

Milk (l)	5	10	15	20	25	30
Curd (kg)	2	4	6	8	10	12

Nonlinear Correlation

→ If the ratio of change between two variables is not constant, then such correlation known as **nonlinear correlation**.

→ If such variables are plotted on a graph paper, a curve is obtained.

→ For Example:

Advertising expenses (₹ in lacs)	3	6	9	12	15
Curd (kg)	10	12	15	15	16

Methods of Studying Correlation

→ There are two different methods of studying correlation:

- Graphical Methods
  - (1) Scatter Diagram
  - (2) Simple Graph
- Mathematical Methods
  - (1) Karl Pearson's coefficient of correlation
  - (2) Spearman's rank coefficient of correlation

Method – 1 ⇝ Correlation Coefficient

Karl Pearson's Coefficient of Correlation

→ The coefficient of correlation is the measure of correlation between two random variables X and Y, and is denoted by r. It is defined as below:

$$\mathbf{r = \frac{cov(X, Y)}{\sigma_x \, \sigma_y}} \qquad \text{... ... (1)}$$

→ Where,

cov(X, Y) is the covariance of variables X and Y.

σ_x & σ_y are standard deviation of X and Y respectively.

→ This expression is known as Karl Pearson's coefficient of correlation.

→ We have,

$$cov(X, Y) = \frac{1}{n} \cdot \sum (x - \bar{x})(y - \bar{y})$$

$$\sigma_x = \sqrt{\frac{\sum (x - \bar{x})^2}{n}}$$

$$\sigma_y = \sqrt{\frac{\sum (y - \bar{y})^2}{n}}$$

→ Substitute the values of cov(X, Y), σ_x and σ_y in equation (1), We get

$$\mathbf{r = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{\sum (x - \bar{x})^2 \cdot \sum (y - \bar{y})^2}}} \qquad \text{... ... (2)}$$

→ Equation (2) can be further reduced to below equation.

$$\mathbf{r = \frac{n \cdot \sum xy - \sum x \sum y}{\sqrt{n \cdot \sum x^2 - (\sum x)^2} \sqrt{n \cdot \sum y^2 - (\sum y)^2}}} \qquad \text{... ... (3)}$$

Example of Method-1: Correlation Coefficient

C 1 Compute the coefficient of correlation between x and y using the following data:

x	2	4	5	6	8	11
y	18	12	10	8	7	5

Answer: $r = -0.9203$

C 2 Calculate Karl Pearson's correlation coefficient between age and playing habits:

Age	20	21	22	23	24	25
No. of students	500	400	300	240	200	160
Regular players	400	300	180	96	60	24

Answer: $r = -0.9823$

C 3 Determine the coefficient of correlation if $n = 10, \bar{x} = 5.5, \bar{y} = 4,$

$\sum x^2 = 385, \sum y^2 = 192, \quad \sum (x+y)^2 = 947.$

Answer: $r = -0.6812$

C 4 Given that $n = 25, \sum x = 125, \sum x^2 = 650, \sum y = 100, \sum y^2 = 460$ and $\sum xy = 508$. Later on, it was found that two of the points $(8, 12)$ **and** $(6, 8)$ **were wrongly entered as** $(6, 14)$ **and** $(8, 6)$. **Prove that r** $= \frac{2}{3}$.

Method – 2 ⇝ Rank Correlation Coefficient

Rank Correlation

→ Let a group of n individuals be arranged in order of merit with respect to some characteristics. The same group would give a different order(rank) for different characteristics.

→ Considering the orders corresponding to two characteristics A and B, the correlation between these n pairs of ranks is known as **rank correlation** in the characteristics A and B for that group of individuals.

→ It is denoted by **ρ**.

Spearman's Rank Correlation Coefficient

→ Edward Spearman's formula for rank correlation coefficient (ρ) is given by

$$\rho = 1 - \frac{6\sum d^2}{n(n^2-1)}$$

Where, d = Difference of Ranks

Tied Rank

→ If there is a tie between the ranks, then it is known as tied rank.

→ Formula for rank correlation coefficient (ρ) is,

$$\rho = 1 - \frac{6\left[\sum d^2 + \frac{1}{12}(m_1{}^3 - m_1) + \frac{1}{12}(m_2{}^3 - m_2) + \cdots\right]}{n\,(n^2-1)}$$

Where, d = Difference of Ranks

m_i = number of times a data repeats = 2, 3, …

→ Note:

- If any data x_i is repeated 2 times, then $m_1 = 2$.
- If any data x_j is repeated 3 times, then $m_2 = 3$.

→ In case of tie between individuals' ranks, **the rank is divided among equal individuals**.

→ For Example:

- If there is tie with two items at 4^{th} rank, then give average rank 4.5 as rank to both items.

$$\text{Average} = \frac{4+5}{2} = 4.5$$

Example of Method-2: Rank Correlation Coefficient

C 1

In a college, IT department has arranged one competition for IT students to develop an efficient program to solve a problem. Ten students took part in the competition and ranked by two judges given in the following table. Find the degree of agreement between the two judges using rank correlation coefficient.

1st judge	3	5	8	4	7	10	2	1	6	9
2nd judge	6	4	9	8	1	2	3	10	5	7

Answer: $\rho = -0.2970$

C 2

The competitions in a beauty contest are ranked by three judges:

1st judge	1	5	4	8	9	6	10	7	3	2
2nd judge	4	8	7	6	5	9	10	3	2	1
3rd judge	6	7	8	1	5	10	9	2	3	4

Use rank correlation to discuss which pair of judges has nearest approach to beauty.

Answer: 2nd and 3rd judges has nearest approach

$[\rho_{12} = 0.5515, \rho_{23} = 0.7333, \rho_{13} = 0.0545]$

C 3

Find the rank correlation coefficient and comment on its value:

Roll no.	1	2	3	4	5	6	7	8	9
Marks in Math.	78	36	98	25	75	82	90	62	65
Marks in Chem.	84	51	91	60	68	62	86	58	53

Answer: $\rho = 0.8333$

C 4

Calculate coefficient of correlation by Spearman's method from following.

Sales	45	56	39	54	45	40	56	60	30	36
Cost	40	36	30	44	36	32	45	42	20	36

Answer: $\rho = 0.7636$

C	5	The coefficient of rank correlation of marks obtained by 10 students in English and Economics was found to be 0.6. It was later discovered that the difference in ranks in the two subjects obtained by one of the students was wrongly taken as 7 instead of 1. Find the correct coefficient of rank correlation. **Answer: 0.8909**

Regression

→ Regression is defined as a method of estimating the value of one variable when the other is known and both are correlated.

→ We use the general form regression line for these algebraic expressions. The algebraic expressions of the regression lines are known as **Regression Equations**.

→ It is highly used in statistical estimation of demand curve, supply curve, production function, cost function, consumption function etc.

Types of Regression

→ Regression is classified into four types:

- Simple Regression
- Multiple Regression
- Linear Regression
- Nonlinear Regression

Simple Regression

→ The regression analysis for studying only two variables at a time is known as **simple regression**.

Multiple Regression

→ The regression analysis for studying more than two variables at a time is known as **multiple regression**.

Linear Regression

→ If the regression curve is a straight line, the regression is known as **linear regression**.

Nonlinear Regression

→ If the regression curve is not a straight line, the regression is known as **nonlinear regression**.

Method of Studying Regression

→ There are two methods of studying regression:

- Method of scatter diagram
- Method of least square

→ We will use method of least square only to find out regression.

Method – 3 ⇝ Linear Regression

Line of Regression (Linear Regression)

→ If the variables, which are highly correlated, are plotted on a graph then the points are around a straight line, the line is known as the **line of regression**.

→ There are two types of line of regression.

- Line of regression of y on x
- Line of regression of x on y

Line of Regression of y on x

→ It is the line which gives the **best estimate for the values of y** for given values of x.

→ The regression equation of y on x is given by

$$\mathbf{y - \bar{y} = b_{yx}\,(x - \bar{x})}$$

$$\text{Where, } b_{yx} = \text{Regression Coefficient} = r\,\frac{\sigma_y}{\sigma_x} = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2}$$

$$= \frac{\sum(x - \bar{x})(y - \bar{y})}{\sum(x - \bar{x})^2}$$

r = Correlation coefficient between x and y

$\bar{x}, \sigma_x$ = Mean & Standard Deviation of all x_i

$\bar{y}, \sigma_y$ = Mean & Standard Deviation of all y_i

Line of Regression of x on y

→ It is the line which gives the **best estimate for the values of x** for given values of y.

→ The regression equation of x on y is given by

$$\mathbf{x - \bar{x} = b_{xy}\,(y - \bar{y})}$$

$$\text{Where, } b_{xy} = \text{Regression Coefficient} = r\,\frac{\sigma_x}{\sigma_y} = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum y^2 - (\sum y)^2}$$

$$= \frac{\sum(x - \bar{x})(y - \bar{y})}{\sum(y - \bar{y})^2}$$

r = Correlation coefficient between x and y

$\bar{x}, \sigma_x$ = Mean & Standard Deviation of all x_i

$\bar{y}, \sigma_y$ = Mean & Standard Deviation of all y_i

Properties of Regression Coefficients

→ The coefficient of correlation is the geometric mean of the coefficients of regression.

i.e., $\mathbf{r = \sqrt{b_{yx} \cdot b_{xy}}}$

→ The product of both b_{xy} and b_{yx} cannot be more than 1.

→ Both the regression coefficients will have the same sign. They are either both positive and both negative. It means,

If $r < 0$, then $b_{yx} < 0$ & $b_{xy} < 0$.

If $r > 0$, then $b_{yx} > 0$ & $b_{xy} > 0$.

→ The arithmetic mean of the regression coefficients is greater than or equal to the correlation coefficient.

i. e. $\frac{b_{xy} + b_{yx}}{2} \geq r$

Example of Method-3: Linear Regression

C 1 Obtain the two lines of regression for the following data:

Sales (No. of tablets)	190	240	250	300	310	335	300
Advertising expense (Rs.)	5	10	12	20	20	30	30

Answer: $y = 0.1766x - 30.4221$; $x = 4.7357y + 189.0807$

C 2 A study of amount of rainfall and quantity of air pollution removed is:

Daily rainfall (0.01 cm)	4.3	4.5	5.9	5.6	6.1	5.2	3.8	2.1	7.5
Particulate removed unit	126	121	116	118	114	118	132	141	108

a. Find the equation of the regression line to predict the particulate removed from the amount of daily rainfall.

b. Find the amount of particulate removed when daily rainfall is 4.8 units.

Answer: a. $y = -6.3240x + 153.1755$; b. 122.8203

C	3	The following data regarding the height(y) and weight(x) of 100 students are given: $\sum x = 15000, \sum y = 6800, \sum x^2 = 2272500, \sum y^2 = 463025,$ $\sum xy = 1022250$. Find the equation of regression line of height on weight. **Answer: $y = 0.1x + 53$**
C	4	The data for advertising and sale given below: Adv. Exp.(x) (Rs. lakh) / Sales(y) (Rs. lakh) Mean: 10 / 90 Standard deviation: 3 / 12 a. Correlation coefficient between prices is 0.8. b. Calculate the two regression lines. c. Find the likely sales when advertising expenditure is 15 lakhs. d. What should be the advertising expenditure if the company wants to attain a sales target of 120 lakhs? **Answer: b. $x = 0.2y - 8$; $y = 3.2x + 58$; c. 106 ; d. 16**
C	5	A study of prices of a certain commodity at Raipur and Kanpur yields the below data: Raipur (Rs) / Kanpur (Rs) Average price/kg: 2.463 / 2.797 Standard deviation: 0.326 / 0.207 Correlation coefficient between prices at Raipur and Kanpur is 0.774. Estimate the most likely price at Raipur corresponding to the price of 3.052 per kilo at Kanpur. **Answer: 2.774**

Method – 4 ⇝ Curve Fitting

Introduction

→ We come across many situations where we often require to find a relationship between two or more variables. For example, weight and height of a person, demand and supply, expenditure depends on income, etc.

→ This relation may be expresses by polynomial or exponential or logarithmic relationship. In order to determine such relationship, first it is requiring to collect the data showing corresponding values of the variables under consideration.

→ Suppose $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ be the data showing corresponding values of the variables x and y under consideration. If we plot the above data points on a coordinate system, then the set of points so plotted form a scatter diagram.

→ From this diagram, it is sometimes possible to visualize a smooth curve approximating the data. Such a curve is known as an **approximating curve**.

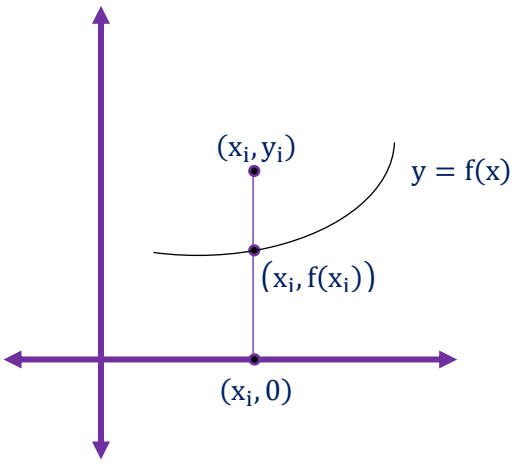
→ In particular, if the data approximate well to a straight line, we say that a linear relationship exists between the variables. It is quite possible that the relationship of the form $y = f(x)$ between two variables x and y, giving the approximating curve and which fit the given data of x and y, is known as **curve fitting**.

Least Square Method

→ Suppose that the data points are $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$, where x is independent and y is dependent variable.

→ Let the fitting curve f(x) has the following deviations (or errors or residuals) from each data points. i.e., $d_1 = y_1 - f(x_1)$.

→ These $d_i = y_i - f(x_I)$ are known as deviation, error or residual. Its value may be positive, negative or zero.



→ To give equal weightage to each error, we square each of these and form their sum.

$$\text{i.e., } D = {d_1}^2 + {d_2}^2 + \cdots + {d_n}^2 = \sum_{i=1}^{n} d_i^2$$

$$= \sum_{i=1}^{n} [y_i - f(x_i)]^2$$

4.1 Fitting of Linear Curves(Fitting a Stright Line)

→ Let, $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ be the set of n values and let the relation between x and y be $\mathbf{y = a + bx}$.

→ We have,

$$D = \sum_{i=1}^{n} [y_i - f(x_i)] = \sum_{i=1}^{n} (y_i - a - bx_i)^2$$

→ If $D = 0$, then all the n points will lie on $y = f(x)$.

→ If $D \neq 0$, $f(x)$ is chosen such that D is minimum.

→ This will be minimum at,

$$\frac{\partial D}{\partial a} = 0 \Longrightarrow -2 \sum_{i=1}^{n} (y_i - a - bx_i) = 0$$

$$\Longrightarrow \sum_{i=1}^{n} y_i - a \sum_{i=1}^{n} 1 - b \sum_{i=1}^{n} x_i = 0$$

$$\Longrightarrow \sum_{i=1}^{n} y_i - na - b \sum_{i=1}^{n} x_i = 0$$

$$\Longrightarrow \sum_{i=1}^{n} y_i = na + b \sum_{i=1}^{n} x_i$$

Similarly, by $\frac{\partial D}{\partial b} = 0 \Longrightarrow \sum_{i=1}^{n} x_i y_i = a \sum_{i=1}^{n} x_i + b \sum_{i=1}^{n} x_i^2$

→ We obtain following **Normal Equations** for the best fitting straight line $y = a + bx$.

$$\mathbf{\sum y = na + b \sum x}$$

$$\mathbf{\sum xy = a \sum x + b \sum x^2}$$

→ These equations can be solved simultaneously to give the best value of a and b such that straight line is the best fit to the data.

Example of Method-4.1: Fitting a Stright Line

<table>
<tr><td>C</td><td>1</td><td>Fit a straight line for the given pairs of (x, y) which are $(1, 5), (2, 7), (3, 9), (4, 10), (5, 11)$.

Answer: $y = 3.9 + 1.5x$</td></tr>
<tr><td>C</td><td>2</td><td>Fit a straight-line $y = ax + b$ to the following data:
<table><tr><td>x</td><td>−2</td><td>−1</td><td>0</td><td>1</td><td>2</td></tr><tr><td>y</td><td>1</td><td>2</td><td>3</td><td>3</td><td>4</td></tr></table>
Answer: $y = 0.7x + 2.6$</td></tr>
<tr><td>C</td><td>8</td><td>By method of least squares, fit a linear relation of the form $P = a + bW$ to the following data, P is the pull required to lift a weight W. Also estimate P, when W is 150.
<table><tr><td>P</td><td>50</td><td>70</td><td>100</td><td>120</td></tr><tr><td>W</td><td>12</td><td>15</td><td>21</td><td>25</td></tr></table>
Answer: $P = -11.8005 + 5.3041W$; $P(150) = 783.8145$</td></tr>
</table>

4.2 Fitting of Quadratic Curves(Fitting a Parabola)

→ Let, $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ be the set of n values and let the relation between x and y be $\mathbf{y = a + bx + cx^2}$.

→ We have,

$$D = \sum_{i=1}^{n} [y_i - f(x_i)] = \sum_{i=1}^{n} (y_i - a - bx_i - cx_i^2)^2$$

→ If $D = 0$, then all the n points will lie on $y = f(x)$. If $D \neq 0$, $f(x)$ is chosen such that D is minimum.

→ Differentiating S with respect to a, b, c and equating with zero (as done while fitting a linear curve). We obtain following **Normal Equations** for the best fitting $\mathbf{y = a + bx + cx^2}$ curve (parabola) of second degree.

$$\mathbf{\sum y = na + b\sum x + c\sum x^2}$$

$$\mathbf{\sum xy = a\sum x + b\sum x^2 + c\sum x^3}$$

$$\mathbf{\sum x^2 y = a\sum x^2 + b\sum x^3 + c\sum x^4}$$

Example of Method-4.2: Fitting a Parabola

C	1	Fit a polynomial of degree two using least square method for the following experimental data. Also, estimate y(2.4). x: 1, 2, 3, 4, 5 y: 5, 12, 26, 60, 97 **Answer: $y = 10.4 - 11.0857x + 5.7143x^2$; $y(2.4) = 16.7087$**
C	2	Fit a second – degree parabola $y = ax^2 + bx + c$ to the following data: x: −3, −2, −1, 0, 1, 2, 3 y: 12, 4, 1, 2, 7, 15, 30 **Answer: $y = 2.1190x^2 + 2.9286x + 1.6667$**

C	3	Fit a relation of the form $R = a + bV + cV^2$ to the following data, where V is the velocity in km/hr. and R is the resistance in km/quintal. Estimate R when $V = 90$.

V	20	40	60	80	100	120
R	5.5	9.1	14.9	22.8	33.3	46.0

Answer: $\mathbf{R = 4.35 + 0.0024V + 0.0029V^2}$; $\mathbf{R(90) = 28.0560}$

Unit – 4.2 Hypothesis Testing – I

Introduction

→ Many problems in engineering required to decide which of two competing claims for statements about parameter is true. Statements are known as **hypothesis**, and the decision-making procedure is known as **hypothesis testing**.

→ This is one of the most useful aspects of statistical inference, because many types of decision-making problems, tests or experiments in the engineering world can be formulated as hypothesis testing problems.

Population OR Universe

→ An aggregate of objects under study is known as **Population OR Universe**.

→ It is a collection of individuals or of their attributes (qualities) or of results of operations which can be numerically specified.

→ A universe containing a finite number of individuals or members is known as a **finite universe**.

- For Example: The universe of the weights of students in a particular class or the universe of smokes in Rothay district.

→ A universe with infinite number of members is known as an **infinite universe**.

- For Example: The universe of pressures at various points in the atmosphere.

→ In some cases, we may be even ignorant whether or not a particular universe is infinite, e.g., the universe of stars.

→ The collection of all possible ways in which a specified event can happen is known as a **hypothetical universe**.

- For Example: The universe of heads and tails obtained by tossing an infinite number of times is a hypothetical universe.

Sampling

→ A finite subset of a universe or population is known as a **sample**.

→ A sample is a small portion of the universe.

→ The number of individuals in a sample is known as **sample size**.

→ Sample size is denoted by **" n "** and population size is denoted by **" N "**.

→ If sample size **$n \geq 30$**, then it is known as **Large Sample,** otherwise it known as **Small Sample.**

→ The process of selecting a sample from a universe is known as **sampling**.

→ The theory of sampling is a study of relationship between a population and samples drawn from the population.

→ The fundamental object of sampling is to get as much information as possible of the whole universe by examining only a part of it.

→ Sampling is quite often used in our day-to-day practical life. For example, in a shop we assess the quality of sugar, rice or any commodity by taking only a handful of it from the bag and then decide whether to purchase it or not.

Parameter and Statistics

→ The statistical constants of **population** are known as **parameters**.

→ The statistical constants of **sample** are known as **statistics.**

→ Notations for statistical constants are as below:

Statistical Constants	For Sample	For Population
No. of elements (Size)	n	N
Proportion	p	P
Mean	$\bar{x}$	μ
Standard Deviation	s	σ
Variance	s^2	σ^2
Correlation Coefficient	r	ρ

Test of Significance

→ An important aspect of the sampling theory is to study the test of significance, which will enable us to decide, on the basis of the results of the sample.

→ For applying the tests of significance, we first set up a hypothesis which is a definite statement about the population parameter known as **null hypothesis,** denoted by $\mathbf{H_0}$.

→ Note that, Null hypothesis $\mathbf{H_0}$ is always in equality.

→ Any hypothesis which is complementary to the null hypothesis is known as an **alternative hypothesis** denoted by $\mathbf{H_1}$.

→ For example, if we want to test the null hypothesis that the population has a specified mean μ_0, then we have $H_0 : \mu = \mu_0$

→ Alternative hypothesis will be

- **Two tailed alternative hypothesis** : $H_1 : \mu \neq \mu_0$.
- **Right (One) tailed alternative hypothesis** : $H_1 : \mu > \mu_0$.
- **Left (One) tailed alternative hypothesis** : $H_1 : \mu < \mu_0$.

→ Hence alternative hypothesis helps to know whether the test is two tailed or one tailed test.

→ Note that, Alternate hypothesis **H_1** is either in the form of **less than** OR **greater than** OR **not equal**.

Errors In Sampling

→ The main aim of the sampling theory is to draw a valid conclusion about the population parameters. In doing this we may commit the following two type of errors.

- Type I error: When H_0 is true, we may reject it.

  $P(\text{Reject } H_0 \text{ when } H_0 \text{ is true}) = \alpha$

  Where, α is called the size of the type I error also referred to as product's risk.

- Type II error: When H_0 is wrong we may accept it.

  $P(\text{Accept } H_0 \text{ when } H_1 \text{ is true}) = \beta$

  where, β is called the size of the type II error, also referred to as consumer's risk.

Level of Significance

→ The maximum probability of making a type I error is known as **level of significance**.

→ It is denoted by **α**.

i.e., **P (Rejecting H_0 when H_0 is true) = α**.

→ The commonly used level of significance in practice are 5% (0.05) and 1% (0.01).

→ For 5% level of significance means that there is a probability of making 5 out of 100 type I error. Similarly, 1% level of significance.

→ If no level of significance is given, α is taken as $5\% = 0.05$.

Critical Region

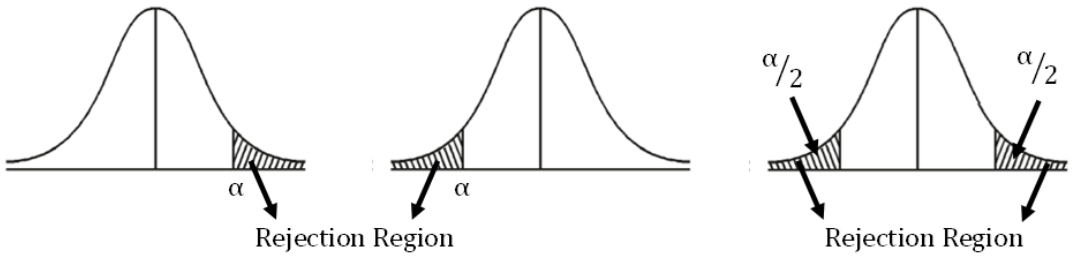
→ The region of the standard normal curve corresponding to predetermined level of significance α is known as **Critical Region**.

→ It is also known as **Rejection Region**.

→ The region under the normal curve which is not covered by the rejection region is known as **Acceptance Region**.

→ Thus, the statistic which leads to rejection of null hypothesis H_0 gives rejection region or critical region.

→ The value of the test statistic calculated to test the null hypothesis H_0 is known as **Critical Value**. Thus, the critical value separates the rejection region from the acceptance region.



Critical value(Z_α)	**Level of significance (α)**		
	1%	**5%**	**10%**
Two Tailed Test	$\lvert Z_\alpha \rvert = 2.58$	$\lvert Z_\alpha \rvert = 1.96$	$\lvert Z_\alpha \rvert = 1.645$
Right Tailed Test	$Z_\alpha = 2.33$	$Z_\alpha = 1.645$	$Z_\alpha = 1.28$
Left Tailed Test	$Z_\alpha = -2.33$	$Z_\alpha = -1.645$	$Z_\alpha = -1.28$

→ Note that, **Null Hypothesis H_0 is rejected when $|z| > 3$** without mentioning any level of significance.

Standard Error

→ The standard deviation of the sampling distribution of a statistic is known as the **Standard Error**.

→ It is denoted by **SE(t)** and read standard error of t.

→ It plays an important role in the theory of large samples and it forms a basis of testing of hypothesis.

Test Statistics

→ If t is any statistic, for large sample, then test statistics

$$\mathbf{Z = \frac{t - E(t)}{SE\ (t)} \sim N(0, 1)}$$

→ is normally distributed with mean 0 and variance 1.

Where, E(t) = Expected value of t

SE(t) = Standard Error of t

Confidence Limits OR Confidence Interval

→ The limits within which a hypothesis should lie with specified probability are known as **confidence limits**.

→ Generally, the confidence limits are set up with 5% or 1% level of significance (α).

→ If the sample value **lies between the confidence limit**, the hypothesis is **accepted**, otherwise **rejected**.

Steps for Testing of Statistical Hypothesis

→ **Step 1:** Set up null hypothesis H_0.

→ **Step 2:** Set up alternative hypothesis H_1.

→ **Step 3:** Set up level of significance α.

→ **Step 4:** Apply test statistic.

→ **Step 5:** Set up critical region. (Given in data OR to find from statistical tabular table).

→ **Step 6:** Conclusion.

- Compare the computed value of Z with critical value Z_α.
- If $|Z| > |Z_\alpha|$, we **reject H_0** and conclude that there is significant difference.
- If $|Z| < |Z_\alpha|$, we **accept H_0** and conclude that there is no significant difference.

Hypothesis Testing for Large Sample – I

Method – 5 ⇝ Test for Single Proportion

Test for Single Proportion

→ Consider a sample X of size n with proportion p taken from population with proportion P of size N.

→ This test is used to find the significant difference between **proportion of sample p & proportion of population P**.

→ In this test,

$$\mathbf{SE(t)} = \begin{cases} \sqrt{\dfrac{P\,Q}{n}} & ; \quad P \text{ is known} \\ \sqrt{\dfrac{p\,q}{n}} & ; \quad P \text{ is unknown} \end{cases}$$

→ **Formula for Test Statistics**

- When population proportion **P is known**

$$\mathbf{Z} = \frac{\mathbf{p - P}}{\sqrt{\dfrac{\mathbf{P\,Q}}{\mathbf{n}}}}$$

where, $Q = 1 - P$

- When population proportion **P is not known**

$$\mathbf{Z} = \frac{\mathbf{p - P}}{\sqrt{\dfrac{\mathbf{p\,q}}{\mathbf{n}}}}$$

where, $q = 1 - p$

→ **Confidence Limits**

$$\textbf{Confidence Limits} = \begin{cases} \mathbf{P \pm Z_\alpha\, SE(t)} & ; \quad P \text{ is known} \\ \mathbf{p \pm Z_\alpha\, SE(t)} & ; \quad P \text{ is unknown} \end{cases}$$

where, Z_α = Critical value at level of significance α

Example of Method-5: Test for Single Proportion

C	1	In a random sample of 160 worker exposed to a certain amount of radiation, 24 experienced some ill effects. Construct a 95% confidence interval for the corresponding true percentage. ($\lvert Z_{0.05} \rvert = 1.96$) **Answer**: $(\mathbf{0.0947, 0.2053})$
C	2	A political party claims that 45% of the voters in an election district prefer its candidate. A sample of 200 voters include 80 who prefer this candidate. Test if the claim is valid at the 5% significance level.($\lvert Z_{0.05} \rvert = 1.96$) **Answer: The claim is valid.**
C	3	A certain cubical die was thrown 9000 times and 5 or 6 was obtained 3240 times. On the assumption of certain throwing, do the data indicate an unbiased die?($\lvert Z_{0.05} \rvert = 1.96$) **Answer: The die is biased.**
C	4	A manufacturer claimed that at least 95% of the equipment which he supplied to a factory conformed to specification. An examination of a sample of 200 pieces of equipment revealed that 18 were faulty. Test this claim at 5% level of significance. ($Z_{0.05} = -1.645$) **Answer: The manufacturer's claim is accepted.**
C	5	The fatality rate of typhoid patients is believed to be 17.26%. In a certain year 640 patients suffering from typhoid were treated in a metropolitan hospital and only 63 patients died. Can you consider the hospital efficient at 1% level of significance? ($Z_{0.05} = -1.645$) **Answer: The hospital is efficient.**
C	6	In a sample of 400 parts manufactured by a factory; the number of defective parts found to be 30. The company, however, claims that at most 5% of their product is defective. Is the claim tenable? (Take level of significance 5%)($Z_{0.05} = 1.645$) **Answer: The claim of manufacturer is not tenable.**

C	7	In a big city, 325 men out of 600 were found to be smokers. Does information support the conclusion that the majority of men in this city are smokers? ($Z_{0.05} = 1.645$) **Answer: Yes, Information supports the conclusion that majority of men are smokers.**

Method – 6 ⇝ Test for Difference of Proportions

Test for Difference of Proportions

→ Consider two samples X_1 and X_2 of sizes n_1 & n_2 and of proportion p_1 & p_2 respectively taken from two different population of sizes N_1 & N_2 and of proportion P_1 & P_2.

→ This test is used to find the significant difference between **sample proportion p_1 & p_2** or **sample proportion P_1 & P_2**.

→ In this test,

$$\mathbf{SE(t)} = \begin{cases} \sqrt{\dfrac{P_1Q_1}{n_1} + \dfrac{P_2Q_2}{n_2}} & ; \text{ P is known} \\ \sqrt{pq\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)} & ; \text{ P is unknown} \end{cases}$$

→ **Formula for Test Statistics**

- When population proportion **P is known**

$$\mathbf{Z = \frac{P_1 - P_2}{\sqrt{\frac{P_1Q_1}{n_1} + \frac{P_2Q_2}{n_2}}}}$$

where, $Q = 1 - P$

- When population proportion **P is not known**

$$\mathbf{Z = \frac{p_1 - p_2}{\sqrt{pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}}$$

where, $p = \dfrac{n_1p_1 + n_2p_2}{n_1 + n_2}$ and $q = 1 - p$

→ **Confidence Limits**

$$\textbf{Confidence Limits} = \begin{cases} \mathbf{(P_1 - P_2) \pm Z_\alpha\, SE(t)} & ; \text{ P is known} \\ \mathbf{(p_1 - p_2) \pm Z_\alpha\, SE(t)} & ; \text{ P is unknown} \end{cases}$$

where, Z_α = Critical value at level of significance α

Example of Method-6: Test for Difference of Proportions

C	1	A random sample of 300 shoppers at a supermarket includes 204 who regularly uses cents off coupons. In another sample of 500 shoppers at a supermarket includes 75 who regularly uses cents off coupons. Obtain 95% confidence limits for the difference in the populations. ($\lvert Z_{0.05} \rvert = 1.96$) **Answer: $(0.4680, 0.5920)$**
C	2	Random samples of 400 men and 600 women were asked whether they would like to have a flyover near their residence. 200 men and 325 women were in favor of the proposal. Test the hypothesis that proportions of men and women in favor of the proposal are same at 5% level of significance. ($\lvert Z_{0.05} \rvert = 1.96$) **Answer: There is no significant difference of opinion between men and women in favor of the proposal.**
C	3	On the basis of their total scores, 200 candidates of a civil service examination are divided into two groups, the upper 30% and the remaining 70%. Consider the first question of examination. Among the first group, 40 had the correct answer, whereas among the second group, 80 had the correct answer. On the basis of these results, can one conclude that the first question is not good at discriminating ability of the type being examined here? ($\lvert Z_{0.05} \rvert = 1.96$) **Answer: Yes, the first question is not good enough at discriminating ability of the type being examined.**
C	4	In two large populations, there are 30% and 25% fair haired people respectively. Is this difference likely to be hidden in samples of 1200 and 900 respectively from the two populations? ($\lvert Z_{0.05} \rvert = 1.96$) **Answer: The difference in population properties is not likely to be hidden in sampling.**

C	5	In a certain city A,100 men in a sample of 400 are found to be smokers. In another city B, 300 men in a sample of 800 are found to be smokers. Does this indicate that there is greater proportion of smokers in B than in A? ($Z_{0.05} = -1.645$) **Answer: The proportion of smokers in city B is greater than in A.**
C	6	A question in a true-false is considered to be smart if it discriminates between intelligent person (IP) and average person (AP). Suppose 205 out of 250 IP's and 137 out of 250 AP's answer a quiz question correctly. Test of 0.01 level of significance whether for the given question, proportion of correct answers can be expected to be at least 15% higher among IP's than among the AP's. ($Z_{0.01} = 2.33$) **Answer: Proportion of correct answer by IP's is 15% more than those by AP's.**

*** * * * * End of the Unit * * * * ***