Subject: Quantitative Analysis

May 2023

1. Answer the following. Marks 20

Q.1. Define Statistics and list the limitations of statistics. (Marks 5)

1) Statistics is defined as collection, compilation analysis and interpretation of numerical data.
2) Statistic is a science of data.
3) Statistics helps in gathering information about the appropriate quantitative data.
4) It depicts the complex data in graphical form, tabular form and in diagrammatic representation to understand it easily
5) It provides the exact description and a better understanding
6) **Limitations of Statistics:**

- For quantitative results, statistical approaches are best applicable.
- On heterogeneous data, statistics cannot be implemented.
- In gathering, analyzing, and interpreting the data, if adequate care is not taken, statistical findings can be misleading.
- Statistical data can be treated effectively only by a person who has professional knowledge of statistics.
- In statistical judgments, certain errors are possible.
- Inferential statistics, in particular, include such errors.

Q.2. Explain sampling and purpose of sampling. (Marks 5)

1) It is a statistical tool in which a fixed no. of observations is taken from a larger population.
2) The behaviour or characteristics of the subset is used to estimate the characteristics of the entire population.
3) **Types of Sampling:**
   1. **Random Sampling:** it is a kind of sampling in which every item in the population has an equal probability of being picked.
   2. **Block Sampling:** Takes a consecutive series of items within the population to use as the sample.
   3. **Judgement Sampling:** an auditor's judgement may be used to select the sample out of the population.
   4. **Systematic Sampling:** begins at a random sampling point within the population itself and this kind of sampling uses a fixed, periodic interval to select items for a sample.
4) **Purpose of Sampling:**
   a) It is to provide information about the statistical information regarding the whole by examining just a few units.
   b) It reduces the time, effort and cost involved.
   c) It allows for minimisation of the loss caused in case of any mishap or failure.
   d) Scientific, observable method of testing a hypothesis.
   e) There is a greater scope for flexibility and probability.

Q.3. What is regression analysis? How does it differ from correlation. (Marks 5)

1) Regression analysis is a set of statistical methods used for the estimation of relationships between a dependent variable and one or more independent variables.
2) It can be utilized to assess the strength of the relationship between variables and for modelling the future relationship between them.
3) Regression analysis has three types:
    1. Linear Regression
    2. Multiple Linear Regression
    3. Non-linear Regression

Parameter	Regression Analysis	Correlation Analysis
1) Purpose	Measures strength and direction of the relationship.	Predicts and models the relationship.
2) Variables	Two variables (Equal roles).	Independent and dependant variables.
3) Calculation	Correlation Coefficient(r)	Regression Equation (y = mx + b)
4) Direction	+1 and -1 (Positive, Negative, No Correlation)	Positive, Negative (Strength and direction)
5) Causation	It does not imply causation.	Can imply causation under Controlled conditions.
6) Data Representation	Single Coefficient.	The equation representing the relationship.
7) Hypothesis Testing	Limited to correlation significance	Tests coefficient's importance in the model.

Q.4. Show the sample variance (S^2) is an unbiased estimator of population variance (σ^2). Also illustrate with an example. (Marks 5)

$$E(X^2) = \sigma^2 + \mu^2$$

$$E(\bar{X}^2) = \frac{\sigma^2}{n} + \mu^2$$

$$\begin{aligned} E[(X_i - \bar{X})^2] &= E\left({X_i}^2\right) - nE(\bar{X}^2) \\ &= \sum(\sigma^2 + \mu^2) - (\frac{\sigma^2}{n} + \mu^2) \\ &= n\sigma^2 + n\mu^2 - \sigma^2 + n\mu^2 \\ &= n\sigma^2 - \sigma^2 \\ &= (n-1)\sigma^2 \end{aligned}$$

Let's prove $E(S^2) = E\left[\frac{(X_i - \bar{X})^2}{(n-1)}\right] = \sigma^2$

$$\begin{aligned} E(S^2) &= E\left[\frac{(X_i - \bar{X})^2}{n-1}\right] \\ &= \frac{1}{n-1} E\left[\sum(X_i - \bar{X})^2\right] \\ &= \frac{1}{n-1}(n-1)\sigma^2 \\ &= \sigma^2 \end{aligned}$$

∴ Hence proved.

Example: Suppose we have a population of 5 individuals with ages: 20, 35, 45, 50, 55
Take sample size first 3.

$$\therefore S^2 = \frac{1}{n-1}\sum(X_i - \bar{X})^2$$

$$\therefore n = 3, \quad \bar{X} = \frac{\sum X}{n} = \frac{100}{3} = 33.33$$

$$\therefore S^2 = \frac{(20-33.33)^2 + (35-33.33)^2 + (45-33.33)^2}{3-1}$$

$$= \frac{177.68 + 2.78 + 136.18}{2}$$

$$= \frac{316.64}{2}$$

$$\therefore S^2 = 158.32$$

Take all Samples,

$$\therefore \sigma^2 = \frac{1}{N}\sum(X_i - \mu)^2$$

$$\therefore N = 5, \quad \mu = \frac{\sum \mu}{n} = \frac{205}{5} = 41$$

$$\therefore \sigma^2 = \frac{(20-41)^2 + (35-41)^2 + (45-41)^2 + (50-41)^2 + (55-41)^2}{5}$$

$$= \frac{441 + 36 + 16 + 81 + 196}{5}$$

$$= \frac{770}{5}$$

$$\therefore \sigma^2 = 154$$

$$E(S^2) = \sigma^2$$

$$E(158.32) = 154$$

$$158.32 = 154$$

Therefore, S^2 is an unbiased estimator of σ^2.

2. Solve the Following: Marks 20

Q.1. In a laboratory experiment on correlation research study, the equations to the two regression lines were found to be $2x - y + 1 = 0$ and $3x - 2y + 7 = 0$. Find the mean of x and y. Also work out the values of regression coefficients and correlation coefficient between the two variables x and y. (Marks 10)

Solving the two regression equations we get mean values of X and Y:

$2x - y = -1$ ……eq(1)

$3x - 2y = -7$ ……eq(2)

Solving equation 1 and 2,

We get, $x = 5, y = 11$.

∴ Regression Line is passed throughs means $\bar{X} = 5$ and $\bar{Y} = 11$.

The regression equation Y on X is $3x - 2y = -7$.

$$2y = 3x + 7$$

$$y = \frac{1}{2}(3x + 7)$$

$$y = \frac{3}{2}x + \frac{7}{2}$$

$$\therefore b_{yx} = \frac{3}{2}(> 1)$$

The regression equation X on Y is $2x - y = -1$.

$$2x = y - 1$$

$$x = \frac{1}{2}(y - 1)$$

$$x = \frac{1}{2}y - \frac{1}{2}$$

$$\therefore b_{xy} = \frac{1}{2}$$

The regression coefficients are positive.

$$r = \pm\sqrt{b_{xy} \cdot b_{yx}} = \pm\sqrt{\frac{1}{2} \times \frac{3}{2}}$$

$$= \sqrt{\frac{1}{2} \times \frac{3}{2}}$$

$$= \sqrt{\frac{3}{4}}$$

$$= 0.8660$$

$$r = 0.8660$$

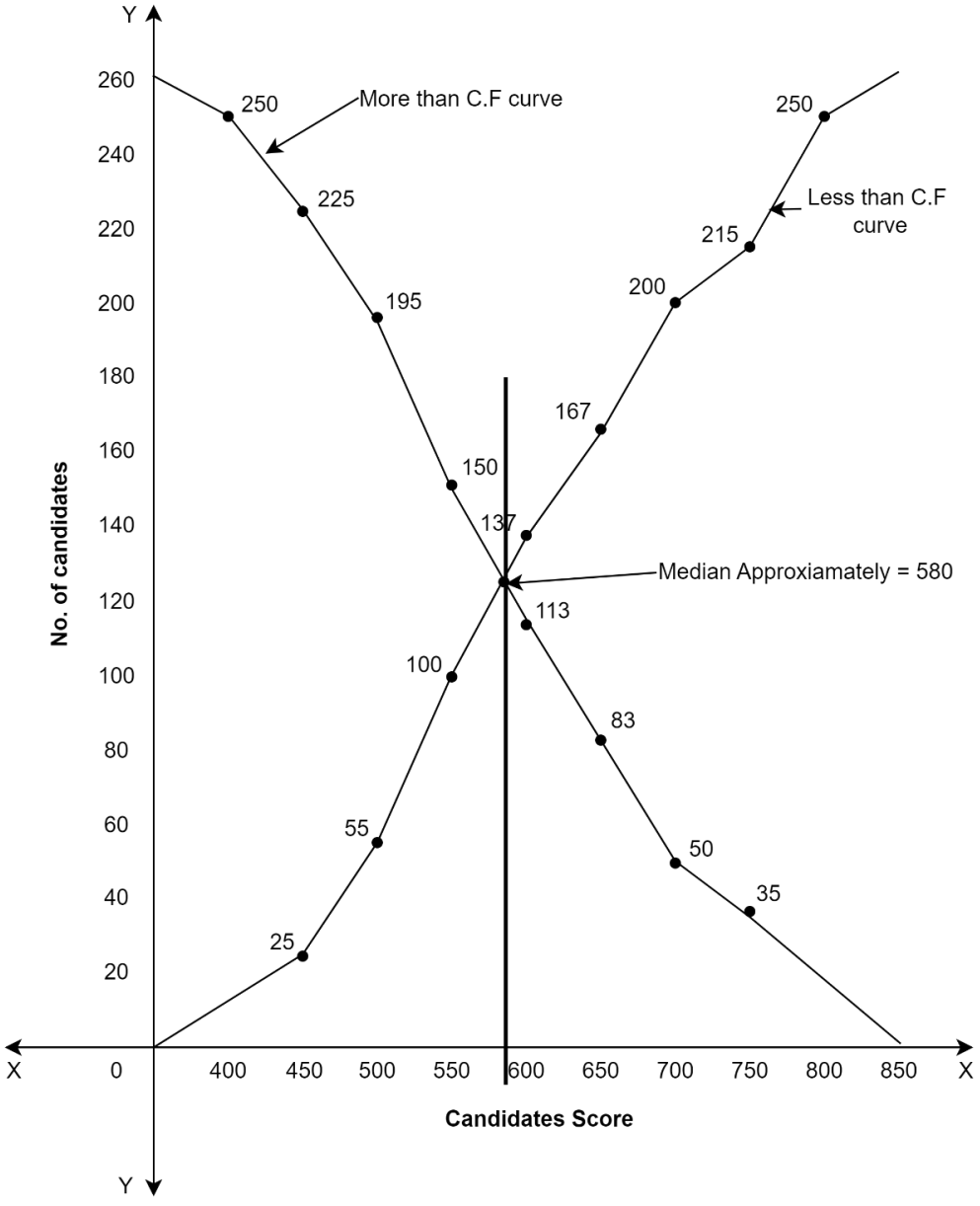
Q.2. The frequency distribution of scores obtained by 250 candidates in an entrance test is as follows. Draw a less than and more than frequency curve(ogive) to represent the given data. also, what is the significance of the point of intersection of the two ogive curves? (Marks 10)

Scores	Number of candidates
400 – 450	**25**
450 – 500	**30**
500 – 550	**45**
550 – 600	**37**
600 – 650	**30**
650 – 700	**33**
700 – 750	**15**
750 – 800	**35**

We make a less than or more than cumulative frequency table:

Scores	Number of candidates	Less than C.F	More than C.F
400 – 450	25	25	225 + 25 = 250
450 – 500	30	25 + 30 = 55	195 + 30 = 225
500 – 550	45	55 + 45 = 100	150 + 45 = 195
550 – 600	37	100 + 37 = 137	113 + 37 = 150
600 – 650	30	137 + 30 = 167	83 + 30 = 113
650 – 700	33	167 + 33 = 200	50 + 33 = 83
700 – 750	15	200 + 15 = 215	35 + 15 = 50
750 – 800	35	215 + 35 = 250	35

Curves:



∴ The significance of the point of intersection of the two ogive curves is approximately 580.

3. Solve the Following: Marks 20

Q.1. The following table gives the age of cars of a certain make and annual maintenance costs. Obtain the regression equation for maintenance costs, taking age of the car as the independent variable. Also, find the maintenance cost for age of the car = 5 years.

Age of Cars (in Years)	Maintenance Cost (In thousands of rupees)
2	**10**
4	**20**
6	**25**
8	**30**

(Marks 10)

Regression equation is $y = bx + a$, where x is an age of cars and y is maintenance cost.

$$\bar{X} = \frac{\sum X}{n} = \frac{20}{4} = 5, \quad \bar{Y} = \frac{\sum Y}{n} = \frac{85}{4} = 21.25, n = 4$$

$$Find\ slope, \qquad b = \frac{\sum XY - n\bar{X}\bar{Y}}{\sum X^2 - nX^2}$$

X	Y	X^2	XY
2	10	4	20
4	20	16	80
6	25	36	150
8	30	64	240
$\sum X = 20$	$\sum Y = 85$	$\sum X^2 = 120$	$\sum XY = 490$

$$\therefore\ b = \frac{490 - 4 \times 5 \times 21.25}{120 - 4 \times 5^2}$$

$$= \frac{65}{20}$$

$$= 3.25$$

$\therefore y = a + bx$

$\therefore y = 5 + 3.25x$ ..... $a = \bar{Y} - b\bar{X}$

Find the maintenance cost for 5 years old car:

Given, $x = 5$

$\therefore y = 5 + 3.25 \times 5 = 21.25$(In hundreds) = 21.25 × 100 = 2125Rs.

∴ The maintenance cost for 5 years old car is: 2125 RS.

Q.2. Explain with illustration the concept of Point Estimation. (Marks 10)

	Sample	Population
Size	n	N
Mean	$\bar{X}$	μ
Standard Deviation	$S.D$	σ
Variance	S^2 or s^2 Capital S for biased Small s for unbiased	σ^2
Proposition	P sampling	π
S_x**(Error)**	$\frac{S}{\sqrt{n}}$	$-$

1) Point estimators are defined as functions that can be used to find the approximate value of a particular point from a given population parameter.
2) In point estimation, we find out the statistic which may use for replace an unknown parameter for all practical purpose.
3) A good estimator is one which is as close to true value the parameter as possible.
4) The sample data of a population is used to find a point estimate or a statistic that can act as the best estimate of an unknown parameter that is given for a population.
5) The maximum likelihood method is a popularly used way to calculate point estimators. This method uses differential calculus to understand the probability function from a given number of sample parameters.
6) Following are the four characteristics of point estimation:
    1. Unbiasedness
    2. Consistency
    3. Efficiency
    4. Sufficiency

4. Solve the Following Marks 20

Q.1. Following is the data about the weights in Kgs of 10 Shipments(X_1), the distances they were moved(X_2) and the damage that was incurred (Y).

Shipment	Damage (thousands of RS) (Y)	Weights in Kgs (X_1)	Distance moved in Km (X_2)
1	**12**	**17**	**10**
2	**15**	**15**	**6**
3	**14**	**15**	**10**
4	**19**	**10**	**21**
5	**8**	**13**	**8**
6	**16**	**15**	**13**
7	**15**	**11**	**9**
8	**25**	**6**	**25**
9	**10**	**15**	**10**
10	**11**	**7**	**8**

i. **Fit the regression $\widehat{Y} = a + b_1X_1 + b_2X_2$** **(Marks 5)**

ii. **Find the coefficient of multiple determination (R_2).** **(Marks 2)**

iii. **Also test the significance of regression (Given the appropriate Table value, F = 9.55, for a Significance level of $\alpha = 0.01$)** **(Marks 3)**

$\therefore \sum Y = 145 \quad \therefore \sum X_1 = 124 \quad \therefore \sum X_2 = 112 \quad \therefore n = 10$

i) $\sum Y = na + b_1 \sum X_1 + b_2 \sum X_2$ ..... eq(1)

$\sum X_1 Y = a \sum X_1 + b_1 \sum {X_1}^2 + b_2 \sum X_1 X_2$ ..... eq(2)

$\sum X_2 Y = a \sum X_2 + b_1 \sum X_1 X_2 + b_2 \sum {X_2}^2$ ..... eq(3)

Y	X_1	X_2	X_1Y	X_2Y	${X_1}^2$	${X_2}^2$	X_1X_2
12	17	10	204	120	289	100	170
15	15	6	225	90	225	36	90
14	15	10	210	140	225	100	150
19	10	21	190	399	100	441	210
8	13	8	104	64	169	64	104
16	15	13	240	208	225	169	195
15	11	9	165	135	121	81	99
25	6	25	150	625	36	625	150
10	15	10	150	100	225	100	150
11	7	8	77	88	49	64	56
$Y = 145$	$X_1 = 124$	$X_2 = 112$	$X_1Y = 1715$	$X_2Y = 1969$	${X_1}^2 = 1664$	${X_2}^2 = 1780$	$X_1X_2 = 1374$

$145 = 10a + 124b_1 + 112b_2$ ..... eq(1)

$1715 = 124a + 1664b_1 + 1374b_2$ ..... eq(2)

$1969 = 112a + 1374b_1 + 1780b_2$ ..... eq(3)

$\therefore a = 14$

$\therefore b_1 = -0.5817$

$\therefore b_2 = 0.6400$

$\widehat{Y} = 14 - 0.58X_1 + 0.64X_2$

ii) $R^2 = \frac{\sum (Y_i - \bar{Y})^2 - \sum (Y_i - \hat{Y})^2}{\sum (Y_i - \bar{Y})^2}$

$\bar{Y} = \frac{\sum Y}{n} = \frac{145}{10} = 14.5$

Y_i	$Y_i - \bar{Y}$	$\widehat{Y}$	$Y_i - \widehat{Y}$	$(Y_i - \bar{Y})^2$	$(Y_i - \widehat{Y})^2$
12	-2.5	10.54	1.46	6.25	2.1316
15	0.5	9.14	5.86	0.25	34.3396
14	-0.5	11.7	2.3	0.25	136.89
19	4.5	21.64	-2.64	20.25	468.2896
8	-6.5	11.58	-3.58	42.25	134.0964
16	1.5	13.62	2.38	2.25	185.5044
15	0.5	13.38	1.62	0.25	179.0244
25	10.5	26.52	-1.52	110.25	703.3104
10	-4.5	11.7	-1.7	20.25	136.89
11	-3.5	15.06	-4.06	12.25	226.8036
$Y_i = 145$	$Y_i - \bar{Y} = 0$	$\hat{Y} = 144.88$	$Y_i - \hat{Y} = 0.12$	$(Y_i - \bar{Y})^2 = 214.5$	$(Y_i - \hat{Y})^2 = 2207.28$

$$R^2 = \frac{214.5 - 2207.28}{214.5}$$

$$\boldsymbol{R^2 = -9.29}$$

iii) $F_\alpha \; at \; 1\% \; level \; of \; significance \; is \; 9.55.$

$F_\alpha = 9.55$

$$F = \frac{\frac{\sum(Y_i - \hat{Y})^2}{p}}{\frac{\sum(Y_i - \bar{Y})^2}{n - p - 1}}$$

Where, P is an independent variable(b's).

$p = 2.$

$$\therefore F = \frac{\frac{2207.28}{2}}{\frac{214.5}{10 - 2 - 1}}$$

$$= \frac{1103.64}{30.64}$$

$$= 36.01$$

$\therefore F = 36.01$

$\therefore F > F_\alpha$, $36.01 > 9.55$ then regression model is significant.

Q.2. Explain Primary data and Secondary data in detail. (Marks 10)

Parameter	Primary Data	Secondary Data
1. Meaning	Data collected by researcher itself.	Data collected by other people.
2. Originality	Original and Unique Information.	Not original and unique information.
3. Adjustment	Does not need adjustment, is focused.	Need adjustment to suit actual aim.
4. Sources	Observations, Surveys, Experiment.	Internal Records, govt. published Data, etc.
5. Type of Data	Qualitative Data	Quantitative Data
6. Methods	Observation, experiment, interview, etc.	Desk research method, searching online, etc.
7. Reliability	More reliable	Less reliable
8. Capability	More capable to solve a problem.	Less capable to solve a problem.
9. Time consumed	More time consuming	Less time consuming
10. Cost-effectiveness	Costly	Economical
11. Suitability	More suitable	May or may not be suitable
12. Need of Investigators	Needs team of trained Investigators.	Does not need of team Investigators.
13. Collected when	Secondary data is inadequate.	Before primary data is collected.

5. Solve the Following **Marks 20**

Q.1. Given $r_{12} = 0.7, r_{13} = 0.61$ and $r_{23} = 0.4$. **(Marks 10)**

Compute:

i. $r_{23.1}$

ii. $r_{13.2}$

iii. $r_{12.3}$

1) $r_{23.1}$

$$r_{23.1} = \frac{r_{23} - r_{21}r_{31}}{\sqrt{(1 - {r_{21}}^2)(1 - {r_{31}}^2)}}$$

$$= \frac{(0.4) - (0.7) \times (0.61)}{\sqrt{(1 - (0.7)^2) \times (1 - (0.61)^2)}}$$

$$= \frac{0.4 - 0.427}{\sqrt{(1 - 0.49) \times (1 - 0.3721)}}$$

$$= \frac{-0.027}{\sqrt{0.51 \times 0.6279}}$$

$$= \frac{-0.027}{0.565}$$

$$\therefore r_{23.1} = -0.047$$

2) $r_{13.2}$

$$r_{13.2} = \frac{r_{13} - r_{12}r_{32}}{\sqrt{(1 - {r_{12}}^2)(1 - {r_{32}}^2)}}$$

$$= \frac{(0.61) - (0.7) \times (0.4)}{\sqrt{(1 - (0.7)^2) \times (1 - (0.4)^2)}}$$

$$= \frac{0.61 - 0.28}{\sqrt{(1 - 0.49) \times (1 - 0.16)}}$$

$$= \frac{0.33}{\sqrt{0.51 \times 0.84}}$$

$$= \frac{0.33}{0.65}$$

$$\therefore r_{13.2} = 0.50$$

3) $r_{12.3}$

$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{(1 - {r_{13}}^2)(1 - {r_{23}}^2)}}$$

$$= \frac{(0.7) - (0.61) \times (0.4)}{\sqrt{(1 - (0.61)^2) \times (1 - (0.4)^2)}}$$

$$= \frac{0.7 - 0.244}{\sqrt{(1 - 0.3721) \times (1 - 0.16)}}$$

$$= \frac{0.456}{\sqrt{0.6279 \times 0.84}}$$

$$= \frac{0.456}{0.726}$$

$$\therefore r_{12.3} = 0.628$$

Q.2. Differentiate between the following pair of concepts: (Marks 10)

i. Critical Region and Region of acceptance (Marks 5)

Sr. no.	Critical Region	Region of acceptance
1.	Represents the range of values of the test statistic where the null hypothesis is rejected.	Represents the range of values of the test statistic where the null hypothesis is not rejected.
2.	Also known as the rejection region.	Also known as the non-rejection region.
3.	Determined based on the chosen significance level (alpha) of the test.	Comprises all values not included in the critical region.
4.	Contains extreme outcomes that provide evidence against the null hypothesis.	Indicates that the data does not provide sufficient evidence to reject the null hypothesis.
5.	If the test statistic falls within this region, the null hypothesis is rejected.	If the test statistic falls within this region, the null hypothesis is retained.
6.	Its size is directly related to the chosen significance level.	Its size is complementary to the size of the critical region.
7.	Helps in identifying statistically significant results.	Helps identify cases where there is insufficient evidence to reject the null hypothesis.
8.	Typically depicted in the tail(s) of the sampling distribution.	Typically represents the bulk of the sampling distribution.

ii. Null Hypothesis and Alternative Hypothesis (Marks 5)

Sr. No.	Null Hypothesis	Alternative Hypothesis
1)	A Null Hypothesis is a statement in which there is no relation between two variables.	An Alternative Hypothesis is a statement in which there is some statistical relation between two variables.
2)	Researcher try to reject or disprove it.	Researcher try to accept or prove it.
3)	Indirect and Implicit.	Direct and Explicit.
4)	P - value: If P value is less than α value, then Null hypothesis is rejected. $\lvert Z_p \rvert < \lvert Z_t \rvert$	P - value: If P value is less than α value, then Alternative hypothesis is accepted. $\lvert Z_p \rvert < \lvert Z_\alpha \rvert$
5)	Null Hypothesis is denoted by H_0.	Alternative Hypothesis is denoted by H_1.
6)	Symbols: Equality =, >=, <=	Symbols: Inequality !=, >, <
7)	Size of sample is $n \geq 30$. Large sample.	Size of sample is $n < 30$. Small sample.
8)	Z - test	T – test

6. Write short note on Marks 20

i. Pie chart and its advantages and Disadvantages (Marks 5)

1) Pie Chart is a pictorial representation of the data.
2) It uses a circle to represent the data and is hence also called a Circle Graph**.**
3) In a Pie Chart, we present the data by dividing the whole circle into smaller slices or sectors, and each slice or sector represents specific data.
4) **Advantages:**
   1. Pie chart is easily understood and comprehended.
   2. Visual representation of data in a pie chart is done as a fractional part of a whole.
   3. Pie chart provides an effective mode of communication to all types of audiences.
   4. Pie chart provides a better comparison of data for the audience.
5) **Disadvantages:**
   1. In the case of too much data, this presentation becomes less effective using a pie chart.
   2. For multiple data sets, we need a series to compare them.
   3. For analyzing and assimilating the data in a pie chart, it is difficult for readers to comprehend.

ii. Method of moments (Marks 5)

1) The method of moments is a technique for estimating the parameters of a statistical model.
2) It works by finding values of the parameters that result in a match between the sample moments and the population moments.
3) The advantage of method of moment is that it is quite easy to use.
4) however, the quality of the result from method of moment is not very good.
5) Suppose a random variable X has density $f(x|\theta)$, and this should be understood as point mass function when the random variable is discrete.
   The $k-th$ theoretical moment of this random variable is defined as
   $$\mu_k = E(X^k) = \int x^k f(x|\theta)\, dx$$
   or
   $$\mu_k = E(X^k) = \sum_x x^k f(x|\theta).$$
6) If $X_1, \cdots, X_n$ are i.i.d. random variables from that distribution, the $k-th$ sample moment is defined as
   $$m_k = \frac{1}{n}\sum_{i=1}^{n} {X_i}^k,$$
   thus, m_k can be viewed as an estimator for μ_k.
   From the law of large number, we have $m_k \to \mu_k$ in probability as $n \to \infty$.
   If we equate μ_k to m_k, usually we will get an equation about the unknown parameter.
7) Solving this equation will help us get the estimator of the unknown parameter.

iii. Multiple Regression (Marks 5)

1) Multiple regression is a statistical technique that can be used to analyze the relationship between a single dependent variable and several independent variables.
2) The objective of multiple regression analysis is to use the independent variables whose values are known to predict the value of the single dependent value.
3) Each predictor value is weighed, the weights denoting their relative contribution to the overall prediction.

$$Y = a + b_1X_1 + b_2X_2 + \cdots + b_nX_n$$

4) Here Y is the dependent variable, and $X_1, \ldots, X_n$ are the n independent variables.
5) In calculating the weights, $a, b_1, \ldots, b_n$, regression analysis ensures maximal prediction of the dependent variable from the set of independent variables.
6) This is usually done by least squares estimation.
7) This approach can be applied to analyze multivariate time series data when one of the variables is dependent on a set of other variables.
8) We can model the dependent variable Y on the set of independent variables.

iv. Neyman Pearson Lemma (Marks 5)

1) The Neyman-Pearson Lemma gives strong guidance about how to choose hypothesis tests.
2) The Neyman-Pearson Lemma is an important result that gives conditions for a hypothesis test to be uniformly most powerful.
3) That is, the test will have the highest probability of rejecting the null hypothesis while maintaining a low false positive rate of α.
4) More formally, consider testing two simple hypotheses:

$$H_0: \theta = \theta_0$$
$$H_1: \theta = \theta_1$$

5) The Neyman-Pearson Lemma says a test is uniformly most powerful test among α-level tests if it rejects H_0 if and only if

$$\frac{fx(x;\theta_1)}{fx(x;\theta_0)} > k$$

for some $k \epsilon R$, were

$$\alpha = P_{\theta_0}\left[\frac{fx(x;\theta_1)}{fx(x;\theta_0)} > k\right]$$

December 2023

1. Answer the Following Marks 20

Q.1. Define "Statistics". Explain Uses and Limitations of Statistics. (Marks 5)

1) Statistics is defined as collection, compilation analysis and interpretation of numerical data.
2) Statistic is a science of data.
3) Statistics helps in gathering information about the appropriate quantitative data.
4) It depicts the complex data in graphical form, tabular form and in diagrammatic representation to understand it easily
5) It provides the exact description and a better understanding.
6) **Uses:**
   1. Forecasting
   2. Financial Analysis
   3. Government
   4. Designing Surveys
   5. Economics
   6. Quality Control
   7. Health Care
   8. Data Analysis
   9. Sports
   10. Probability
   11. Politics
   12. Research
7) **Limitations of Statistics:**

- For quantitative results, statistical approaches are best applicable.
- On heterogeneous data, statistics cannot be implemented.
- In gathering, analyzing, and interpreting the data, if adequate care is not taken, statistical findings can be misleading.
- Statistical data can be treated effectively only by a person who has professional knowledge of statistics.
- In statistical judgments, certain errors are possible.

Q.2. A random sample of size 100 has a standard deviation of 5. What can you say about the maximum error with 95% confidence is 1.96. (Marks 5)

$\therefore n = 100,\ \sigma = 5$, confidence = 1.96.

Maximum error =

$$E(Error) = Z_{Confidence} \times \frac{\sigma}{\sqrt{n}}$$

$$= 1.96 \times \frac{5}{\sqrt{100}}$$

$$= 1.96 \times 0.5$$

$$= 0.98$$

$\therefore$ Maximum error with 95% confidence level is equal to plus or minus 0.98 from the mean.

Q.3. What are assumptions of Multiple Linear Regression? (Marks 5)

There are a number of assumptions that should be assessed before performing a multiple regression analysis:

1) The dependant variable (the variable of interest) needs to be using a continuous scale.
2) There are two or more independent variables. These can be measured using either continuous or categorical means.
3) The three or more variables of interest should have a linear relationship, which you can check by using a scatterplot.
4) The data should have homoscedasticity. In other words, the line of best fit is not dissimilar as the data points move across the line in a positive or negative direction. Homoscedasticity can be checked by producing standardised residual plots against the unstandardized predicted values.
5) The data should not have two or more independent variables that are highly correlated. This is called multicollinearity which can be checked using Variance-inflation-factor or VIF values. High VIF indicates that the associated independent variable is highly collinear with the other variables in the model.
6) There should be no spurious outliers.
7) The residuals (errors) should be approximately normally distributed. This can be checked by a histogram (with a superimposed normal curve) and by plotting the of the standardised residuals using either a P-P Plot, or a Normal Q-Q Plot.

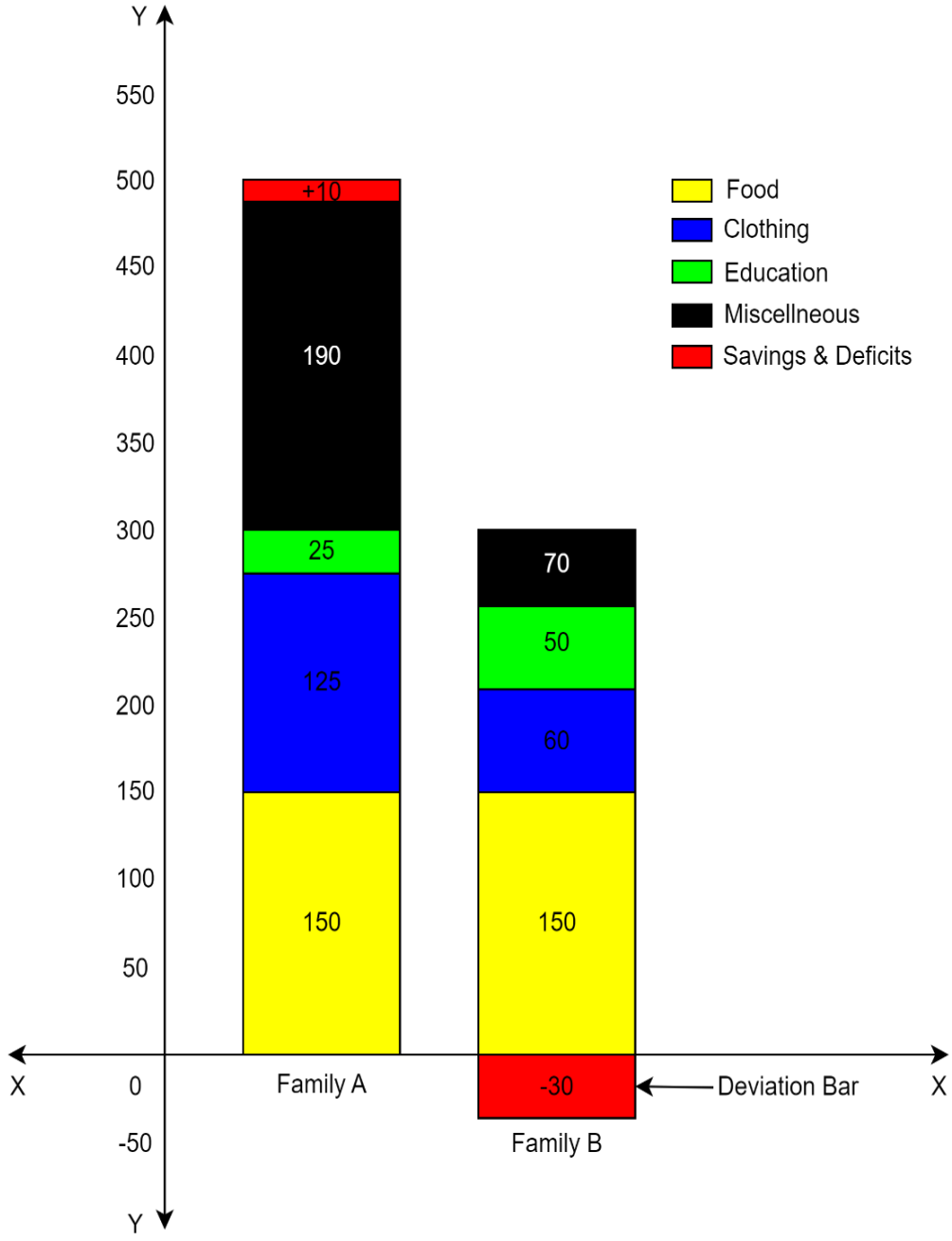
Q.4. Distinguish between Null and Alternative hypothesis. (Marks 5)

Sr. No.	Null Hypothesis	Alternative Hypothesis
1)	A Null Hypothesis is a statement in which there is no relation between two variables.	An Alternative Hypothesis is a statement in which there is some statistical relation between two variables.
2)	Researcher try to reject or disprove it.	Researcher try to accept or prove it.
3)	Indirect and Implicit.	Direct and Explicit.
4)	P - value: If P value is less than α value, then Null hypothesis is rejected. $\left\|Z_p\right\| < \|Z_t\|$	P - value: If P value is less than α value, then Alternative hypothesis is accepted. $\left\|Z_p\right\| < \|Z_\alpha\|$
5)	Null Hypothesis is denoted by H_0.	Alternative Hypothesis is denoted by H_1.
6)	Symbols: Equality =, >=, <=	Symbols: Inequality !=, >, <
7)	Size of sample is $n \geq 30$. Large sample.	Size of sample is $n < 30$. Small sample.
8)	Z - test	T – test

2. Answer the Following — Marks 20

Q.1. Represent the following data by a percentage sub-divided bar diagram. (Marks 10)

Item of Expenditure	Family A	Family B
	Income Rs 500	Income Rs 300
Food	150	150
Clothing	125	60
Education	25	50
Miscellaneous	190	70
Savings or Deficits	+10	-30



Q.2. Distinguish between primary data and secondary. What precautions should be taken in the use of secondary data. (Marks 10)

Parameter	Primary Data	Secondary Data
1. Meaning	Data collected by researcher itself.	Data collected by other people.
2. Originality	Original and Unique Information.	Not original and unique information.
3. Adjustment	Does not need adjustment, is focused.	Need adjustment to suit actual aim.
4. Sources	Observations, Surveys, Experiment.	Internal Records, govt. published Data, etc.
5. Type of Data	Qualitative Data	Quantitative Data
6. Methods	Observation, experiment, interview, etc.	Desk research method, searching online, etc.
7. Reliability	More reliable	Less reliable
8. Capability	More capable to solve a problem.	Less capable to solve a problem.
9. Time consumed	More time consuming	Less time consuming
10. Cost-effectiveness	Costly	Economical
11. Suitability	More suitable	May or may not be suitable
12. Need of Investigators	Needs team of trained Investigators.	Does not need of team Investigators.
13. Collected when	Secondary data is inadequate.	Before primary data is collected.

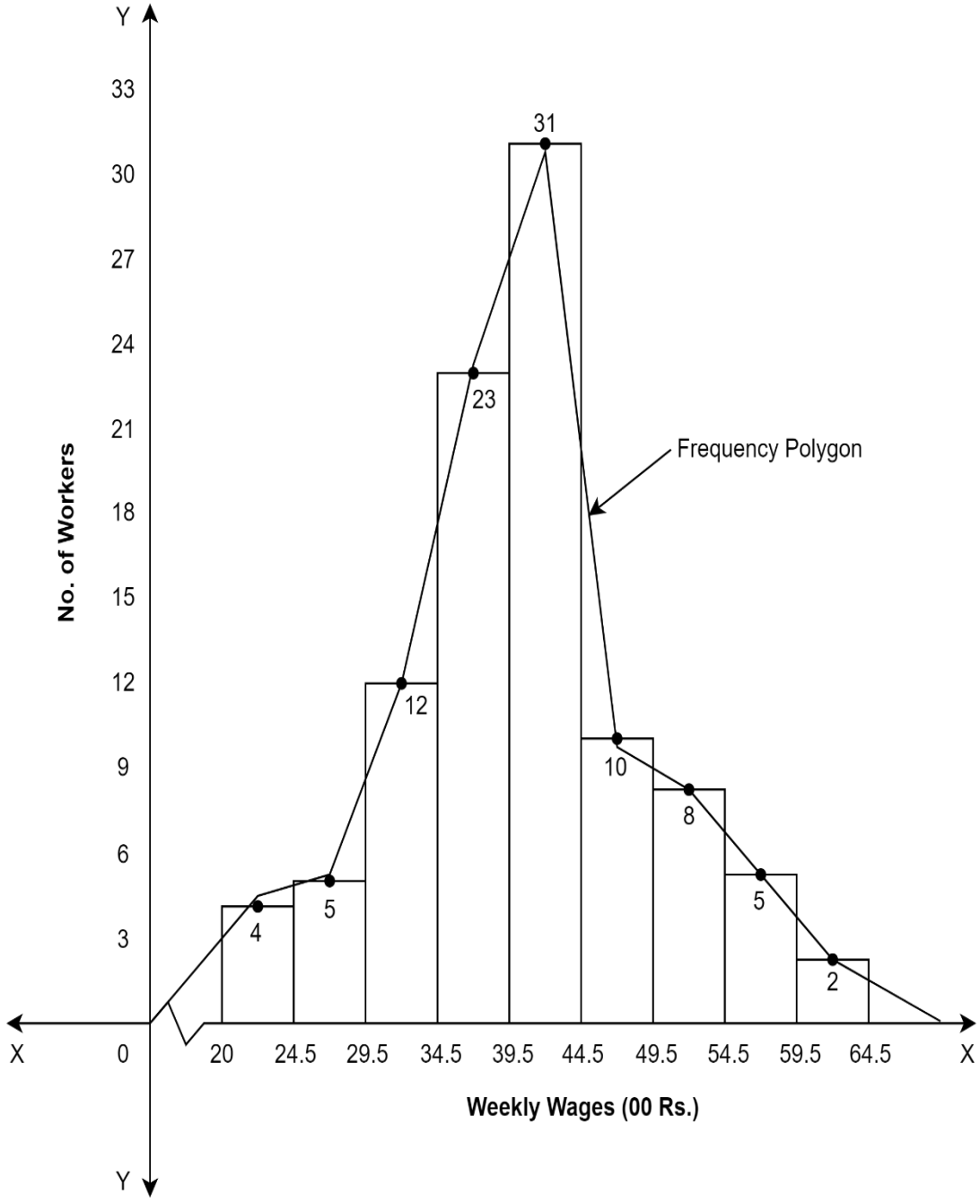
Following some precautions should be taken in the use of secondary data:

- Suitable purpose of investigation.
- Inadequate data.
- Definition of units.
- Degree of accuracy.
- Time and condition of collection of facts.
- Homogeneous conditions.
- Comparison.

3. Answer the Following Marks 20

Q.1. The following Table gives the frequency distribution of the weekly wages (in '00RS.) of 100 workers in factory. Draw the Histogram and frequency polygon of the distribution. (Marks 10)

Weekly wages ('00 RS.')	20-24	25-29	30-34	35-39	40-44	45-49	50-54	55-59	60-64	Total
No. of Workers	4	5	12	23	31	10	8	5	2	100



Q.2. The equation of two lines of regression obtained in correlation analysis are given below: $2X = 8 - 3Y$ and $2Y = 5 - X$. Obtain the value of the correlation coefficient. (Marks 10)

Let Regression line of X on Y be,

$2X = 8 - 3Y$

$X = \frac{8}{2} - \frac{3}{2}Y$

$X = -1.5Y + 4$

$\therefore b_{xy} = -1.5 \; or \; -\frac{3}{2}$

Let Regression line of Y on X be,

$2Y = 5 - X$

$Y = \frac{5}{2} - \frac{1}{2}X$

$Y = -0.5X + 2.5$

$\therefore b_{yx} = -0.5 \; or \; -\frac{1}{2}$

The regression coefficients are negative,

Coefficient correlation is,

$r = \pm\sqrt{b_{xy} \times b_{yx}}$

$= \pm\sqrt{-\frac{3}{2} \times -\frac{1}{2}}$

$= \pm 0.866$

$\therefore r = -0.866$

4. Answer the Following — Marks 20

Q.1. From the data given below find: (Marks 10)

a) The Two regression coefficients

b) The Two regression equations

c) The coefficient of correlation between the marks in Economics and Statistics

d) The most likely marks in Statistics if marks in Economics are 30.

Marks in Economics	**25**	**28**	**35**	**32**	**31**	**36**	**29**	**38**	**34**	**32**
Marks in Statistics	**43**	**46**	**49**	**41**	**36**	**32**	**31**	**30**	**33**	**39**

$\therefore \bar{X} = \frac{\sum X}{N} = \frac{320}{10} = 32$

$\therefore \bar{Y} = \frac{\sum Y}{N} = \frac{380}{10} = 38$

X	Y	$x = X - 32$	$y = Y - 38$	x^2	y^2	xy
25	43	-7	5	49	25	-35
28	46	-4	8	16	64	-32
35	49	3	11	9	121	33
32	41	0	3	0	9	0
31	36	-1	-2	1	4	2
36	32	4	-6	16	36	-24
29	31	-3	-7	9	49	21
38	30	6	-8	36	64	-48
34	33	2	-5	4	25	-10
32	39	0	1	0	1	0
$\sum X = 320$	$\sum Y = 380$	$\sum x = 0$	$\sum y = 0$	$\sum x^2 = 140$	$\sum y^2 = 398$	$\sum xy = -93$

a) Two regression coefficients:

1) Regression coefficients X on Y:

$$b_{xy} = \frac{N\sum xy - (\sum x)(\sum y)}{N\sum y^2 - \sum(y)^2}$$
$$= \frac{10 \times -93 - 0 \times 0}{10 \times 398 - 0^2}$$
$$= \frac{-930 - 0}{3980 - 0}$$
$$= \frac{-930}{3980}$$
$$\therefore b_{xy} = -0.2336$$

2) Regression coefficients Y on X:

$$b_{yx} = \frac{N\sum xy - (\sum x)(\sum y)}{N\sum x^2 - \sum(x)^2}$$
$$= \frac{10 \times -93 - 0 \times 0}{10 \times 140 - 0^2}$$
$$= \frac{-930 - 0}{1400 - 00}$$
$$= \frac{-930}{1400}$$
$$\therefore b_{yx} = -0.6642$$

b) Two regression equations:

1) Regression equation of X on Y:

$$X - \bar{X} = b_{xy}(Y - \bar{Y})$$
$$X - 32 = -0.2336(Y - 38)$$
$$X - 32 = -0.2336Y + 8.8768$$
$$X = -0.2336Y + 40.8768$$

∴ The regression equation of X on Y is $X = -0.2336Y + 40.8768$

2) Regression equation of Y on X:

$Y - \bar{Y} = b_{yx}(X - \bar{X})$

$Y - 38 = -0.6642(X - 32)$

$Y - 38 = -0.6642X + 21.2544$

$Y = -0.6642X + 59.2544$

∴ The regression equation of Y on X is $Y = -0.6642X + 59.2544$

c) coefficient of correlation between the marks in Economics and Statistics:

$r = \pm\sqrt{b_{xy} \times b_{yx}}$

$= \pm\sqrt{-0.2336 \times -0.6642}$

$= \pm 0.3938$

∴ Both regression coefficients are negative, so take negative sign,

$r = -0.3938$

d) most likely marks in Statistics if marks in Economics are 30:

$X = 30, Y = ?$

$Y = -0.6642X + 59.2544$

$= -0.6642(30) + 59.2544$

$= -19.929 + 59.2544$

$\therefore Y = 39.3254$

Q.2. Explain the following point Estimation Properties with Example (Marks 10)

i) Consistency:

1) It states that the estimator stays close to the parameter's value as the population's size increases.
2) Thus, a large sample size is required to maintain its consistency level.
3) When the expected value moves towards the parameter's value, we state that the estimation is consistent.
4) **Example:**
    1. Suppose you're estimating the mean height of students in a school.
    2. You take random samples of increasing sizes, say 10 students, 50 students, 100 students, and so on.
    3. With each increase in sample size, the mean height calculated from the sample should approach the true mean height of all students in the school.
    4. If this happens, the estimator for the mean height is considered consistent.

ii) Unbiasedness:

1) The most efficient estimator is considered the one which has the least unbiased and consistent variance among all the estimators considered.
2) The variance considers how dispersed the estimator is from the estimate.
3) The smallest variance should deviate the least when different samples are brought into place.
4) But, of course, this also depends on the distribution of the population.

5) Example:

1. Let's say you're estimating the average score of students in a class.
2. You take multiple random samples and calculate the mean score for each sample.
3. If, on average, these sample means are equal to the true average score of the entire class, then the estimator is unbiased.
4. This means that sometimes the estimate might be higher than the true value, and sometimes it might be lower, but on average, it hits the mark.

5. Answer the Following — Marks 20

Q.1. The data with regard to the cost of production of 8 different drugs and cost of ingredients and packaging cost, are as given below: (Marks 10)

Sr. no.	Cost of production (Rs.) (Y)	Cost of ingredients (In thousands of Rs.) (X_1)	Packaging Cost (Rs.) (X_2)
1	100	17	19
2	79	50	54
3	100	90	75
4	129	30	36
5	158	15	16
6	106	20	25
7	58	20	24
8	78	50	53

a) Fit the regression $\widehat{Y} = a + b_1X_1 + b_2X_2$ (Marks 5)

b) Find the coefficient of multiple determination (R_2). (Marks 2)

c) Also test the significance of regression (Given F = 5.786, for a Significance level of $\alpha = 0.05$) (Marks 3)

$$\therefore \sum Y = 808 \qquad \therefore \sum X_1 = 292 \qquad \therefore \sum X_2 = 302 \quad \therefore n = 8$$

i)

$$\sum Y = na + b_1 \sum X_1 + b_2 \sum X_2 \quad \text{..... eq(1)}$$

$$\sum X_1 Y = a \sum X_1 + b_1 \sum {X_1}^2 + b_2 \sum X_1 X_2 \quad \text{..... eq(2)}$$

$$\sum X_2 Y = a \sum X_2 + b_1 \sum X_1 X_2 + b_2 \sum {X_2}^2 \quad \text{..... eq(3)}$$

Y	X_1	X_2	X_1Y	X_2Y	${X_1}^2$	${X_2}^2$	X_1X_2
100	17	19	1700	1900	289	361	323
79	50	54	3950	4266	2500	2916	2700
100	90	75	9000	7500	8100	5625	6750
129	30	36	3870	4644	900	1296	1080
158	15	16	2370	2528	225	256	240
106	20	25	2120	2650	400	625	500
58	20	24	1160	1392	400	576	480
78	50	53	3900	4134	2500	2809	2650
$Y = 808$	$X_1 = 292$	$X_2 = 302$	$X_1Y = 28070$	$X_2Y = 29014$	${X_1}^2 = 15314$	${X_2}^2 = 14464$	$X_1X_2 = 14723$

$808 = 8a + 292b_1 + 302b_2$ ..... eq(1)

$28070 = 292a + 15314b_1 + 14723b_2$ ..... eq(2)

$29014 = 302a + 14723b_1 + 14464b_2$ ..... eq(3)

$\therefore a = 137.55$

$\therefore b_1 = 2$

$\therefore b_2 = -2.90$

$\boldsymbol{\hat{Y} = 137.55 + 2X_1 - 2.90X_2}$

ii) **R^2 =**

$$R^2 = \frac{\sum(Y_i - \bar{Y})^2 - \sum(Y_i - \hat{Y})^2}{\sum(Y_i - \bar{Y})^2}$$

$$\bar{Y} = \frac{\sum Y}{n} = \frac{808}{8} = 101$$

$\boldsymbol{Y_i}$	$\boldsymbol{Y_i - \bar{Y}}$	$\boldsymbol{\hat{Y}}$	$\boldsymbol{Y_i - \hat{Y}}$	$\boldsymbol{(Y_i - \bar{Y})^2}$	$\boldsymbol{(Y_i - \hat{Y})^2}$
100	-1	131.84	-31.84	1	1013.78
79	-22	124.69	-24.69	484	609.59
100	-1	160.8	-60.8	1	3696.64
129	28	122.31	-22.31	784	497.73
158	57	134.11	-34.11	3249	1163.49
106	5	125.3	-25.3	25	640.09
58	-43	127.39	-27.39	1849	750.21
78	-23	126.78	-26.78	529	717.16
$Y_i = 808$	$Y_i - \bar{Y} = 0$	$\hat{Y} = 1053.22$	$Y_i - \hat{Y} = -253.22$	$(Y_i - \bar{Y})^2 = 6922$	$(Y_i - \hat{Y})^2 = 9088.69$

$$\therefore R^2 = \frac{6922 - 9088.69}{6922}$$

$\therefore \boldsymbol{R^2 = -0.3130}$

iii) F_α *at* 0.05 *level of significance is* 5.786.

$F_\alpha = 5.786$

$$F = \frac{\frac{\sum(Y_i - \hat{Y})^2}{p}}{\frac{\sum(Y_i - \bar{Y})^2}{n - p - 1}}$$

Where, P is an independent variable(b's).

$p = 2.$

$$\therefore F = \frac{\frac{9088.69}{2}}{\frac{6922}{8 - 2 - 1}}$$

$$= \frac{4544.345}{1384.4}$$

$$= 3.282$$

$\therefore F = 3.282$

$\therefore F < F_\alpha$, $3.282 < 5.786$ then regression model is not significant.

Q.2. What is hypothesis testing? (Marks 10)

i) Z-Test for Single Mean

ii) Z-Test for Difference of Mean

1) Hypothesis Testing is a type of statistical analysis in which you put your assumptions about a population parameter to the test.
2) It is used to estimate the relationship between 2 statistical variables.
3) **Z-Test for single Mean:**
   i) The z-test is a statistical test used to determine if a sample mean is significantly different from a known population mean.
   ii) It is used when the population standard deviation is known.
   iii) The formula for the single mean z-test is:

$$|\boldsymbol{Z}| = \frac{\bar{\boldsymbol{X}} - \boldsymbol{\mu}}{\boldsymbol{S} \cdot \boldsymbol{E}(\bar{\boldsymbol{X}})} \; \boldsymbol{or} \; |\boldsymbol{Z}| = \frac{\bar{\boldsymbol{X}} - \boldsymbol{\mu}}{\frac{\boldsymbol{\sigma}}{\sqrt{\boldsymbol{n}}}}$$

$$\boldsymbol{S} \cdot \boldsymbol{E}(\bar{\boldsymbol{X}}) = \frac{\boldsymbol{\sigma}}{\sqrt{\boldsymbol{n}}}$$

Were,

$|Z| =$ The Z-Statistic
$n =$ The Sample Size
$\bar{X} =$ The Sample Mean
$\mu =$ The Population Mean
$\sigma =$ The Population Standard Deviation
$S \cdot E =$ The Standard Error

4) **Z-Test for Difference of Mean:**
   i) A z-test is a statistical test to determine whether two population means are different or to compare one mean to a hypothesized value when the variances are known and the sample size is large.
   ii) A z-test is a hypothesis test for data that follows a normal distribution.
   iii) The formula for the double mean z-test is:

$$|\boldsymbol{Z}| = \frac{\bar{\boldsymbol{X}}_{\boldsymbol{1}} - \bar{\boldsymbol{X}}_{\boldsymbol{2}}}{\sqrt{\frac{{\boldsymbol{\sigma}_{\boldsymbol{1}}}^{\boldsymbol{2}}}{\boldsymbol{n}_{\boldsymbol{1}}} + \frac{{\boldsymbol{\sigma}_{\boldsymbol{2}}}^{\boldsymbol{2}}}{\boldsymbol{n}_{\boldsymbol{2}}}}}$$

Were,

$|Z| =$ The Z-Statistics for the two groups
$\bar{X}_1 \; or \; \bar{X}_2 =$ The sample means of the two groups
$n_1 \; or \; n_2 =$ The Samples sizes of the two groups
$\sigma_1 \; or \; \sigma_2 =$ The Population Standard Deviation of the two groups

6. Answer the Following Marks 20

Q.1. Explain the method of maximum likelihood estimation. (Marks 10)

1) Maximum Likelihood Estimation (MLE) is a statistical method used to estimate the parameters of a probability distribution.
2) Here's how the method of maximum likelihood estimation works:

1. Formulate the Likelihood Function:

1) Given a statistical model with parameters θ, the likelihood function, denoted as $L(\theta \mid x)$,measures the probability of observing the given sample data x for different values of the parameters θ.
2) For independent and identically distributed (i.i.d.) data, the likelihood function is often represented as the product of the probability density functions (pdf) or probability mass functions (pmf) of the individual data points:

$$L(\theta \mid x) = f(x_1, \theta) \times f(x_2, \theta) \times \ldots\ldots\ldots\ldots \times f(x_n, \theta)$$

3) Alternatively, it can be expressed as the joint probability density function (pdf) or probability mass function (pmf) of the entire sample x:

$$L(\theta \mid x) = f(x, \theta)$$

2. Maximize the Likelihood Function:

1) The goal of MLE is to find the values of the parameters θ, that maximize the likelihood function $L(\theta \mid x)$.
2) This is typically done by taking the derivative of the likelihood function with respect to each parameter, setting the derivatives equal to zero, and solving for the parameter values.
3) In some cases, it might be more convenient to maximize the log-likelihood function $(In(L(\theta \mid x)))$ instead, as it simplifies the computations and does not change the location of the maximum.

3. Estimate the Parameters:

1) Once the maximum likelihood estimates of the parameters are obtained, they are used as point estimates for the true parameter values.
2) These estimates are denoted as $\widehat{\theta}\, MLE$ and are often accompanied by standard errors or confidence intervals to quantify the uncertainty associated with the estimates.

4. Assess the Model:

1) After obtaining the parameter estimates, it's essential to assess the goodness-of-fit of the model to the data.
2) This can be done using various diagnostic tools, such as residual analysis, goodness-of-fit tests, and graphical methods.

5. Interpretation:

1) The maximum likelihood estimates provide the parameter values that make the observed data the most likely under the assumed statistical model.
2) These estimates are asymptotically efficient, meaning that as the sample size increases, they approach the true parameter values with high probability.

Q.2. Explain the Neyman Pearson Lemma. (Marks 10)

1) The Neyman-Pearson Lemma gives strong guidance about how to choose hypothesis tests.
2) The Neyman-Pearson Lemma is an important result that gives conditions for a hypothesis test to be uniformly most powerful.
3) That is, the test will have the highest probability of rejecting the null hypothesis while maintaining a low false positive rate of α.
4) More formally, consider testing two simple hypotheses:

$$H_0: \theta = \theta_0$$

$$H_1: \theta = \theta_1$$

5) The Neyman-Pearson Lemma says a test is uniformly most powerful test among α-level tests if it rejects H_0 if and only if

$$\frac{fx(x;\theta_1)}{fx(x;\theta_0)} > k$$

for some $k \epsilon R$, were

$$\alpha = P_{\theta_0}\left[\frac{fx(x;\theta_1)}{fx(x;\theta_0)} > k\right]$$

December 2022

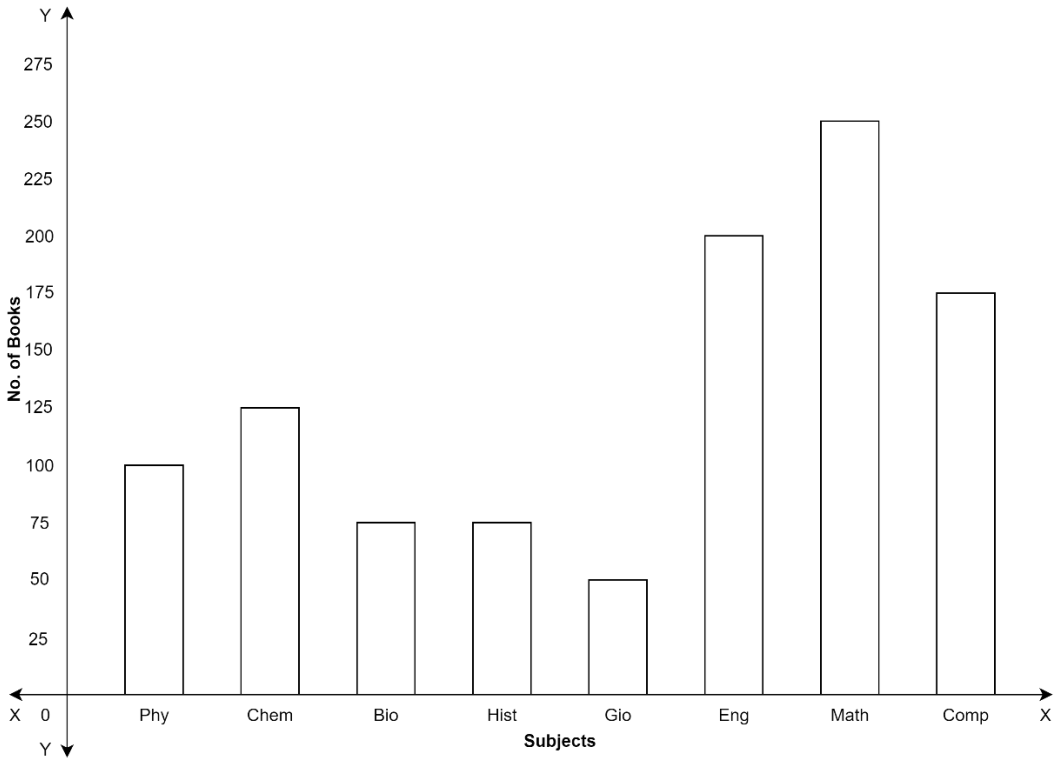
Q.1. Solve the Following: Marks 20

a) Explain Bar chart with following Example: (Marks 5)

Ex. The following table shows the number of books of different subject in library.

Subject	Phy	Chem	Bio	Hist	Gio	Eng	Math	Comp
No. of Books	100	125	75	75	50	200	250	175

1) A bar chart or bar graph is a chart or graph that presents categorical data with rectangular bars with heights or lengths proportional to the values that they represent. The bars can be plotted vertically or horizontally.
2) There are some types of a Bar Chart or Graph:
   - Simple bar diagram
   - Percentage bar diagram
   - Multiple bar diagram
   - Subdivided/Component bar diagram
   - Deviation bar diagram
   - Broken bar diagram



b) Equations of the two lines of regression are: $x + 6y = 6$ and $3x + 2y = 10$
Find: (Marks 5)

i) Mean of X and Man of Y
ii) Regression coefficients b_{yx} and b_{xy}
iii) Correlation coefficient between X and Y

1) Mean of X and Y:

$x + 6y = 6$ ……..eq(1)

$3x + 2y = 10$ ……..eq(2)

Solving equation 1 and 2, we get,

$\therefore x = 3 \; or \; y = \frac{1}{2}$

$\therefore$ The mean of x and y is $\bar{X} = 3, \bar{Y} = \frac{1}{2}$

2) Regression Coefficients:

1. Regression coefficient of Y on X:

$x + 6y = 6$

$6y = -x + 6$

$y = -\frac{1}{6}x + \frac{1}{6}$

$y = -\frac{1}{6}$

$\therefore$ The regression coefficient of Y on X is $b_{yx} = -\frac{1}{6}$

2. Regression coefficient of X on Y:

$3x + 2y = 10$

$3x = 10 - 2y$

$x = \frac{1}{3}(10 - 2y)$

$x = \frac{10}{3} - \frac{2}{3}y$

$\therefore x = -\frac{2}{3}$

$\therefore$ The regression coefficient of X on Y is $b_{xy} = -\frac{2}{3}$

3) Correlation coefficient between X and Y:

$r = \pm\sqrt{b_{yx} \times b_{xy}}$

$= \pm\sqrt{-\frac{1}{6} \times -\frac{2}{3}}$

$= \pm\sqrt{\frac{1}{3}}$

$= \pm 0.3333$

The Both regression coefficients are negative. So, take negative sign:

$\therefore r = -0.3333$

c) In a certain trivariate distribution: $r_{12} = 0.7, r_{23} = 0.6, r_{31} = 0.6$ find the partial correlation coefficient $r_{12.3}$. (Marks 5)

$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{(1 - {r_{13}}^2)(1 - {r_{23}}^2)}}$$

$$= \frac{(0.7) - (0.6) \times (0.6)}{\sqrt{(1 - (0.6)^2)(1 - (0.6)^2)}}$$

$$= \frac{0.7 - 0.36}{\sqrt{(1 - 0.36)(1 - 0.36)}}$$

$$= \frac{0.34}{\sqrt{0.64 \times 0.64}}$$

$$= \frac{0.34}{0.64}$$

$$= 0.53125$$

$\therefore r_{12.3} = 0.53125$

d) A survey conducted over the last 25 years indicated the in 10 years the winter was mild, in 8 years it was cold and int the remaining 7 years it was very cold. A company sells 1000 woollen coats in a mild year, 1300 in a cold year and 2000 in a very cold year. You are required to find the yearly expected profit of the company if a woollen coat costs Rs. 1730 and it is sold to stores for Rs. 2480. (Marks 5)

1) Find expected profit of the company if a woollen coat profit costs Rs. 1730 and it is sold to stores for Rs. 2480:

   Profit = woollen coat cost – sold for stores cost

   = 2480 – 1730

   = 750

   $\therefore Profit = 750\ Rs.$

2) Calculate total no. of coats sell in each year:
   - Mild Year: $10 \times 1000 = 10000\ coats.$
   - Cold Year: $8 \times 1300 = 10400\ coats.$
   - Very Cold Year: $7 \times 2000 = 14000\ coats.$
3) Calculate Total Coats sells in 25 Years:

   $Total\ Coats: 10000 + 10400 + 14000 = 34400\ Coats$
4) Calculate Total profit of the coats sell in 25 years:

   $Total\ Cost = 34400\ \times 750 = 25800000\ Rs.$
5) Calculate expected profit for each year:

   $Per\ Year\ Profit = 25800000\ \div 25 = 1032000\ Rs.$

   ∴ So, the Yearly expected profit of the company is $1032000\ Rs.$

Q.2. Solve the Following: Marks 20

a) Define the term "Statistics" and discuss its use in business and trade. Also point out its limitations. (Marks 10)

1) Statistics is defined as collection, compilation analysis and interpretation of numerical data.
2) Statistic is a science of data.
3) Statistics helps in gathering information about the appropriate quantitative data.
4) It depicts the complex data in graphical form, tabular form and in diagrammatic representation to understand it easily.
5) It provides the exact description and a better understanding.
6) **Use of statistics in a business or trade:**

Business:

1. Market Research
2. Demand Forecasting
3. Financial Analysis
4. Quality Control
5. Risk Management
6. Performance Measurement

Trade:

1. Market Analysis
2. Risk Assessment
3. Technical Analysis
4. Algorithmic trading
5. Performance Evaluation

7) **Limitations of Statistics:**

- For quantitative results, statistical approaches are best applicable.
- On heterogeneous data, statistics cannot be implemented.
- In gathering, analyzing, and interpreting the data, if adequate care is not taken, statistical findings can be misleading.
- Statistical data can be treated effectively only by a person who has professional knowledge of statistics.
- In statistical judgments, certain errors are possible.

b) What are the various methods of collecting statistical data? Which of these is most reliable and why? (Marks 10)

Parameter	Primary Data	Secondary Data
1. Meaning	Data collected by researcher itself.	Data collected by other people.
2. Originality	Original and Unique Information.	Not original and unique information.
3. Adjustment	Does not need adjustment, is focused.	Need adjustment to suit actual aim.
4. Sources	Observations, Surveys, Experiment.	Internal Records, govt. published Data, etc.
5. Type of Data	Qualitative Data	Quantitative Data
6. Methods	Observation, experiment, interview, etc.	Desk research method, searching online, etc.
7. Reliability	More reliable	Less reliable
8. Capability	More capable to solve a problem.	Less capable to solve a problem.
9. Time consumed	More time consuming	Less time consuming
10. Cost-effectiveness	Costly	Economical
11. Suitability	More suitable	May or may not be suitable
12. Need of Investigators	Needs team of trained Investigators.	Does not need of team Investigators.
13. Collected when	Secondary data is inadequate.	Before primary data is collected.

Q.3. Solve the Following: Marks 20

a) Find the Mean Deviation from the Median for the following data. (Marks 10)

Age of Workers	**20-25**	**25-30**	**30-35**	**35-40**	**40-45**	**45-50**	**50-55**	**55-60**
No. of Workers	**120**	**125**	**175**	**160**	**150**	**140**	**100**	**30**

Age of Workers	No. of Workers f_i	Cumulative Frequency $(c.f.)$	Midpoint x_i	$x_i - median$	$f_i \cdot (x_i - median)$
20-25	120	120	22.5	15	1800
25-30	125	245	27.5	10	1250
30-35	175	420	32.5	5	875
35-40	160	580	37.5	0	0
40-45	150	730	42.5	5	750
45-50	140	870	47.5	10	1400
50-55	100	970	52.5	15	1500
55-60	30	1000	57.5	20	600
Total	1000				8175

$\therefore N = 1000$

$\therefore \frac{N}{2} = \frac{1000}{2} = 500$

$\therefore$ Therefore, 35 – 40 is the medial class.

$$\therefore Median = l + \frac{\frac{N}{2} - C}{f} \times h$$

Here,

$N = 1000$

$l = 35\ (lowest\ age\ of\ median\ class)$

$C = 420\ (previous\ cumulative\ frequency\ of\ median\ class)$

$f = 160\ (current\ frequency\ of\ median\ class)$

$h = 5\ (range\ of\ per\ class)$

$$\therefore Median = 35 + \frac{\frac{1000}{2} - 420}{160} \times 5$$

$$= 35 + \frac{500 - 420}{160} \times 5$$

$$= 35 + 2.5$$

$$= 37.5$$

$\therefore Median = 37.5$

Let's, find Mean deviation from the median:

$$M.D.M = \frac{1}{N}\sum_{i=1}^{8} f_i x_i - M$$

$$= \frac{1}{1000} \times 8175$$

$$= 8.175$$

$\therefore M.D.M = 8.175$

b) A survey of 370 students from Commerce faculty and 130 students from Science Faculty revealed that 180 students were studying for only C.A. Examinations, 140 for only Costing Examinations and 80 for both C.A. and Costing Examinations. The rest had offered part-time Management Courses, of those studying for Costing only, 13 were girls and 90 boys belonged to Commerce faculty. Out of 80 studying for both C.A. and Costing, 72 were from Commerce Faculty amongst which 70 were boys. Amongst those who offered part-time Management Courses, 50 boys were from Science Faculty and 30 boys and 10 girls form Commerce faculty. In all three were 110 boys in Science Faculty. Present the above information in a tabular form. Find the number of students form Science Faculty studying for part-time Management Courses. (Marks 10)

	Commerce			Science			Total		
	Boys	Girls	Total	Boys	Girls	Total	Boys	Girls	Total
C.A	130	25	155	21	4	25	151	29	180
Costing	90	13	103	35	2	37	125	15	140
Both	70	2	72	4	4	8	74	6	80
Management	30	10	40	50	10	60	80	20	100
Total	320	50	370	110	20	130	430	70	500

Q.4. Solve the Following: Marks 20

a) A department store gives in-service training to its salesmen which is followed by a test. It is considering whether it should terminate the service of any salesmen who does not do well in the test. The following data give the test scores and sales made by nine salesmen during a certain period:

Test Scores	**14**	**19**	**24**	**21**	**26**	**22**	**15**	**20**	**19**
Sales ("00 Rs.")	**31**	**36**	**48**	**37**	**50**	**45**	**33**	**41**	**39**

Calculate the coefficient of correlation between the test scores and the sales. Does it indicate that the termination of services of low-test scores is justified? If the firm wants a minimum sales volume of Rs. 30,000, what is the minimum test score that will ensure continuation of service? Also estimate that most probable sales volume of a salesmen making a score of 28. (Marks 10)

$\therefore \bar{X} = \frac{\sum X}{n} = \frac{180}{9} = 20 \qquad \therefore \bar{X} = 20$

$\therefore \bar{Y} = \frac{\sum Y}{n} = \frac{360}{9} = 40 \qquad \therefore \bar{Y} = 40$

X	Y	$x = X - 20$	$y = Y - 40$	x^2	y^2	xy
14	31	- 6	-9	36	81	54
19	36	- 1	-4	1	16	4
24	48	4	8	16	64	32
21	37	1	-3	1	9	-3
26	50	6	10	36	100	60
22	45	2	5	4	25	10
15	33	- 5	-7	25	49	35
20	41	0	1	0	1	0
19	39	- 1	-1	1	1	1
$X = 180$	$Y = 360$	$x = 0$	$y = 0$	$x^2 = 120$	$y^2 = 346$	$xy = 193$

1) Regression Coefficients:

1. Regression Coefficient X on Y:

$$b_{xy} = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - \sum(x)^2}$$
$$= \frac{9 \times 193 - 0 \times 0}{9 \times 120 - 0^2}$$
$$= \frac{1737}{1080}$$
$$= 1.60$$
$$\therefore b_{xy} = 1.60$$

2. Regression Coefficient Y on X:

$$b_{yx} = \frac{n\sum xy - \sum x \sum y}{n\sum y^2 - \sum(y)^2}$$
$$= \frac{9 \times 193 - 0 \times 0}{9 \times 346 - 0^2}$$

$$= \frac{1737}{3114}$$
$$= 0.55$$
$$\therefore b_{yx} = 0.55$$

2) Regression Equations:

1. Regression Equation X on Y:

$\therefore X - \bar{X} = b_{xy}(Y - \bar{Y})$

$X - 20 = 1.60(Y - 40)$

$X - 20 = 1.60Y - 64$

$X = 1.60Y + 44$

∴ The regression equation of X on Y is $X = 1.60Y + 44$

2. Regression Equation Y on X:

$\therefore Y - \bar{Y} = b_{yx}(X - \bar{X})$

$Y - 40 = 0.55(X - 20)$

$Y - 40 = 0.55X - 11$

$Y = 0.55X + 29$

∴ The regression equation of Y on X is $Y = 0.55X + 29$

3) Coefficient Correlation:

$\therefore r = \pm\sqrt{b_{xy} \times b_{yx}}$

$= \pm\sqrt{1.60 \times 0.55}$

$= \pm\sqrt{0.88}$

$= \pm 0.93$

Both regression equation is positive. So, take positive sign:

$\therefore r = 0.93$

∴ Hence, the termination of services for low test scores is justified.

4) Find Test score:

Sales = $1.60 \times Test\ Score + 44$

$30000 = 1.60 \times Test\ Score + 44$

$1.60 \times Test\ Score = 30000 - 14$

$1.60 \times Test\ Score = 29986$

$\frac{29986}{1.60} = Test\ Score$

$\therefore Test\ Score = 18.741 \approx 18$

5) When $X = 28, Y =$?

$Y = 0.55X + 29$

$Y = 0.55(28) + 29$

$\therefore Y = 44.4$

b) Define a random variable and its mathematical expectation. (Marks 10)

1) The various outcomes of a random experiment is denoted with the help of a variable which is called a random variable.
2) For example: In case of throwing a die, we may use a variable X for representing the outcome of the throw. Thus, X will take the values 1, 2, 3, 4, 5 and 6.
3) But in some cases, the outcomes may be qualitative e.g. tossing of a coin which may be head or tail, the colours of balls drawn from an urn may be red, yellow, white etc.
4) But for mathematical convenience the qualitative outcomes may be expressed in quantitative forms. For example, in tossing of a coin we may denote the outcome 'Head' by 1 and 'Tail' by 0.
5) In this way each outcome of a random experiment, whether it is qualitative or quantitative, can be expressed by a real number.
6) There are two types of random variables:
   (a) Discrete random variable
   (b) (b) Continuous random variable
7) **Mathematical Expectation:**

Let us consider a discrete random variable X which assumes the values $x_1, x_2, \ldots, x_n$ with respective probabilities $p_1, p_2, \ldots, p_n$, such that $\sum p_i = 1$, then the mathematical expectation of the random variable X is given by the sum of the products of the different values of X with their corresponding probabilities. The expectation of a random variable is generally denoted by $E(X)$.

Thus, $E(X) = \sum_{i=1}^{n} x_i \times P(X = x_i) = \sum_{i=1}^{n} p_i x_i$ provided the series is convergent and $\sum p_i = 1$.

In case the discrete random variable takes countably infinite number of values then we have

$$E(X) = \sum_{i=1}^{n} x_i \times P(X = x_i) = \sum_{i=1}^{n} p_i x_i$$

If X is a continuous random variable with probability density function $f(x), -\infty < x < \infty$ Then the mathematical expectation of the random variable X is given by

$E(X) = \int_{-\infty}^{\infty} x\, f\,(x)dx$ provided $\int_{-\infty}^{\infty} f\,(x)dx = 1$

The expectation of the random variable X serves as the measure of central tendency of the probability distribution of X.

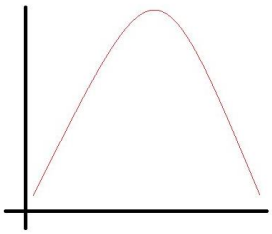
Q.5. Solve the Following: Marks 20

a) Write a detailed note on least square regression. (Marks 10)

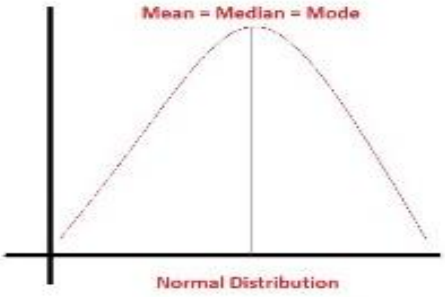
1) Least square regression is a technique that helps you draw a line of best fit depending on your data points.
2) The line is called the least square regression line, which perfectly depicts the changes in your y (response) variables and their corresponding x (explanatory) variable.
3) The line that we draw through the scatterplots does not have to pass through all the plotted points, provided there is a perfect linear relationship between the variables.
4) Equation of least square regression line: $\hat{y} = a + bx$

b) What is the test of skewness? (Marks 10)

1) Skewness is a measure of lack of symmetry, i.e. it measures the deviation of the given distribution of a random variable from a symmetric distribution.
2) **Normal Distribution:**
    1. A Normal Distribution is a probability distribution that is symmetric about the mean.
    2. It is also known as a Gaussian Distribution.
    3. The distribution appears as a Bell-shaped curve**,** which means the mean is the most frequent data in the given data set.

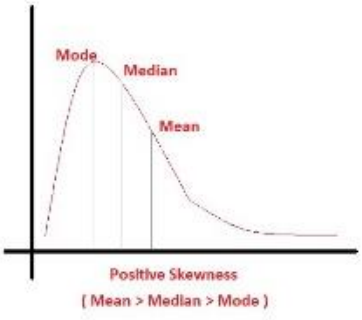


    4. In Normal Distribution:

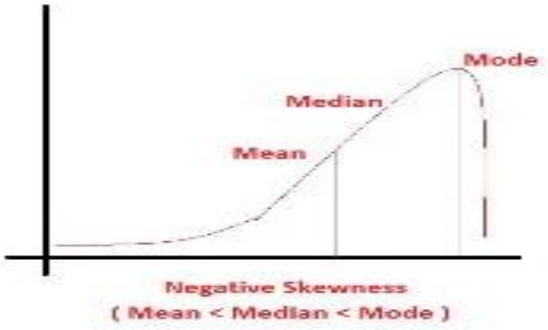


Mean = Median = Mode

3) **Standard Normal Distribution:** When the mean in a Normal Distribution is 0 and the Standard Deviation is 1, then the Normal Distribution is called a Standard Normal Distribution.
4) **Types of Skewness:**
    - **Positive Skewness:**
        1. In positive skewness, the extreme data values are larger, which in turn increases the mean value of the data set.
        2. In Positive Skewness:
           Mean > Median > Mode



- **Negative Skewness:**
  1. In negative skewness, the extreme data values are smaller, which decreases the mean value of the dataset.
  2. In Negative Skewness:
     Mean < Median < Mode



5) Unlike the Normal Distribution (mean = median = mode), in positive and negative skewness, the mean, median, and mode are all different.

Q.6. Solve the Following: Marks 20

a) Explain the following point Estimation Properties with example. (Marks 10)

i) Consistency

1) It states that the estimator stays close to the parameter's value as the population's size increases.
2) Thus, a large sample size is required to maintain its consistency level.
3) When the expected value moves towards the parameter's value, we state that the estimation is consistent.
4) **Example:**
   i. Suppose you're estimating the mean height of students in a school.
   ii. You take random samples of increasing sizes, say 10 students, 50 students, 100 students, and so on.
   iii. With each increase in sample size, the mean height calculated from the sample should approach the true mean height of all students in the school.
   iv. If this happens, the estimator for the mean height is considered consistent.

ii) Unbiasedness

1) The most efficient estimator is considered the one which has the least unbiased and consistent variance among all the estimators considered.
2) The variance considers how dispersed the estimator is from the estimate.
3) The smallest variance should deviate the least when different samples are brought into place.
4) But, of course, this also depends on the distribution of the population.
5) **Example:**
   i. Let's say you're estimating the average score of students in a class.
   ii. You take multiple random samples and calculate the mean score for each sample.

iii. If, on average, these sample means are equal to the true average score of the entire class, then the estimator is unbiased.

iv. This means that sometimes the estimate might be higher than the true value, and sometimes it might be lower, but on average, it hits the mark.

b) What is Hypothesis testing? For large samples explain (Marks 10)

i) Test of Significance for a single mean

ii) Test of Significance of difference between two means

1) Hypothesis Testing is a type of statistical analysis in which you put your assumptions about a population parameter to the test.
2) It is used to estimate the relationship between 2 statistical variables.
3) **Z-Test for single Mean:**
   1. The z-test is a statistical test used to determine if a sample mean is significantly different from a known population mean.
   2. It is used when the population standard deviation is known.
   3. The formula for the single mean z-test is:

$$|\boldsymbol{Z}| = \frac{\bar{\boldsymbol{X}} - \boldsymbol{\mu}}{\boldsymbol{S} \cdot \boldsymbol{E}(\bar{\boldsymbol{X}})} \; \boldsymbol{or} \; |\boldsymbol{Z}| = \frac{\bar{\boldsymbol{X}} - \boldsymbol{\mu}}{\frac{\boldsymbol{\sigma}}{\sqrt{\boldsymbol{n}}}}$$

$$\boldsymbol{S} \cdot \boldsymbol{E}(\bar{\boldsymbol{X}}) = \frac{\boldsymbol{\sigma}}{\sqrt{\boldsymbol{n}}}$$

Were,

$|Z| =$ The Z-Statistic

$n =$ The Sample Size

$\bar{X} =$ The Sample Mean

$\mu =$ The Population Mean

$\sigma =$ The Population Standard Deviation

$S \cdot E =$ The Standard Error

4) **Z-Test for Difference of Mean:**
   1. A z-test is a statistical test to determine whether two population means are different or to compare one mean to a hypothesized value when the variances are known and the sample size is large.
   2. A z-test is a hypothesis test for data that follows a normal distribution.
   3. The formula for the double mean z-test is:

$$|\boldsymbol{Z}| = \frac{\bar{\boldsymbol{X}}_{\boldsymbol{1}} - \bar{\boldsymbol{X}}_{\boldsymbol{2}}}{\sqrt{\frac{{\boldsymbol{\sigma}_{\boldsymbol{1}}}^{\boldsymbol{2}}}{\boldsymbol{n}_{\boldsymbol{1}}} + \frac{{\boldsymbol{\sigma}_{\boldsymbol{2}}}^{\boldsymbol{2}}}{\boldsymbol{n}_{\boldsymbol{2}}}}}$$

Were,

$|Z| =$ The Z-Statistics for the two groups

$\bar{X}_1 \; or \; \bar{X}_2 =$ The sample means of the two groups

$n_1 \; or \; n_2 =$ The Samples sizes of the two groups

$\sigma_1 \; or \; \sigma_2 =$ The Population Standard Deviation of the two groups

May 2022

Q.1. In a simple study about coffee habits in two towns A and B the following information is given
Town A: Females were 40%, total coffee drinkers were 45% and female non coffee drinkers were 20%.
Town B: Males were 55%, male non coffee drinkers were 30% and female coffee drinkers were 15%.
Present the data into a table format.

	Town A			Town B			Total
	Male	Female	Total	Male	Female	Total	
Coffee drinkers	25	20	45	25	15	40	95
Non coffee drinkers	35	20	55	30	30	60	115
Total	60	40	100	55	45	100	200

Q.2. Perform simple linear regression, Determine slope and intercept. (Marks 10)

X	1	2	3	3	4	5
y	8	4	5	2	2	0

x	y	x^2	y^2	xy
1	8	1	64	8
2	4	4	16	8
3	5	9	25	15
3	2	9	4	6
4	2	16	4	8
5	0	25	0	0
$x = 18$	$y = 21$	$x^2 = 64$	$y^2 = 113$	$xy = 45$

Let find a slope:

$$b(slope) = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2}$$
$$= \frac{6 \times 45 - 18 \times 21}{6 \times 64 - 18^2}$$
$$= \frac{270 - 378}{384 - 324}$$
$$= \frac{-108}{60}$$
$$= -1.8$$

$\therefore$ Slope is $b = -1.66$.

Let find an intercept:

$$a(intercept) = \frac{\sum y \sum x^2 - \sum x \sum xy}{n(\sum x^2) - (\sum x)^2}$$
$$= \frac{21 \times 64 - 18 \times 45}{6 \times 64 - 18^2}$$
$$= \frac{1344 - 810}{384 - 324}$$
$$= \frac{534}{60}$$

$$= 8.9$$

∴ An intercept is $a = 8.9$

Simple linear regression formula is given by,

$$\boldsymbol{y = a + bx}$$

∴ Slope $b = -1.8$ and intercept $a = 8.9$

∴ $y = 8.9 - 1.8x$

∴ The simple linear regression is $y = 8.9 - 1.66x$.

Q.3. What do you mean by a questionnaire? What is the difference between a questionnaire and a schedule? State the essential points to be remembered in drafting a questionnaire.

1) A questionnaire is a list of questions that ask respondents about themselves or others.

Basis	Questionnaire	Schedule
Meaning	A questionnaire is a research instrument used by any researcher as a tool to collect data or gather information from any source or subject of his or her interest from the respondents.	A schedule is a formalized arrangement of inquiries, proclamations, statements, and spaces for replies given to the enumerators who pose inquiries to the respondents and note down the responses.
Filled by	A questionnaire is filled by the respondents.	A schedule is filled by an enumerator.
Response Rate	The response rate of a questionnaire is low.	The response rate of a schedule is high.
Cost	It is economical in terms of time, effort, and money.	It is expensive in terms of time, effort, and money.
Coverage	A large area can be covered through a questionnaire.	Comparatively small areas can be covered through a schedule.
Respondent's Identity	The identity of the respondent is unknown.	As the enumerator visits the informant personally, his identity is known.
Dependency of Success	The success of a questionnaire depends upon its quality.	The success of a schedule depends upon the honesty and competence of the enumerator.
Usage	A questionnaire is used only when the people are literate and cooperative.	A schedule can be used in both cases when people are literate and illiterate.

Following few essential points to be remembered in drafting a questionnaire:

- The questionnaire flows.
- Keep it short & simple.
- Be neutral in questioning.
- Avoid Double barrelled questions.
- Don't assume respondents are experts.
- Avoid negatives or double negatives.

Q.4. What is Stratified sampling? Explain the merits and limitations of Stratified sampling.

1) Stratified sampling is a method of collecting data that involves dividing a large population into smaller subgroups, and there are various pros and cons of the stratified sampling method.
2) It's commonly used when conducting surveys or gathering statistical data.
3) It allows people to survey a large population but in a more manageable way.
4) For example, if you're surveying a university population about how satisfying their school experience has been, you may use this method to divide the students up by program.
5) This not only provides more in-depth data, but it also makes the large task easier to do.
6) **Merits of Stratified Sampling:**
    1. **More accurate data:**
       This method of data collection can allow for more accurate information.
    2. **More diverse data:**
       Having multiple subgroups within your sample population also allows you to collect a more diverse range of data.
    3. **More manageable:**
       Having subgroups within your population can also make the data, and the work of collecting the data, more manageable.
    4. **More cost-effective:**
       If you use stratified sampling, it can be a most cost-effective method of conducting a survey.
    5. **Prevents sample bias:**
       Stratified sampling allows researchers to examine their sample and build groups of participating who are free of bias.
7) **Limitations of Stratified Sampling:**
    a) **Lacks versality:**
       This method only works for studies that require sample populations and surveys
    b) **Difficult Data Analysis:**
       With more and more subgroups, there also comes a larger input of information. The information can be specific and intricate, and it can take a long time to analyze.
    c) **Requires more planning:**
       Creating a study using a stratified sampling method can require a significant amount of planning.

Q.5. What is diagrammatic representation of data? Explain its advantages.

1) Diagrammatic presentation is the visual form of presentation of data in which facts are highlighted in the language of diagrams.
2) It consists in presenting statistical material in interesting and attractive geometrical figures (Bars, Circle, Rectangle, Squares, Graphs, etc.), pictures, maps and charts etc.
3) It will attract the attention of a large number of persons.
4) It facilitates comparison between two or more sets of data.
5) **Advantages of Diagrammatic Representation:**
    1. Diagrams are attractive and impressive
    2. Diagrams facilitate comparison
    3. Diagrams simplify data
    4. Universal applicability
    5. Easy to remember

6. Diagrams save time for understanding
7. Diagrams provides more information

Q.6. The manufacturer of a certain make of electric bulbs claims that his bulbs have mean life of 25 months with standard deviation of 5 months. A random sample of 6 bulbs gave the following value:
Life of bulbs in months: 24, 26, 30, 20, 20, 18
Is the manufacturer's claim valid at 1% level of significance? (Given that the table values of the appropriate test statistics at said level are 4.032, 3.707, and 3.499 for 5, 6 and 7 degrees of freedom respectively). (Marks 10)

$\therefore n = 6, \mu = 25, \sigma = 5, \bar{X} = ?$

$$\bar{X} = \frac{\sum X}{n} = \frac{24 + 26 + 30 + 20 + 20 + 18}{6} = \frac{138}{6} = 23 \qquad \therefore \bar{X} = 23$$

Life of bulbs in months (X)	$x = X - 23$	x^2
24	1	1
26	3	9
30	7	49
20	-3	9
20	-3	9
18	-5	25
$X = 138$	$x = 0$	$x^2 = 102$

$$S = \sqrt{\frac{\sum x^2}{n}}$$

$$= \sqrt{\frac{102}{6}}$$

$\therefore S = 4.12$

1. **Hypothesis:**
   H_0 : Null Hypothesis
   $\mu = 25$
   Checks the manufacturer's claim is valid.
   H_1 : Alternative Hypothesis
   $\mu \neq 25$
   Checks the manufacturer's claim is not valid.
2. **Computation of test statistics:**

$$S \cdot E(\bar{X}) = \frac{S}{\sqrt{n-1}}$$

$$= \frac{4.12}{\sqrt{6-1}}$$

$$= \frac{4.12}{\sqrt{5}}$$

$$= 1.84$$

$$|t| = \frac{\bar{X} - \mu}{S \cdot E(\bar{X})}$$

$$= \frac{23 - 25}{1.84}$$

$$|t| = -1.08$$

$\therefore t = 1.08$

3. **Level of significance:**
   $\alpha = 0.01$
4. **Critical Value:**
   t_α at 1% Level of significance or degrees of freedom $n = n - 1 = 6 - 1 = 5$ is 4.032
   $t_\alpha = 4.032$
5. **Decision:**
   $\therefore t < t_\alpha$
   $\therefore 1.08 < 4.032$
   ∴ Null hypothesis is accepted and Alternative hypothesis is rejected.
   ∴ The manufacturer's claims is Valid.

Extra Questions:

Q.1. A random sample of 900 items is taken from a normal population who's the mean and variance are 4. Can the sample with mean 4.5 be regarded as truly random one at 1% level of significance? (Table value at 1% is 2.58). (Marks 5)

$\therefore n = 900, \bar{X} = 4.5, \mu = 4, S^2 = 4, LOS = 1\%$

$\therefore \sigma(S) = S^2 = \sqrt{4} = 2 \quad \therefore \sigma = 2$

1. **Hypothesis:**

   H_0 : Null Hypothesis

   $\mu = 4$

   H_1 : Alternative Hypothesis

   $\mu \neq 4$

2. **Test statistics:**

   $$\boldsymbol{S \cdot E(\bar{X}) = \frac{\sigma}{\sqrt{n}}}$$
   $$= \frac{2}{\sqrt{900}}$$
   $$= \frac{2}{30}$$
   $$= 0.06$$
   $$|\boldsymbol{z}| = \frac{\boldsymbol{\bar{X} - \mu}}{\boldsymbol{S \cdot E(\bar{X})}}$$
   $$= \frac{4.5 - 4}{0.063}$$
   $$|z| = 7.936$$
   $$\therefore z = 7.936$$

3. **Level of significance:**

   $\alpha = 0.01$

4. **Critical Value:**

   Z_α at 1% Level of significance is 2.58

   $Z_\alpha = 2.58$

5. **Decision:**

   $\therefore z > z_\alpha$

   $\therefore 7.936 > 2.58$

   $\therefore$ Null hypothesis is rejected and Alternative hypothesis is accepted.

Q.2. The height of 10 children selected at random from a given locality had a mean 63.2cms and variance 6.25cms. Test at 5% level of significance the hypothesis that the children of the given locality are on the average less than 65cms in all. Given for 9 degrees of freedom $(\boldsymbol{t > 1.83}) = \boldsymbol{0.5}$ (Marks 5)

$\therefore n = 10, \bar{X} = 63.5, \mu = 65, s^2 = 6.25, LOS = 5\%$

$\therefore s = s^2 = \sqrt{6.25} = 2.5 \quad \therefore s = 2.5$

1. **Hypothesis:**

   H_0 : Null Hypothesis

   $\mu \geq 65$

   H_1 : Alternative Hypothesis

   $\mu < 65$

2. **Test statistics:**

$$\boldsymbol{S \cdot E(\bar{X})} = \frac{\boldsymbol{s}}{\sqrt{\boldsymbol{n-1}}}$$
$$= \frac{2.5}{\sqrt{10-1}}$$
$$= \frac{2.5}{3}$$
$$= 0.833$$
$$|\boldsymbol{t}| = \frac{\boldsymbol{\bar{X} - \mu}}{\boldsymbol{S \cdot E(\bar{X})}}$$
$$= \frac{63.2 - 65}{0.833}$$
$$|t| = -2.16$$
$$\therefore t = 2.16$$

3. **Level of significance:**

$\alpha = 0.05$

4. **Critical Value:**

t_α at 5% Level of significance or degrees of freedom $n = n - 1 = 10 - 1 = 9$ is 1.83

$t_\alpha = 1.83$

5. **Decision:**

$\therefore t > t_\alpha$

$\therefore 2.16 > 1.83$

$\therefore$ Null hypothesis is rejected and Alternative hypothesis is accepted.

Q.3. Find y when $x_1 = 3700$ kg and $x_2 = 260$ km from least square regression equation of y in x_1 and x_2 for the following:

Y	160	112	69	90	123	186
x_1 (1000 kg)	4.0	2.0	1.6	1.2	3.4	4.8
x_2 (100 km)	1.5	2.2	1.0	2.0	0.8	1.6

$\therefore \sum Y = 740, \sum x_1 = 17, \sum x_2 = 9.1, n = 6$

Let, solve:

$\therefore \boldsymbol{\sum Y = na + b_1 \sum x_1 + b_2 \sum x_2}$ .....eq(1)

$\therefore \boldsymbol{\sum x_1 y = a \sum x_1 + b_1 \sum {x_1}^2 + b_2 \sum x_1 x_2}$ .....eq(2)

$\therefore \boldsymbol{\sum x_2 y = a \sum x_2 + b_1 \sum x_1 x_2 + b_2 \sum {x_2}^2}$ .....eq(3)

Y	x_1	x_2	${x_1}^2$	${x_2}^2$	$x_1 y$	$x_2 y$	$x_1 x_2$
160	4	1.5	16	2.25	640	240	6
112	2	2.2	4	4.84	224	246.4	4.4
69	1.6	1	2.56	1	110.4	69	1.6
90	1.2	2	1.44	4	108	180	2.4
123	3.4	0.8	11.56	0.64	418.2	98.4	2.72
186	4.8	1.6	23.04	2.56	892.8	297.6	7.68
$Y = 740$	$x_1 = 17$	$x_2 = 9.1$	${x_1}^2 = 58.6$	${x_2}^2 = 15.29$	$x_1 y = 2393.4$	$x_2 y = 1131.4$	$x_1 x_2 = 24.8$

$\therefore 740 = 6a + 17b_1 + 9.1b_2$ .....eq(1)

$\therefore 2393.4 = 17a + 58.6b_1 + 24.8b_2$ .....eq(2)

$\therefore 1131.4 = 9.1a + 24.8b_1 + 15.29b_2$ .....eq(3)

$\therefore a = -4.57$
$\therefore b_1 = 30.94$
$\therefore b_2 = 26.53$
$\therefore$ The regression line equation is $Y = -4.57 + 30.94X_1 + 26.53X_2$

When $X_1 = 3700$ and $X_2 = 260$, $Y =$?

$$Y = -4.57 + 30.94X_1 + 26.53X_2$$
$$= -4.57 + 30.94(3700) + 26.53(260)$$
$$= 121371.23$$
$$\therefore Y = 121371.23$$

Q.4.