Q.1)

Write a short note on:-

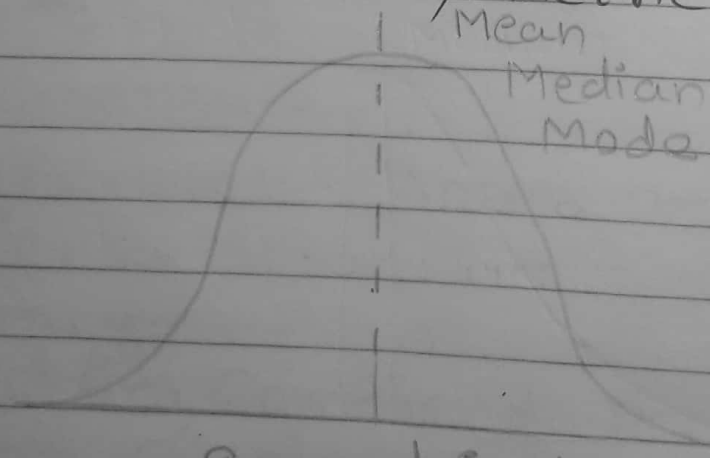
Q.1) a) Skewness and Kurtosis

→ i) The extent to which a distribution of data points is concentrated at one end or the other is known as 'Skewness'.

ii) The degree of peakedness of a distribution of points is called as 'Kurtosis'.

iii) Based on the three Central Tendency measures we can decide the skewness of the distribution.

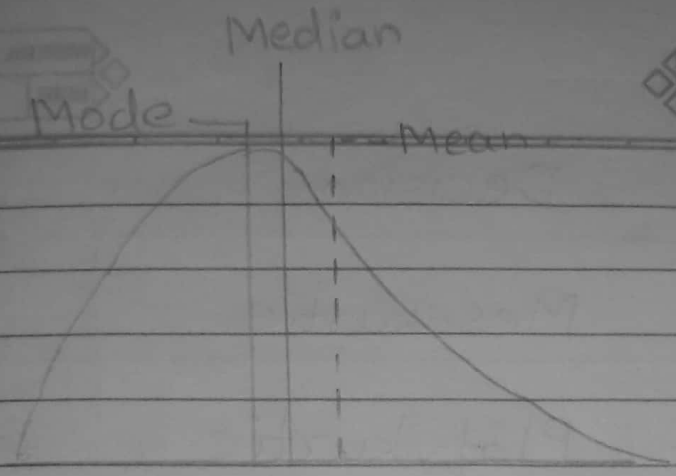
a) If mean = median = mode, then we can ensure that the shape of the distribution is Symmetric.



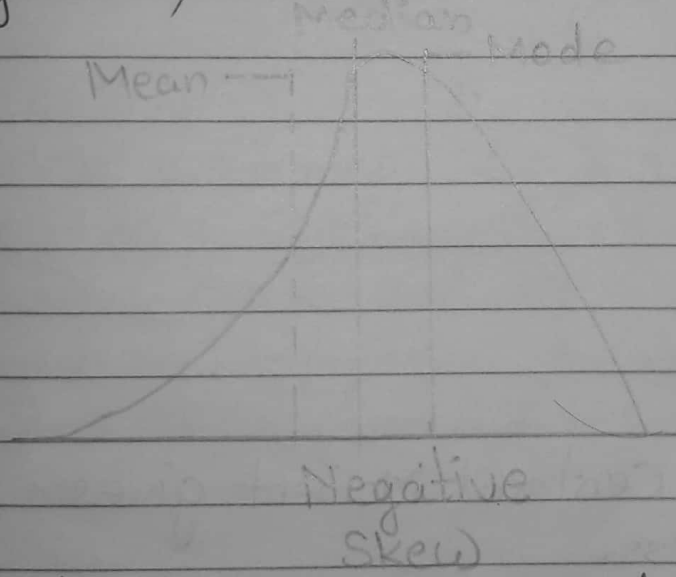
Symmetrical Distribution.

b) If mode < median < mean, then we can ensure that, the shape of the distribution trails to the right, is positively skewed.

Scanned by CamScanner



c) If mean < median < mode, the shape of the distribution trails to the left, is negatively skewed.



iv) Kurtosis is measured Based on the Moments

Formula's are as follows:-

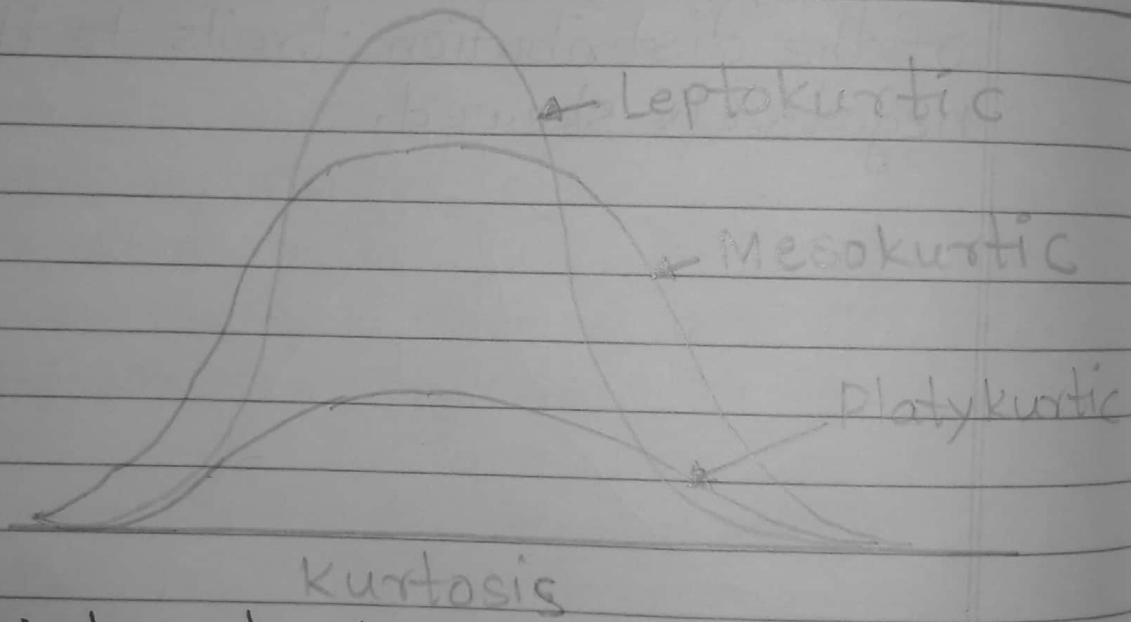
$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} \qquad \beta_2 = \frac{\mu_4}{\mu_2^2}$$

$\gamma_1 = \sqrt{\beta_1}$ (Gamma$_1$) Coefficient of Skewness

$\gamma_2 = \beta_2 - 3$ (Gamma$_2$) Coefficient of Kurtosis.

Scanned by CamScanner

If,	Decision	If,
a) $\beta_2 = 3$	Mesokurtic	$\gamma_2 = 0$
b) $\beta_2 < 3$	Platykurtic	$\gamma_2 < 0$
c) $\beta_2 > 3$	leptokurtic	$\gamma_2 > 0$



v) Third Central Moment gives us skeeoness.

vi) Whereas, Fourth central Moment gives us kurtosis.

vii) Two more popular measures of skeeoness are

a) Karl Pearson's Coefficient of skeeoness
b) Bowley's Coefficient of skeeoness.

a) Karl Pearson's Coefficient of Skeeoness

Formula:-

Scanned by CamScanner

$$S = \frac{\text{mean} - \text{mode}}{\text{S.D}} = \frac{\bar{x} - Z}{\sigma}$$

(Standard Deviation)

Note:- If mode is ill-defined, we can go for following formula;

$$S = \frac{3(\text{mean} - \text{median})}{\text{S.D}}$$

(Standard Deviation)

$$= \frac{3(\bar{x} - M)}{\sigma}$$

Since we know, $Z = 3M - 2\bar{x}$

b) Bowley's Coefficient of skewness:-
Bowley's coefficient of skewness is based on quartiles. Also called as Galton skewness.

Formula:-

$$S = \frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{(Q_3 - Q_2) + (Q_2 - Q_1)}$$

Q.1) b) Type I and Type II Error

→ Statisticians use specific definitions and Symbols for signifiying the Error in the Hypothesis.

→ Rejecting a null hypothesis when it is true is called a Type I error, and it's probablity (which, as we have seen, is also the significance level of the test) is Symbolized α (Alpha.)

Scanned by CamScanner

→ Alternatively, accepting a null hypothesis, when it is false is called a Type II Error and it's probablity is symbolized as β (Beta.)

→ There is a trade off between these two Errors:-

i) The Probablity of making one type of error can be reduced only if we are willing to increase the probablity of making the other type of error.

→ To deal with this trade off, in personal and professional situations, decision makers decide the appropriate level of significance by examining the costs or penalties attached to both type of errors.

→ If we do not want to occur Type I- Error as well as Type -II Error we need to increase the number of samples.

Q.1) c) p-value

→ The Computed probablity of getting the observed result or any result at least as extreme in it's difference from what the null hypothesis would imply - is called the p-value.

→ A p-value of 0.05 is the de-facto standard cut off between significant

Scanned by CamScanner

and non-significant results.

→ If this de facto value is used as the critical value, it will result in wrong result 5% of the time.

→ p value tells us the probablity that the results occurred by chance. It is the probablity of wrongly rejecting null hypothesis. If p-value less than 0.05, we say that the test result is significant at 5% level of significance.

→ The observed significance level is a criterion that can be computed from a sample data. It is a probability

i) If p-value $< \alpha$; reject the null hypothesis. There is statistical significant difference between two treatments.

ii) If p-value $> \alpha$; accept the Null Hypothesis.

Q-1) d) ANOVA vs T-test

T-test	ANOVA
i) T-test is a hypothesis test that is used to Compare the means of two Samples.	i) ANOVA (Analysis of Variance) is a statistical techniaue that is used to Compare means of more than two samples.

Scanned by CamScanner

ii) The probable distribution is student's t distribution.

ii) The Probable distribution is F-Distribution.

iii) Null Hypothesis always differs based on the tailedness of the test.

iii) Null Hypothesis i.e. H_0
$H_0: \mu_1 = \mu_2 = \mu_3$ or
H_1: At least two population means are different.

iv) Test statistic used for T-test

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

(for one Sample Hypothesis & It will changed as we moved from using one Sample to two Sample.

iv) Test statistic for ANOVA is

$$F = \frac{S_b}{S_w}$$

where S_b = between Column Variance, S_w = within Column Variance.

Q.1) e) Nature of Normal Curve

→ A normal curve is a bell-shaped curve which shows the probablity distribution of a continuous random variable.

→ Moreover, the normal curve represents a normal distribution... Also

Scanned by CamScanner

the Standard normal curve represents a normal curve with mean 0 and standard deviation 1.

- Here we have four Characteristics of a Normal Distribution.

i) Normal distributions are symmetric, unimodal, and asymptotic.

ii) and the mean, median and mode are all equal.

iii) A normal distribution is perfectly Symmetrical around it's center.

iv) That is, the right side of the center is a mirror image of the left side.

- In normal distribution curve, The Curve is symmetrical and bell shaped, showing that trials will usually give a result near the average, but will occasionally deviate by large amounts.

Q.1) f) Probablity

→ In everything, we guess the chances of successful outcomes, from business to medicine to ~~medicine~~ the weather.

→ A Probablity provides a quantitative description of the chances or likelihoods associated with various outcomes.

Scanned by CamScanner

→ It provides a bridge between descriptive and inferential Statistics

→ It is a numerical measure which indicates the chance of occurence.

→ If an experiment has n equally likely simple events and if m be the favorable number of outcome to an event A. Then the probablity of A, P(A) is,

$$P(A) = \frac{\text{Number of favourable outcomes}}{\text{Total number of outcomes}}$$

$$= \frac{m}{n}$$

Results from Probablity / Rules :-

i) $0 \leq P(A) \leq 1$

ii) $P(A) = 1 - P(A')$

iii) $P(\phi) = 0$ where ϕ is null event.

→ Probablity of an Event must have following properties

i) P(A) must be between 0 and 1.

ii) If event A can never occur, P(A) = 0. If event A always occurs when the experiment is performed, P(A) = 1

iii) The Sum of the probablities for

all simple events in S equals 1.

Q.1) g) Cluster vs stratified Sampling

Cluster Sampling	Stratified Sampling
i) cluster Sampling refers to a Sampling method where in the members of the population are selected at random, from naturally occurring groups Called 'cluster.'	i) Stratified Sampling is one in which the population is divided into homogeneous Segments, and then, the Sample is randomly taken from the Segments.
ii) Randomly selected clusters are taken from all the individuals.	ii) Randomly Selected individuals are taken from all the strata.
iii) Selection of population elements are performed Collectively.	iii) Selection of population elements are performed individually.
iv) Between groups their might	iv) Homogeneous elements

Scanned by CamScanner

Contain similarity of the data/ samples.	are found within group of each strata.
v) within groups their might be different Categories of data.	vi) Heterogeneous elements are found ~~with~~ Between groups.
vi) Bifurcation of data has been done through Naturally occuring groups.	vi) Bifurcation of data has been imposed by the researcher.
vii) Objective of Performing cluster Sampling is to reduce cost and improve efficiency.	vii) objective of performing Stratified Sampling is to increase precision and representation.

Scanned by CamScanner

Q.1) b) Sampling error vs non-Sampling error

	Sampling error	Non-Sampling error
Meaning	Sampling error is a type of error, occurs due to the sample selected does not perfectly represents the population of interest.	An error occurs due to Sources other than Sampling, while Conducting Survey activities is known as non-Sampling error.
Cause:-	Deviation between Sample mean and population mean	Deficiency and analysis of data.
Type	Random	Random or Non-random
Occurs	only when Sample is selected.	Both in Sample and census.
Sample size	Possibility of error reduced with the increase in Sample size	It has nothing to do with the Sample size.

Scanned by CamScanner

Q.1) i) Binomial vs Poisson Distribution

Binomial Distribution	Poisson Distribution
1) Binomial distribution is biparametric. i.e. It is featured by two parameters 'n' and 'p'.	1) Where as Poisson distribution is uniparametric. i.e. Characterized by a single parameter 'λ'.
2) In a Binomial distribution, there are only two possible outcomes. i.e. Success or Failure.	2) Conversely, there are an unlimited number of possible outcomes in the case of Poisson Distribution.
3) In binomial distribution mean(np) > Variance(npq)	3) While in Poisson distribution mean = Variance
4) For Discrete Variables	4) For Continuous Variables

Scanned by CamScanner

Q.1) j) Sampling Distribution of the Mean

→ The mean of the sampling distribution of the mean is the mean of the population from which the scores were sampled.

→ Therefore, if a population has a mean μ, then the mean of the sampling distribution of the mean is also μ.

→ The symbol μ_M is used to refer to the mean of the sampling distribution of the mean.

→ Therefore, the formula for the mean of the sampling distribution of the mean can be written as:

$$\mu_M = \mu$$

→ VARIANCE:-

The variance of the sampling distribution of the mean is computed as follows:-

$$\sigma^2_M = \frac{\sigma^2}{N}$$

That is the variance of the sampling distribution of the mean is the population variance divided by N, the

Scanned by CamScanner

Sample size (the number of scores used to Compute a mean.) Thus, the larger the Sample size the smaller the Variance of the Sampling distribution of the mean.

→ The standard error of the mean is the standard deviation of the Sampling distribution of the mean. It is therefore the square root of the Variance of the Sampling distribution of the mean and can be written as:

$$\sigma_M = \frac{\sigma}{\sqrt{14}}$$

→ The standard error is represented by a σ because it is a standard deviation.

→ The subscript (M) indicates that the standard error in question is the standard error of the mean.

Scanned by CamScanner

Q.1) K) Central Limit Theorem

→ The Central Limit Theorem states that the sampling distribution of the sample means approaches a normal distribution as the sample size gets larger

Scanned by CamScanner

-no matter what the shape of the population distribution.

→ Central Limit Theorem (CLT) is an important because all the samples will follow a normal (approximately) distribution pattern with ~~follow a normal~~ all the variances being approximately equal to the variance of the population, divided by each sample's size.

Q.1) L) Correlation Coefficient:-

→ A correlation coefficient is a numerical measure of same type of correlation (meaning a statistical relation between two variables.

→ The two columns of a given data set can be the two variables.

→ Several types of correlation coefficient exist, each with their own definition and own range of usability and characteristics.

→ They all assume the values in the range of -1 to 1, where ± 1 indicates the strongest possible agreement and 0 indicating the strongest possible disagreement.

Scanned by CamScanner

→ Certain problems include the possibility of incorrectly being used to infer causal relationship between the variables.

Properties:-

1] The absolute values of both sample and population (Pearson Correlation) are less than or equal to 1.
2] Correlation Coefficients equal to -1 or 1 correspond to data points lying exactly on a Line.
3] The Correlation Coefficient (Pearson) is Symmetric than Corr $(x, y) = Corr(y, x)$

Q.1) m) Outlier detection:-

→ Outlier are extreme values that deviate from other observations on data, they may indicate a variability in measurements, experimental error.

→ They are of 2 kinds: univariate and multivariate.

→ Most Common causes of outlier on a data set:

1] Data entry errors (human)
2] Measurement errors (instrument)
3] Experiment errors (data extraction)

Scanned by CamScanner

4) Data Processing errors (data manipulation
5) Sampling errors (extracting or mixing data.
6] Natural (novelties in data)
7] Intentional (dummy to test detection methods)

→ Some of the most popular methods of outlier detection:-

- Z-Score/Extreme Value analysis
- Statistical modelling
- Linear regression models
- Proximity based models.

Q.1) N) Properties of normal distribution

→ The normal distribution is also referred to as Gaussian or Gauss distribution.

→ It is made relevant by the Central limit theorem, which states that the averages obtained from independent, identically distributed random variables with almost the same normal distributions, regardless of the type of distribution they are sampled from.

→ Normal distribution follows the following characteristics.

Scanned by CamScanner

1) It is Symmetric
- The distribution curve can be divided in the middle to produce two equal halves.

2) The mean, median, mode
- The middle point of a normal distribution is the point with the maximum frequency, which means it possess the most number of variables/observations.

3) Empirical rule
- There is a Constant proportion of distance lying under the curve between the mean and specific number of standard deviation from the mean.

Example: 68.25% of all cases falls within 1 standard deviation.
95% of all cases fall within 2 standard deviation.
99% of all cases fall within 3 standard deviation.

4) Skewness and Kurtosis
- They are Coefficients that measure how different a distribution is from a normal distribution.

Scanned by CamScanner

- Skewness measures the symmetry which kurtosis measures the thickness of the tail ends.

Q.1) Q) Assumptions of Linear regression:-

→ There are 5 basic assumptions of linear Regression: -

→ Linear relationship between the features and target:

- we assume that there is a linear relationship and thus capture that.

→ Little or no multicollinearity between the features:

- Multicollinearity is a state of very high inter-Correlations among the independent Variables.
- It weakens the statistical power of the regression model.

→ Homoscedacitity:

- It describes a situation in which the errors term (that is the noise or random distribution in the relationship) is the same across all values of the independent Variable.

→ Normal distribution of error terms:

- The assumption is that the error (residuals) follows a

Scanned by CamScanner

normal curve.

→ little or no auto correlation in the residuals.

- Auto correlation occurs when the residual error are dependent on each.
- The presence of correlation in error terms drastically reduces model's accuracy.

Q.1) f) Assumptions of Chi-Square goodness of fit test

Assumption 1:- one categorical variable.

→ dichotomous, nominal or ordinal

→ Examples:-

Dichotomous:- Gender (male, female)
Treatment (medication, no medication)
Education (undergrad, Post Grad)

Nominal:- Ethnicity (Caucasian, Indian, Historian)
Profession (Doctor, teacher, nurse, dentist)

ordinal:- Scale (pointer scale)
Physical activity (sedentary, low, moderate, high)

Scanned by CamScanner

Assumption 2: You should have independence of observation
→ which means you should have there is no relationship between any of the cases.

Assumption 3:- The group of the Categorical variable must be mutually exclusive.

Assumption 4:- There must be at least 5 expected frequencies in each group of your categorical variable.

Q.1) a) Correlation vs Covariance

Covariance:-

→ It is the relationship between a pair of random variables where change in one variable causes change in another variable.

→ It can take any value between $-\infty$ to $+\infty$, where the negative value represents the negative relationship whereas the positive value represent the positive relationship.

→ It is used for linear relationship.

Scanned by CamScanner

- It gives the direction of relationship between variables

$$Covri(x,y) = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{n}$$

Correlation:-

- It shows weather and how strongly pairs of variable are related to each other.

→ Correlation takes values between +1 and −1, where in value close to +1 is strongly positive Correlation and value close to −1 is strongly negative Correlation.

→ It gives direction and strength of relationship between variables.

$$Corr(x,y) = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2 \sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

Scanned by CamScanner

Q.1) r) How to check the Normality of data.

→ ~~Look Later~~ Value of the Shapiro-Wilk Test is greater than 0.05, the data is normal. If it is below 0.05, the data significantly deviate from a normal distribution. If you need to use skewness and kurtosis values to determine normality, rather the Shapiro-Wilk test, you will find these in our enhanced testing for normality guide.

Scanned by CamScanner

Q.1) 8) Point Estimate vs Interval Estimate

→ A point estimate is a single number that is used to estimate an unknown population parameter.

Scanned by CamScanner

Ex. While watching a football match, you say "I bet their line must weigh average 250 pounds"

→ A point estimation is often in sufficient, because it is either right or wrong.

→ You do not know how right or how wrong is your answer (i.e. you do not know the margin of error)

→ It is much more useful if it is accompanied by an estimate of error.

→ An interval Estimate is a range of value used to estimate a population parameter.

→ It indicates the error in two ways:-

- By the extent of it's range
- By the probablity of true population parameter lying within that range.

Ex. The dept head would say something like "I estimate that true enrollment in this course will fall between 330 and 380.

Scanned by CamScanner

Q.1) → t) F Test Vs ANOVA

→ The Analysis of Variance (ANOVA) is the process of assigning the total Variation into it's Components, in an experimental setup and each component is tested separately for it's significance.

→ The test is a test for Variance, a function of the corresponding sum of Squares (SS). Each Component SS is tested with respect to the Error sum of Squares (ESS).

— The ratio of these two sums of squares is the same as the ratio of two variances and therefore is a F_Test.

— A simple example is the ANOVA for a two-way classification experimental set up.

— The resultant Total sum of Squares (TSS) is split up into two components as the Between sum Squares (BSS) and within Sum of Squares (WSS).

— The remainder is the ESS which is used for testing each of the other two.

— The F-Test used for testing

Scanned by CamScanner

any Null hypothesis based on the equality of any two variances, that may not be the results of an ANOVA.

Scanned by CamScanner

Q. 2) Write the steps involved in testing of hypothesis. Also explain the critical value approach and p-value approach.

→ Test Procedure:- (Classical Method)

1) Determine and state H_0 (i.e. Null Hypothesis) and H_1 (i.e. Alternate Hypothesis)
2) Decide the Significance level α and the critical region

Scanned by CamScanner

3) Based on the parameter, choose the test statistic
4) Using available data Compute the test statistic
5) Make the statistical Accept or Reject Decision based on
a) Computed value of the test statistic
b) The critical region identified in step 2

P-value method
1) Determine and state H_0 (i.e. Null Hypothesis) and H_1 (i.e. Alternate Hypothesis)
2) Decide the significance level α
3) Based on the parameter, choose the test statistic
4) Using available data Compute the test statistic and p value
~~p value~~
5) Make the statistical Accept or Reject decision based on
• α & p-value
a) p-value less than α should reject H_0.
b) p-value greater than α should not reject H_0.

Scanned by CamScanner