STATISTICS ASSIGNMENT

Q1to Q15 are descriptive types. Answer in brief.

1. (a) What is central limit theorem

   The theorem states that the distribution of a sample mean that approximates the normal distribution, as the sample size becomes larger, assuming that all the samples are similar, and no matter what the shape of the population distribution

   (b) why is it important?

   Its significance can be seen with the distribution of our population. This theorem allows you to simplify problems in statistics by allowing you to work with a distribution that is approximately normal

2. (a) What is sampling?
   This involves the process where are predetermined number of observations from a larger population. It's a method that allows us to get information about the population based on the statistics from a subset of the population without having to investigate every individual

   (b) How many sampling methods do you know?
   1. **Simple Random**
   2. **Systematic Random**
   3. **Stratified**
   4. **Cluster**
   5. **Convenience**
   6. **Quota**
   7. **Judgement**
   8. **Snowball**

3. What is the difference between type1 and typeII error?

TYPE I	**TYPE II**
A type 1 error is also known as a false positive and occurs when a researcher incorrectly rejects a true null hypothesis. i.e. if Ho is True and you reject Ho	A type II error does not reject the null hypothesis, even though the alternative hypothesis is the true state of nature. In other words, a false finding is accepted as true. i.e. if Ho is False and you failed to reject Ho
A type I error also known as False positive**.**	A type II error also known as False negative. It is also known as false null hypothesis.
The probability that we will make a type I error is designated 'α' (alpha). Therefore, type I error is also known as alpha error.	Probability that we will make a type II error is designated 'β' (beta). Therefore, type II error is also known as beta error.
The probability of type I error is equal to the level of significance.	The probability of type II error is equal to one minus the power of the test.
Type I errors are generally considered more serious.	Type II errors are given less preference.
It can be reduced by decreasing the level of significance.	It can be reduced by increasing the level of significance.

4. What do you understand by the term Normal distribution?

This connotes that Distribution is always normal irrespective of sample size. That is a Normal Distribution will always remain a Normal Distribution no matter the subset of samples collected and re-sampled. The curve will still remain a bell shape no matter what. A Normal Distribution is the proper term for a probability bell curve. In a norm al distribution the mean is Zero and the Standard Deviation is 1

5. (a) What is correlation

As the name implies this has to do with to what extent or measure does a variable change due to the change of another variable. Correlation acknowledges the existence of a covariance but wants to know more about the extent of variation. It can take any value between -1 to -1. Wherein values close to +1 insinuate a strong positive correlation and values close to -1 is an indicator of strong negative correlation.

(b) What is covariance

It's a systematic relationship between a pair of random variables wherein a change i n one variable is reciprocated by equivalent change in another variable. It can take a ny value between -∞ to +∞, wherein a negative value(see formular) is an indicato r of a negative relationship whereas a positive value represents a positive relations hip and when the value is zero, it means no relationship

KEY	COVARIANCE	CORRELATION
Meaning	Measure indicating the extent to which two random variable chang e in tandem.	Statistical measure that indicates how strongly two variables are rel ated
What is it?	Measure of Correlation (has dime nsion)	Scaled version of covariance(dim ensionless)
Possible Values	Lies between **-∞ to +∞**	Lies between **-1 to +1**
Change in scale	Affects covariance	Does not affect correlation
Unit Free measure	No(has dimension)	Yes(dimensionless)

6. Differentiate between univariate, Biavariate and multivariate analysis

UNIVARIATE	BIVARIATE	MULTIVARIATE
Data contains only one variable and doesn't deal with a causes or effect relationships	Data set contains two variables and the aim is to undertake comparisons between the two data set	Data set contains more than two variables and the aim is to undertake comparisons between more than two data set

7. (a) What do you understand by sensitivity

This mean from the Total number of Actual positive results, how many positives were correctly predicted by the model.

(b) how would you calculate it?

Sensitivity = (TP)/(TP+FN)

Where TP= True Positive

FN= False Negative

8. (a) What is hypothesis testing?
   It's a formal procedure for investigating our ideas about the world using statistics. It is most often used by scientists to test specific predictions, called hypotheses, that arise from theories.
   (b) What is H0 and H1?
   Ho – Null Hypothesis
   H1 – Alternate Hypothesis
   (c) What is H0 and H1 for two-tail test?
   Ho – Null Hypothesis
   H1 – Alternate Hypothesis

9. What is quantitative data and qualitative data?
   Quantitative data are measures of values or counts and are expressed as numbers. Quantitative data are data about numeric variables (e.g. how many; how much; or how often). Qualitative data are measures of 'types' and may be represented by a name, symbol, or a number code

10. How to calculate range and interquartile range?
    The IQR describes the middle 50% of values when ordered from lowest to highest. To find the interquartile range (IQR), first find the median (middle value) of the lower and upper half of the data. These values are quartile 1 (Q1) and quartile 3 (Q3). The IQR is the difference between Q3 and Q1
    i.e.
    IQR(Interquantile range)
    IQR=Q3-Q1(Where Q1 and Q3 are the 1^{st} and 3^{rd} Quantile)

11. What do you understand by bell curve distribution?
    A bell curve is a type of graph that is used to visualize the distribution of a set of chosen values across a specified group that tends to have central, normal values, as peaks with low and high extremes tapering off relatively symmetrically on either side.

12. Mention one method to find outliers.
    By finding the Z-score

13. What is p-value in hypothesis testing?
    The p-value, or probability value, tells you how likely it is that your data could have occurred under the null hypothesis. It does this by calculating the likelihood of your test statistic, which is the number calculated by a statistical test using your data.

14. What is the Binomial Probability Formula?
    Binomial probability refers to the probability of exactly x successes on n repeated trials in an experiment that has two possible outcomes (commonly called a binomial experiment). If the probability of success on an individual trial is p, then the binomial probability is $nCx \cdot px \cdot (1-p)n-x$

15. Explain ANOVA and it's applications.
    ANOVA is used to compare differences of means among more than 2 groups. It doe s this by looking at variation in the data and where that variation is found(hence its name). Specifically, ANOVA compares the amount of variation between groups wit h the amount of variation within groups