STATISTICS WORKSHEET-4

Q1to Q15 are descriptive types. Answer in brief.

1. What is central limit theorem and why is it important?

Ans.) The Central Limit Theorem is a statistical theory which states that when the large sample size has a finite variance, the samples will be normally distributed and the mean of samples will be approximately equal to the mean of the whole population.

The central limit theorem is useful when analyzing large data sets because it allows one to assume that the sampling distribution of the mean will be normally-distributed in most cases. This allows for easiest statistical analysis and inference.

2. What is sampling? How many sampling methods do you know?

Ans.) Sampling is a process in statistical analysis where researchers take a predetermined number of observations from a larger population. The method of sampling depends on the type of analysis being performed, but it may include simple random sampling or systematic sampling.

In Statistics, there are different sampling techniques available to get relevant results from the population. These are categorized into two different types of sampling methods. They are:

- Probability Sampling Methods
- Non-probability Sampling methods

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

Type I Error	**Type II Error**
It is the acceptance of hypothesis which ought to be accepted or incorrect rejection of true null hypothesis	It is the acceptance of hypothesis which ought to be rejected or incorrect acceptance of false null hypothesis
Is equivalent to false positive	Is equivalent to false negative
Represents a false hit	Represents a miss
Is same as the level of significance	Is same as the power of test
Denoted by alpha	Denoted by beta

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

Ans.) A normal distribution is an arrangement of a data set in which most values cluster in the middle of the range and the rest taper off symmetrically toward either extreme.

5. What is correlation and covariance in statistics?

Ans.) Correlation is a statistical measure that expresses the extent to which two variables are linearly related (meaning they change together at a constant rate). It's a common tool for describing simple relationships without making a statement about cause and effect.

Covariance is an indicator of the extent to which 2 random variables are dependent on each other. A higher number denotes higher dependency. Correlation is a statistical measure that indicates how strongly two variables are related.

6. Differentiate between univariate ,Biavariate,and multivariate analysis.

Univariate analysis	Biavariate analysis	Multivariate analysis

STATISTICS WORKSHEET-4

Summarizes only one variable at a time.	Compares two variables.	Compares more than two variables.
It has the simplest form of analysis.	The analysis of this type of data deals with causes and relationships and the analysis is done to find out the relationship among the two variables.	The ways to perform analysis on this data depends on the goals to be achieved.Some of the techniques are regression analysis,path analysis,factor analysis and multivariate analysis of variance (MANOVA).
Example of a univariate data can be height.	Example of bivariate data can be temperature and ice cream sales in summer season.	Example of this type of data is suppose an advertiser wants to compare the popularity of four advertisements on a website, then their click rates could be measured for both men and women and relationships between variables can then be examined.

7. What do you understand by sensitivity and how would you calculate it?

Ans.) Sensitivity is the percentage of true positives (e.g. 90% sensitivity = 90% of people who have the target disease will test positive).
The formula to calculate is,
Sensitivity = (True positive)/(True positive + False negative)

8. What is hypothesis testing? What is H0 and H1? What is H0 and H1 for two-tail test?

Ans.) Hypothesis testing is a form of statistical inference that uses data from a sample to draw conclusions about a population parameter or a population probability distribution. First, a tentative assumption is made about the parameter or distribution. This assumption is called the null hypothesis and is denoted by H0. An alternative hypothesis (denoted Ha), which is the opposite of what is stated in the null hypothesis, is then defined. The hypothesis-testing procedure involves using sample data to determine whether or not H0 can be rejected. If H0 is rejected, the statistical conclusion is that the alternative hypothesis Ha is true.

Two-tailed Test H0 : μ = k H1 : μ 6= k P-value = 2P(z > |t|) If P-value □ α, we reject H0. If P-value > α, we do not reject H0. Note: For each formula to find z-scores, if you can assume that x has a normal distribution, then any sample size n will work. If you cannot assume this, use a sample size n 30

STATISTICS WORKSHEET-4

9. What is quantitative data and qualitative data?

Ans.) Quantitative data refers to any information that can be quantified. If it can be counted or measured, and given a numerical value, it's quantitative data. Quantitative data can tell you "how many," "how much," or "how often", etc.

Unlike quantitative data, qualitative data cannot be measured or counted. It's descriptive, expressed in terms of language rather than numerical values.

10. How to calculate range and interquartile range?

Ans.) Range is calculated by subtracting the lowest value from the highest value.

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 range is calculated by subtracting the lowest value from the highest value. While a large range means high variability, a small range means low variability in a distribution. The IQR is the difference between Q3 and Q1.

11. What do you understand by bell curve distribution?

Ans.) 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 tend to have a central, normal values, as peak with low and high extremes tapering off relatively symmetrically on either side.

12. Mention one method to find outliers.

Ans.) Statistical outlier detection which involves applying statistical tests or procedures to identify extreme values. One can convert extreme data points into z scores that tell you how many standard deviations away they are from the mean.
If a value has a high enough or low enough *z* score, it can be considered an outlier. As a rule of thumb, values with a z score greater than 3 or less than –3 are often determined to be outliers.

13. What is p-value in hypothesis testing?

Ans.) The p-value is a number, calculated from a statistical test, that describes how likely you are to have found a particular set of observations if the null hypothesis were true. P-values are used in hypothesis testing to help decide whether to reject the null hypothesis.

14. What is the Binomial Probability Formula?

Ans.) The binomial probability formula is:

$P_x = (^n{}_x)\, p^x q^{n-x}$,
P= binomial probability
x= no. of trials for a specific outcome within n trials
$(^n{}_x)$= no. of combinations
p= probability of success on a single trial
q= probability of failure on a single trial
n= no. of trials

15. Explain ANOVA and it's applications.

Ans.) Analysis of variance (ANOVA) is an analysis tool used in statistics that splits an

STATISTICS WORKSHEET-4

observed aggregate variability found inside a data set into two parts: systematic factors and random factors. The systematic factors have a statistical influence on the given data set, while the random factors do not. Analysts use the ANOVA test to determine the influence that independent variables have on the dependent variable in a regression study.

The ANOVA test is the initial step in analyzing factors that affect a given data set. Once the test is finished, an analyst performs additional testing on the methodical factors that measurably contribute to the data set's inconsistency. The analyst utilizes the ANOVA test results in an f-test to generate additional data that aligns with the proposed regression models.
The ANOVA test allows a comparison of more than two groups at the same time to determine whether a relationship exists between them. The result of the ANOVA formula, the F statistic (also called the F-ratio), allows for the analysis of multiple groups of data to determine the variability between samples and within samples.
If no real difference exists between the tested groups, which is called the null hypothesis, the result of the ANOVA's F-ratio statistic will be close to 1. The distribution of all possible values of the F statistic is the F-distribution. This is actually a group of distribution functions, with two characteristic numbers, called the numerator degrees of freedom and the denominator degrees of freedom.

****