FLIP ROBO

STATISTICS WORKSHEET-4

Q1to Q15 are descriptive types. Answer in brief.

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

   Answer - The central limit theorem (CLT) states that the distribution of sample means approximates a normal distribution as the sample size gets larger, regardless of the population's distribution.

   The Central Limit Theorem is important for statistics because it allows us to safely assume that the sampling distribution of the mean will be normal in most cases.

2. What is sampling? How many sampling methods do you know?
   Answer - Sampling is a technique of selecting individual members or a subset of the population to make statistical inferences from them and estimate the characteristics of the whole population.
   Two types of sampling methods –
   1.Probability sampling: Probability sampling is a sampling technique where a researcher sets a selection of a few criteria and chooses members of a population randomly. All the members have an equal opportunity to be a part of the sample with this selection parameter.
   2.Non-probability sampling: In non-probability sampling, the researcher chooses members for research at random. This sampling method is not a fixed or predefined selection process. This makes it difficult for all elements of a population to have equal opportunities to be included in a sample.

3. What is the difference between type1 and typeII error?
   Answer –
   type1 –
   1.Type I error refers to non-acceptance of hypothesis which ought to be accepted.
   2.Equivalent to False positive.
   3. It is incorrect rejection of true null hypothesis.
   4. Equals the level of significance.
   5.Indicated by 'α'
   typeII –
   1.Type II error is the acceptance of hypothesis which ought to be rejected.
   2. Equivalent to False negative.
   3. It is incorrect acceptance of false null hypothesis.
   4. Equals the power of test.
   5. Indicated by β'

4. What do you understand by the term Normal distribution?
   Answer - 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?
   Covariance -
   1. Covariance is an indicator of how two random variables are dependent on each other.
   2. We can deduct correlation from a covariance.
   3. The value of covariance lies in the range of -∞ and +∞.
   4. Covariance is affected.
   5. Covariance has a definite unit as deduced by the multiplication of two numbers and their units.

   Correlation –
   1. Correlation indicates how strongly these two variables are related, provided other conditions are constant.
   2. Correlation provides a measure of covariance on a standard scale.
   3. Correlation is limited to values between the range -1 and +1.
   4. Correlation is not affected by a change in scales or multiplication by a constant.
   5. Correlation is a unitless absolute number between -1 and +1, including decimal values.
6. Differentiate between univariate ,Biavariate,and multivariate analysis.
   Answer –
   1. Univariate data –
      This type of data consists of only one variable. The analysis of univariate data is thus the simplest form of analysis since the information deals with only one quantity that changes. It does not deal with causes or relationships and the main purpose of the analysis is to describe the data and find patterns that exist within it. The example of a univariate data can be height.
   2. Bivariate data –
      This type of data involves two different variables. 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.Example of bivariate data can be temperature and ice cream sales in summer season.
   3. Multivariate data –
      When the data involves three or more variables, it is categorized under multivariate. 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?
   Answer - Sensitivity analysis determines how different values of an independent variable affect a particular dependent variable under a given set of assumptions. Sensitivity analysis allows for forecasting using historical, true data.

   Sensitivity analysis is often performed in analysis software, and Excel has built in functions to help perform the analysis. For example, a company may perform NPV analysis using a discount rate of 6%. Sensitivity analysis can be performed by analyzing scenarios of 5%, 8%, and 10% discount rates as well by simply maintaining the formula but referencing the different variable values.

8. What is hypothesis testing? What is H0 and H1? What is H0 and H1 for two-tail test?
   Answer - Hypothesis Testing is a type of statistical analysis in which you put your assumptions about a population parameter to the test.
   In hypothesis testing there are two mutually exclusive hypotheses –
   1. Null Hypothesis (H0) – A statement about the value of population parameter that is assumed to be true for the purpose of testing.

2.Alternative Hypothesis (H1) - A statement about the value of population parameter that is assumed to be true if the Null Hypothesis is rejected during testing.
Tn two tailed test –
H0 – The mean income of females is equal to the mean income of females.
H1 – The mean income of females is not equal to the mean income of the males.

9. What is quantitative data and qualitative data?
Answer –
Quantitative data: The data collected on the grounds of the numerical variables are quantitative data. Quantitative data are more objective and conclusive in nature. It measures the values and is expressed in numbers. The data collection is based on “how much” is the quantity. The data in quantitative analysis is expressed in numbers so it can be counted or measured.
Qualitative data: The data collected on grounds of categorical variables are qualitative data. Qualitative data are more descriptive and conceptual in nature. It measures the data on basis of the type of data, collection, or category. The data collection is based on what type of quality is given. Qualitative data is categorized into different groups based on characteristics.

10. How to calculate range and interquartile range?
Answer –
To calculate the range, you need to find the largest observed value of a variable (the maximum) and subtract the smallest observed value (the minimum). The range only takes into account these two values and ignore the data points between the two extremities of the distribution.

The interquartile range and semi-interquartile range give a better idea of the dispersion of data. To calculate these two measures, you need to know the values of the lower and upper quartiles. The lower quartile, or first quartile (Q1), is the value under which 25% of data points are found when they are arranged in increasing order. The upper quartile, or third quartile (Q3), is the value under which 75% of data points are found when arranged in increasing order. The median is considered the second quartile (Q2). The interquartile range is the difference between upper and lower quartiles. The semi-interquartile range is half the interquartile range.

11. What do you understand by bell curve distribution ?
Answer - A bell curve is a graph depicting the normal distribution, which has a shape reminiscent of a bell. The top of the curve shows the mean, mode, and median of the data collected. Its standard deviation depicts the bell curve's relative width around the mean.

12. Mention one method to find outliers.
An outlier is a piece of data that is an abnormal distance from other points. In other words, it’s data that lies outside the other values in the set.
Using visualizations -
To visualize the data using with a box plot so you can see the data distribution at a glance. This type of chart highlights minimum and maximum values (the range), the median, and the interquartile range for your data.Many computer programs highlight an outlier on a chart with an asterisk, and these will lie outside the bounds of the graph.

13. What is p-value in hypothesis testing?
Answer –
The p-value is defined as the probability of obtaining the result at least as extreme as the observed result of a statistical hypothesis test, assuming that the null hypothesis is true.

14. What is the Binomial Probability Formula?
    Answer –
    In probability theory and statistics, the binomial distribution is the discrete probability distribution that gives only two possible results in an experiment, either Success or Failure.
    Binomial probability distribution is $P(r) = nCr \cdot pr\ (1 - p)n{-}r$

15. Explain ANOVA and it’s applications.
    Answer –
    Analysis of variance (ANOVA) is an analysis tool used in statistics that splits an 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.
    ANOVA test to compare different suppliers and select the best available.
    The Formula for ANOVA is:
    $F=MST/MSE$
    where F=ANOVA coefficient
    MST=Mean sum of squares due to treatment MSE=Mean sum of squares due to error