STATISTICS WORKSHEET-4

Q1to Q15 are descriptive types. Answer in brief.

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

The central limit theorem states that the sampling distribution of the mean approaches a normal distribution, as the sample size increases. This fact holds especially true for sample sizes over 30. Therefore, as a sample size increases, the sample mean and standard deviation will be closer in value to the population mean and standard deviation .

This is useful, as the research never knows which mean in the sampling distribution is the same as the population mean, but by selecting many random samples from a population the sample means will cluster together, allowing the research to make a very good estimate of the population mean.

Thus, as the sample size increases the sampling error will decrease.

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

Sampling is a process in statistical analysis where researchers take a predetermined number of observations from a larger population.

Simple random smpling- In simple random sampling technique, every item in the population has an equal and likely chance of being selected in the sample. Since the item selection entirely depends on the chance, this method is known as “Method of chance Selection”. As the sample size is large, and the item is chosen randomly, it is known as “Representative Sampling”

Systemic sampling- In the systematic sampling method, the items are selected from the target population by selecting the random selection point and selecting the other methods after a fixed sample interval. It is calculated by dividing the total population size by the desired population size.
Stratified - In a stratified sampling method, the total population is divided into smaller groups to complete the sampling process. The small group is formed based on a few characteristics in the population. After separating the population into a smaller group, the statisticians randomly select the sample.
cluster- In the clustered sampling method, the cluster or group of people are formed from the population set. The group has similar significatory characteristics. Also, they have an equal chance of being a part of the sample. This method uses simple random sampling for the cluster of population.

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

Type I Error-- A type I error occurs when the null hypothesis is true but is rejected. In other words, if a true null hypothesis is incorrectly rejected, type I error occurs.

Type II Error --occurs when the null hypothesis is false but invalidly fails to be rejected. In other words, failure to reject a false null hypothesis results in type II error.

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

The normal distribution is a symmetrical, bell-shaped distribution in which the mean, median and mode are all equal. It is a central component of inferential statistics.
A normal distribution is a type of continuous probability distribution in which most data points cluster toward the mean.

5. What is correlation and covariance in statistics?

Both covariance and correlation measure the relationship and the dependency between two variables. Covariance indicates the direction of the linear relationship between variables. Correlation measures both the strength and direction of the linear relationship between two variables.

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

Univariate-
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.
Biavariate-
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.
Multivariate-
When the data involves three or more variables, it is categorized under multivariate. It is similar to bivariate but contains more than one dependent variable. 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).

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

Sensitivity analysis is an analysis technique that works on the basis of what-if analysis like how independent factors can affect the dependent factor and is used to predict the outcome when analysis is performed under certain conditions. It is commonly used by investors who takes into consideration the conditions that affect their potential investment to test, predict and evaluate result.

The two main types of sensitivity analysis are local sensitivity analysis and global sensitivity analysis. Local sensitivity analysis assesses the effect of a single parameter at a time while holding all other parameters constant, while global sensitivity analysis is a more broad analysis used in more complex modeling scenarios such as Monte Carlo techniques.

Methods of Sensitivity Analysis
There are different methods to carry out the sensitivity analysis:
Modeling and simulation techniques
Scenario management tools through Microsoft excel
Key takeaways-

- Sensitivity analysis determines how different values of an independent variable affect a particular dependent variable under a given set of assumptions.
- This model is also referred to as a what-if or simulation analysis.
- Sensitivity analysis can be used to help make predictions in the share prices of publicly traded companies or how interest rates affect bond prices.
- Sensitivity analysis allows for forecasting using historical, true data.
- While sensitivity analysis determines how variables impact a single event, scenario analysis is more useful to determine many different outcomes for more broad situations.

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

Hypothesis testing is a statistical test used to determine the relationship between two data sets, between two or more independent and dependent variables.
Hypothesis testing is a scientific process to conclude whether to reject or not reject the null hypothesis.
Statistical methods are used for hypothesis testing.
Null hypothesis or H0-- Null hypothesis is a statement which states no relationship between two variables or two phenomena.
Ex. of stating null hypothesis: There is no significant relationship between smoking and lung cancer.
Alternate hypothesis or H1— Alternative hypothesis/alternate hypothesis is a statement which states some statistical significance between two phenomena.
Ex. of stating alternative hypothesis: There is statistical significant relationship between smoking and lung cancer.
Two tailed test--A hypothesis test that is designed to show whether the mean of a sample is significantly greater than and significantly less than the mean of a population is referred to as a two tailed test.

9. What is quantitative data and qualitative data?

Quantitative data are measures of values or counts and are expressed as numbers.
Qualitative data are measures of 'types' and may be represented by a name, symbol, or a number code.
Data collected about a numeric variable will always be quantitative and data collected about a categorical variable will always be qualitative.

10. How to calculate range and interquartile range?

The interquartile range is a measure of where the “middle fifty” is in a data set.
Where a range is a measure of where the beginning and end are in a set, an interquartile range is a measure of where the bulk of the values lie.
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 formula is the first quartile subtracted from the third quartile: IQR = Q3 – Q1.

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 tend to have a central, normal values, as peak with low and high extremes tapering off relatively symmetrically on either side.
Bell curves are visual representations of normal distribution, also called Gaussian distribution.
Bell curves are useful for quickly visualizing a data set's mean, mode and median because when the distribution is normal, the mean, median and mode are all the same.

12. Mention one method to find outliers.

Percentile method or using boxplot we can find the presence ofoutliers.

Step 1: Put the numbers in order.
1, 2, 5, 6, 7, 9, 12, 15, 18, 19, 27.
Step 2: Find the median.
1, 2, 5, 6, 7, 9, 12, 15, 18, 19, 27.
Step 3: Place parentheses around the numbers above and below the median.
Not necessary statistically, but it makes Q1 and Q3 easier to spot.
(1, 2, 5, 6, 7), 9, (12, 15, 18, 19, 27).

Step 4: Find Q1 and Q3
Think of Q1 as a median in the lower half of the data and think of Q3 as a median for the upper half of data.
(1, 2, 5, 6, 7), 9, (12, 15, 18, 19, 27). Q1 = 5 and Q3 = 18.
Step 5: Subtract Q1 from Q3 to find the interquartile range- IQR = Q3 – Q1.
18 – 5 = 13.

Low outliers = Q1 – 1.5(Q3 – Q1) = Q1 – 1.5(IQR)
High outliers = Q3 + 1.5(Q3 – Q1) = Q3 + 1.5(IQR)
Anything outside of the value we got for low outliers and high outliers will be our outliers.

13. What is p-value in hypothesis testing?

The P-value is known as the probability value. It is defined as the probability of getting a result that is either the same or more extreme than the actual observations. The P-value is known as the level of marginal significance within the hypothesis testing that represents the probability of occurrence of the given event.
P-value > 0.05 means the result is not statistically significant and hence don't reject the null hypothesis.
P-value < 0.05 means the result is statistically significant. Generally, reject the null hypothesis in favour of the alternative hypothesis

14. What is the Binomial Probability Formula?

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**. For example, if we toss a coin, there could be only two possible outcomes: heads or tails, and if any test is taken, then there could be only two results: pass or fail. This distribution is also called a binomial probability distribution.
In binomial probability distribution, the number of 'Success' in a sequence of n experiments, where each time a question is asked for yes-no, then the boolean-valued outcome is represented either with success/yes/true/one (probability p) or failure/no/false/zero (probability q = 1 − p). A single success/failure test is also called a Bernoulli trial or Bernoulli experiment, and a series of outcomes is called a Bernoulli process. For n = 1, i.e. a single experiment, the binomial distribution is a Bernoulli distribution.

15. Explain ANOVA and it's applications.
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.
Analysis of variance (ANOVA) is a statistical technique that is used to check if the means of two or more groups are significantly different from each other. ANOVA checks the impact of one or more factors by comparing the means of different samples.