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 if we have a population with mean μ and standard deviation σ and take sufficiently large random samples from the population with replacement, then the distribution of the sample mean is asymptotically normal. We can calculate the mean of the sample means for the random samples we choose from the population:

$\mu\bar{X}=\mu$ As well as the standard deviation of sample means:

$\sigma\bar{X}=\sigma n$

According to the central limit theorem, the form of the sampling distribution will approach normalcy as the sample size is sufficiently large (usually n>30). regardless of the population distribution.

Importance of Central Limit Theorem :

This is useful since the research never knows which mean in the sampling distribution corresponds to the population mean, but by taking numerous random samples from population, the sample means will cluster together ,allowing the researcher to obtain a very accurate estimate of the population mean.

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

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. Different sampling methods are widely used by researchers in market research so that they do not need to research the entire population to collect actionable insights.

It is also a time-convenient and a cost-effective method and hence forms the basis of any research design. Sampling techniques can be used in a research survey software for optimum derivation.

sampling methods:

1. Probability sampling
2. Non-probability sampling

3. **What is the difference between type1 and type II error?**

Sl No	Type I Error	Type II Error
1	Type I error refers to non acceptance of hypothesis which ought to be accepted.	Type II error is the acceptance of hypothesis which ought to be rejected.
2	False Positive	False Negative
3	It is incorrect rejection of true null hypothesis.	It is incorrect acceptance of False null hypothesis.
4	A False hit .	A miss.
5	Equals the level of significant.	Equals the power of test.
6	Greek Letter Alpha	Greek Letter Beta.

The graph of the pdf (probability density function) is a bell shaped curve
The normal random variable takes values from – to +
It is symmetric and centered around the mean (which is also the median and mode)
Any normal distribution can be specified with just now parameters - the Mean (μ) and the standard deviation (o)
We write this as X N(μ, o)

The probability associated with any single value of the random variable is always Zero
Probability of values being in a range = Area under the pdf curve in that range

normal distribution, also called Gaussian distribution

Differentiate between univariate , Biavariate,and multivariate analysis.

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

The sensitivity is calculated by dividing the percentage change in output by the percentage change in input.

Calculation :

Sensitivity analysis is used **to** identify how much variations in the input values for a given variable impact the results for a mathematical model. Sensitivity analysis can identify the best data to be collected for analyses to evaluate a project's return on investment (ROI).

5.**What is correlation and covariance in statistics?**

Covariance is nothing but a measure of correlation. Correlation refers to the scaled form of covariance. Covariance indicates the direction of the linear relationship between variables. Correlation on the other hand measures both the strength and direction of the linear relationship between two variables

Sl No	Covariance	Correlation
1	Covariance is a measure to indicate the extent to which two random variables change in tandem	Correlation is a measure used to represent how strongly two random variables are related to each other.
2	Covariance is nothing but a measure of correlation.	Correlation refers to the scaled form of covariance.
3	Covariance indicates the direction of the linear relationship between variables.	Correlation on the other hand measures both the strength and direction of the linear relationship between two variables
4	Covariance can vary between -∞ and +∞	Correlation ranges between -1 and +1
5	Covariance is affected by the change in scale. If all the values of one variable are multiplied by a constant and all the values of another variable are multiplied, by a similar or different constant, then the covariance is changed.	Correlation is not influenced by the change in scale.
6	Covariance assumes the units from the product of the units of the two variable	Correlation is dimensionless, i.e. It's a unit-free measure of the relationship between variables.
7	Covariance of two dependent variables measures how much in real quantity (i.e. cm, kg, liters) on average they co-vary.	Correlation of two dependent variables measures the proportion of how much on average these variables vary w.r.t one another
8	Covariance is zero in case of independent variables (if one variable moves and the other doesn't) because then the variables do not necessarily move together.	Independent movements do not contribute to the total correlation. Therefore, completely independent variables have a zero correlation.

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

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.

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

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.
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 determines how different values of an independent variable affect a particular dependent variable under a given set of assumptions. In other words, sensitivity analyses study how various sources of uncertainty in a mathematical model contribute to the model's overall uncertainty

Find the percentage change in the output and the percentage change in the input. Find sensitivity by dividing the percentage change in output by the percentage change in input

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

hypothesis testing:

- Is based on the population parameters**.**
- Must be clearly stated for correct decision-making.
- Is proved based on the evidence from Statistical test

Null HypothesisH0 : is a statement (assumption) about population(s) parameters.
It is the one assume to be true unless started. Generally describe the present status.

Alternative Hypothesis H1 : Is the negation or compliment of the null hypothesis.

H0: defendant is innocent; • H1: defendant is guilty. H0 (innocent) is rejected if H1 (guilty) is supported by evidence beyond “reasonable doubt.” Failure to reject H0 (prove guilty) does not

that the evidence is insufficient to reject it.

Two-tailed test:

The level of significance which is selected in Step 1 (e.g., $\alpha = 0.05$) dictates the critical value. For example, in an upper tailed Z test, if $\alpha = 0.05$ then the critical value is $Z = 1.645$.

9.What is quantitative data and qualitative data?

Qualitative Data

Qualitative data is a set of information which can not be measured using numbers. It generally consist of words, subjective narratives. Result of an qualitative data analysis can come in form of highlighting key words, extracting information and concepts elaboration. For example, a study on parents perception about the current education system for their kids. The resulted information collected from them might be in narrative form and you need to deduce the analysis that they are satisfied, un-satisfied or need improvement in certain areas and so on.

Quantitative Data

Quantitative data is a set of numbers collected from a group of people and involves statistical analysis.For example if you conduct a satisfaction survey from participants and ask them to rate their experience on a scale of 1 to 5. You can collect the ratings and being numerical in nature, you will use statistical techniques to draw conclusions about participants satisfaction.

10. How to calculate range and interquartile range?

Range: the difference between the highest and lowest values. Interquartile range: the range of the middle half of a distribution

So, there are 3 quartiles. First Quartile is denoted by Q_1 known as the lower quartile, the second Quartile is denoted by Q_2 and the third Quartile is denoted by Q_3 known as the upper quartile. Therefore, the interquartile range is equal to the upper quartile minus lower quartile.

11. What do you understand by bell curve distribution ?

A bell curve is the informal name of a graph that depicts a normal probability distribution. The term obtained its name due to the bell-shaped curve of the normal probability distribution graph.
the term is not quite correct because the normal probability distribution is not the only probability distribution whose graph shows a bell-shaped curve. For example, the graphs of the Cauchy and logistic distributions also demonstrate a bell-shaped curve.

Characteristics of a Bell Curve

The bell curve is perfectly symmetrical. It is concentrated around the peak and decreases on either side. In a bell curve, the peak represents the most probable event in the dataset while the other events are equally distributed around the peak. The peak of the curve corresponds to the mean of the dataset (note that the mean in a normal probability distribution also equals the median and the mode).

The dispersion of the data on the bell curve is measured by the standard deviation. The probabilities of the bell curve and the standard deviation share a few important relationships, including:

12. Mention one method to find outliers

There are four ways to identify outliers:

- Sorting method.
- Data visualization method.
- Statistical tests (z scores)
- Interquartile range method.

The “single outlier” tests (**Grubbs and Dixon**) are designed to detect one outlier only, and should not be repeated for several outliers. The “multiple outlier” test (generalized ESD) attempts to detect multiple outliers, if present.

13. What is p-value in hypothesis testing?

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?

The binomial distribution formula is used in statistics to find the probability of the specific outcome-success or failure in a discrete distribution.

Binomial probability refers to the probability of exactly x successes on n repeated trials in an experiment which 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 $_nC_x \cdot p^x \cdot (1-p)^{n-x}$.

15. Explain ANOVA and it’s applications.

Analysis of variance, or ANOVA, is a statistical method that separates observed variance data into different components to use for additional tests. A one-way ANOVA is used for three or more groups of data, to gain information about the relationship between the dependent and independent variables.

One-Way Analysis of Variance (ANOVA) tells you if there are any statistical differences between the means of three or more independent groups.