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 asserts that the distribution of the sample mean is asymptotically normal if we have a population with mean and standard deviation and take sufficiently large random samples from the population with replacement.

For the random samples we select from the population, we may compute the sample mean as follows:

$\mu X^{-}=\mu$

Along with sample means' standard deviation:

$\sigma X^{-}=\sigma n$

The central limit theorem predicts that once sample sizes go high enough (often n>30), the form of the sampling distribution will start to resemble normality. whatever the distribution of the population.

Value of the Central Limit Theorem:

This is advantageous because the researcher never knows which mean in the sampling distribution corresponds to the population mean. However, by selecting numerous random samples from a population, the sample means will cluster together, enabling the researcher to obtain a very precise estimate of the population mean.

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

Ans. In statistics, sampling is a method when researchers determine a representative segment of a larger population that is then used to conduct a study. Sampling generally comes in two forms — probability sampling and non-probability sampling.

The two different types of sampling methods are::

- Probability Sampling
- Non-probability Sampling

Probability Sampling: - The probability sampling technique makes use of a random selection technique. In this strategy, every eligible person has a chance to choose a sample from the entire sample space. This approach takes longer and costs more money than the non-probability sampling approach. The advantage of probability sampling is that it ensures the sample will accurately reflect the population.

Probability Sampling methods are further classified into different types, such as simple random sampling, systematic sampling, stratified sampling, and clustered sampling.

Let us discuss the different types of probability sampling methods along with illustrative examples here in detail.

Non-probability Sampling: - The non-probability sampling method is a technique in which the researcher selects the sample based on subjective judgment rather than the random selection. In this method, not all the members of the population have a chance to participate in the study.

Non-probability Sampling methods are further classified into different types, such as convenience sampling, consecutive sampling, quota sampling, judgmental sampling, snowball sampling. Here, let us discuss all these types of non-probability sampling in detail.

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

Ans.

BASIS FOR COMPARISON	TYPE I ERROR	TYPE II ERROR
Meaning	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.
Equivalent to	False positive	False negative
What is it?	It is incorrect rejection of true null hypothesis.	It is incorrect acceptance of false null hypothesis.
Represents	A false hit	A miss
Probability of committing error	Equals the level of significance.	Equals the power of test.

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

Ans. An example of a continuous probability distribution is the normal distribution, in which the majority of data points cluster in the middle of the range while the remaining ones taper off symmetrically toward either extreme. The distribution's mean is another name for the centre of the range.

5. What is correlation and covariance in statistics?

Ans. Covariance :- Covariance is an indicator of the extent to which 2 random variables are dependent on each other. A higher number denotes higher dependency.

Correlation :- Correlation is a statistical measure that indicates how strongly two variables are related.

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

Ans. 6. 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 are 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?

Ans. Sensitivity analysis is a type of study that uses what-if analysis to determine how independent variables may have an impact on the dependent variable. It is used to forecast results when analysis is conducted under certain circumstances. Investors who study the factors that could effect a possible investment frequently utilise it to test, forecast, and assess results.

FORMULA

$Z = X2 + Y2$

Calculation of the Sensitivity Analysis (Step by Step)

1) The analyst must first create the fundamental formula, which will serve as the output formula. The output formula, for instance, can be the NPV formula.

2) The analyst must next decide which variables must be made sensitive because they are essential to the output formula. The initial investment and the cost of capital can be the independent variables in the NPV formula in Excel.

3) Next, establish the independent variables' likely range.

4) The range of one independent variable should be placed along the rows of an Excel sheet, and the other set should be placed along the columns.

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

Ans. 8. HYPOTHESIS TESTING :

Hypothesis testing is an act in statistics whereby an analyst tests an assumption regarding a population parameter. The methodology employed by the analyst depends on the nature of the data used and the reason for the analysis.

Hypothesis testing is used to assess the plausibility of a hypothesis by using sample data. Such data may come from a larger population, or from a data-generating process. The word "population" will be used for both of these cases in the following descriptions.

Hypothesis Testing is a type of statistical analysis in which you put your assumptions about a population parameter to the test. It is used to estimate the relationship between 2 statistical variables.

H0: The null hypothesis, H0, is a statistical proposition stating that there is no significant difference between a hypothesized value of a population parameter and its value estimated from a sample drawn from that population.

H1: is a statistical proposition stating that there is a significant difference between a hypothesized value of a population parameter and its estimated value. When the null hypothesis is tested, a decision is either correct or incorrect.

H0 & H1 FOR TWO-TAIL TESTING:

Null hypothesis (H0): The null hypothesis here is what currently stated to be true about the population. In our case it will be the average height of students in the batch is 100.

H0 : $\mu = 100$

Alternate hypothesis (H1): The alternate hypothesis is always what is being claimed. “In our case, Tedd believes (Claims) that the actual value has changed”. He doesn’t know whether the average has gone up or down, but he believes that it has changed and is not 100 anymore.

H1: $\mu \neq 100$

9. What is quantitative data and qualitative data?

Ans. QUALITATIVE DATA:- qualitative data is non-statistical and is typically unstructured or semi-structured. This data isn't necessarily measured using hard numbers used to develop graphs and charts. Instead, it is categorized based on properties, attributes, labels, and other identifiers.

Qualitative data can be used to ask the question "why." It is investigative and is often open-ended until further research is conducted. Generating this data from qualitative research is used for theorizations, interpretations, developing hypotheses, and initial understandings.

Qualitative data can be generated through:

- Texts and documents
- Audio and video recordings
- Interview transcripts and focus groups
- Observations and notes

QUANTITATIVE DATA: Contrary to qualitative data, quantitative data is statistical and is typically structured in nature – meaning it is more rigid and defined. This data type is measured using numbers and values, making it a more suitable candidate for data analysis.

Whereas qualitative is open for exploration, quantitative data is much more concise and close-ended. It can be used to ask the questions "how much" or "how many," followed by conclusive information.

Quantitative data can be generated through:

- Tests
- Experiments
- Surveys
- Market reports
- Metrics

10. How to calculate range and interquartile range?

Ans. CALCULATION OF RANGE: 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

CALCULATION OF INTERQUARTILE RANGE: 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.D

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 values selected from a defined group that usually has a median and normal value, with lower and upper limits that decrease relatively n 'each side.

A bell curve is a visual representation of a normal distribution, also called a Gaussian distribution. The normal distribution curve, when drawn, follows a bell curve, hence the name.

Although the exact shape may vary depending on the distribution of the population, the peak is always in the middle and the pattern is constant.

12. Mention one method to find outliers.

Ans. Sorting method

You can sort quantitative variables from low to high and scan for extremely low or extremely high values. Flag any extreme values that you find.

This is a simple way to check whether you need to investigate certain data points before using more sophisticated methods.

Your dataset for a pilot experiment consists of 8 values.

180 156 9 176 163 1827 166 171

You sort the values from low to high and scan for extreme values.

9 156 163 166 171 176 180 1872

13. What is p-value in hypothesis testing?

Ans. The probability that you would have discovered a specific collection of observations if the null hypothesis were true is expressed as a number called the p value, which is determined from a statistical test. In order to determine whether to reject the null hypothesis, P values are utilised in hypothesis testing.

14. What is the Binomial Probability Formula?

Ans. Formula

P_{x} = {n \choose x} p^{x} q^{n-x}

P = binomial probability

x = number of times for a specific outcome within n trials

{n \choose x} = number of combinations

p = probability of success on a single trial

q = probability of failure on a single trial

n = number of trials

15. Explain ANOVA and it's applications.

Ans. Analysis of variance, or ANOVA, is a measurable technique that isolates noticed difference information into various parts to use for extra tests. A one-way ANOVA is utilized for at least three gatherings of information, to acquire data about the connection between the reliant and free factors.