STATISTICS WORKSHEET-4

ANSWER KEY

1. The CLT is a statistical theory that states that - if you take a sufficiently large sample size from a population with a finite level of variance, the mean of all samples from that population will be roughly equal to the population mean.

   Importance of Central Limit Theorem:

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

2. Probability sampling methods include simple random sampling, systematic sampling, stratified sampling, and cluster sampling. What is non-probability sampling? In non-probability sampling, the sample is selected based on non-random criteria, and not every member of the population has a chance of being included.

   Types of Sampling Method

   In Statistics, there are different sampling techniques available to get relevant results from the population. The two different types of sampling methods are**::** Probability Sampling, Non-probability Sampling.

3. **Difference between Type 1 and Type 2 Error :**

   Type -1 Error (Error of the first kind)

   It is also known as a false-positive. It occurs if the researcher rejects a correct null hypothesis in the population. i.e., incorrect rejection of the null hypothesis. Measured by alpha (significance level). If the significance level is fixed at 5%, It means there are about five chances of type – 1 error out of 100.

   Type -2 Error (Error of the second kind)

   It is also known as a false negative. It occurs if a researcher fails to reject a null hypothesis that is actually a false hypothesis. Measured by beta (the power of test). The probability of committing a type -2 error is calculated by 1 – beta (the power of test).

4. A normal distribution is a type of continuous probability distribution in which most data points cluster toward the middle of the range, while the rest taper off symmetrically toward either extreme. The middle of the range is also known as the mean of the distribution.

5. In statistics, correlation is a measure that determines the degree to which two or more random variables move in sequence. When an equivalent movement of another variable reciprocates the movement of one variable in some way or another during the study of two variables, the variables are said to be correlated.

   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. Univariate statistics summarize only one variable at a time. Bivariate statistics compare two variables. Multivariate statistics compare more than two variables.

7. The technique used to determine how independent variable values will impact a particular dependent variable under a given set of assumptions is defined as sensitive analysis. It's usage will depend on one or more input variables within the specific boundaries, such as the effect that changes in interest rates will have on a bond's price. It is also known as the what – if analysis. Sensitivity analysis can be used for any activity or system. All from planning a family vacation with the variables in mind to the decisions at corporate levels can be done through sensitivity analysis.

   Firstly the base case output is defined; say the NPV at a particular base case input value (V1) for which the sensitivity is to be measured. All the other inputs of the model are kept constant. Then the value of the output at a new value of the input (V2) while keeping other inputs constant is calculated. Find the percentage change in the output and the percentage change in the input. The sensitivity is calculated by dividing the percentage change in output by the percentage change in input.

8. 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.
   In hypothesis testing there are two mutually exclusive hypotheses; the Null Hypothesis (H0) and the Alternative Hypothesis (H1). One of these is the claim to be tested and based on the sampling results (which infers a similar measurement in the population), the claim will either be supported or not. The claim might be that the population proportion (or mean) has increased, decreased, stayed the same, or that it has changed. According to the words used in the problem, the claim will be either H0 or H1. Note that the Null Hypothesis, H0, ALWAYS contains the condition of equality.

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.
Alternate hypothesis (H1): The alternate hypothesis is always what is being claimed.

9. Quantitative data are measures of values or counts and are expressed as numbers. Quantitative data are data about numeric variables (e.g. how many; how much; or how often). Qualitative data are measures of 'types' and may be represented by a name, symbol, or a number code.

10. To calculate the range, we 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. It's used as a supplement to other measures, but it is rarely used as the sole measure of dispersion because it's sensitive to extreme values. The interquartile range and semi-interquartile range give a better idea of the dispersion of data. To calculate these two measures, we 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. 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. **Sorting method :** We 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.

13. 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. 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 $nCx \cdot px \cdot (1-p)n-x$. Here nCx indicates the number of different combinations of x objects selected from a set of n objects. Some textbooks use the notation (nx) instead of nCx . Note that if p is the probability of success of a single trial, then (1−p) is the probability of failure of a single trial.

15. 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.

    We can use the ANOVA test to compare different suppliers and select the best available. ANOVA (Analysis of Variance) is used when we have more than two sample groups and determine whether there are any statistically significant differences between the means of two or more independent sample groups.