STATISTICS WORKSHEET-4

Answers the Following Questions-

Ans 1-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. This means that we can take advantage of statistical techniques that assume a normal distribution, as we will see in the next section.

Ans 2-Sampling is the process of selecting a number of cases from all the cases in a particular group or universe. Context: Sampling is the research strategy of collecting data from a part of a population with a view to drawing inferences about the whole. The “population” in this sense is often termed the “universe".

There are two primary types of sampling methods that you can use in your research:

Probability sampling involves random selection, allowing you to make strong statistical inferences about the whole group.

Non-probability sampling involves non-random selection based on convenience or other criteria, allowing you to easily collect data

Ans 3-A type I error (false-positive) occurs if an investigator rejects a null hypothesis that is actually true in the population; a type II error (false-negative) occurs if the investigator fails to reject a null hypothesis that is actually false in the population.

Ans 4-The normal distribution is a continuous probability distribution that is symmetrical around its mean, most of the observations cluster around the central peak, and the probabilities for values further away from the mean taper off equally in both directions. Extreme values in both tails of the distribution are similarly unlikely

Ans 5-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

Ans 6-Univariate analysis looks at one variable, Bivariate analysis looks at two variables and their relationship. Multivariate analysis looks at more than two variables and their relationship

Ans 7-Sensitivity is the percentage of true positives (e.g. 90% sensitivity = 90% of people who have the target disease will test positive). Specificity is the percentage of true negatives (e.g. 90% specificity = 90% of people who do not have the target disease will test negative).

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

Ans 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.

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

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.

Ans 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

Ans 10-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 IQR describes the middle 50% of values when ordered from lowest to highest. 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

Ans 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

Ans 12-Using visualizations

You can use software to visualize your data with a box plot, or a box-and-whisker 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

Ans 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

Ans 14-Binomial probability distribution is

$$P(r) = nCr \cdot pr\ (1 - p)n-r$$

Ans 15-ANOVA, is a statistical method that separates observed variance data into different components to use for additional tests

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.