STATISTICS WORKSHEET SOLUTION-4

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.

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.

2.Sampling means selecting the group that you will actually collect data from in your research. For example, if you are researching the opinions of students in your university, you could survey a sample of 100 students. In statistics, sampling allows you to test a hypothesis about the characteristics of a population.

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.

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.

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. 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. Sensitivity analysis is a financial model that determines how target variables are affected based on changes in other variables known as input variables. It is a way to predict the outcome of a decision given a certain range of variables.

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

8. Hypothesis testing is a form of statistical inference that uses data from a sample to draw conclusions about a population parameter or a population probability distribution. First, a tentative assumption is made about the parameter or distribution.

In a jury trial the hypotheses are: 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 imply innocence, only that the evidence is insufficient to reject it.

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. 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.

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. 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.

Many computer programs highlight an outlier on a chart with an asterisk, and these will lie outside the bounds of the graph.

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. The binomial distribution formula is for any random variable X, given by; $P(x:n,p) = {}^{n}Cxx\ p^{x}\ (1-p)^{n-x}$ Or $P(x:n,p) = {}^{n}C_{x}\ p^{x}\ (q)^{n-x}$

where,

n = the number of experiments

x = 0, 1, 2, 3, 4, …

p = Probability of success in a single experiment

q = Probability of failure in a single experiment (= 1 – p)

The binomial distribution formula is also written in the form of n-Bernoulli trials, where ${}^{n}C_{x} = n!/x!(n-x)!$. Hence, $P(x:n,p) = n!/[x!(n-x)!].p^{x}.(q)^{n-x}$

15. 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.

ANOVA is helpful for testing three or more variables. It is similar to multiple two-sample t-tests. However, it results in fewer type I errors and is appropriate for a range of issues. ANOVA groups differences by comparing the means of each group and includes spreading out the variance into diverse sources.