**Statistics Worksheet 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.

1. When you conduct research about a group of people, it’s rarely possible to collect data from every person in that group. Instead, you select a sample. The sample is the group of individuals who will actually participate in the research.

To draw valid conclusions from your results, you have to carefully decide how you will select a sample that is representative of the group as a whole. This is called a sampling method.

There are two methods:

1. Probability sampling involves random selection, allowing you to make strong statistical inferences about the whole group.
2. Non-probability sampling involves non-random selection based on convenience or other criteria, allowing you to easily collect data.

3) **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.

**Cause of Type – 1 Error:**

The significance level is decided before testing the hypothesis

Sample size is not considered. This may occur due to chance

It can be reduced by decreasing the level of significance.

**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).

**Cause of Type – 2 Error:**

A statistical test is not powerful enough. It is caused by a smaller sample size. It may hide the significance level of the items being tested.

It can be reduced by increasing the level of significance.

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.

The value of covariance lies in the range of -∞ and +∞.

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

Correlation is limited to values between the range -1 and +1

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.

**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.

7) Sensitivity is a measure of how well a machine learning model can detect positive instances. It is also known as the true positive rate (TPR) or recall. Sensitivity is used to evaluate model performance because it allows us to see how many positive instances the model was able to correctly identify.

Mathematically, sensitivity or true positive rate can be calculated as the following: Sensitivity = (True Positive)/(True Positive + False Negative)

8) 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 : μ = 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: μ ≠100

9) 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.

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.

10) The interquartile range formula is the first quartile subtracted from the third quartile:

IQR = Q3 – Q1.

The range is calculated by subtracting the lowest value from the highest value.

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. Bell curves are visual representations of normal distribution, also called Gaussian distribution. Bell curves are useful for quickly visualizing a data set's mean, mode and median because when the distribution is normal, the mean, median and mode are all the same.

12) **Statistical outlier detection:** Statistical outlier detection involves applying statistical tests or procedures to identify extreme values.

You can convert extreme data points into z scores that tell you how many standard deviations away they are from the mean.

If a value has a high enough or low enough z score, it can be considered an outlier. As a rule of thumb, values with a z score greater than 3 or less than –3 are often determined to be outliers.

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. The smaller the p value, the more likely you are to reject the null hypothesis.

14) 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) 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.

These are applications of ANOVA:

1. Quality and cost comparison
2. Product safety tests
3. Optimize production
4. Understanding Data Sets