Answers for Statistics Worksheet-4

1. Central Limit Theorem states that the distribution of the mean of the random sample collected from any population tends to normal distribution when the sample size increases. Also, the population mean can be approximated as the mean of the sample means, if large random samples are collected.

Central Limit Theorem is important as the distribution of sample means will approach normal distribution for any population distribution, which is therefore, useful for estimating the population mean or conducting any statistical inferences.

2. Sampling is a part of statistical analysis, in which a fixed number of samples or observations are randomly taken from the population.

3. Type-I error (also known as false positive) occurs during hypothesis testing, when the hull hypothesis is falsely rejected, whereas, **Type-II** error (also known as false negative) occurs when we falsely reject the alternative hypothesis.

4. Normal distribution is a probability distribution a random variable that follows **'bell-shaped'** symmetric curve, where the most probable events in a series of data occurs towards centre of the curve.

5. Correlation in a value that is able to quantify the strength and direction of a variable with respect to other variable, in terms of how related they are. It can take values between -1 to 1, where negative, zero and positive values represents negative correlation, zero correlation and positive correlation respectively.

On the other hand, **covariance** also measures how two variables are related together, but its magnitude is unbounded, so it is difficult to interpret it, therefore, by dividing covariance by product of the standard deviations of two variables, one can calculate the normalized version of the statistic.

6.

S.No.	**Univariate Analysis**	**Bivariate Analysis**	**Multi-variate Analysis**
1.	It involves in analysing the one dimensional or single variable.	It involves in analysing two dimensional or two variables.	It involves in analysing three or more variables.
2.	It can be done using histograms, boxplots, etc	It can be done using scatter-plots, correlation, etc.	It can be done using 3-D plots, multi-variate regression analysis, cluster analysis, etc

7. Sensitivity of a test (also called as true positive rate) is the proportion of samples that are truly positive given a positive result in a statistical test.

Mathematically represented as:

$$\textbf{\textit{Sensitivity}} = \frac{\boldsymbol{Number\ of\ true\ Positives}}{\boldsymbol{Number\ of\ true\ Positives + Number\ of\ false\ Negatives}}$$

8. Hypothesis testing is a statistical testing approach to answer meaningful questions from the sample data, in which the answers are statistically significant rather than just mere chance.

Ho is called **null hypothesis**, which implies no statistical significance, whereas, **H1** is called **alternative hypothesis**, which implies statistical significance from the observed data.

For a two tailed test, **Ho** represent that the mean of the sample is not significantly different from the value we want to compare, whereas, **H1** represents that the mean of the sample is either significantly lower or higher than the value we want to compare with.

9. Quantitative data are usually numerical data that can be counted like, number of people, temperature, area of a house, etc. Whereas, **qualitative** data are usually categorical data that has inherent qualities within a particular category, for example, gender (like male or female), countries, colours, etc.

10. Range of any numerical variables is the difference between maximum and minimum values.

Interquartile range (IQR) is the difference between 75^{th} percentile and 25^{th} percentile.

11. Bell curve distribution is probability distribution also known as Gaussian or Normal curve that is symmetrical in shape about the central value. Here, the highest point on the '**bell-shaped'** curve implies, most probable events in a series of data occurs towards centre of the curve

12. In **Interquartile range (IQR)** method of outlier detection, the data points that are either below IQR*1.5 or above IQR*1.5 are considered as outliers.

13. A **p-value** is used in hypothesis testing to help us support or reject the null hypothesis. It is a probability value under the assumption of null hypothesis. For example, a p-value of 0.045, implies that, there is 4.5% chance that the result of a hypothesis test could be random, i.e., by chance.

14. The binomial distribution formula can calculate the probability of success for binomial distributions. Mathematically, it is given as:

$$P(x) = \binom{n}{x} p^x q^{n-x} = \frac{n!}{(n-x)!\,x!} p^x q^{n-x}$$

Here, ***x =*** number of success desired

n = the number of trials

p = probability of getting success in one trial

q = probability of getting success in one trial

15. ANOVA (Analysis of Variance) is a statistical technique to analyse the variation of a target continuous variable measured with respect to the conditions defined by categorical features. It is used to check whether the means of two or more groups are significantly different from each other in comparison.

ANOVA can be applied and not limited to the following cases:

- Analysing gas mileage of different vehicles or same vehicles under different fuel types.
- Understanding the impact of temperature, pressure or chemical concentration on some chemical reaction.
- Studying whether advertisements of different kinds affect the numbers of customer responses.
- Recommendation of a fertilizer against others for the improvement of a crop yield.
- Understanding the performance, quality or speed of manufacturing processes based on number of cells or steps they're divided into.