STATISTICS WORKSHEET-4

1. What is central limit theorem and why is it important?

Answer: The central limit theorem states that the distribution of sample means approximates a normal distribution as the sample size gets larger (assuming that all the samples are identical in size ), regardless of population distribution shape.

Central limit theorem is a statistical theory that states that given a sufficiently large sample size from a population with a finite level of variance, the mean of all samples from the same population will be approximately equal to the mean of the population.

All the samples will follow an approximate normal distribution pattern , with all variances being approximately equal to the variance of population divided by each sample’s size.

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. What is sampling? How many sampling methods do you know?

Answer: **Sampling**, means selecting a group (a sample) from a population from which we will collect data for our research. Sampling is an important aspect of a research study as the results of the study majorly depend on the sampling technique used. So, in order to get accurate results or the results that can estimate the population well, the sampling technique should be chosen wisely.

There are two types of sampling methods:

* [Probability sampling](https://www.scribbr.com/methodology/probability-sampling/) involves random selection, allowing you to make strong statistical inferences about the whole group.
* [Non-probability sampling](https://www.scribbr.com/methodology/non-probability-sampling/) involves non-random selection based on convenience or other criteria, allowing you to easily collect data.

3. What is the difference between type1 and typeII error?

Answer: In Type I error would be to reject the Null Hypothesis and say the machine is not running properly when in fact it was operating properly. Since the company does not want to needlessly stop production and recalibrate the machine, the satisfaction chooses to limit the probability of Type I error by setting the level of significance to 5%.

In Type II error is when we fail to reject the null hypothesis when it is actually false. In the prior example, the satisfaction failed to reject the Null Hypothesis because the probability of making Type I error (rejecting a true Null Hypothesis) exceeded the significance level of 5%. However, the satisfaction could have made Type II error if the machine is really operating improperly. One of the important and often overlooked tasks is to analyse the probability of making Type II error.

4. What do you understand by the term Normal distribution?

Answer: In normal distribution data are symmetrically distributed around mean, median, mode or bell-shaped distribution. The form of a normal distribution is determined by it’s mean and standard deviation. Mean = 0 and standard deviation =1. In normal distribution, approximately 68% of the data collected will fall within +/- one standard deviation of the mean; approximately 95% within +/- two standard deviation; and 99.7% within three standard deviations.

5. What is correlation and covariance in statistics?

Answer: **Correlation** gives a better understanding of covariance. It is normalized covariance.

Correlation tells us how correlated the variables are to each other. It is also called as Pearson correlation co-efficient. The value of correlation ranges from -1 to +1. -1 indicates negative correlation that is with an increase in 1 variable independent, there is a decrease in the other dependent variable. +1 indicates positive correlation that is with an increase in one variable independent, there is an increase in the other dependent variable. 0 indicates that the variables are independent of each other.

**Covariance** is a measure of the relationship between the variability of 2 variables that is it measures the degree of change in the variables. When one variable changes, will there be the same/ a similar change in the other variable.

6. Differentiate between univariate ,Biavariate,and multivariate analysis.

Answer: Univariate data contains only one variable. The purpose of the univariate analysis is to describe the data and find patterns that exist within it. Example: height of students. Bivariate data involves two different variables. The analysis of this type of data deals with causes and relationships and the analysis is done to determine the relationship between the two variables. Example: temperature and ice cream sales in the summer season Multivariate data involves three or more variables, it is categorized under multivariate. It is similar to a bivariate but contains more than one dependent variable. Example: data for house price prediction.

7. What do you understand by sensitivity and how would you calculate it?

Answer: Sensitivity is commonly used to validate the accuracy of a classifier (Logistic, SVM, Random Forest etc.). It can be calculated as: TN/(TN+FP).

8. What is hypothesis testing? What is H0 and H1? What is H0 and H1 for two-tail test?

Answer: A hypothesis test evaluates two mutually exclusive statements about a population to determine which statement is best supported by the sample data. Ex : you have a coin and you don’t know whether the coin is fair or not. First, decide null and alternate hypothesis and alpha=0.05. H0 : that is a fair coin. H1 : the coin is not fair.

9. What is quantitative data and qualitative data?

Answer: Quantitative data are anything that can be expressed as a number, or quantified. Examples of quantitative data are scores on achievement tests, number of hours of study, or weight of a subject. Qualitative data cannot be expressed as a number. Data that represent nominal scales such as gender, social, economic status, and religious preference are usually considered to be qualitative data.

10. How to calculate range and interquartile range?

Answer: The interquartile range is a measure of where the middle is in a data set. Where a range is a measure of where the beginning and end are in a set, an interquartile range is a measure of where the bulk of the values lie. That’s why it’s preferred over many other measures of spread when reporting things like school performance or SAT scores. The interquartile range formula is the first quartile subtracted from the third quartile: IQR = Q3 – Q1 .

11. What do you understand by bell curve distribution ?

Answer: In a bell curve, the centre contains the greatest number of a value and, therefore, it is the highest point on the arc of the line. This point is referred to the mean, but in simple terms, it is the highest number of occurrences of an element.

12. Mention one method to find outliers.

Answer: Outlier values can be identified by using boxplot or any other graphical analysis method. If the number of outlier values is few then they can be assessed individually but for large number of outliers the values can be substituted with either the 99th or the 1st percentile values. All extreme values are not outlier values. The most common ways to treat outlier values are: a. To change the value and bring in within a range b. To just remove the value.

13. What is p-value in hypothesis testing?

Answer: 13. p-value is defined as the probability that the data would be at least as extreme as those observed, if the null hypothesis were true. The p-value reflects the strength of evidence against the null hypothesis.

14. What is the Binomial Probability Formula?

Answer: The binomial distribution consists of the probabilities of each of the possible numbers of successes in N trials for independent events that each have a probability of π. Here, the possible outcomes are two.

15. Explain ANOVA and it’s applications.

Answer: Analysis of Variance (ANOVA) is a technique which is used to compare the means of multiple samples. Whether there is a significant difference between the mean of 2 samples, can be evaluated using z-test or t-test but in case of more than 2 samples, t-test cannot be applied as it accumulates the error and it will be more difficult as the number of sample will increase (for example: for 4 samples — 12 t-test will have to be performed). The ANOVA technique enables us to perform this simultaneous test.