Statistics Worksheet 4

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

Ans : The central limit theorem 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 holds for the sample of size greater than or equal to 30. This theorem is very important for testing hypotheses in statistical analysis. In other words, the central limit theorem states that for any population with mean and standard deviation, the distribution of the sample mean for sample size N has mean μ and standard deviation σ / √n . The CLT can be applied to almost all types of probability distributions. But there are some exceptions. For example, if the population has a finite variance. Also, this theorem applies to independent, identically distributed variables. It can also be used to answer the question of how big a sample you want. Remember that as the sample size grows, the standard deviation of the sample average falls because it is the population standard deviation divided by the square root of the sample size. This theorem is an important topic in statistics. In many real-time applications, a certain random variable of interest is a sum of a large number of independent random variables. In these situations, we can use the CLT to justify using the normal distribution.

Applications of Central Limit Theorem

1] The sample distribution is assumed to be normal when the distribution is unknown or not normally distributed according to Central Limit Theorem. This method assumes that the given population is distributed normally. It helps in data analysis.

2] The sample mean deviation decreases as we increase the samples taken from the population, which helps in estimating the mean of the population more accurately.

3] The sample mean is used in creating a range of values which likely includes the population mean.

4] The concept of the Central Limit Theorem is used in election polls to estimate the percentage of people supporting a particular candidate as confidence intervals.

5] CLT is used in calculating the mean family income in a particular country.

2) What is sampling? How many sampling methods do you know?

Ans : Sampling is a method that allows us to get information about the population based on the statistics from a subset of the population (sample), without having to investigate every individual. Sampling is done to draw conclusions about populations from samples, and it enables us to determine a population's characteristics by directly observing only a portion (or sample) of the population.Selecting a sample requires less time than selecting every item in a population.Sample selection is a cost-efficient method.Analysis of the sample is less cumbersome and more practical than an analysis of the entire population.

Sampling Methods are as:

Sampling methods consists of two type they are as:

1. Probability Sampling
2. Non-Probability Sampling

Probability Sampling

In probability sampling, every element of the population has an equal chance of being selected. Probability sampling gives us the best chance to create a sample that is truly representative of the population.

Types of probability sampling

1. Simple Random Sampling

This is a type of sampling technique you must have come across at some point. Here, every individual is chosen entirely by chance and each member of the population has an equal chance of being selected.

2. Systematic Sampling

In this type of sampling, the first individual is selected randomly and others are selected using a fixed 'sampling interval'.

3. Stratified Sampling

In this type of sampling, we divide the population into subgroups (called strata) based on different traits like gender, category, etc. We use this type of sampling when we want representation from all the subgroups of the population. However, stratified sampling requires proper knowledge of the characteristics of the population.

4. Cluster Sampling

In a clustered sample, we use the subgroups of the population as the sampling unit rather than individuals. The population is divided into subgroups, known as clusters, and a whole cluster is randomly selected to be included in the study

Non-Probability Sampling

In non-probability sampling, all elements do not have an equal chance of being selected. Consequently, there is a significant risk of ending up with a non-representative sample which does not produce generalizable results

Types of Non-Probability Sampling

1.Convenience Sampling

This is perhaps the easiest method of sampling because individuals are selected based on their availability and willingness to take part.

2.Quota Sampling

In this type of sampling, we choose items based on predetermined characteristics of the population.

3. Judgment Sampling

It is also known as selective sampling. It depends on the judgment of the experts when choosing whom to ask to participate.

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

Ans : In statistics, type I error is defined as an error that occurs when the sample results cause the rejection of the null hypothesis, in spite of the fact that it is true. In simple terms, the error of agreeing to the alternative hypothesis, when the results can be ascribed to chance.

Also known as the alpha error, it leads the researcher to infer that there is a variation between two observances when they are identical. The likelihood of type I error, is equal to the level of significance, that the researcher sets for his test. Here the level of significance refers to the chances of making type I error.

Type I error is an error that takes place when the outcome is a rejection of null hypothesis which is, in fact, true.

When on the basis of data, the null hypothesis is accepted, when it is actually false, then this kind of error is known as Type II Error. It arises when the researcher fails to deny the false null hypothesis. It is denoted by Greek letter 'beta (β)' and often known as beta error.

Type II error is the failure of the researcher in agreeing to an alternative hypothesis, although it is true. It validates a proposition; that ought to be refused. The researcher concludes that the two observances are identical when in fact they are not.

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

Ans :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.The normal distribution is also known as a Gaussian distribution or probability bell curve. It is symmetric about the mean and indicates that values near the mean occur more frequently than the values that are farther away from the mean.

The normal distribution is one of the most important probability distributions for independent random variables for three main reasons.

First, normal distribution describes the distribution of values for many natural phenomena in a wide range of areas, including biology, physical science, mathematics, finance and economics. It can also represent these random variables accurately.

Second, the normal distribution is important because it can be used to approximate other types of probability distribution, such as binomial, hypergeometric, inverse (or negative) hypergeometric, negative binomial and Poisson distribution.

Third, normal distribution is the key idea behind the central limit theorem, or CLT, which states that averages calculated from independent, identically distributed random variables have approximately normal distributions. This is true regardless of the type of distribution from which the variables are sampled, as long as it has finite variance.

5) What is correlation and covariance in statistics?

Covariance is an indicator of how two random variables are dependent on each other. A higher number denotes higher dependency. We can deduct correlation from a covariance. The value of covariance lies in the range of -∞ and +∞.Covariance is affected. Covariance has a definite unit as deduced by the multiplication of two numbers and their units.

Correlation indicates how strongly these two variables are related, provided other conditions are constant. The maximum value is +1, representing a perfect dependent relationship. Correlation provides a measure of covariance on a standard scale. It is deduced by dividing the calculated covariance by standard deviation. Correlation is limited to values between the range -1 and +1. Correlation is not affected by a change in scales or multiplication by a constant. Correlation is a unitless absolute number between -1 and +1, including decimal values.

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

Ans : 1. 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. It does not deal with causes or relationships and the main purpose of the analysis is to describe the data and find patterns that exist within it. The example of a univariate data can be height.

2. 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.Example of bivariate data can be temperature and ice cream sales in summer season.

3. Multivariate data –

When the data involves three or more variables, it is categorized under multivariate. Example of this type of data is suppose an advertiser wants to compare the popularity of four advertisements on a website, then their click rates could be measured for both men and women and relationships between variables can then be examined.

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

Ans : Sensitivity analysis is an analysis technique that works on the basis of what-if analysis like how independent factors can affect the dependent factor and is used to predict the outcome when analysis is performed under certain conditions. It is commonly used by investors who takes into consideration the conditions that affect their potential investment to test, predict and evaluate result.

The formula for sensitivity analysis is basically a financial model in excel where the analyst is required to identify the key variables for the output formula and then assess the output based on different combinations of the independent variables.Mathematically, the dependent output formula is represented as,Z = X2 + Y2.

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

Ans : Hypothesis testing can be defined as a statistical tool that is used to identify if the results of an experiment are meaningful or not. It involves setting up a null hypothesis and an alternative hypothesis. These two hypotheses will always be mutually exclusive. This means that if the null hypothesis is true then the alternative hypothesis is false and vice versa. An example of hypothesis testing is setting up a test to check if a new medicine works on a disease in a more efficient manner.

9) What is quantitative data and qualitative data?

Ans : 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.Qualitative data can be used to ask the question "why." It is investigative and is often open-ended until further research is conducted. Generating this data from qualitative research is used for theorizations, interpretations, developing hypotheses, and initial understandings.

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.Whereas qualitative is open for exploration, quantitative data is much more concise and close-ended. It can be used to ask the questions "how much" or "how many," followed by conclusive information.

10)How to calculate range and interquartile range?

Ans :

10) What do you understand by bell curve distribution ?

Ans : 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. A normal distribution curve, when graphed out, typically follows a bell-shaped curve, hence the name. While the precise shape can vary according to the distribution of the population, the peak is always in the middle and the curve is always symmetrical.

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.The long tail refers to the part of the bell curve that stretches out in either direction. If the diagram above represents a population under study, the fat area under the bell curve .

12)Mention one method to find outliers.

Ans : Outliers are values at the extreme ends of a dataset.Some outliers represent true values from natural variation in the population. Other outliers may result from incorrect data entry, equipment malfunctions, or other measurement errors.An outlier isn't always a form of dirty or incorrect data, so you have to be careful with them in data cleansing. What you should do with an outlier depends on its most likely cause.

The one of the method of finding outliers is as below

Sorting method

You can sort quantitative variables from low to high and scan for extremely low or extremely high values. Flag any extreme values that you find.This is a simple way to check whether you need to investigate certain data points before using more sophisticated methods.

13) What is p-value in hypothesis testing?

Ans : 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.

The p-value serves as an alternative to rejection points to provide the smallest level of significance at which the null hypothesis would be rejected. A smaller p-value means that there is stronger evidence in favor of the alternative hypothesis.

P-value is often used to promote credibility for studies or reports by government agencies. For example, the U.S. Census Bureau stipulates that any analysis with a p-value greater than 0.10 must be accompanied by a statement that the difference is not statistically different from zero. The Census Bureau also has standards in place stipulating which p-values are acceptable for various publications.

14)What is the Binomial Probability Formula?

Ans : Binomial probability refers to the probability of exactly x successes on n repeated trials in an experiment which has two possible outcomes (commonly called a binomial experiment).If the probability of success on an individual trial is p , then the binomial probability is $nCx \cdot px \cdot (1-p)n-x$ .Here nCx indicates the number of different combinations of x objects selected from a set of n objects. Some textbooks use the notation (nx) instead of nCx .Note that if p is the probability of success of a single trial, then (1−p) is the probability of failure of a single trial.

15)Explain ANOVA and it's applications.

Ans : Analysis of variance, or ANOVA, is a statistical method that separates observed variance data into different components to use for additional tests. A one-way ANOVA is used for three or more groups of data, to gain information about the relationship between the dependent and independent variables. The ANOVA test allows a comparison of more than two groups at the same time to determine whether a relationship exists between them. The result of the ANOVA formula, the F

statistic (also called the F-ratio), allows for the analysis of multiple groups of data to determine the variability between samples and within samples.

Applications of ANOVA

One- Way ANOVA:-It is hypothesis test in which only one categorical variable or single factor is taken into consideration.

With the help of F-distribution it enables us to compare the means of three or more samples.

Null hypothesis is “All population means should be equal” whereas Alternate hypothesis is “There should be the difference in at least one mean”

Assumptions:-Populations from which the samples are drawn are approximately normally distributed.

The populations from which the samples are drawn have the same variance.

The samples drawn from different populations are random and independent.

Applications:-Gender as categorical variable impacting state wise sales of ecomm sites e.g. Filpkart or Amazon.

Different level of Blood pressure in 3 groups of populations

Measure glycogen content for multiple samples of heart, liver, kidney, lungs etc

Two-way ANOVA:-

Examines the effect of two independent factors on dependent variable

Also studies the inter-relationship between independent variables influencing the values of the dependent variables, if any.

Assumptions:-Populations from which the samples are drawn are approximately normally distributed.The categorical independent group should have the same size.

Two or more than two categorical independent groups in two factors.

Measurement of dependent variables at continuous level.

Applications:-Analyzing the test score of a class based on gender and age. Here test score is the dependent variable and gender and age are the independent variables.

Measure response to three different drugs in both men and women. Drug treatment is one factor and gender is the other.