**STATISTICS WORKSHEET-4**

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

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

CLT 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 the population divided by each sample’s size.

**Importance of Central Limit Theoram**

* The central limit theorem tells us that no matter what the distribution of the population is, the shape of the sampling distribution will approach normality as the sample size (N) increases.
* This is useful, as the research never knows which mean in the sampling distribution is the same as the population mean, but by selecting many random samples from a population the sample means will cluster together, allowing the research to make a very good estimate of the population mean.
* Thus, as the sample size (N) increases the sampling error will decrease.

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

Ans- Sampling is a technique of selecting individual members or a subset of the population to make statistical inferences from them and estimate characteristics of the whole population. Different sampling methods are widely used by researchers in market research so that they do not need to research the entire population to collect actionable insights. It is also a time-convenient and a cost-effective method and hence forms the basis of any research design. Sampling techniques can be used in a research survey software for optimum derivation.

**Types of Sampling :**

* **Probability sampling:** Probability sampling is a sampling technique where a researcher sets a selection of a few criteria and chooses members of a population randomly. All the members have an equal opportunity to be a part of the sample with this selection parameter.
* **Non-probability sampling:**In non-probability sampling, the researcher chooses members for research at random. This sampling method is not a fixed or predefined selection process. This makes it difficult for all elements of a population to have equal opportunities to be included in a sample.
* **Types of Probability Sampling** **:**
* **Simple random sampling:** One of the best probability sampling techniques that helps in saving time and resources, is the Simple random sampling method. It is a reliable method of obtaining information where every single member of a population is chosen randomly, merely by chance. Each individual has the same probability of being chosen to be a part of a sample.
* **Cluster sampling:** Cluster sampling is a method where the researchers divide the entire population into sections or clusters that represent a population. Clusters are identified and included in a sample based on demographic parameters like age, sex, location, etc. This makes it very simple for a survey creator to derive effective inference from the feedback.
* **Systematic sampling:** Researchers use the Systematic sampling method to choose the sample members of a population at regular intervals. It requires the selection of a starting point for the sample and sample size that can be repeated at regular intervals. This type of sampling method has a predefined range, and hence this sampling technique is the least time-consuming.
* **Stratified random sampling:** Stratified random sampling is a method in which the researcher divides the population into smaller groups that don’t overlap but represent the entire population. While sampling, these groups can be organized and then draw a sample from each group separately.
* **Types of non-probability sampling :**
* **Convenience sampling:** This method is dependent on the ease of access to subjects such as surveying customers at a mall or passers-by on a busy street. It is usually termed as Convenience sampling, because of the researcher’s ease of carrying it out and getting in touch with the subjects. Researchers have nearly no authority to select the sample elements, and it’s purely done based on proximity and not representativeness. This non-probability sampling method is used when there are time and cost limitations in collecting feedback. In situations where there are resource limitations such as the initial stages of research, convenience sampling is used.
* **Judgmental or purposive sampling:** Judgmental or purposive sampling are formed by the discretion of the researcher. Researchers purely consider the purpose of the study, along with the understanding of the target audience. For instance, when researchers want to understand the thought process of people interested in studying for their master’s degree. The selection criteria will be: “Are you interested in doing your masters in …?” and those who respond with a “No” are excluded from the sample.
* **Snowball sampling:** Snowball sampling is a sampling method that researchers apply when the subjects are difficult to trace. For example, it will be extremely challenging to survey shelterless people or illegal immigrants. In such cases, using the snowball theory, researchers can track a few categories to interview and derive results. Researchers also implement this sampling method in situations where the topic is highly sensitive and not openly discussed—for example, surveys to gather information about HIV Aids. Not many victims will readily respond to the questions. Still, researchers can contact people they might know or volunteers associated with the cause to get in touch with the victims and collect information.
* **Quota sampling:**  In Quota sampling, the selection of members in this sampling technique happens based on a pre-set standard. In this case, as a sample is formed based on specific attributes, the created sample will have the same qualities found in the total population. It is a rapid method of collecting samples.

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

Ans- Difference between type 1 and type II error :

**Definition** – Type 1 error, in statistical hypothesis testing, is the error caused by rejecting a null hypothesis when it is true.

Type II error is the error that occurs when the null hypothesis is accepted when it is not true.

**Also termed** – Type 1 error is equivalent to false positive.

Type II error is equivalent to a false negative.

**Meaning** – Type 1 error is a false rejection of a true hypothesis.

Type II error is the false acceptance of as incorrect hypothesis.

**Symbol** – Type 1 error is denoted by α

Type II error is denoted by β

**Probability** – The probability of type 1 error is equal to the level of significance.

The probability of type II error is equal to one minus the power of test.

**Reduced** – Type 1 error can be reduced by decreasing the level of significance.

Type II error can be reduced by increasing the level of significance.

**Cause** - Type 1 error is caused by luch or chance.

Type II error is caused by a smaller sample size or a less powerful test.

**What Is It** – Type 1 error is similar to a false hit.

Type II error is similar to a miss.

**Hypothesis** – Type 1 error is associated with rejecting the null hypothesis.

Type II error is associated with rejecting the alternative hypothesis.

**When does it happen** – Type 1 error happens when the acceptance levels are set too lenient.

Type II error happens when the acceptance level are set too stringent.

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

Ans- Normal distribution, also known as the Gaussian distribution, is a probability distribution that is symmetric about the mean, showing that data near the mean are more frequent in occurrence than data far from the mean. In graph form, normal distribution will appear as a bell curve.

* A normal distribution is the proper term for a probability bell curve.
* In a normal distribution the mean is zero and the standard deviation is 1. It has zero skew and a kurtosis of 3.
* Normal distributions are symmetrical, but not all symmetrical distributions are normal.
* In reality, most pricing distributions are not perfectly normal.

1. What is correlation and covariance in statistics?

Ans**- Correlation** between two random variables, ρ(X,Y) is the covariance of the two variables normalized by the variance of each variable. This normalization cancels the units out and normalizes the measure so that it is always in the range [0, 1]:

* + - * + ρ(X,Y) = Cov(X,Y) **/**√ Var(X) Var(Y)

Correlation measures linearity between X and Y.

* + - * + ρ(X,Y) = 1 Y = aX + b where a = σy/σx
        + ρ(X,Y) = −1 Y = aX + b where a = −σy/σx
        + ρ(X,Y) = 0 absence of linear relationship

If ρ(X,Y) = 0 we say that X and Y are “uncorrelated.” If two variables are independent, then their correlation will be 0. However, like with covariance. it doesn’t go the other way. A correlation of 0 does not imply independence.

**Covariance** is a quantitative measure of the extent to which the deviation of one variable from its mean matches the deviation of the other from its mean. It is a mathematical relationship that is defined as:

* + - * Cov(X,Y) = E[(X − E[X])(Y − E[Y])]

The meaning of this mathematical definition may not be obvious at a first glance. If X and Y are both above their respective means, or if X and Y are both below their respective means, the expression inside the outer expectation will be positive. If one is above its mean and the other is below, the term is negative. If this expression is positive on average, the two random variables will have a positive correlation. We can rewrite the above equation to get an equivalent equation:

* + - * Cov(X,Y) = E[XY] − E[Y]E[X]

Using this equation (and the fact that the expectation of the product of two independent random variables is equal to the product of the expectations) is it easy to see that if two random variables are independent their covariance is 0. The reverse is not true in general: if the covariance of two random variables is 0, they can still be dependent.

1. Differentiate between univariate ,Biavariate,and multivariate analysis?

## Ans- Univarate Analysis

Univariate analysis is the simplest form of data analysis where the data being analyzed contains only one variable. Since it's a single variable it doesn’t deal with causes or relationships.  The main purpose of univariate analysis is to describe the data and find patterns that exist within it.

You can think of the variable as a category that your data falls into. One example of a variable in univariate analysis might be "age". Another might be "height". Univariate analysis would not look at these two variables at the same time, nor would it look at the relationship between them.

Some ways you can describe patterns found in univariate data include looking at mean, mode, median, range, variance, maximum, minimum, quartiles, and standard deviation. Additionally, some ways you may display univariate data include frequency distribution tables, bar charts, histograms, frequency polygons, and pie charts.

**Bivarate Analysis**

Bivariate analysis is used to find out if there is a relationship between two different variables. Something as simple as creating a scatterplot by plotting one variable against another on a Cartesian plane (think X and Y axis) can sometimes give you a picture of what the data is trying to tell you. If the data seems to fit a line or curve then there is a relationship or correlation between the two variables.  For example, one might choose to plot caloric intake versus weight.

**Multivariate Analysis**

Multivariate analysis is the analysis of three or more variables.  There are many ways to perform multivariate analysis depending on your goals.  Some of these methods include:

* Additive Tree
* Canonical Correlation Analysis
* Cluster Analysis
* Correspondence Analysis / Multiple Correspondence Analysis
* Factor Analysis
* Generalized Procrustean Analysis
* MANOVA
* Multidimensional Scaling
* Multiple Regression Analysis
* Partial Least Square Regression
* Principal Component Analysis / Regression / PARAFAC
* Redundancy Analysis.

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

Ans- The technique used to determine how independent variable values will impact a particular dependent variable under a given set of assumptions is defined as **sensitive analysis**. It’s usage will depend on one or more input variables within the specific boundaries, such as the effect that changes in interest rates will have on a bond’s price.

It is also known as the what – if analysis. Sensitivity analysis can be used for any activity or system. All from planning a family vacation with the variables in mind to the decisions at corporate levels can be done through sensitivity analysis.

Below are mentioned the steps used to conduct sensitivity analysis:

* Firstly the base case output is defined; say the NPV at a particular base case input value (V1) for which the sensitivity is to be measured. All the other inputs of the model  are kept constant.
* Then the value of the output at a new value of the input (V2) while keeping other inputs constant is calculated.
* Find the percentage change in the output and the percentage change in the input.
* The sensitivity is calculated by dividing the percentage change in output by the percentage change in input.
* This process of testing sensitivity for another input (say cash flows growth rate) while keeping the rest of inputs constant is repeated until the sensitivity figure for each of the inputs is obtained. The conclusion would be that the higher the sensitivity figure, the more sensitive the output is to any change in that input and vice versa.

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

Ans- Hypothesis testing is a statistical method that is used in making statistical decision using experimental data. Hypothesis testing is basically an assumption that we make about the population parameter.

EX : you say average student in class is 40 or a boy is taller that girls.

All those example we assume need some statistic way to prove those. We need some mathematical conclusion what ever we are assuming is true.

Hypothesis testing is an essential procedure in statistics. A hypothesis test evaluates two mutually exclusive statement about a population to determine which statement is best supported by the sample data . When we say that a finding is statistically significant , it’s thanks to a hypothesis test.

**Null hypothesis (H0) –** In inferential statistics, the null hypothesis is a general statement or default position that ther is no relationship between two measured phenomena, or no association among groups.

In other words it is a basic assumption or made based on domain or problem knowledge.

Example: a company production is = 50 unit/per day etc.

**Alternative hypothesis (H1)** – The alternative hypothesis is the hypothesis used in hypothesis testing that is contrary to the null hypothesis. It is usually taken to be that the observations are the result of a real effect (with some amount of chance variation superposed)

Example: a company production is != 50 unit/per day etc.

Two-tailed test: A two-tailed test is a statistical test in which the critical area of a distribution is two-sided and tests whether a sample is greater than or less than a certain range of values. If the sample being tested falls into either of the critical areas, the alternative hypothesis is accepted instead of the null hypothesis.

Example: a college! = 4000 student or data science != 80% org adopted.

1. What is quantitative data and qualitative data?

Ans**- Qualitative data** is non-statistical and is typically unstructured or semi-structured in nature. 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 type of data is measured using numbers and values, which makes 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

1. How to calculate range and interquartile range?

Ans- The range of a variable equals the difference between the maximum and minimum values. The range of income is:

Range(income) = max (income) – min (income) = 48000 – 24000 = 24000

Range only reflects the difference between largest and smallest observation, but it fails to reflect how data is centralized.

**Interquartile Range (IQR)**

The first quartile (Q1) is the 25th percentile of a data set, the second quartile (Q2) is the 50th percentile (median), and the third quartile (Q3) is the 75th percentile.

The IQR measures the difference between 75th and 25th observation using the formula:

IQR = Q3 – Q1

1. What do you understand by bell curve distribution ?

Ans- The term **bell curve** is used to describe the mathematical concept called normal distribution, sometimes referred to as Gaussian distribution. "Bell curve" refers to the bell shape that is created when a line is plotted using the data points for an item that meets the criteria of normal distribution.

In a bell curve, the center 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,](https://www.thoughtco.com/the-mean-median-and-mode-2312604) but in simple terms, it is the highest number of occurrences of an element (in statistical terms, the mode).

1. Mention one method to find outliers.

Ans- We can use the interquartile range (IQR), several quartile values, and an adjustment factor to calculate boundaries for what constitutes minor and major outliers. Minor and major denote the unusualness of the outlier relative to the overall distribution of values. Major outliers are more extreme. Analysts also refer to these categorizations as mild and extreme outliers.

The IQR is the middle 50% of the dataset. It’s the range of values between the third quartile and the first quartile (Q3-Q1). We can take the IQR, Q1 and Q3 values to calculate the following outlier fences for our dataset: lower outer, lower inner, upper inner, and upper outer. These fences determine whether data points are outliers and whether they are mild or extreme. Values that fall inside the two fences are not outliers.

1. What is p-value in hypothesis testing?

Ans- A p value is used in hypothesis testing to help you support or reject or reject the null hypothesis. The p value is the evidence against a null hypothesis. The smaller the p value, the stronger the evidence that you should reject the null hypothesis.

P values are expressed as decimal although it may be easier to understand what they are if you convert them to a percentage. For example, a p value of 0.0254 is 2.54%. This means there is a 2.54% chance your result could be random (i.e happened by chance). That’s pretty tiny. On the other hand , a large p-value of .9(90%) means your result have a 90% probability of being completely random and not due to anything in your experiment. Therefore the smaller the p-value, the more important (“significant”) your results.

When you run a hypothesis test, you compare the p value from your test to the alpha level you selected when you ran the test. Alpha levels can also be written as percentages.

Graphically, the p value is the area in the tail of a probability distribution. It’s calculated when you run hypothesis test and is the area to the right of the test statistic (if you are running a two-tailed test, it’s the area to the left and to the right).

1. What is the Binomial Probability Formula?

Ans-

**The binomial distribution formula is** :

**b(x; n, P) = nCx \* Px \* (1 – P)n – x**

**Where:  
 b = binomial probability  
 x = total number of “successes” (pass or fail, heads or tails etc.)  
 P = probability of a success on an individual trial  
 n = number of trials**

1. . Explain ANOVA and it’s applications?

Ans- In some decision-making situations, the sample data may be divided into various groups i.e. the sample may be supposed to have consisted of k-sub samples. There are interest lies in examining whether the total sample can be considered as homogenous or there is some indication that sub-samples have been drawn from different populations. So, in these situations, we have to compare the mean values of various groups, with respect to one or more criteria.

The total variation present in a set of data may be partitioned into a number of non-overlapping components as per the nature of the classification. The systematic procedure to achieve this is called **Analysis of Variance (ANOVA)**. With the help of such a partitioning, some testing of hypothesis may be performed.

Initially, Analysis of Variance (ANOVA) had been employed only for the experimental data from the Randomized Designs but later they have been used for analyzing survey and secondary data from the Descriptive Research.

Analysis of Variance may also be visualized as a technique to examine a dependence relationship where the response (dependence) variable is metric (measured on interval or ratio scale) and the factors (independent variables) are categorical in nature with a number of categories more than two.

There are two main types of ANOVA:

one-way (or unidirectional) and two-way. There also variations of ANOVA. For example, MANOVA (multivariate ANOVA) differs from ANOVA as the former tests for multiple dependent variables simultaneously while the latter assesses only one dependent variable at a time. One-way or two-way refers to the number of independent variables in your analysis of variance test. A one-way ANOVA evaluates the impact of a sole factor on a sole response variable. It determines whether all the samples are the same. The one-way ANOVA is used to determine whether there are any statistically significant differences between the means of three or more independent (unrelated) groups.

A two-way ANOVA is an extension of the one-way ANOVA. With a one-way, you have one independent variable affecting a dependent variable. With a two-way ANOVA, there are two independents. For example, a two-way ANOVA allows a company to compare worker productivity based on two independent variables, such as salary and skill set. It is utilized to observe the interaction between the two factors and tests the effect of two factors at the same time.