STATISTICS WORKSHEET-4

1Q. Central Limit Theorem : Central limit theorem is a statistical theory which states that when the large sample size is having a finite variance, the samples will be normally distributed and the mean of samples will be approximately equal to the mean of the whole population.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 .

As the sample size gets bigger and bigger, the mean of the sample will get closer to the actual population mean. If the sample size is small, the actual distribution of the data may or may not be normal, but as the sample size gets bigger, it can be approximated by a normal distribution. This statistical theory is useful in simplifying analysis while dealing with stock index and many more.

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.

2Q. 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.
Samples Methods:
Probability sampling : Involves random selection, allowing you to make strong statistical inferences about the whole group.
-Simple random sampling
-Cluster sampling
-Systematic sampling
-Stratified random sampling

Uses of probability sampling :

There are multiple uses of probability sampling:
. Reduce sample bias
. Diverse Population
. Create an Accurate Sample

Non-probability sampling : Involves non-random selection based on convenience or other criteria, allowing you to easily collect data

-Convenience sampling
-Judgmental or purposive sampling
-Snowball sampling
-Quota sampling

Uses of non-probability sampling :

.Create a hypothesis
.Exploratory research
.Budget and time constraints

3Q.

BASIS OF COMPARISON	**TYPE I ERROR**	**TYPE II ERROR**
Description	A type 1 error is also known as a false positive and occurs when a researcher incorrectly rejects a true null hypothesis.	A type II error does not reject the null hypothesis, even though the alternative hypothesis is the true state of nature. In other words,

		a false finding is accepted as true.
Alternative Name	A type I error also known as False positive**.**	A type II error also known as False negative. It is also known as false null hypothesis.
Other Names	The probability that we will make a type I error is designated 'α' (alpha). Therefore, type I error is also known as alpha error.	Probability that we will make a type II error is designated 'β' (beta). Therefore, type II error is also known as beta error.
Equivalence	The probability of type I error is equal to the level of significance.	The probability of type II error is equal to one minus

		the power of the test.
Associated With	Type I error is associated with rejecting the null hypothesis.	Type II error is associated with rejecting the alternative hypothesis.
Cause	It is caused by luck or chance.	It is caused by a smaller sample size or a less powerful test.
Probability	The probability of Type I error reduces with lower values of (*a)* since the lower value makes it difficult to reject null hypothesis.	The probability of Type II error reduces with higher values of (*a)* since the higher value makes it easier to reject the null hypothesis.
Preference	Type I errors are generally	Type II errors are

	considered more serious.	given less preference.
Reduction	It can be reduced by decreasing the level of significance.	It can be reduced by increasing the level of significance.
Occurence	It happens when the acceptance levels are set too lenient.	It happens when the acceptance levels are set too stringent.

4Q. Normal Distribution is a bell-shaped frequency distribution curve which helps describe all the possible values a random variable can take within a given range with most of the distribution area is in the middle and few are in the tails, at the extremes. This distribution has two key parameters: the mean (μ) and the standard deviation (σ) which plays key role in assets return calculation and in risk management strategy.

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 comes with a perfectly symmetrical shape. This means that the distribution curve can be divided in the middle to produce two equal halves. The symmetric shape occurs when one-half of the observations fall on each side of the curve.

5Q. CORRELATION : A correlation matrix is used to study the strength of a relationship between two variables. It not only shows the direction of the relationship, but also shows how strong the relationship is. The correlation formula can be represented as:

$$COR(X,Y)=COV(X,Y)/sqrt[VAR(X),VAR(Y)]$$

COVARIANCE : A covariance matrix is used to study the direction of the linear relationship between variables. Suppose we have two variables X and Y, then the covariance between these two variables is represented as cov(X,Y). If Σ(X) and Σ(Y) are the expected values of the variables, the covariance formula can be represented as:

$$COV(X,Y)=(1/n-1)\ \Sigma(xi-E(X))(yi-E(Y))$$

6Q.

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.

7Q. 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 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.

Sensitivity = True Positive/True Diseased * 100

8Q. A statistical hypothesis is an assertion or conjecture concerning one or more populations. To prove that a hypothesis is true, or false, with absolute certainty, we would need absolute knowledge. That is, we would have to examine the entire population. Instead, hypothesis testing concerns on how to use a random sample to judge if it is evidence that supports or not the hypothesis. Hypothesis testing is formulated in terms of two hypotheses:

- H0: the null hypothesis
- H1: the alternate hypothesis.

H0 is a null hypothesis while H1 is an alternative hypothesis. Research studies and testing usually formulate two hypotheses. One will describe the prediction while the other will describe all other possible outcomes. For example, you predict that A is related to B (null hypothesis).

Two-tailed hypothesis :

Two-tailed hypothesis tests are also known as nondirectional and two-sided tests because you can test for effects in both directions. When you perform a two-tailed test, you split the significance level percentage between both tails of the distribution.

9Q. Quantitative data : 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.

Qualitative data : 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.

10Q. In Statistics, the interquartile range is the smallest of all the measures of dispersion. It is the difference between the two extreme conclusions of the distribution. In other words, the range is the difference between the maximum and the minimum observation of the distribution It is defined by

Range = Xmax - Xmin

Where Xmax is the largest observation and Xmin is the smallest observation of the variable values.

The interquartile range defines the difference between the third and the first quartile. Quartiles are the partitioned values that divide the whole series into 4 equal parts. So, there are 3 quartiles. First Quartile is denoted by Q_1 known as the lower quartile, the second Quartile is denoted by Q_2 and the third Quartile is denoted by Q_3 known as the upper quartile. Therefore, the interquartile range is equal to the upper quartile minus lower quartile.

Interquartile range = Upper Quartile – Lower Quartile = $Q_3 – Q_1$

where Q_1 is the first quartile and Q_3 is the third quartile of the series

11Q. A bell curve is a common type of distribution for a variable, also known as the normal distribution. The term "bell curve" originates from the fact that the graph used to depict a normal distribution consists of a symmetrical bell-shaped curve.

The highest point on the curve, or the top of the bell, represents the most probable event in a series of data its mean , mode and median in this case, while all other possible occurrences are symmetrically distributed around the mean, creating a downward-sloping curve on each side of the peak. The width of the bell curve is described by its standard deviation.

The term "bell curve" is used to describe a graphical depiction of a normal probability distribution, whose underlying standard deviations from the mean create the curved bell shape. A standard deviation is a measurement used to quantify the variability of data dispersion, in a set of given values around the mean. The mean, in turn, refers to the average of all data points in the data set or sequence and will be found at the highest point on the bell curve.

Financial analysts and investors often use a normal probability distribution when analyzing the returns of a security or of overall market sensitivity. In finance, standard deviations that depict the returns of a security are known as volatility.

12Q. Z-scores can quantify the unusualness of an observation when your data follow the normal distribution. Z-scores are the number of standard deviations above and below the mean that each value falls. For example, a Z-score of 2 indicates that an observation is two standard deviations above the average while a Z-score of two signifies it is two standard deviations below the mean. A Z-score of zero represents a value that equals the mean.To calculate the Z-score for an observation, take the raw measurement, subtract the mean, and divide by the standard deviation.

13Q. In Statistics, the researcher checks the significance of the observed result, which is known as test static. For this test, a hypothesis test is also utilized. The P-value or probability value concept is used everywhere in the statistical analysis. It determines the statistical significance and the measure of significance testing. In this article, let us discuss its definition, formula, table, interpretation and how to use P-value to find the significance level etc. in detail.

The P-value is known as the probability value. It is defined as the probability of getting a result that is either the same or more extreme than the actual observations. The P-value

is known as the level of marginal significance within the hypothesis testing that represents the probability of occurrence of the given event. The P-value is used as an alternative to the rejection point to provide the least significance at which the null hypothesis would be rejected. If the P-value is small, then there is stronger evidence in favour of the alternative hypothesis.
We Know that P-value is a statistical measure, that helps to determine whether the hypothesis is correct or not. P-value is a number that lies between 0 and 1. The level of significance(ɑ) is a predefined threshold that should be set by the researcher. It is generally fixed as 0.05.

14Q. A binomial distribution can be thought of as simply the probability of a success or failure outcome in an experiment or survey that is repeated multiple times. The binomial is a type of distribution that has two possible outcomes the prefix bi means two, or twice. For example, a coin toss has only two possible outcomes: heads or tails and taking a test could have two possible outcomes: pass or fail.

The binomial distribution formula is:
$b(x; n, P) = {}_{n}C_{x} * P^{x} * (1 - P)^{n - x}$
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

15Q. 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.

ANOVA is a statistical technique that assesses potential differences in a scale-level dependent variable by a nominal-level variable having 2 or more categories. For example, an ANOVA can examine potential differences in IQ scores by Country US vs. Canada vs. Italy vs. Spain. Developed by Ronald Fisher in 1918, this test extends the *t* and the *z* test which have the problem of only allowing the nominal level variable to have two categories. This test is also called the Fisher analysis of variance.

The use of ANOVA depends on the research design. Commonly, ANOVAs are used in three ways: one-way ANOVA, two-way ANOVA, and N-way ANOVA.

One-Way ANOVA

A one-way ANOVA has just one independent variable. For example, difference in IQ can be assessed by Country, and

County can have 2, 20, or more different categories to compare.

Two-Way ANOVA

A two-way ANOVA (are also called factorial ANOVA) refers to an ANOVA using two independent variables. Expanding the example above, a 2-way ANOVA can examine differences in IQ scores (the dependent variable) by Country (independent variable 1) and Gender (independent variable 2). Two-way ANOVA can be used to examine the interaction between the two independent variables. Interactions indicate that differences are not uniform across all categories of the independent variables. For example, females may have higher IQ scores overall compared to males, but this difference could be greater (or less) in European countries compared to North American countries.

N-Way ANOVA

A researcher can also use more than two independent variables, and this is an n-way ANOVA (with n being the number of independent variables you have). For example, potential differences in IQ scores can be examined by Country, Gender, Age group, Ethnicity, etc, simultaneously.

General Purpose and Procedure

Omnibus ANOVA test:

The null hypothesis for an ANOVA is that there is no significant difference among the groups. The alternative hypothesis assumes that there is at least one significant difference among the groups. After cleaning the data, the researcher must test the assumptions of ANOVA. They must then calculate the *F*-ratio and the associated probability value (*p*-value). In general, if the *p*-value associated with the *F* is smaller than .05, then the null hypothesis is rejected and the alternative hypothesis is supported. If the null hypothesis is rejected, one concludes that the means of all the groups are not equal. Post-hoc tests tell the researcher which groups are different from each other.

So what if you find statistical significance? Multiple comparison tests

When you conduct an ANOVA, you are attempting to determine if there is a statistically significant difference among the groups. If you find that there is a difference, you will then need to examine where the group differences lay. At this point you could run post-hoc tests which are *t* tests examining mean differences between the groups. There are several multiple comparison tests that can be conducted that will control for Type I error rate, including the Bonferroni, Scheffe, Dunnet, and Tukey tests.