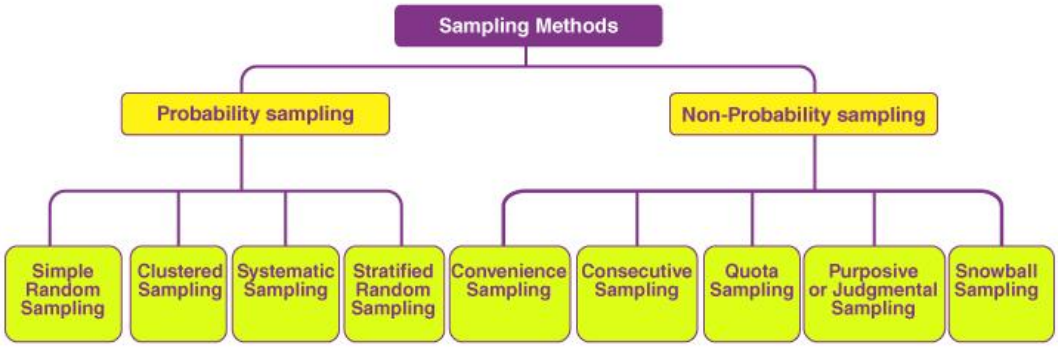
STATISTICS WORKSHEET-4

Ques1. What is central limit theorem and why is it important?
Ans: The Central Limit Theorem (CLT) is a statistical theory that posits that the mean and standard deviation derived from a sample, will accurately approximate the mean and standard deviation of the population the sample was taken from as the size of the sample increases. The minimum number of members of a population needed in order for a sample to adequately represent the population it was pulled from, is 30 according to the central limit theorem.

Purpose: In finance, the central limit theorem can be used to expedite analysis. Since indices often have hundreds, sometimes thousands of stocks contained within them an analyst doesn't have enough time in a month, much less a day to go through them all. But by putting the CLT to work, an analyst can take just 30 stocks out of an index and be able to approximate the quality of the index as a whole and thereby make a confident assessment.

Importance of Central Limit Theorem:
This is useful since the researcher never knows which mean in the sampling distribution corresponds to the population mean, but by taking numerous random samples from a population, the sample means will cluster together, allowing the researcher to obtain a very accurate estimate of the population mean.

Ques2. What is sampling? How many sampling methods do you know?
Ans: Sampling is the first step in the process of converting a continuous analog signal to a sequence of digital numbers. This article provides an insight into time and frequency domains of sampled signals. The concept of the spectral window, defined by the sampling process, helps understand digital signals and signal processing.



- **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
- **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

Ques3. What is the difference between type1 and typeII error?
Ans: **Key Differences Between Type I and Type II Error**

The points given below are substantial so far as the differences between type I and type II error is concerned:

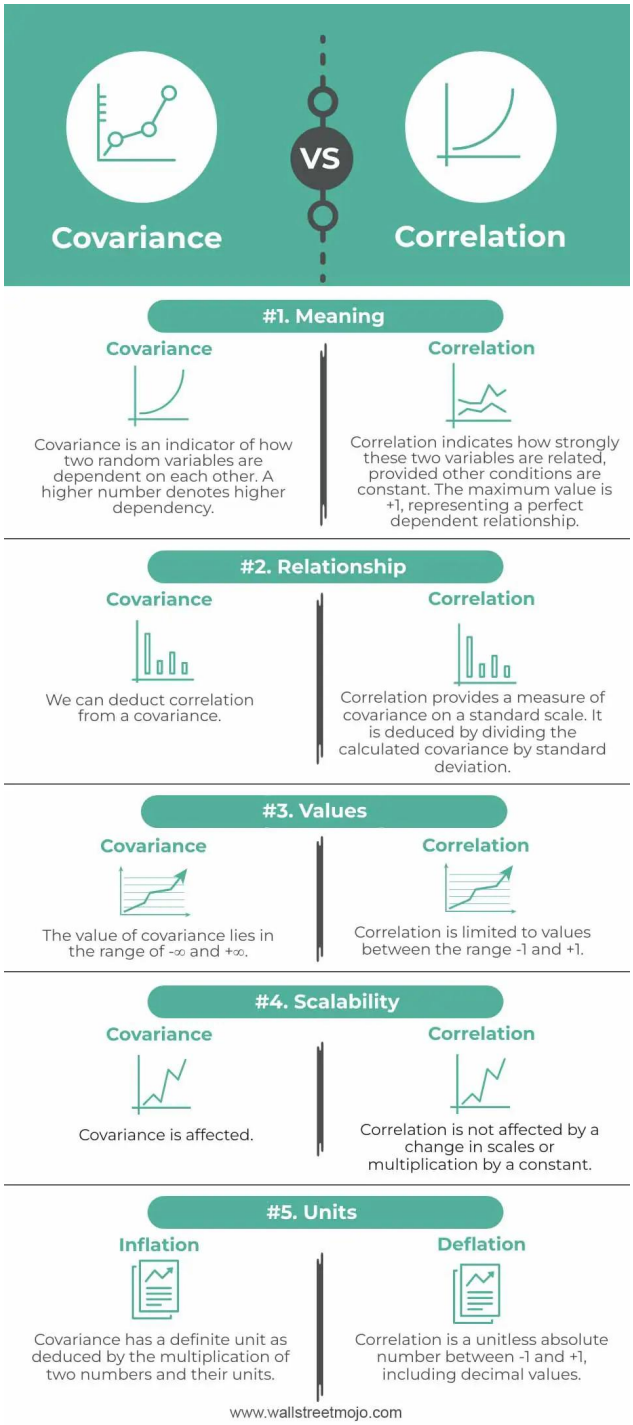
1. Type I error is an error that takes place when the outcome is a rejection of null hypothesis which is, in fact, true. Type II error occurs when the sample results in the acceptance of null hypothesis, which is actually false.
2. Type I error or otherwise known as false positives, in essence, the positive result is equivalent to the refusal of the null hypothesis. In contrast, Type II error is also known as false negatives, i.e. negative result, leads to the acceptance of the null hypothesis.
3. When the null hypothesis is true but mistakenly rejected, it is type I error. As against this, when the null hypothesis is false but erroneously accepted, it is type II error.
4. Type I error tends to assert something that is not really present, i.e. it is a false hit. On the contrary, type II error fails in identifying something, that is present, i.e. it is a miss.
5. The probability of committing type I error is the sample as the level of significance. Conversely, the likelihood of committing type II error is same as the power of the test.
6. Greek letter 'α' indicates type I error. Unlike, type II error which is denoted by Greek letter 'β'.

Ques4. What do you understand by the term Normal distribution?
Ans: Normal distribution is a continuous 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.

Ques5. What is correlation and covariance in statistics?

Ans: ***Covariance and correlation*** *are two terms that are exactly opposite to each other. However, they both are used in statistics and regression analysis. Covariance shows us how the two variables vary, whereas correlation shows us the relationship and how they are related.*

Correlation and covariance are two statistical concepts used to determine the relationship between two random variables. Correlation defines how a change in one variable will impact the other, while covariance defines how two items vary together.



Ques6. 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.

Ques7. 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.*

Calculation of the Sensitivity Analysis (Step by Step)

Lets start.

- Firstly, the analyst is required to design the basic formula, which will act as the output formula. For instance, say NPV formula can be taken as the output formula.
- Next, the analyst needs to identify which are the variables that are required to be sensitized as they are key to the output formula. In the NPV formula in excel, the cost of capital and the initial investment can be the independent variables.
- Next, determine the probable range of the independent variables.
- Next, open an excel sheet and then put the range of one of the independent variable along the rows and the other set along with the columns.

Ques8. What is hypothesis testing? What is H0 and H1? What is H0 and H1 for two-tail test?
Ans: *Hypothesis testing ascertains whether a particular assumption is true for the whole population. It is a statistical tool. It determines the validity of inference by evaluating sample data from the overall population.*

H1: The hypothesis that we are interested in proving. Null hypothesis:
H0: The complement of the alternative hypothesis.

Hypothesis testing is formulated in terms of two hypotheses: H0: the null hypothesis; • H1: the alternate hypothesis.

Ques9. What is quantitative data and qualitative data?
Ans: Qualitative data is a set of information which can not be measured using numbers. It generally consist of words, subjective narratives. Result of an qualitative data analysis can come in form of highlighting key words, extracting information and concepts elaboration. For example, a study on parents perception about the current education system for their kids. The resulted information collected from them might be in narrative form and you need to deduce the analysis that they are satisfied, un-satisfied or need improvement in certain areas and so on.

Quantitative data is a set of numbers collected from a group of people and involves statistical analysis.For example if you conduct a satisfaction survey from participants and ask them to rate their experience on a scale of 1 to 5. You can collect the ratings and being numerical in nature, you will use statistical techniques to draw conclusions about participants satisfaction.

Ques10. How to calculate range and interquartile range?
Ans: In Statistics, the **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

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 Q1 known as the lower quartile, the second Quartile is denoted by Q2 and the third Quartile is denoted by Q3 known as the upper quartile. Therefore, the interquartile range is equal to the upper quartile minus lower quartile.

The difference between the upper and lower quartile is known as the interquartile range. The formula for the interquartile range is given below

Interquartile range = Upper Quartile – Lower Quartile = Q3 – Q1

where Q1 is the first quartile and Q3 is the third quartile of the series.

Ques11. What do you understand by bell curve distribution ?
Ans: 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.

Understanding a Bell Curve

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.

Ques12. 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.

True outliers

True outliers should always be retained in your dataset because these just represent natural variations in your sample. True outliers are also present in variables with skewed distributions where many data points are spread far from the mean in one direction. It's important to select appropriate statistical tests or measures when you have a skewed distribution or many outliers.

Other outliers

Outliers that don't represent true values can come from many possible sources:

- Measurement errors
- Data entry or processing errors
- Unrepresentative sampling

This type of outlier is problematic because it's inaccurate and can distort your research results.

In practice, it can be difficult to tell different types of outliers apart. While you can use calculations and statistical methods to detect outliers, classifying them as true or false is usually a subjective process.

Ques13. What is p-value in hypothesis testing?

Ans: The P value, or calculated probability, is the probability of finding the observed, or more extreme, results when the **null hypothesis (H_0)** of a study question is true – the definition of 'extreme' depends on how the hypothesis is being tested. P is also described in terms of rejecting **H_0** when it is actually true, however, it is not a direct probability of this state. The null hypothesis is usually an hypothesis of "no difference" e.g. no difference between blood pressures in group A and group B. Define a null hypothesis for each study question clearly before the start of your study. The only situation in which you should use a **one sided** P value is when a large change in an unexpected direction would have absolutely no relevance to your study. This situation is unusual; if you are in any doubt then use a **two sided** P value. The term **significance level (alpha)** is used to refer to a pre-chosen probability and the term "P value" is used to indicate a probability that you calculate after a given study. The **alternative hypothesis (H_1)** is the opposite of the null hypothesis; in plain language terms this is usually the hypothesis you set out to investigate. For example, question is "is there a significant (not due to chance) difference in blood pressures between groups A and B if we give group A the test drug and group B a sugar pill?" and alternative hypothesis is " there is a difference in blood pressures between groups A and B if we give group A the test drug and group B a sugar pill". If your P value is less than the chosen significance level then you reject the null hypothesis i.e. accept that your sample gives reasonable evidence to support the alternative hypothesis. It does NOT imply a "meaningful" or "important" difference; that is for you to decide when considering the real-world relevance of your result. The choice of significance level at which you reject H_0 is arbitrary. Conventionally the 5% (less than 1 in 20 chance of being wrong), 1% and 0.1% ($P < 0.05$, 0.01 and 0.001) levels have been used. These numbers can give a false sense of security.

Ques14. What is the Binomial Probability Formula?

Ans: The binomial distribution formula helps to check the probability of getting "x" successes in "n" independent trials of a binomial experiment. To recall, the binomial distribution is a type of probability distribution in statistics that has two possible outcomes. In probability theory, the binomial distribution comes with two parameters n and p.

The probability distribution becomes a binomial probability distribution when it meets the following requirements.

- Each trial can have only two outcomes or the outcomes that can be reduced to two outcomes. These outcomes can be either a success or a failure.
- The trails must be a fixed number.
- The outcome of each trial must be independent of each others.
- And the success of probability must remain the same for each trial.

Ques15. Explain ANOVA and it's applications.

Ans: ANOVA stands for Analysis of Variance. One-Way Analysis of Variance tells you if there are any statistical differences between the means of three or more independent groups.

The one-way ANOVA can help you know whether or not there are significant differences between the means of your independent variables (such as the first example: age, gender, income). When you understand how each independent variable's mean is different from the others, you can begin to understand which of them has a connection to your dependent variable and begin to learn what is driving that behaviour.

Application of ANOVA

- ANOVA is designed to detect differences among means from populations subject to different treatments.
- ANOVA is a joint test - The equality of several population means is tested simultaneously or jointly.
- ANOVA tests for the equality of several population means by looking at two estimators of the population variance (hence,analysis of variance).