FLIP ROBO

STATISTICS WORKSHEET-4

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

- The central limit theorem (CLT) states that the distribution of sample means approximates a normal distribution as the sample size gets larger, regardless of the population's distribution.
- Sample sizes equal to or greater than 30 are often considered sufficient for the CLT to hold.
- A key aspect of CLT is that the average of the sample means and standard deviations will equal the population mean and standard deviation.
- A sufficiently large sample size can predict the characteristics of a population more accurately.
- CLT is useful in finance when analyzing a large collection of securities to estimate portfolio distributions and traits for returns, risk, and correlation.

Important

- 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.

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

- Sampling is a technique of selecting individual members or a subset of the population to make statistical inferences from them and estimate the 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 Methods

- **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.

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

Basis of Difference	**Type I Error**	**Type II Error**
Occurrence	A type I error occurs when the null **hypothesis** is true but is rejected. In other words, if a true null hypothesis is incorrectly rejected, type I error occurs.	A type II error occurs when the **null hypothesis** is false but invalidly fails to be rejected. In other words, failure to reject a false null hypothesis results in type II error.
Comparison	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.**

Designation	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.
Probability of committing error	Type I error equals to the level of significance (α) ‘α’ is the **so-called p-value.**	Type II error equals to the statistical power of a test. The probability 1- ‘β’ is **called the statistical power of the study.**
Represents	Type I error represents ‘a false hit’.	Type II error represents ‘a miss’.
Nature	We may reject the null hypothesis when the null hypothesis is true is known as Type I error.	We may accept the null hypothesis, when in fact null hypothesis is not true is known as Type II error.
Importance	Type I errors are generally considered more serious.	Type II errors are given less preference.
Acceptance	It refers to non-acceptance of hypothesis, which ought to be accepted.	It refers to the acceptance of hypothesis, which ought to be rejected.
Consequence	The probability of Type I error reduces with lower values of *(α)* since the lower value makes it difficult to reject null hypothesis.	The probability of Type II error reduces with higher values of *(α)* since the higher value makes it easier to reject the null hypothesis.

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

- The normal distribution, also known as the Gaussian distribution, is the most important probability distribution in statistics for independent, random variables. Most people recognize its familiar bell-shaped curve in statistical reports.
- The normal distribution is a continuous probability distribution that is symmetrical around its mean, most of the observations cluster around the central peak, and the probabilities for values further away from the mean taper off equally in both directions.

- Extreme values in both tails of the distribution are similarly unlikely. While the normal distribution is symmetrical, not all symmetrical distributions are normal.
- For example, the Student's t, Cauchy, and logistic distributions are symmetric.

5. What is correlation and covariance in statistics?

- Covariance measures how the two variables move concerning each other and is an extension of the concept of variance. It can take any value from -∞ to +∞.
    1. The higher this value, the more dependent the relationship is. A positive number signifies positive covariance and denotes a direct connection. Effectively this means that an increase in one variable would also lead to a corresponding increase in the other variable, provided other conditions remain constant.
    2. On the other hand, a negative number signifies negative covariance, which denotes an inverse relationship between the two variables. Though covariance is perfect for defining the type of relationship, it is not good for interpreting its magnitude.
- Correlation is a step ahead of covariance as it quantifies the relationship between two random variables. In simple terms, it is a unit measure of how these variables change concerning each other .
    1. The correlation has an upper and lower cap on a range, unlike covariance. It can only take values between +1 and -1. A correlation of +1 indicates that random variables have a direct and strong relationship.
    2. On the other hand, the correlation of -1 indicates a strong inverse relationship, and an increase in one variable will lead to an equal and opposite decrease in the other variable. 0 means that the two numbers are independent.

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

- **Univariate Analysis**

    - Univariate analysis is the simplest of the three analyses where the data you are analyzing is only one variable. There are many different ways people use univariate analysis. The most common univariate analysis is checking the central tendency (mean, median and mode), the range, the maximum and minimum values, and standard deviation of a variable.
    - Common visual technique used for univariate analysis is a histogram, which is a frequency distribution graph. You could also use a box plot or violin plot to compare the spread of the variables and provides an insight into outliers. Using any of the above mentioned to compare the "sepal_length" in the iris dataset across species is only comparing one variable, therefore a Univariate analysis.

- **Bivariate Analysis**
    - Bivariate analysis is where you are comparing two variables to study their relationships. These variables could be dependent or independent to each other. In Bivariate analysis is that there is always a Y-value for each X-value.
    - The most common visual technique for bivariate analysis is a scatter plot, where one variable is on the x-axis and the other on the y-axis. In addition to the scatter plot, regression plot and correlation coefficient are also frequently used to study the relationship of the variables. For example, continuing with the iris dataset, you can compare "*sepal_length*" vs "*sepal_width*" or "*sepal_length*" vs the "*petal_length*"to see if there is a relationship.

- **Multivariate Analysis**
    - Multivariate analysis is similar to Bivariate analysis but you are comparing more than two variables. For three variables, you can create a 3-D model

to study the relationship (also known as Trivariate Analysis). However, since we cannot visualize anything above the third dimension, we often rely on other softwares and techniques for us to be able to grasp the relationship in the data.

  - In terms of visualization, Seaborn library in Python allows for pairplots where it generates one large chart of selected variables against one another in a series of scatter plots and histograms depending on the type of variable, also known as scatter plot matrix. Again, in the series to come, I will provide the code and examples of this.
  - Depending on the dataset and the depth of analysis required, there are other techniques that you could deploy, such as Principal Component Analysis or logistic regression, linear regression, cluster analysis, etc. Again, in the series to come, I will provide the code and examples of this and dive deeper into PCA and its importance in data.

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

- Sensitivity analysis determines how different values of an independent variable affect a particular dependent variable under a given set of assumptions.
- This model is also referred to as a what-if or simulation analysis.
- Sensitivity analysis can be used to help make predictions in the share prices of publicly traded companies or how interest rates affect bond prices.
- Sensitivity analysis allows for forecasting using historical, true data.
- While sensitivity analysis determines how variables impact a single event, scenario analysis is more useful to determine many different outcomes for more broad situations.

Calculate Sensitivity Analysis

Sensitivity analysis is often performed in analysis software, and Excel has built in functions to help perform the analysis. In general, sensitivity analysis is calculated by leveraging formulas that reference different input cells. For example, a company may perform NPV analysis using a discount rate of 6%. Sensitivity analysis can be performed by analyzing scenarios of 5%, 8%, and 10% discount rates as well by simply maintaining the formula but referencing the different variable values.

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

- Hypothesis testing is used to assess the plausibility of a hypothesis by using sample data.
- The test provides evidence concerning the plausibility of the hypothesis, given the data.
- Statistical analysts test a hypothesis by measuring and examining a random sample of the population being analyzed.

Ho And H1

In hypothesis testing Ho is the null hypothesis and H1 (someteomes written Ha) is the alternative hypothesis.

The idea is that if the null hypothesis is true then the data are not likely to be far from satisfying it. For example, if the mean of the population is 3, the mean of the sample should not be too far from 3. If it's a long way from 3 (in the direction of the alternative) we would say that the null hypothesis is unlikely to be true and reject the null hypothesis at some level..

The logic is the reverse of what many students expect. If we reject the null hypothesis at the 5%*5%* level, that doesn't mean that the chance the Ho*Ho* is true is 0.05*0.05*, it means that the probability of getting a result in the critical region* **if** Ho is true is 0.05*0.05*.

*The critical region is a region far (in the direction of H1) from where we would expect the data to lie **if** Ho*Ho* is true. (That could be in any direction if H1 is two tailed.)

Two-Tailed Test

- In statistics, a two-tailed test is a method in which the critical area of a distribution is two-sided and tests whether a sample is greater or less than a range of values.
- It is used in null-hypothesis testing and testing for statistical significance.
- If the sample being tested falls into either of the critical areas, the alternative hypothesis is accepted instead of the null hypothesis.
- By convention two-tailed tests are used to determine significance at the 5% level, meaning each side of the distribution is cut at 2.5%.

9. What is quantitative data and qualitative data?

- **Quatitative** data are anything that can be expressed as a number, orquantified. Examples of quantitative data are scores on achievement tests,number of hours of study, or weight of a subject. These data may berepresented by ordinal, interval or ratio scales and lend themselves to moststatistical manipulation.
- **Qualitative** data cannot be expressed as a number. Data thatrepresent nominal scales such as gender, socieo economic status, religiouspreference are usually considered to be qualitative data.

10. How to calculate range and interquartile range?

The **interquartile range** is a measure of where the "middle fifty" 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.

Calculate range

To calculate the range, you need to find the largest observed value of a variable (the maximum) and subtract the smallest observed value (the minimum). The range only takes into account these two values and ignore the data points between the two extremities of the distribution. It's used as a supplement to other measures, but it is rarely used as the sole measure of dispersion because it's sensitive to extreme values.

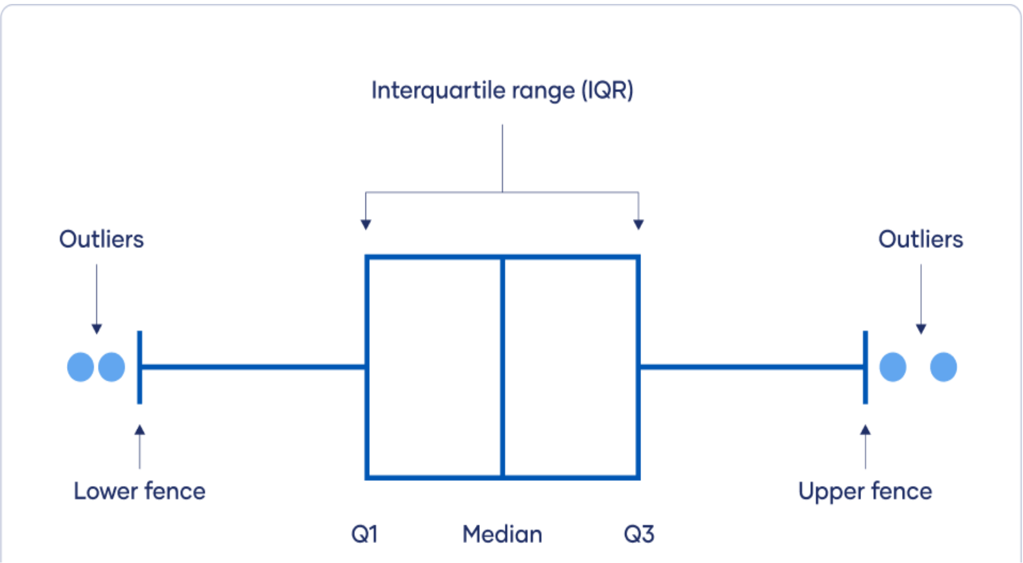
The interquartile range and semi-interquartile range give a better idea of the dispersion of data. To calculate these two measures, you need to know the values of the lower and upper quartiles. The lower quartile, or first quartile (Q1), is the value under which 25% of data points are found when they are arranged in increasing order. The upper quartile, or third quartile (Q3), is the value under which 75% of data points are found when arranged in increasing order. The median is considered the second quartile (Q2). The interquartile range is the difference between upper and lower quartiles. The semi-interquartile range is half the interquartile range.

11.What do you understand by bell curve distribution ?

- 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.

12. Mention one method to find outliers.

Using the interquartile range
The **interquartile range** (IQR) tells you the range of the middle half of your dataset. You can use the IQR to create "fences" around your data and then define outliers as any values that fall outside those fences.



This method is helpful if you have a few values on the extreme ends of your dataset, but you aren't sure whether any of them might count as outliers.

Interquartile range method

a. Sort your data from low to high
b. Identify the first quartile (Q1), the median, and the third quartile (Q3).
c. Calculate your IQR = Q3 – Q1
d. Calculate your upper fence = Q3 + (1.5 * IQR)
e. Calculate your lower fence = Q1 – (1.5 * IQR)
f. Use your fences to highlight any outliers, all values that fall outside your fences.

Your outliers are any values greater than your upper fence or less than your lower fence.

13. What is p-value in hypothesis testing?

- A p-value is a statistical measurement used to validate a hypothesis against observed data.
- A p-value measures the probability of obtaining the observed results, assuming that the null hypothesis is true.
- The lower the p-value, the greater the statistical significance of the observed difference.

- A p-value of 0.05 or lower is generally considered statistically significant.
- P-value can serve as an alternative to or in addition to preselected confidence levels for hypothesis testing.

14. What is the Binomial Probability Formula?

Binomial probability refers to the probability of exactly x*x* successes on n*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·px·(1−p)n−x*nCx·px·(1−p)n−x*

Here *nCx* indicates the number of different combinations of x*x* objects selected from a set of n*n* objects. Some textbooks use the notation *(nx)* instead of *nCx* .

Note that if p*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.

- 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.
- If no true variance exists between the groups, the ANOVA's F-ratio should equal close to 1.

There are two main types of ANOVA viz. one–way ANOVA and two–way ANOVA.

a. **One Way ANOVA –** It is also known as one factor ANOVA. Here, we are using one criterion variable (or called as a factor) and analyze the difference between more than two sample groups. Suppose in glass industry, we want to compare the variation of three batches (glass) for their average weight (factor).

b. **Two Way ANOVA –** Here, we are using two independent variables (factors) and analyze the difference between more than two sample groups. Similarly, we want to compare the variation of three batches of glass w.r.t weight and hardness (two factors).

We have discussed the basic concepts of ANOVA when we have one factor, we use one-way ANOVA and when we have two factors, we use two-way ANOVA. I guess this concept is clear and understandable.

WORKSHEET