Statistics Worksheet 4

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

   The CLT is 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.

   Consider there are 15 sections in class X, and each section has 50 students. Our task is to calculate the average marks of students in class X.

   The standard approach will be to calculate the average simply:

   - Calculate the total marks of all the students in Class X
   - Add all the marks
   - Divide the total marks by the total number of students

   But what if the data is extremely large? Is this a good approach? No way, calculation marks of all the students will be a tedious and time-consuming process. So, what are the alternatives? Let's take a look at another approach.

   To begin, select groups of students from the class at random. This will be referred to as a sample. Create several samples, each with 30 students.

   - Calculate each sample's individual mean.
   - Calculate the average of these sample means.
   - The value will give us the approximate average marks of the students in Class X.
   - The histogram of the sample means marks of the students will resemble a bell curve or normal distribution.

2. What is sampling? How many sampling methods do you know?
   Sample is a subset of the population. It is the specific group from which you collect data. The number of elements or individuals in a sample is called the sample size. The process of selecting a sample is called sampling. For example, the sample of sheep in Rajasthan, India; the sample of elementary school students in New York, US; the sample of data science blogging websites on the internet. The size of the sample is always less than the size of the population.

   Sampling methods :
   - Simple Random Sampling (SRS) :
   - Stratified Sampling
   - Cluster Sampling

- Systematic Sampling
- Convenience Sampling

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.

As with any probability distribution, the normal distribution describes how the values of a variable are distributed. It is the most important probability distribution in statistics because it accurately describes the distribution of values for many natural phenomena. Characteristics that are the sum of many independent processes frequently follow normal distributions. For example, heights, blood pressure, measurement error, and IQ scores follow the normal distribution

5. What is correlation and covariance in statistics?
Covariance and Correlation are very helpful in understanding the relationship between two continuous variables. Covariance tells whether both variables vary in the same direction (positive covariance) or in the opposite direction (negative covariance). There is no meaning of covariance numerical value only sign is useful. Whereas Correlation explains the change in one variable leads how much proportion change in the second variable. Correlation varies between -1 to +1. If the correlation value is 0 then it means there is no Linear Relationship between variables however other functional relationship may exist.
Covariance:
In the study of covariance only sign matters. A positive value shows that both variables vary in the same direction and negative value shows that they vary in the opposite direction.

Covariance between two variables x and y can be calculated as follows:

$$s_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

Where:

- x̄ is the sample mean of x
- ȳ is sample mean of y
- x_i and y_i are the values of x and y for ith record in the sample.
- n is the no of records in the sample

Correlation:

As covariance only tells about the direction which is not enough to understand the relationship completely, we divide the covariance with a standard deviation of x and y respectively and get correlation coefficient which varies between -1 to +1.

-1 and +1 tell that both variables have a perfect linear relationship.
Negative means they are inversely proportional to each other with the factor of correlation coefficient value.
Positive means they are directly proportional to each other mean vary in the same direction with the factor of correlation coefficient value.
if the correlation coefficient is 0 then it means there is no linear relationship between variables however there could exist other functional relationship.
if there is no relationship at all between two variables then correlation coefficient will certainly be 0 however if it is 0 then we can only say that there is no linear relationship but there could exist other functional relationship.
Correlation between x and y can be calculated as follows:

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

Where:

- S_xy is the covariance between x and y.
- S_x and S_y are the standard deviations of x and y respectively.
- r_xy is the correlation coefficient.

The correlation coefficient is a dimensionless quantity. Hence if we change the unit of x and y then also the coefficient value will remain the same.

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

Heaight (cm)	164	151	160	159	163	158	155

Suppose that the heights of seven students of a class is recorded(above table), there is only one variable that is height and it is not dealing with any cause or relationship. The description of patterns found in this type of data can be made by drawing conclusions using central tendency measures (mean, median and mode), dispersion or spread of data (range, minimum, maximum, quartiles, variance and standard deviation) and by using frequency distribution tables, histograms, pie charts, frequency polygon and bar charts.

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.

Temperature (Celsius)	Ice cream sales
20	2000
25	2500
35	5000
43	7800

Suppose the temperature and ice cream sales are the two variables of a bivariate data(above table). Here, the relationship is visible from the table that temperature and sales are directly proportional to each other and thus related because as the temperature increases, the sales also increase. Thus bivariate data analysis involves comparisons, relationships, causes and explanations. These variables are often plotted on X and Y axis on the graph for better understanding of data and one of these variables is independent while the other is dependent.

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.

   It is similar to bivariate but contains more than one dependent variable. The ways to perform analysis on this data depends on the goals to be achieved. Some of the techniques are regression analysis, path analysis, factor analysis and multivariate analysis of variance (MANOVA).

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

The sensitivity of a test is also called the true positive rate (TPR) and is the proportion of samples that are genuinely positive that give a positive result using the test in question. For example, a test that correctly identifies all positive samples in a panel is very sensitive. Another test that only detects 60 % of the positive samples in the panel would be deemed to have lower sensitivity as it is missing positives and giving higher a false negative rate (FNR). Also referred to as type II errors, false negatives are the failure to reject a false null hypothesis (the null hypothesis being that the sample is negative).

The following equation is used to calculate a test’s sensitivity:

$$\text{Sensitivity} = \frac{\text{Number of true positives}}{\text{(Number of true positives + Number of false negatives)}}$$

$$= \frac{\text{Number of true positives}}{\text{Total number of individuals with the illness}}$$

8. What is hypothesis testing? What is H0 and H1? What is H0 and H1 for two-tail test?
   Hypothesis Testing is a type of statistical analysis in which you put your assumptions about a population parameter to the test. It is used to estimate the relationship between 2 statistical variables.

   The Null Hypothesis is the assumption that the event will not occur. A null hypothesis has no bearing on the study's outcome unless it is rejected.
   H0 is the symbol for it, and it is pronounced H-naught.

   The Alternate Hypothesis is the logical opposite of the null hypothesis. The acceptance of the alternative hypothesis follows the rejection of the null hypothesis.
   H1 is the symbol for it.

   In two tails, the test sample is checked to be greater or less than a range of values in a Two-Tailed test, implying that the critical distribution area is two-sided.
   If the sample falls within this range, the alternate hypothesis will be accepted, and the null hypothesis will be rejected.

9. What is quantitative data and qualitative data?
   Quantitative data: The data collected on the grounds of the numerical variables are quantitative data. Quantitative data are more objective and conclusive in nature. It measures the values and is expressed in numbers. The data collection is based on “how much” is the quantity. The data in quantitative analysis is expressed in numbers so it can be counted or measured. The data is extracted from experiments, surveys, market reports, matrices, etc.
   Qualitative data: The data collected on grounds of categorical variables are qualitative data. Qualitative data are more descriptive and conceptual in nature. It measures the data on basis of the type of data, collection, or category. The data collection is based on what type of quality is given. Qualitative data is categorized into different groups based on characteristics. The data obtained from these kinds of analysis or research is used in theorization, perceptions, and developing hypothetical theories. These data are collected from texts, documents, transcripts, audio and video recordings, etc.

10. How to calculate range and interquartile range?
    Let us see how to calculate range and interquartile range by considering an example data set : 6, 47, 49, 15, 43, 41, 7, 39, 43, 41, 36
    - You first need to arrange the data points in increasing order. As you do so, you can give them a rank to indicate their position in the data set. Rank 1 is the data point with the smallest value, rank 2 is the data point with the second-lowest value, etc.

Rank	Value

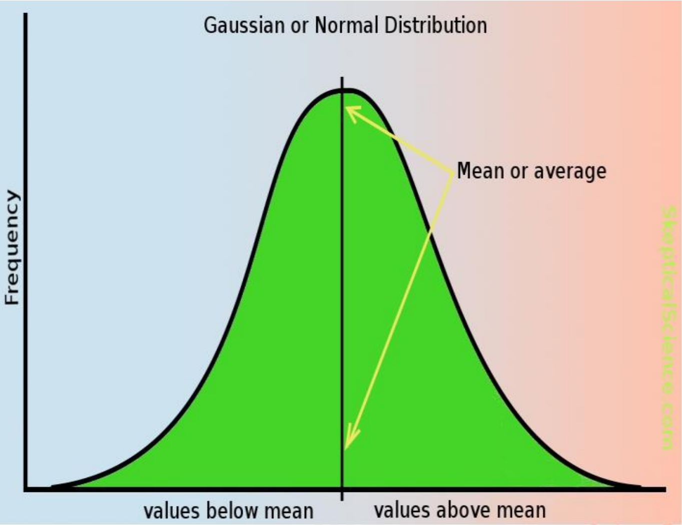
1	6
2	7
3	15
4	36
5	39
6	41
7	41
8	43
9	43
10	47
11	49

- Then you need to find the rank of the median to split the data set in two. As we have seen in the section on the median, if the number of data points is an uneven value, the rank of the median will be

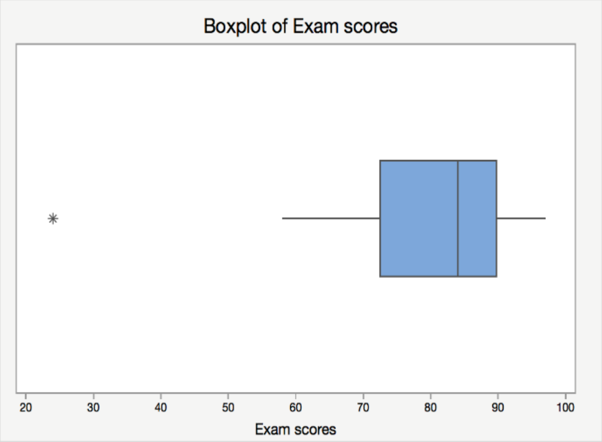
  (n + 1) ÷ 2 = (11 + 1) ÷ 2 = 6

- The rank of the median is 6, which means there are five points on each side.
- Then you need to split the lower half of the data in two again to find the lower quartile. The lower quartile will be the point of rank (5 + 1) ÷ 2 = 3. The result is Q1 = 15. The second half must also be split in two to find the value of the upper quartile. The rank of the upper quartile will be 6 + 3 = 9. So Q3 = 43.
- Once you have the quartiles, you can easily measure the spread. The interquartile range will be Q3 - Q1, which gives 28 (43-15). The semi-interquartile range is 14 (28 ÷ 2) and the range is 43 (49-6).

11. What do you understand by bell curve distribution ?
    The normal distribution is often called the bell curve because the graph of its probability density looks like a bell. It is also known as called Gaussian distribution, after the German mathematician Carl Gauss who first described it.
    The normal distribution is a continuous probability distribution that is symmetrical on both sides of the mean, so the right side of the centre is a mirror image of the left side.
    The area under the normal distribution curve represents probability and the total area under the curve sums to one.
    Most of the continuous data values in a normal distribution tend to cluster around the mean, and the further a value is from the mean, the less likely it is to occur. The tails are asymptotic, which means that they approach but never quite meet the horizon (i.e. x-axis).
    For a perfectly normal distribution the mean, median and mode will be the same value, visually represented by the peak of the curve.



12. Mention one method to find outliers.
    One of the easiest method that can be used to find out outliers is by graphing the data. Boxplots, histograms, and scatterplots can highlight outliers.
    Boxplots display asterisks or other symbols on the graph to indicate explicitly when datasets contain outliers. These graphs use the interquartile method with fences to find outliers, which I explain later. The boxplot below displays an example, It's clear that the outlier is quite different than the typical data value.



13. What is p-value in hypothesis testing?
    In statistical hypothesis testing, P-Value or probability value can be defined as the measure of the probability that a real-valued test statistic is at least as extreme as the value actually obtained. P-value shows how likely it is that your set of observations could have occurred under the null hypothesis. P-Values are used in statistical hypothesis testing to determine whether to reject the null hypothesis. The smaller the p-value, the stronger the likelihood that you should reject the null hypothesis.

P-values are expressed as decimals and can be converted into percentage. For example, a p-value of 0.0237 is 2.37%, which means there's a 2.37% chance of your results being random or having happened by chance. The smaller the P-value, the more significant your results are.

14. What is the Binomial Probability Formula?
    In the binomial probability, the number of successes X in ‘n’ trials of a binomial experiment is called a binomial random variable. The probability distribution of the random variable X is called a binomial distribution, and is given by the formula as below:
    P(X)=Cnxpxqn−xP(X)=Cxnpxqn−x
    Where n is the number of trials, x is 0, 1, 2..., n, p is the probability of success in a single trial, q is the probability of failure in a single trial and the value of q is 1-p. P(X) gives the probability of successes in n binomial trials.
    The combination formula is Cnx=n!x!(n−x)!Cxn=n!x!(n−x)!.

15. Explain ANOVA and it’s applications
    ANOVA is to test for differences among the means of the population by examining the amount of variation within each sample, relative to the amount of variation between the samples. Analysing variance tests the hypothesis that the means of two or more populations are equal.
    In a regression study, analysts use the ANOVA test to determine the impact of independent variables on the dependent variable.
    Applications :
    - Test students from multiple colleges to see if students from one of the colleges consistently outperform students from the other colleges.
    - In a business application, an R&D researcher might test two different processes of creating a product to see if one process is better than the other in terms of cost efficiency.
    - Suppose, there is a group of patients who are suffering from fever.
      They are being given three different medicines that have the same functionality i.e. to cure fever. To understand the effectiveness of each medicine and choose the best among them, the ANOVA test is used.