Name of the Assignment	: Statistics (Worksheet – 4)
Submitted by	: Shalini Joshi
Designation	: Data Science Intern
Date of Submission	: 7th Jan,2023

Q1to Q15 are descriptive types. Answer in brief.

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

Ans: The **central limit theorem** states that the sampling distribution of a sample mean is approximately normal if the sample size is large enough, *even if the population distribution is not normal*.

The central limit theorem also states that the sampling distribution will have the following properties:

- The mean of the sampling distribution will be equal to the mean of the population distribution:

$$x = \mu$$

- The variance of the sampling distribution will be equal to the variance of the population distribution divided by the sample size:

$$s2 = \sigma2 / n$$

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.

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

Ans: Sampling is a process in statistical analysis where researchers take a predetermined number of observations from a larger population.

The chosen sample should be a fair representation of the entire population. When taking a sample from a larger population, it is important to consider how the sample is chosen.

To draw valid conclusions from your results, you have to carefully decide how you will select a sample that is representative of the group as a whole. This is called a *sampling method*.

There are two primary types of sampling methods that you can use in your research:

- ***Probability sampling*** involves random selection, allowing you to make strong statistical inferences about the whole group.
- ***Non-probability sampling*** involves non-random selection based on convenience or other criteria, allowing you to easily collect data.

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

S.No.	Basis for Comparison	Type 1 Error	TypeII Error
1	What is it?	It is incorrect rejection of true null hypothesis.	It is incorrect acceptance of false null hypothesis.
2	Represents	A false hit	A miss
3	Probability of committing error	Equals the level of significance	Equals the power of test
4	Indicated by	Greek letter 'α'	Greek letter 'β'

4. 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 Normal distribution, the mean(average)= mode(most frequent observation) = median(mid-point).

KEY TAKEAWAYS

- The 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.
- Many naturally-occurring phenomena tend to approximate the normal distribution.
- In finance, most pricing distributions are not, however, perfectly normal.

5. What is correlation and covariance in statistics?

Ans: A **correlation** is a statistical measure of the relationship between two variables. The measure is best used in variables that demonstrate a linear relationship between each other.

The correlation coefficient is a value that indicates the strength of the relationship between variables. The coefficient can take any values from -1 to 1. The interpretations of the values are:

- -1: Perfect negative correlation. The variables tend to move in opposite directions (i.e., when one variable increases, the other variable decreases).
- 0: No correlation. The variables do not have a relationship with each other.
- 1: Perfect positive correlation. The variables tend to move in the same direction (i.e., when one variable increases, the other variable also increases).

Covariance is a measure of the relationship between two random variables and to what extent, they change together.

Types of Covariance

Covariance can have both positive and negative values. Based on this, it has two types:

- Positive Covariance
- Negative Covariance

Positive Covariance

If the covariance for any two variables is positive, that means, both the variables move in the same direction. Here, the variables show similar behaviour. That means, if the values (greater or lesser) of one variable corresponds to the values of another variable, then they are said to be in positive covariance.

Negative Covariance

If the covariance for any two variables is negative, that means, both the variables move in the opposite direction. It is the opposite case of positive covariance, where greater values of one variable correspond to lesser values of another variable and vice-versa.

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

Ans:

- Univariate statistics summarize only one variable at a time.
- Bivariate statistics compare two variables.
- Multivariate statistics compare more than two variables.

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

Ans: Sensitivity measures how often a test correctly generates a positive result for people who have the condition that's being tested for.

Sensitivity analysis is a method for predicting the outcome of a decision if a situation turns out to be different compared to the key predictions.

Formula to calculate sensitivity:

$$\textbf{Sensitivity} = \frac{\textbf{Total Positive Tests}}{\textbf{Total True Positives + Total False Negatives}}$$

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

Ans: Hypothesis testing in statistics refers to analyzing an assumption about a population parameter. It is used to make an educated guess about an assumption using statistics. With the use of sample data, hypothesis testing makes an assumption about how true the assumption is for the entire population from where the sample is being taken.

H0 is a false hypothesis, whereas H1 is an alternative one. Two hypotheses are usually formed in research studies and testing. Although the alternative hypothesis may be negative, it is not always a rejection of the null hypothesis, but rather a test of whether or not the null hypothesis is correct.

H 1: μ ≠ μ 0, where a difference is hypothesized and this is called a two-tailed test.

9. What is quantitative data and qualitative data?

Ans: Quantitative data refers to any information that can be quantified. If it can be counted or measured, and given a numerical value, it's quantitative data. For eg. How much revenue did the company make in 2019? It is analyzed using statistical analysis.

Unlike quantitative data, qualitative data cannot be measured or counted. It’s descriptive, expressed in terms of language rather than numerical values. For eg. Product reviews and customer testimonials. It is analyzed by grouping it in terms of meaningful categories or themes

10. How to calculate range and interquartile range?

Ans: To calculate range:

Range = Highest value - lowest value (H - L = R)

To calculate **Interquartile Range (IQR):**

IQR = Q3-Q1

Q3: third quartile

Q1: first quartile

11. 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.

12. Mention one method to find outliers.

Ans: One of the methods to find outliers is: **‘Sorting method’**

You can **sort** quantitative variables from low to high and scan for extremely low or extremely high values. Flag any extreme values that you find.

13. What is p-value in hypothesis testing?

Ans: The p in p-value stands for **probability**. The p-value method is used in Hypothesis Testing to check the significance of the given Null Hypothesis. Then, deciding to reject or support it is based upon the specified significance level or threshold.

A p-value is calculated in this method which is a test statistic. This statistic can give us the probability of finding a value (Sample Mean) that is as far away as the population mean. Based on that probability and a significance level, we Reject or Fail to Reject the Null

Hypothesis. Generally, the lower the p-value, the higher the chances are for Rejecting the Null Hypothesis and vice versa.

14. What is the Binomial Probability Formula?

Ans: The binomial probability formula for any random variable x is given by

P (x : n, p) = nCx px qn-x

n = the number of trials

x varies from 0, 1, 2, 3, 4, …

p = probability of success

q = probability of failure = 1 – p

15. Explain ANOVA and its applications.

Ans: 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.

ANOVA has it's applications in the following areas:

- Used to design an area; With ANOVA, you can get designs like; Randomized complete block design (RCBD) and Latin square design (LSD).
- Used in identifying gender age differences
- Used in identifying how far a person can throw javelin
- Used in analyzing variance between samples
- Used to determine the best materials to build products for your customers
- Used in healthcare and food industries

**************************The End**