Statistics
Answers Sheet

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

   In simple terms, the theorem states that the sampling distribution of the mean approaches a normal distribution as the size of the sample increases, regardless of the shape of the original population distribution.

   The Central Limit Theorem is important for statistics because it allows us to safely assume that the sampling distribution of the mean will be normal in most cases. This means that we can take advantage of statistical techniques that assume a normal distribution, as we will see in the next section.

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

   Sampling is a process in statistical analysis where researchers take a predetermined number of observations from a larger population. The method of sampling depends on the type of analysis being performed, but it may include simple random sampling or systematic sampling.

   There are two types of sampling methods: 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 type II error?**

   In statistical hypothesis testing, a type I error is caused by disapproving a null hypothesis that is otherwise correct while in contrast, Type II error occurs when the null hypothesis is not rejected even though it is not true.
   A type I error (false-positive) occurs if an investigator rejects a null hypothesis that is actually true in the population; a type II error (false-negative) occurs if the investigator fails to reject a null hypothesis that is actually false in the population.

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

   A normal distribution is an arrangement of a data set in which most values cluster in the middle of the range and the rest taper off symmetrically toward either extreme. 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.

5. **What is correlation and covariance in statistics?**

   In statistical terms we use correlation to denote association between two quantitative variables. We also assume that the association is linear, that one variable increases or decreases a fixed amount for a unit increase or decrease in the other.

   However, Covariance is a statistical tool that is used to determine the relationship between the movements of two random variables. When two stocks tend to move together, they are seen as having a positive covariance; when they move inversely, the covariance is negative.

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

   Univariate statistics summarize only one variable at a time. Univariate analysis is the simplest of the three analyses where the data you are analysing 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.

   Bivariate statistics compare two variables. 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. Multivariate statistics compare more than two variables.

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. However, since we cannot visualize anything above the third dimension, we often rely on other software and techniques for us to be able to grasp the relationship in the data.

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

   Sensitivity Analysis is a tool used in financial modelling to analyse how the different values of a set of independent variables affect a specific dependent variable under certain specific conditions.

   The sensitivity is calculated by dividing the percentage change in output by the percentage change in input.
   Sensitivity= A / (A+C) X100

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

   The theory, methods, and practice of testing a hypothesis by comparing it with the null hypothesis. The null hypothesis is only rejected if its probability falls below a predetermined significance level, in which case the hypothesis being tested is said to have that level of significance.

   In hypothesis testing there are two mutually exclusive hypotheses; the Null Hypothesis (H0) and the Alternative Hypothesis (H1). One of these is the claim to be tested and based on the sampling results (which infers a similar measurement in the population), the claim will either be supported or not.

   Specify the Null(H0) and Alternate(H1) hypothesis. Choose the level of Significance(α) Find Critical Values. Find the test statistic.

9. **What is quantitative data and qualitative data?**

   Quantitative data is data expressing a certain quantity, amount or range. Usually, there are measurement units associated with the data, e.g. metres, in the case of the height of a person. It makes sense to set boundary limits to such data, and it is also meaningful to apply arithmetic operations to the data.

Qualitative data is the descriptive and conceptual findings collected through questionnaires, interviews, or observation. Analysing qualitative data allows us to explore ideas and further explain quantitative results.

10. How to calculate range and interquartile range?

The range is calculated by subtracting the lowest value from the highest value. While a large range means high variability, a small range means low variability in a distribution.

To find the interquartile range (IQR), first find the median (middle value) of the lower and upper half of the data. These values are quartile 1 (Q1) and quartile 3 (Q3). The IQR is the difference between Q3 and Q1.

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.

The interquartile range (IQR) measures the dispersion of the data points between the first and third quartile marks. The general rule for using it to calculate outliers is that a data point is an outlier if it is over 1.5 times the IQR below the first quartile or 1.5 times the IQR above the third quartile.

To find the IQR, you subtract the first quartile from the third quartile:

IQR = Q3-Q1

Where

Q3 = the third quartile = the median of the upper half of the data set

Q1 = the first quartile = the median of the lower half of the data set

13. What is p-value in hypothesis testing?

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.

14. What is the Binomial Probability Formula?

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

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.

We can use the ANOVA test to compare different suppliers and select the best available. ANOVA (Analysis of Variance) is used when we have more than two sample groups and determine whether there are any statistically significant differences between the means of two or more independent sample groups.