Answer1-

The Central Limit Theorem tells us that as sample sizes get larger, the sampling distribution of the mean will become normally distributed, even if the data within each sample are not normally distributed.
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.

Answer2-

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 cost-effective method and hence forms the basis of any research design. Sampling techniques can be used in research survey software for optimum derivation.
There are two 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.

Answer3-

Type I error:
• A type I error is also known as a false positive and occurs when a researcher incorrectly rejects a true null hypothesis.
• A type I error also known as False positive.
• The probability that we will make a type I error is designated 'α' (alpha). Therefore, type I error is also known as alpha error.
• Type I error equals to the level of significance (α), 'α' is the so-called p-value.
• Type I error is associated with rejecting the null hypothesis.
• It is caused by luck or chance.
• The probability of Type I error reduces with lower values of (α) since the lower value makes it difficult to reject null hypothesis.
• Type I errors are generally considered more serious.
• It can be reduced by decreasing the level of significance.

• The probability of type I error is equal to the level of significance.
• It happens when the acceptance levels are set too lenient.

Type II Error:
• A type II error does not reject the null hypothesis, even though the alternative hypothesis is the true state of nature. In other words, a false finding is accepted as true.
• A type II error also known as False negative. It is also known as false null hypothesis.
• Probability that we will make a type II error is designated 'β'(beta). Therefore, type II error is also known as beta error.
• Type II error equals to the statistical power of a test.
• Type II error equals to the statistical power of a test. The probability 1- 'β' is called the statistical power of the study.
• Type II error is associated with rejecting the alternative hypothesis.
• It is caused by a smaller sample size or a less powerful test.
• The probability of Type II error reduces with higher values of (α)since the higher value makes it easier to reject the null hypothesis.
• Type II errors are given less preference.
• It can be reduced by increasing the level of significance.
• The probability of type II error is equal to one minus the power of the test.
• It happens when the acceptance levels are set too stringent.

Answer4

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 graph form, normal distribution will appear as a bell curve.

- A 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 kurtosis of 3.
- Normal distributions are symmetrical, but not all symmetrical distributions are normal.
- In reality, most pricing distributions are not perfectly normal.

Normal Distribution Formula:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

Where,
• x is the variable
• μ is the mean

• σ is the standard deviation

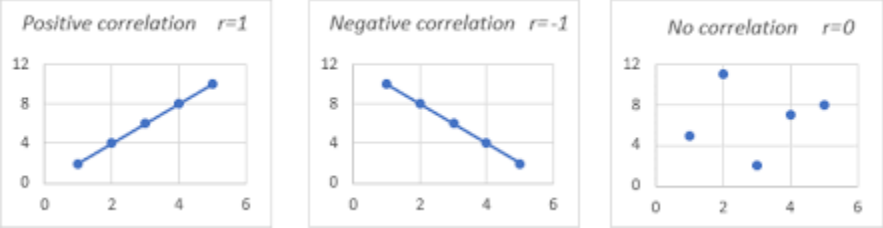
Answer5-

Correlation:

Correlation is a statistical measure that expresses the extent to which two variables are linearly related. It's a common tool for describing simple relationships without making a statement about cause and effect. It is used to test relationships between quantitative variables or categorical variables. In other words, it's a measure of how things are related. The study of how variables are correlated is called correlation analysis.

A correlation coefficient is a way to put a value to the relationship. Correlation coefficients have a value of between -1 and 1. A "0" means there is no relationship between the variables at all, while -1 or 1 means that there is a perfect negative or positive correlation.

The graph of Correlation is:



Correlation is of three types:

• **Simple Correlation**: In simple correlation, a single number expresses the degree to which two variables are related.

• **Partial Correlation**: When one variable's effects are removed, the correlation between two variables is revealed in partial correlation.

• **Multiple correlation:** A statistical technique that uses two or more variables to predict the value of one variable

Correlation coefficient Formula:

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[\, n\Sigma x^2 - (\Sigma x)^2 \,][\, n\Sigma y^2 - (\Sigma y)^2 \,]}}$$

Where,

n: Quantity of Information

Σx: Total of the First Variable Value

Σy: Total of the Second Variable Value

Σxy: Sum of the Product of & Second Value

Σx2: Sum of the Squares of the First Value
Σy2: Sum of the Squares of the Second Value

Covariance:
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.
Covariance is different from the correlation coefficient, a measure of the strength of a correlative relationship. It is a significant tool in modern portfolio theory used to ascertain what securities to put in a portfolio. Risk and volatility can be reduced in a portfolio by pairing assets that have a negative covariance.

Types of Covariance:
• **Positive Covariance:** If the covariance for any two variables is positive, that means, both the variables move in the same direction.
• **Negative Covariance:** If the covariance for any two variables is negative, that means, both the variables move in the opposite direction.

Formula:
Population Covariance

$$Cov(X,Y) = \frac{\Sigma(x_i-\bar{x})(y_i-\bar{y})}{N}$$

Sample Covariance

$$Cov(X,Y) = \frac{\Sigma(x_i-\bar{x})(y_i-\bar{y})}{N-1}$$

These are the formula for finding Population and Sample Covariance.

where,

- x_i = data value of x
- y_i = data value of y
- $\bar{x}$ = mean of x
- $\bar{y}$ = mean of y
- N = number of data values.

Answer6-
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 the univariate analysis is to describe the data and find patterns that exist within it. Univariate statistics summarize only one variable at a time. **Bivariate statistics** compare two variables. **Multivariate statistics** compare more than two variables.
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 is the analysis of three or more variables. There are many ways to perform multivariate analysis depending on your goals.

Answer7-

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. It can be used to help make predictions in the share prices of publicly traded companies or how interest rates affect bond prices. It 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.

Answer8-

Hypothesis Testing:
Hypothesis Testing is a type of statistical analysis in which we put our assumptions about a population parameter to the test. It is used to estimate the relationship between 2 statistical variables. Example of statistical hypothesis is:
A doctor believes that 3D (Diet, Dose, and Discipline) is 90% effective for diabetic patients.
Types of Hypothesis Testing:

- **Null Hypothesis:**
  It is denoted by symbol H0. 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.
- **Alternate Hypothesis:**
  It is denoted by symbol H1. The Alternate Hypothesis is the logical opposite of the null hypothesis. The acceptance of the alternative hypothesis follows the rejection of the null hypothesis**.**

Two-Tailed Hypothesis Testing:
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.
Example 1:
Suppose,
H0: mean = 50 and
H1: mean ≠ 50 (the mean can be greater than or less than 50 but not equal to)

Answer9-

Quantitative data is defined as the value of data in the form of counts or numbers where each data set has a unique numerical value associated with it. This data is any quantifiable information that can be used for mathematical calculations and statistical analysis, such that real-life decisions can be made based on these mathematical derivations. Quantitative data is used to answer questions such as “How many?”, “How often?”, “How much?”. This data can be verified and can also be conveniently evaluated using mathematical techniques.

Qualitative data is defined as the data that approximates and characterizes. Qualitative data can be observed and recorded. This data type is non-numerical in nature. This type of data is collected through methods of observations, one-to-one interviews, conducting focus groups, and similar methods. Qualitative data in statistics is also known as categorical data – data that can be arranged categorically based on the attributes and properties of a thing or a phenomenon.

Answer10-

To calculate the range, we need to find the maximum value of a variable and subtract the minimum value. 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 (IQR) give a better idea of the dispersion of data. To calculate it, we 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.

Formula: **IQR = Q3 - Q1**

Answer11-

A bell curve is a type of graph that is used to visualize the distribution of a set of chosen values across a specified group that tend to have a central, normal values, as peak with low and high extremes tapering off relatively symmetrically on either side. Bell curves are visual representations of normal distribution, also called Gaussian distribution.

A normal distribution curve, when graphed out, typically follows a bell-shaped curve, hence the name. While the precise shape can vary according to the distribution of the population, the peak is always in the middle and the curve is always symmetrical.

Bell curves are useful for quickly visualizing a data set's mean, mode and median because when the distribution is normal, the mean, median and mode are all the same. The long tail refers to the part of the bell curve that stretches out in either direction. If the diagram above represents

a population under study, the fat area under the bell curve is where most of the population falls.

Answer12

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. All data points beyond the IQR limit are considered outliers.

Answer13-

In statistical hypothesis testing, the p-value or probability value is, for a given statistical model, the probability that, when the null hypothesis is true, the statistical summary (such as the absolute value of the sample mean the difference between two compared groups) would be greater than or equal to the actual observed results.

Answer14-

The binomial distribution forms the base for the famous binomial test of statistical importance. A test that has a single outcome such as success/failure is also called a Bernoulli trial or Bernoulli experiment, and a series of outcomes is called a Bernoulli process. Consider an experiment where each time a question is asked for a yes/no with a series of n experiments. Then in the binomial probability distribution, the Boolean-valued outcome the success/yes/true/one is represented with probability p and the failure/no/false/zero with probability q (q = 1 − p). In a single experiment when n = 1, the binomial distribution is called a Bernoulli distribution.

Formula for binomial distribution:

$$P(X) = {}_nC_x\, p^x q^{n-x}$$

Where,

- n = the number of experiments
- x = 0, 1, 2, 3, 4, …
- p = Probability of success in a single experiment
- q = Probability of failure in a single experiment (= 1 – p)

Answer15-

Analysis of variance (ANOVA) is an analysis tool used in statistics that splits an observed aggregate variability found inside a data set into two parts: systematic factors and random factors. The systematic factors have a statistical influence on the given data set, while the random factors do not. Analysts use the ANOVA test to determine the influence that independent variables have on the dependent variable in a regression study.

Analysis of variance, or ANOVA, is a statistical method that separates observed variance data into different components to use for additional tests.

Types of ANOVA:

- **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).
- **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).

The Formula for ANOVA is:

$$F = MST/MSE$$

where:
F=ANOVA coefficient
MST=Mean sum of squares due to treatment
MSE=Mean sum of squares due to error