

STATISTICS WORKSHEET-4

Q1 to Q15 are descriptive types. Answers were written just under the questions.

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

The central limit theorem states that if we have a population with mean and standard deviation and take sufficiently large random samples from the population with replacement, then the distribution of the sample mean is asymptotically normal.

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?

Sampling, in simple terms, means selecting a group (a sample) from a population from which we will collect data for our research. Sampling is an important aspect of a research study as the results of the study majorly depend on the sampling technique used.

Sampling techniques

- Simple Random
- Systematic
- Stratified
- Cluster
- Convenience
- Quota
- Judgement
- Snowball

3. What is the difference between type I and type II error?

Type I Error

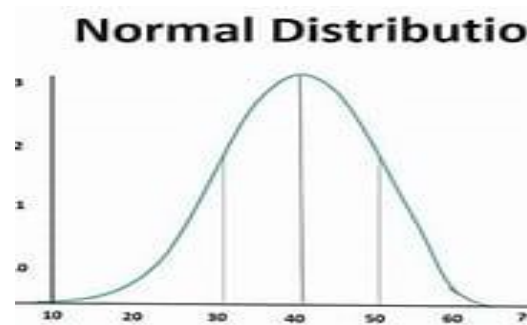
- A type I error occurs when the null hypothesis is true but is rejected. In other words, if a true null hypothesis is incorrectly rejected, type I error occurs.
- A type I error also known as False positive.
- The probability that we will make a type I error is designated ' α ' (alpha). Therefore, type I error is also known as alpha error

Type II Error

- 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.
- 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.

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

Normal 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.



5. What is correlation and covariance in statistics?

Covariance signifies the direction of the linear relationship between the two variables. By direction we mean if the variables are directly proportional or inversely proportional to each other. (Increasing the value of one variable might have a positive or a negative impact on the value of the other variable).

Correlation analysis is a method of statistical evaluation used to study the strength of a relationship between two, numerically measured, continuous variables. It not only shows the kind of relation (in terms of direction) but also how strong the relationship is.

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

Univariate Analysis

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 univariate analysis is to describe the data and find patterns that exist within it.

Bivariate Analysis

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

Multivariate Analysis

Multivariate analysis is the analysis of three or more variables. There are many ways to perform multivariate analysis depending on your goals. Some of these methods include:

Additive Tree

Canonical Correlation Analysis

Cluster Analysis

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

Sensitivity analysis is an analysis technique that works on the basis of what-if analysis like how independent factors can affect the dependent factor and is used to predict the outcome when analysis is performed under certain conditions. It is commonly used by investors who takes into consideration the conditions that affect their potential investment to test, predict and evaluate result.

The formula for sensitivity analysis is basically a financial model in excel where the analyst is required to identify the key variables for the output formula and then assess the output based on different combinations of the independent variables.

Mathematically, the dependent output formula is represented as,

$$[Z = X^2 + Y^2]$$

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. In hypothesis testing there are two mutually exclusive hypotheses; the Null Hypothesis (H0) and the Alternative Hypothesis (H1). H1 is a statistical proposition stating that there is a significant difference between a hypothesized value of a population parameter and its estimated value. The null hypothesis (H0) is a statement of “no difference,” “no association,” or “no treatment effect.” The alternative hypothesis, Ha is a statement of “difference,” “association,” or “treatment effect.”

9. What is quantitative data and qualitative data?

Quantitative data is anything that can be counted or measured; it refers to numerical data. Qualitative data is descriptive, referring to things that can be observed but not measured—such as colors or emotions.

10. How to calculate range and interquartile range?

- **Range**

In arithmetic, the **range** of a set of data is the difference between the largest and smallest values. However, in descriptive statistics, this concept of range has a more complex meaning. The range is the size of the smallest interval which contains all the data and provides an indication of statistical dispersion. It is measured in the same units as the data.

- **Interquartile range**

After calculating quartiles, it's rather easy to calculate an interquartile range. Simply remove the first quartile from the third one or in simple ones Simply subtract the first quartile from the third quartile.

[Representation of IQR = $(q_3 - q_1)$]

11. What do you understand by bell curve distribution ?

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

12. Mention one method to find outliers.

Statistical outlier detection

Statistical outlier detection involves applying **statistical tests** or procedures to identify extreme values.

You can convert extreme data points into **z scores** that tell you how many standard deviations away they are from the mean.

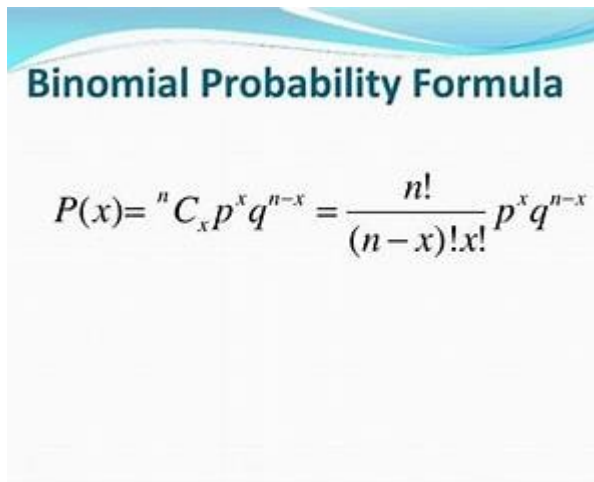
If a value has a high enough or low enough z score, it can be considered an outlier. As a rule of thumb, values with a z score greater than 3 or less than -3 are often determined to be outliers.

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.

14. What is the Binomial Probability Formula?

The binomial probability formula can be used to calculate the probability of success for binomial distributions. Binomial probability distribution along with normal probability distribution are the two probability distribution types. To recall, the binomial distribution is a type of distribution in statistics that has two possible outcomes.



A presentation slide titled "Binomial Probability Formula" with a blue and white wavy header. The formula is displayed in the center:
$$P(x) = {}^n C_x p^x q^{n-x} = \frac{n!}{(n-x)!x!} p^x q^{n-x}$$

15. Explain ANOVA and it's applications.

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.

If no real difference exists between the tested groups, which is called the [null hypothesis](#), the result of the ANOVA's F-ratio statistic will be close to 1.

The Formula for ANOVA is:

$$F = MST/MSE$$

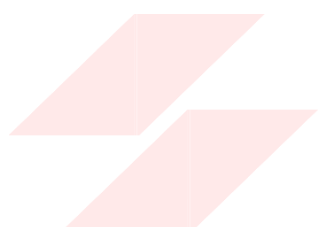
Where F=ANOVA coefficient

MST=Mean sum of squares due to treatment

MSE=Mean sum of squares due to error

It has various applications but some are:-

- **Farming:** It used in farming industry in which we can test on which gives maximum crop yields.
- **Retail:** Store are often interested in understanding whether different types of promotions, store layouts, advertisement tactics, etc. lead to different sales. This is the exact type of analysis that ANOVA is built for.
- **Medical:** Researchers are often interested in whether or not different medications affect patients differently, which is why they often use one-way or two-way ANOVA's in these situations.
- **Environmental Sciences:** Researchers are often interested in understanding how different levels of factors affect plants and wildlife. Because of the nature of these types of analyses, ANOVA's are often used.



FLIP ROBO