WORKSHEET 4

STATISTICS

INTERNSHIP 32

FLIP ROBO TECHNOLOGIES



Submitted by: Shahla M

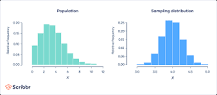
SME : Khushbooo Garg

**Q1to Q15 are descriptive types. Answer in brief.**

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

ANSWER :

The central limit theorem says that the sampling distribution of the mean will always be normally distributed, as long as the sample size is large enough. Regardless of whether the population has a normal, Poisson, binomial, or any other distribution, the sampling distribution of the mean will be normal.



- 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.

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.

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

ANSWER :

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.

Eg : Simple Random Sample, Systematic Sample, Stratified Sample , Cluster Sample

**Non-probability sampling** involves non-random selection based on convenience or other criteria, allowing you to easily collect data.

Eg : Convenience Sample , Purposive Sample, Snowball Sample , Quota Sample

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

ANSWER :

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?**

ANSWER :

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.

The normal distribution is the most important probability distribution in statistics because it fits many natural phenomena. For example, heights, blood pressure, measurement error, and IQ scores follow the normal distribution

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

ANSWER :

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.

**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 ,Biavariate,and multivariate analysis.**

ANSWER :

**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.

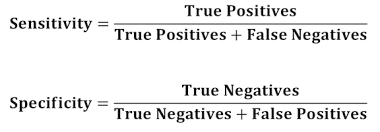
**2. Bivariate data –**  
This type of data involves **two different variables**. The analysis of this type of data deals with causes and relationships and the analysis is done to find out the relationship among the two variables.Example of bivariate data can be temperature and ice cream sales in summer season.

**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.

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

ANSWER :

Sensitivity analysis is used to identify how much variations in the input values for a given variable impact the results for a mathematical model. Sensitivity analysis can identify the best data to be collected for analyses to evaluate a project's return on investment (ROI).



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

ANSWER :

Hypothesis testing concerns on how to use a random sample to judge if it is evidence that supports or not the hypothesis. Hypothesis testing is formulated in terms of two hypotheses:

**Null hypothesis (H0):** The null hypothesis here is what currently stated to be true about the population. In our case it will be the average height of students in the batch is 100.

**Alternate hypothesis (H1):** The alternate hypothesis is always what is being claimed...

**10. How to calculate range and interquartile range?**

To calculate the range, you need to find the largest observed value of a variable (the maximum) and subtract the smallest observed value (the minimum**)**. The range only takes into account these two values and ignore the data points between the two extremities of the distribution

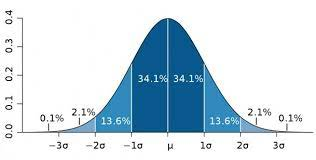
The IQR describes the middle 50% of values when ordered from lowest to highest. To find the interquartile range (IQR), ​first find the median (middle value) of the lower and upper half of the data. These values are quartile 1 (Q1) and quartile 3 (Q3). The IQR is the difference between Q3 and Q1.

**IQR = Q3 – Q1.**

**11. What do you understand by bell curve distribution ?**

ANSWER :

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.**

ANSWER :

**Z-scores** to Detect Outliers:- Z-scores can quantify the unusualness of an observation when your data follow the normal distribution. Z-scores are the number of standard deviations above and below the mean that each value falls. For example, a Z-score of 2 indicates that an observation is two standard deviations above the average while a Z-score of -2 signifies it is two standard deviations below the mean. A Z-score of zero represents a value that equals the mean. To calculate the Z-score for an observation, take the raw measurement, subtract the mean, and divide by the standard deviation.

**13. What is p-value in hypothesis testing?**

ANSWER :

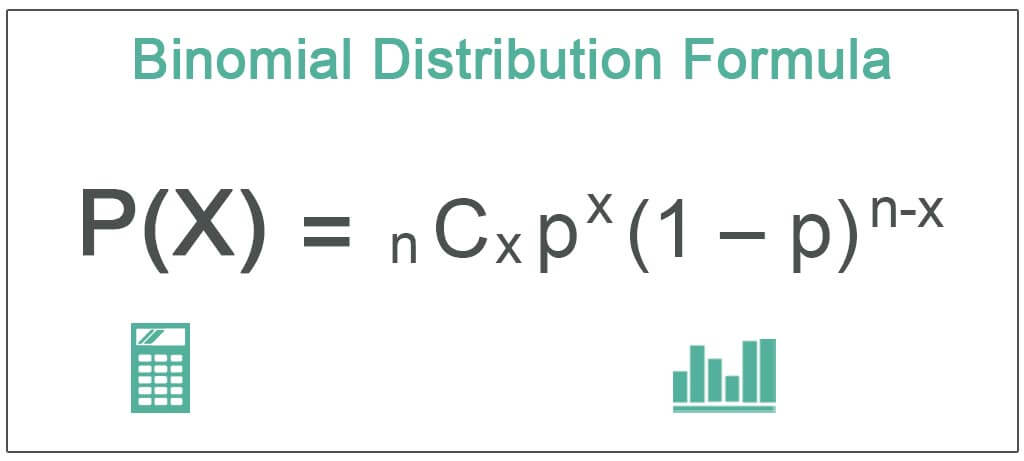
In statistics, the p-value is the probability of obtaining results at least as extreme as the observed results of a statistical hypothesis test, assuming that the null hypothesis is correct. The p-value is used as an alternative to rejection points to provide the smallest level of significance at which the null hypothesis would be rejected. A smaller p-value means that there is stronger evidence in favor of the alternative hypothesis.

**14. What is the Binomial Probability Formula?**

ANSWER :

The binomial distribution formula is used in statistics to find the probability of the specific outcome-success or failure in a discrete distribution.

Binomial probability refers to the probability of exactly x successes on n repeated trials in an experiment which has two possible outcomes (commonly called a binomial experiment). If the probability of success on an individual trial is p , then the binomial probability is



n = Total number of trials

x = Total number of successful trials

p = probability of success in a single trial

1-p = probability of failure in a single trial =

**15. Explain ANOVA and it’s applications**

ANSWER :

Analysis of Variance (ANOVA) is a statistical formula used to compare variances across the means (or average) of different groups. A range of scenarios use it to determine if there is any difference between the means of different groups.

A common approach to figure out a reliable treatment method would be to analyse the days it took the patients to be cured. We can use a statistical technique which can compare these three treatment samples and depict how different these samples are from one another. Such a technique, which compares the samples on the basis of their means, is called ANOVA.

Analysis of variance (ANOVA) is a statistical technique that is used to check if the means of two or more groups are significantly different from each other.

ANOVA checks the impact of one or more factors by comparing the means of different samples. One real-life application of analysis of variance is the recommendation of a fertilizer against others for the improvement of a crop yield. Anova can be in different fields of sciences, i.e. all the problems of testing more than three groups.

