Assignment 4

Statistics

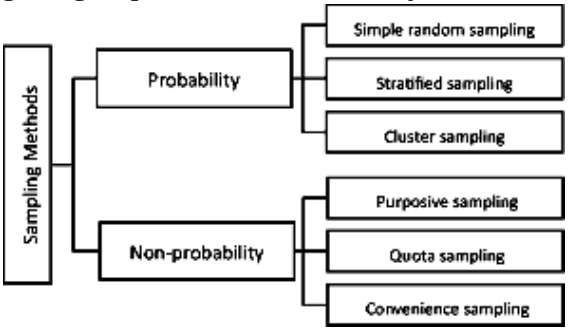
1. What is central limit theorem and why is it important?
   Answer:

   The central limit theorem states that if you have a population with mean μ and standard deviation σ and take sufficiently large random samples from the population with replacement, then the distribution of the sample means will be approximately 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

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

   Sampling means selecting the group that one will actually collect data from in research.

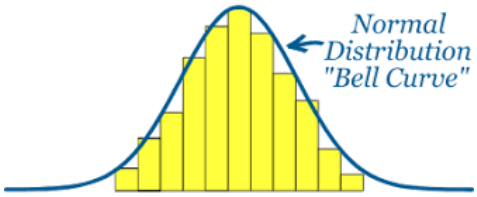


3. What is the difference between type1 and typeII error?
   Answer:

   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?

   A normal distribution is a type of continuous probability distribution in which most data points cluster toward the middle of the range, while the rest taper off symmetrically toward either extreme. The middle of the range is also known as the mean of the distribution.



5. What is correlation and covariance in statistics?
   Answer:

   Correlation is a statistical measure that indicates how strongly two variables are related.

   Covariance is an indicator of the extent to which 2 random variables are dependent on each other. A higher number denotes higher dependency.

6. Differentiate between univariate ,Biavariate,and multivariate analysis.
   Answer:
   - Univariate analysis looks at one variable.
   - Bivariate analysis looks at two variables and their relationship.
   - Multivariate analysis looks at more than two variables and their relationship.

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

   The sum of sensitivity (true positive rate) and false negative rate would be 1. The higher the true positive rate, the better the model is in identifying the positive cases in the correct manner.

   The sensitivity is calculated by dividing the percentage change in output by the percentage change in input.

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 evident that supports the hypothesis or not.

   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.

9. What is quantitative data and qualitative data?
   Answer:

   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. Analyzing qualitative data allows us to explore ideas and further explain quantitative results.

10. How to calculate range and inter-quartile range?
    Answer:

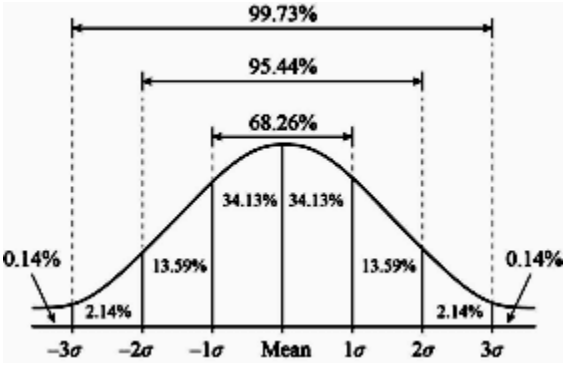
    Range - find the largest observed value of a variable (the maximum) and subtract the smallest observed value (the minimum).

    IQR - The formula for finding the inter-quartile range takes the third quartile value and subtracts the first quartile value. Equivalently, the inter-quartile range is the region between the 75th and 25th percentile (75 – 25 = 50% of the data). i.e. 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-score

13. What is p-value in hypothesis testing?

Answer:

The p value is a number, calculated from a statistical test, that describes how likely you are to have found a particular set of observations if the null hypothesis were true. P values are used in hypothesis testing to help decide whether to reject the null hypothesis.

14. What is the Binomial Probability Formula?

Answer:

$P(r) = {}^{n}C^{r} \cdot p^{r} (1 - p)^{n-r}$.

15. Explain ANOVA and it's applications.

Answer:

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.

ANOVA is helpful for testing three or more variables. It is similar to multiple two-sample t-tests. However, it results in fewer, type I errors and is appropriate for a range of issues. ANOVA groups differences by comparing the means of each group and includes spreading out the variance into diverse sources.

It provides the overall test of equality of group means. It can control the overall type I error rate (i.e. false positive finding) It is a parametric test so it is more powerful, if normality assumptions hold true.