ASSIGNMENT-4

STATISTICS

Q1to Q15 are descriptive types. Answer in brief.

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

   Answer: The CLT is a statistical theory that states that - 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 (CLT) states that the distribution of sample means approximates a normal distribution as the sample size gets larger, regardless of the population's distribution. Sample sizes equal to or greater than 30 are often considered sufficient for the CLT to hold. A key aspect of CLT is that the average of the sample means and standard deviations will equal the population mean and standard deviation. A sufficiently large sample size can predict the characteristics of a population more accurately. CLT is useful in finance when analyzing a large collection of securities to estimate portfolio distributions and traits for returns, risk, and correlation.

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

   A sample is a subset of individuals from a larger population. Sampling means selecting the group that you will actually collect data from in your research. Sampling methods are:

   1. Simple random sampling
   2. Systematic sampling
   3. Stratified sampling
   4. Cluster sampling
   5. Accidental sampling

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: In statistics, a normal distribution (also known as Gaussian, Gauss, or Laplace– Gauss distribution) is a type of continuous probability distribution for a real-valued random variable. The general form of its probability density function is

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

The parameter μ is the mean or expectation of the distribution (and also its median and mode), while the parameter σ is its standard deviation. The variance of the distribution is σ 2 . A random variable with a Gaussian distribution is said to be normally distributed, and is called a normal deviate.

5. What is correlation and covariance in statistics?

Answer: Correlation is a statistical measure that expresses the extent to which two variables are linearly related (meaning they change together at a constant rate). It's a common tool for describing simple relationships without making a statement about cause and effect. Covariance is an indicator of the extent to which 2 random variables are dependent on each other. A higher number denotes higher dependency. Correlation is a statistical measure that indicates how strongly two variables are related.

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

Answer: Explorative data analysis has different analysis like: Univariate statistics summarize only one variable at a time. Bivariate statistics compare two variables. Multivariate statistics compare more than two variables.

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

   Answer: Sensitivity is a measure of how well a machine learning model can detect positive instances. It is also known as the true positive rate (TPR) or recall. Sensitivity is used to evaluate model performance because it allows us to see how many positive instances the model was able to correctly identify. A model with high sensitivity will have few false negatives, which means that it is missing a few of the positive instances. In other words, sensitivity measures the ability of a model to correctly identify positive examples. This is important because we want our models to be able to find all of the positive instances in order to make accurate predictions. 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. Mathematically, sensitivity or true positive rate can be calculated as the following: Sensitivity = (True Positive)/(True Positive + False Negative) A high sensitivity means that the model is correctly identifying most of the positive results, while a low sensitivity means that the model is missing a lot of positive results. The following are the details in relation to True Positive and False Negative used in the above equation. True Positive: Persons predicted as suffering from the disease (or unhealthy) are actually suffering from the disease (unhealthy); In other words, the true positive represents the number of persons who are unhealthy and are predicted as unhealthy. False Negative: Persons who are actually suffering from the disease (or unhealthy) are actually predicted to be not suffering from the disease (healthy). In other words, the falsenegative represents the number of persons who are unhealthy and got predicted as healthy. Ideally, we would seek the model to have low false negatives as it might prove to be lifethreatening or business threatening

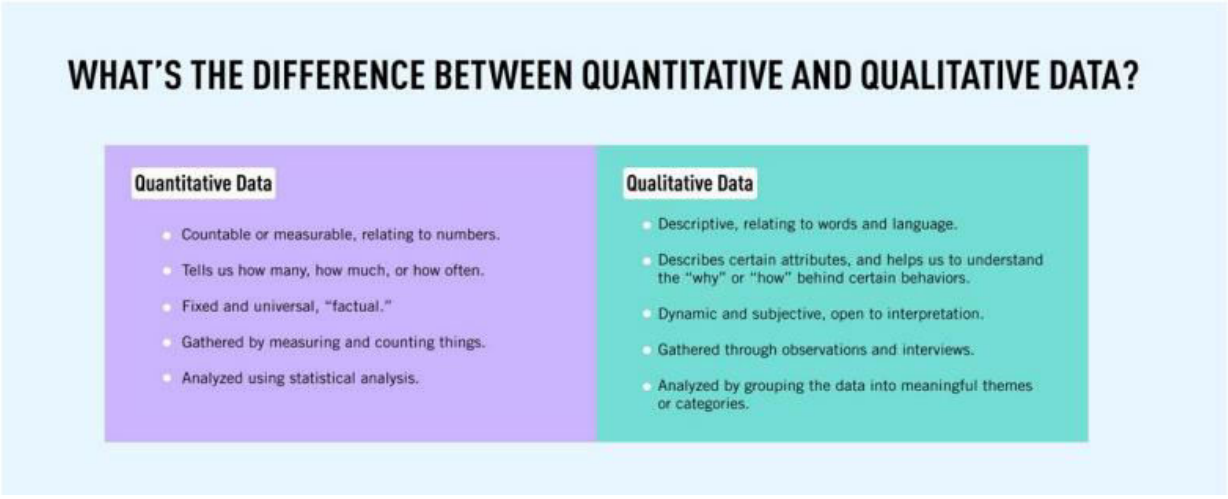
8. What is hypothesis testing? What is H0 and H1? What is H0 and H1 for two-tail test?

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**.** "In our case, Tedd believes(Claims) that the actual value has changed". He doesn't know whether the average has gone up or down, but he believes that it has changed and is not 100 anymore.

9. What is quantitative data and qualitative data?

Answer: The main differences between quantitative and qualitative data lie in what they tell us, how they are collected, and how they are analyzed. Let's summarize the key differences before exploring each aspect in more detail: Quantitative data is countable or measurable, relating to numbers. Qualitative data is descriptive, relating to language. Quantitative data is gathered by measuring and counting. Qualitative data is collected by interviewing and observing. Quantitative data is analyzed using statistical analysis, while qualitative data is analyzed by grouping it in terms of meaningful categories or themes.



10. How to calculate range and interquartile range?

Answer: The Interquartile range formula helps in finding the difference between the third quartile and the first quartile. The Interquartile range formula measures the variability, based on dividing an ordered set of data into quartiles. Quartiles are three values or cuts that divide each respective part as the first, second, and third quartiles, denoted by Q1Q1, Q2Q2, and Q3Q3, respectively. Q1Q1 is the cut in the first half of the rank-ordered data set Q2Q2 is the median value of the set Q3Q3 is the cut in the second half of the rank-ordered data set. The Interquartile Range (IQR) formula is a measure of the middle 50% of a data set. The smallest of all the measures of dispersion in statistics is called the Interquartile Range. The difference between the upper and lower quartile is known as the interquartile range. Interquartile range = Upper Quartile – Lower Quartile Q2=Q3–Q1Q2=Q3–Q1 where, IQR = Interquartile range (IQR = Q2Q2) Q1Q1 = (1/4)[(n + 1)]th term) Q3Q3 = (3/4)[(n + 1)]th term) n = number of data points The following steps help us to find the IQR: The simple trick is to arrange the data points in ascending order. Q2Q2 is the median of the data. If the number of data points is odd, the middle term is (n+1)/2 and if the number of data points is even, the median is the mean of the two middle points. Q1Q1 is the median of the data points to the left of the median found in

step 2. Q3Q3 is the median of the data points to the right of the median found in step 2. IQR = Q2=Q3–Q1

11. What do you understand by bell curve distribution ?

Answer: A bell curve is a graph depicting the normal distribution, which has a shape reminiscent of a bell. The top of the curve shows the mean, mode, and median of the data collected. Its standard deviation depicts the bell curve's relative width around the mean

12. Mention one method to find outliers.

Answer: Outliers are extreme values that differ from most other data points in a dataset. They can have a big impact on your statistical analyses and skew the results of any hypothesis tests. It’s important to carefully identify potential outliers in your dataset and deal with them in an appropriate manner for accurate results. There are four ways to identify outliers: Sorting method Data visualization method Statistical tests (z scores) Interquartile range method

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?

n probability theory, one of the important discrete distributions is the binomial distribution. It consists of n and p as parameters. It is used to find the number of successes in a sequence of n independent experiments. It is associated with the outcome on Boolean values namely success (denoted with the probability p) or failure (denoted with the probability q = 1 – p). An experiment consisting of 1 success/failure is a Bernoulli trial. If n = 1, then binomial distribution becomes a Bernoulli distribution. The binomial distribution must satisfy the following criteria.

- The trial number is fixed.
- Every trial or observation is independent. No trial will have an effect on the probability of the upcoming trial.
- The success probability is the same from one trial to the trial.

The binomial probability formula for any random variable x is given by

$P(x : n, p) = {}^{n}C_{x}\, p^{x}\, q^{n-x}$

n = the number of trials

x varies from 0, 1, 2, 3, 4, …

p = probability of success

q = probability of failure = 1 – p

The binomial distribution can be converted into the Bernoulli distribution as follows.

For n-bernoulli trials, ${}^{n}C_{x} = n! / x!\,(n – x)!$.

$P(x : n, p) = n! / [x! (n – x)!] . p^x . (q)^{n-x}$

15. Explain ANOVA and it's applications.

Answer: 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. An ANOVA test is a type of statistical test used to determine if there is a statistically significant difference between two or more categorical groups by testing for differences of means using variance. The ANOVA test allows a comparison of more than two groups at the same time to determine whether a relationship exists between them. The result of the ANOVA formula, the F statistic (also called the F-ratio), allows for the analysis of multiple groups of data to determine the variability between samples and within samples. Application of ANOVA in Quality and cost comparison, Product safety tests, Optimize production