Statistics assignment 4

Answer-1

The **Central Limit Theorem (CLT)** is a mainstay of **statistics** and **probability**. The theorem expresses that as the size of the sample expands, the distribution of the mean among multiple samples will be like a **Gaussian distribution**.

The CLT gives us a certain distribution over our estimations. We can utilize this to pose an inquiry about the probability of an estimate that we make. For example, assume we are attempting to think about how an election will turn out.

The CLT works from the center out. That implies on the off chance that you are presuming close to the center, for example, that around two-thirds of future totals will fall inside one standard deviation of the mean, you can be secure even with little samples.

IMPORTANCE

This is useful, as the research never knows which mean in the sampling distribution is the same as the population mean, but by selecting many random samples from a population the sample means will cluster together, allowing the research to make a very good estimate of the population mean.

Answer-2

Probability sampling methods include **simple random sampling, systematic sampling, stratified sampling, and cluster sampling.**

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.

Types of sampling: sampling methods

1. **Probability sampling- It** 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.
2. **Non-probability sampling:** In this 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.

Answer-3

Type -1 Error (Error of the first kind)

- It is also known as a false-positive.
- It occurs if the researcher rejects a correct null hypothesis in the population.
  - i.e., incorrect rejection of the null hypothesis.
- Measured by alpha (significance level).
- If the significance level is fixed at 5%,
  - It means there are about five chances of type – 1 error out of 100.
- **Cause of Type – 1 Error**
  - The significance level is decided before testing the hypothesis
  - Sample size is not considered
  - This may occur due to chance
- It can be reduced by decreasing the level of significance.

Type -2 Error (Error of the second kind)

- It is also known as a false negative.
- It occurs if a researcher fails to reject a null hypothesis that is actually a false hypothesis.
- Measured by beta (the power of test).
- The probability of committing a type -2 error is calculated by 1 – beta (the power of test).
- **Cause of Type – 2 Error:**
  - A statistical test is not powerful enough.
  - It is caused by a smaller sample size.
    - It may hide the significance level of the items being tested.
- It can be reduced by increasing the level of significance.

Examples of Type -1 and Type – 2 Errors

Now, let's take an example to understand better type – 1 and type – 2 errors:

Example -1: A man goes to test the coronavirus (COVID -19). So, the possible errors are

Type -1 Error (False Positive): Test results are positive, but you don't.

Type – 2 Error (False Negative): Test results are negative, but you do.

Example – 2 A man goes to trial and is being tried for murder.

Null Hypothesis: Man is innocent until proven guilty.

Alternative Hypothesis: Man is guilty.

Type -1 Error (False Positive): Found guilty, but you are innocent.

True Positive: Found guilty and being guilty.

True Negative: Found Innocent and being innocent

Type -2 Error (False Negative): Found innocent, but you are guilty.

Answer-4

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.

In statistics, a normal distribution or Gaussian distribution is a type of continuous probability distribution for a real-valued random variable.

The standard normal distribution (z distribution) is a normal distribution with a mean of 0 and a standard deviation of 1. Any point (x) from a normal distribution can be converted to the standard normal distribution (z) with the formula **z = (x-mean) / standard deviation**.

Answer-5

Correlation is a statistical measure (expressed as a number) that describes the size and direction of a relationship between two or more variables. A correlation between variables, however, does not automatically mean that the change in one variable is the cause of the change in the values of the other variable.

Types of Correlation

- Positive Linear Correlation. There is a positive linear correlation when the variable on the x -axis increases as the variable on the y -axis increases. ...
- Negative Linear Correlation. ...
- Non-linear Correlation (known as curvilinear correlation) ...
- No Correlation.

Covariance is a measure to indicate the extent to which two random variables change in tandem. Correlation is a measure used to represent how strongly two random variables are related to each other. Covariance is nothing but a measure of correlation. Correlation refers to the scaled form of covariance.

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 measure of the relationship between two random variables and to what extent, they change together**. Or we can say, in other words, it defines the changes between the two variables, such that change in one variable is equal to change in another variable.

Answer-6

Univariate statistics summarize only one variable at a time. Bivariate statistics compare two variables. Multivariate statistics compare more than two variables.

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.

3. **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.
It is similar to bivariate but contains more than one dependent variable. The ways to perform analysis on this data depends on the goals to be achieved.Some of the techniques are regression analysis,path analysis,factor analysis and multivariate analysis of variance (MANOVA)

Answer-7

Sensitivity analysis **allows for forecasting using historical, true data**. By studying all the variables and the possible outcomes, important decisions can be made about businesses, the economy, and making investments.

Sensitivity=**[a/(a+c)]×100**

Answer-8

Hypothesis testing is the process used **to evaluate the strength of evidence from the sample** and provides a framework for making determinations related to the population, ie, it provides a method for understanding how reliably one can extrapolate observed findings in a sample under study to the larger population from.

The 5 steps of hypothesis testing

- Step 1: State your null and alternate hypothesis. ...
- Step 2: Collect data. ...
- Step 3: Perform a statistical test. ...

- Step 4: Decide whether to reject or fail to reject your null hypothesis. ...
- Step 5: Present your findings.

(H0) is null hypothesis

(H1) is alternate hypothesis

Our null hypothesis is that **the mean is equal to x**. A two-tailed test will test both if the mean is significantly greater than x and if the mean significantly less than x.
If you are using a significance level of 0.05, a two-tailed test allots half of your alpha to testing the statistical significance in one direction and half of your alpha to testing statistical significance in the other direction. This means that .025 is in each tail of the distribution of your test statistic. When using a two-tailed test, regardless of the direction of the relationship you hypothesize, you are testing for the possibility of the relationship in both directions. For example, we may wish to compare the mean of a sample to a given value *x* using a t-test. Our null hypothesis is that the mean is equal to *x*. A two-tailed test will test both if the mean is significantly greater than *x* and if the mean significantly less than *x*. The mean is considered significantly different from *x* if the test statistic is in the top 2.5% or bottom 2.5% of its probability distribution, resulting in a p-value less than 0.05.

Answer-9

Quantitative data are measures of values or counts and are expressed as numbers. Quantitative data are data about numeric variables (e.g. how many; how much; or how often).
Example: **age, weight, GPA, income**.

Qualitative data are measures of 'types' and may be represented by a name, symbol, or a number code.
Example: race, gender, class (freshman, sophomore, etc.), major.

Answer-10

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.

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.

Answer-11
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.

The bell curve is perfectly symmetrical. It is concentrated around the peak and decreases on either side. In a bell curve, the peak represents the most probable event in the dataset while the other events are equally distributed around the peak.

Answer-12

There are four ways to identify outliers:

1. Sorting method.
2. Data visualization method.
3. Statistical tests (z scores)

4. Interquartile range method.

   IQR is **used to measure variability by dividing a data set into quartiles**. The data is sorted in ascending order and split into 4 equal parts. Q1, Q2, Q3 called first, second and third quartiles are the values which separate the 4 equal parts

Answer-13

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.

P-value is often used to promote credibility for studies or reports by government agencies. For example, the U.S. Census Bureau stipulates that any analysis with a p-value greater than 0.10 must be accompanied by a statement that the difference is not statistically different from zero. The Census Bureau also has standards in place stipulating which p-values are acceptable for various publications.

Answer-14

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

Binomial distribution is calculated by multiplying the probability of success raised to the power of the number of successes and the probability of failure raised to the power of the difference between the number of successes and the number of trials.

The binomial distribution formula is calculated as:

$P_{(x:n,p)} = {}_nC_x \, x \, p^x(1-p)^{n-x}$

where:

- n is the number of trials (occurrences)
- x is the number of successful trials
- p is probability of success in a single trial
- nCx is the combination of n and x. A combination is the number of ways to choose a sample of x elements from a set of n distinct objects where order does not matter and replacements are not allowed. Note that nCx=n!/(r!(n−r)!), where ! is factorial (so, 4! = 4 × 3 × 2 × 1).

Answer-15

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

The Anova test is performed by comparing two types of variation, the variation between the sample means, as well as the variation within each of the samples. The below mentioned formula represents one-way Anova test statistics: Alternatively, **F = MST/MSE**. **MST = SST/ p-1**.

Application of anova-

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 used in a business context **to help manage budgets by comparing your budget to costs to help manage revenue and inventory**,

for example. ANOVA can also be used to forecast trends by analyzing patterns in data to better understand the future performance of sales.