STATISTICS WORKSHEET-4

Q1to Q15 are descriptive types. Answer in brief.

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

Ans. The central limit theorem states that if we 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 mean is asymptotically normal.

Normal Distribution = Distrbution is always normal irrespective of of sample size
Non-Normal distribution - If sample size is adequate (appr> 30 sample), distribution starts looking normal

$$\mu_{\bar{x}} = \mu \text{ (Mean of sample mean = Population mean)}$$

$$\sigma_{\bar{x}} = \sigma/\sqrt{n} \text{ (Population std / sqrt of sample size)}$$

Where,
μ = Population mean
σ = Population standard deviation
μx = Sample mean
σx= Sample standard deviation
n = Sample size
x̄ = Sample mean

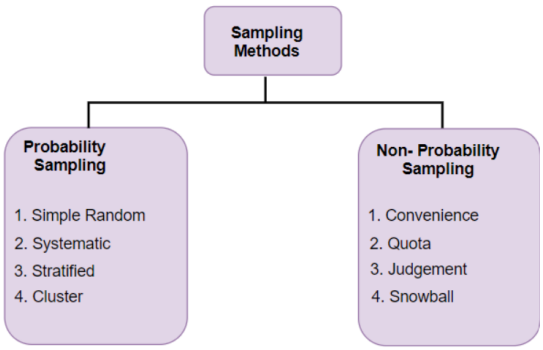
Importance of Central Limit Theorem:
This is useful since the researcher never knows which mean in the sampling distribution corresponds to the population mean, but by taking numerous random samples from a population, the sample means will cluster together, allowing the researcher to obtain a very accurate estimate of the population mean.

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

Ans. **Sampling** : It is a method that allows us to get information about the population based on the statistics from a subset of the population (sample), without having to investigate every individual.

There are two type of sampling methods

- **Probability Sampling:** In probability sampling, every element of the population has an equal chance of being selected. Probability sampling gives us the best chance to create a sample that is truly representative of the population
- **Non-Probability Sampling:** In non-probability sampling, all elements do not have an equal chance of being selected. Consequently, there is a significant risk of ending up with a non-representative sample which does not produce generalizable results



3. What is the difference between type1 and type II error?

Ans. **Type 1 error:-**

- Type 1 error, in statistical hypothesis testing, is the error caused by rejecting a null hypothesis when it is true.
- Type 1 error is caused when the hypothesis that should have been accepted is rejected.
- Type I error is denoted by α (alpha) known as an error, also called the level of significance of the test.
- This type of error is a false negative error where the null hypothesis is rejected based on some error during the testing.
- The null hypothesis is set to state that there is no relationship between two variables and the cause-effect relationship between two variables, if present, is caused by chance.
- Type 1 error occurs when the null hypothesis is rejected even when there is no relationship between the variables. As a result of this error, the researcher might end up believing that the hypothesis works even when it doesn't.

Type 2 error:-

- Type II error is the error that occurs when the null hypothesis is accepted when it is not true.
- In simple words, Type II error means accepting the hypothesis when it should not have been accepted.
- The type II error results in a false negative result.
- In other words, type II is the error of failing to accept an alternative hypothesis when the researcher doesn't have adequate power.
- The Type II error is denoted by β (beta) and is also termed as the beta error.
- The null hypothesis is set to state that there is no relationship between two variables and the cause-effect relationship between two variables, if present, is caused by chance.
- Type II error occurs when the null hypothesis is acceptable considering that the relationship between the variables is because of chance or luck, and even when there is a relationship between the variables.
- As a result of this error, the researcher might end up believing that the hypothesis doesn't work even when it should.

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

Ans. 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?

Ans. **Covariance** :

- A systematic relationship between a pair of random variables wherein a change in one variable reciprocated by an equivalent change in another variable.
- Covariance can take any value between -∞ to +∞, wherein the negative value is an indicator of negative relationship whereas a positive value represents the positive relationship and when the value is zero, it indicates no relationship.
- In addition to this, when all the observations of the either variable are same, the covariance will be zero.
- When we change the unit of observation on any or both the two variables, then there is no change in the strength of the relationship between two variables but the value of covariance is changed.
-

Correlation:

- A measure which determines the change in one variable due to change in other variable.
- Correlation is of two types, i.e. positive correlation or negative correlation.
- Correlation can take any value between -1 to +1, wherein values close to +1 represents strong positive correlation and values close to -1 is an indicator of strong negative correlation.

  There are four measures of correlation:

  1. Scatter diagram
  2. Product-moment correlation coefficient
  3. Rank correlation coefficient
  4. Coefficient of concurrent deviations

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

Ans. **Univarate 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.

You can think of the variable as a category that your data falls into. One example of a variable in univariate analysis might be "age". Another might be "height". Univariate analysis would not look at these two variables at the same time, nor would it look at the relationship between them.

Some ways you can describe patterns found in univariate data include looking at mean, mode, median, range, variance, maximum, minimum, quartiles, and standard deviation. Additionally, some ways you may display univariate data include frequency distribution tables, bar charts, histograms, frequency polygons, and pie charts.

Bivarate 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. For example, one might choose to plot caloric intake versus weight.

Multivariate Analysis:-

Multivariate analysis is the analysis of three or more variables. There are many ways to perform multivariate analysis depending on our goals.

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

Ans. A sensitivity analysis determines how different values of an independent variable affect a particular dependent variable under a given set of assumptions. In other words, sensitivity analyses study how various sources of uncertainty in a mathematical model contribute to the model's overall uncertainty. This technique is used within specific boundaries that depend on one or more input variables.

Sensitivity is calculated as :- A/(A+C) × 100 where, A = True positives C = False negatives

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

Ans. 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.

H0 and H1

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.

Null Hypothesis = Ho = > Decisions always leads to status quo.Current status/assumption doesn't change

Alternate Hypothesis = Ha > Decisions leads to opposite of Ho.

H0 and H1 for two-tail test

Two-tailed hypothesis tests are also known as nondirectional and two-sided tests because you can test for effects in both directions. When you perform a two-tailed test, you split the significance level percentage between both tails of the distribution.

9. What is quantitative data and qualitative data?

Ans. **Qualitative data**: It is non-statistical and is typically unstructured or semi-structured. This data isn't necessarily measured using hard numbers used to develop graphs and charts. Instead, it is categorized based on properties, attributes, labels, and other identifiers.

Quantitative data: It is statistical and is typically structured in nature – meaning it is more rigid and defined. This data type is measured using numbers and values, making it a more suitable candidate for data analysis.

10. How to calculate range and interquartile range?

Ans. **calculate range**: 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.

Interquartile range :To calculate these two measures, need to know the values of the lower and upper quartiles. The lower quartile, or first quartile (Q1), is the value under which 25% of data points are found when they are arranged in increasing order. The upper quartile, or third quartile (Q3), is the value under which 75% of data points are found when arranged in increasing order. The median is considered the second quartile (Q2). The interquartile range is the difference between upper and lower quartiles. The semi-interquartile range is half the interquartile range.

11. What do you understand by bell curve distribution ?

Ans. The term "bell curve" is used to describe a graphical depiction of a normal probability distribution, whose underlying standard deviations from the mean create the curved bell shape. A standard deviation is a measurement used to quantify the variability of data dispersion, in a set of given values around the mean.

12. Mention one method to find outliers.

Ans. Using 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?

Ans.In statistical hypothesis testing, P-Value or probability value can be defined as the measure of the probability that a real-valued test statistic is at least as extreme as the value actually obtained. P-value shows how likely it is that your set of observations could have occurred under the null hypothesis. P-Values are used in statistical hypothesis testing to determine whether to reject the null hypothesis. The smaller the p-value, the stronger the likelihood that you should reject the null hypothesis.

14. What is the Binomial Probability Formula?

Ans. The Binomial Probability distribution of exactly x successes from n number of trials is given by the below formula-

P (X) = nCx px qn – x Where,

n = Total number of trials

x = Total number of successful trials

p = probability of success in a single trial

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

15. Explain ANOVA and it's applications.

Ans. ANOVA is used to compare differences of means among more than 2 groups. It does this by looking at variation in the data and where that variation is found (hence its name). Specifically, ANOVA compares the amount of variation between groups with the amount of variation within groups.

- Null hypothesis, typically is that, all means are equal.
- The independent variables are categorical.
- Dependent variables are continues.

WORKSHEET

FLIP ROBO