STATISTICS WORKSHEET-4

Answers:

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

Answer: The Central Limit Theorem is a statistical concept that states that the sample mean distribution of a random variable will assume a near-normal or normal distribution if the sample size is large enough.

- The theorem states that the sampling distribution of the mean approaches a normal distribution as the size of the sample increases, regardless of the shape of the original population distribution
- The sample size must be 30 or higher for the central limit theorem to hold

Importance: 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

Central limit theorem helps us to make inferences about the sample and population parameters and construct better machine learning models using them

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

Answer: Sampling is the process of selecting a number of cases from all the cases in a particular group or universe. It is the most crucial part of inferential statistics. Inference or conclusions to be drawn based on sample of a whole population.

1) Probability sampling: involves random selection

- Simple random sampling.
- Systematic sampling.
- Stratified sampling.
- Cluster sampling.

2) Non-probability sampling: involves non random selection

- convenience sampling
- voluntary response sampling
- purposive sampling
- snowball sampling
- quota sampling

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

Answer:

Type 1 error	Type 2 error
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
Ex: the test result says you have diabetes, but you actually don't	Ex: the test result says you don't have diabetes, but you actually do

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

Answer: 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

It will be appearing as a bell curve in graphical form. The mean, median and mode are all equal. This is seen as perfect/ideal distribution of data and more preferred for statistical analysis.

If the mean, median and mode are unequal, the distribution will be either positively or negatively skewed

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. Used for describing simple relationships without making a statement about cause and effect.

Correlation coefficient ranges between -1 to +1 and is unit less

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.

Covariance ranges between -∞ to +∞ and has unit of variance

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

Answer:

Univariate analysis	**Bivariate analysis**	**Multivariate analysis**
This type of data consists of only one variable	This type of data involves two different variables	When the data involves three or more variables
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 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	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
The example of a univariate data can be height.	Example of bivariate data can be temperature and ice cream sales in summer season	Example flight fare prediction by duration, season, distance, weather, destination,…. data

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

Answer: Sensitivity is known as the True Positive Rate or Recall. It informs us about the proportion of actual positive cases that have gotten predicted as positive by our model.

Sensitivity = True Positive / (True Positive + False Negative)

Popular and critical model performance parameter in medical field, sensitivity may describe how well a test can detect a specific disease or condition in people who actually have the disease or condition.

No test has 100% sensitivity because some people who have the disease or condition will not be identified by the test

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

Answer: Hypothesis testing is a form of statistical inference that uses data from a sample to draw conclusions about a population parameter

H0: Null hypothesis is a type of statistical hypothesis that proposes that no statistical significance exists in a set of given observations

H1: Alternative hypothesis is a type of statistical hypothesis that proposes there is statistical significance exists in a set of given observations and is contradictory for null hypothesis

2 tailed test with example:

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 weight of students in the batch is x.

Alternate hypothesis (H1): The alternate hypothesis is always what is being claimed. It will be the average weight of students in the batch is less than x or greater than x.

9. What is quantitative data and qualitative data?
Answer: Qualitative data is information that cannot be counted, measured or easily expressed using numbers. They can be collected in form of text, audio, images, etc.
Quantitative data is data that can be expressed as certain quantity, amount or range. Example: Length that can be measured in meters.

10. How to calculate range and interquartile range?
Answer: 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)
Range = max(observed values) – min(observed values)
Interquartile range (IQR) is a measure of statistical dispersion, which is the spread of the data. Can also be called the midspread (middle 50% or fourth spread or H-spread). It is defined as the difference between the 75th and 25th percentiles of the data.
Q1 = First quartile = 25th percentile
Q3 = Third quartile = 75th percentile
IQR = Q3 – Q1
That is the data points between Q3 aand Q1

11. What do you understand by bell curve distribution ?
Answer: The term "bell curve" originates from the fact that the graph used to depict a normal distribution consists of a symmetrical bell-shaped curve.
The highest point on the curve, or the top of the bell, represents the most probable event in a series of data (its mean = mode = median), while all other possible occurrences are symmetrically distributed around the mean, creating a downward-sloping curve on each side of the peak. The width of the bell curve is described by its standard deviation.

12. Mention one method to find outliers.
Answer: An outlier is an observation that lies an abnormal distance from other values in a random sample from a population.

1) Interquartile range (IQR) method:
   Q1 = First quartile = 25th percentile
   Q3 = Third quartile = 75th percentile
   The data points which fall below (Q1 – 1.5 IQR) or above (Q3 + 1.5 IQR) are outliers

2) Zscore method:
The difference between any value in a set of data and the mean of that data, when measured in standard deviation units is simply called the Z-score, also known as the standard score.
$z = (x - \mu) / \sigma$
x is the raw score
μ is the mean of the population
σ is the standard deviation of the population

Threshold zscore value is used to know the relevant values as outliers. For example, if the threshold value is assumed to be 3, then the z-score above +3 and the z-score below -3 are considered outliers.

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.

- $P > 0.05$ is the probability that the null hypothesis is true.
- 1 minus the P value is the probability that the alternative hypothesis is true.
- A statistically significant test result ($P \leq 0.05$) means that the test hypothesis is false or should be rejected.
- A P value greater than 0.05 means that no effect was observed.

14. What is the Binomial Probability Formula?

Answer: Binomial probability describes the likelihood of getting a specific number of success outcomes when conducting an experiment with a set number of trials. There are only two possible results or outcomes in a binomial experiment, success or failure

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

n = the number of experiments

x = 0, 1, 2, 3, 4, …

p = Probability of success in a single experiment

q = Probability of failure in a single experiment (= 1 – p)

15. Explain ANOVA and it’s applications.

Answer: ANOVA - Analysis of Variance is a statistical formula used to compare variances across the means (or average) of different groups.

Tells you if there are any statistical differences between the means of three or more independent groups.

Applications:

1) To prove or disprove if medication were equally effective or not

2) To help manage budgets by comparing your budget to costs to help manage revenue and inventory

3) ANOVA can also be used to compare the performance of different athletes and identify trends in performance over time

4) A real-time use case of ANOVA in the e-commerce industry would be to track customer satisfaction levels with different types of products.