STATISTICS WORKSHEET-4

Q1to Q15 are descriptive types. Answer in brief.

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

Answer: The central limit theorem is a statistical theorem that states that given a sufficiently large sample size from a population with a finite level of variance, the mean of all samples from the same population will be approximately equal to the mean of the population. Additionally, the distribution of the sample means will be approximately normal, even if the distribution of the original population is not normal.

The central limit theorem is important because it provides a way to make statistical inferences about a population based on a sample drawn from that population. This is especially useful when it is not practical to measure all members of a population, as is often the case in many real-world situations. For example, a pollster might use the central limit theorem to make inferences about the opinions of a whole country based on a sample of a few thousand people.

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

Answer: Sampling is the process of selecting a subset of individuals from a larger population to study. The subset of individuals that is selected is called a sample. Sampling is used in research and data analysis to make inferences about the characteristics of a population based on the characteristics of the sample.

There are several different types of sampling methods, including:

1. Simple random sampling: This is a method of sampling where every member of the population has an equal chance of being selected for the sample.
2. Stratified sampling: This is a method of sampling where the population is divided into strata (homogeneous subgroups) and a sample is drawn from each stratum.

3. Cluster sampling: This is a method of sampling where the population is divided into clusters, and a sample of clusters is selected. The individuals within the selected clusters are then included in the sample.
4. Systematic sampling: This is a method of sampling where the members of the population are selected at regular intervals.
5. Convenience sampling: This is a method of sampling where the individuals in the sample are selected based on their availability or accessibility.
6. Quota sampling: This is a method of sampling where the sample is selected to match the proportions of certain characteristics in the population.

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

Answer: In statistical hypothesis testing, a type I error is a error that occurs when the null hypothesis is rejected, even though it is true. A type I error is also known as a false positive.

On the other hand, a type II error is a error that occurs when the null hypothesis is not rejected, even though it is false. A type II error is also known as a false negative.

The probability of making a type I error is denoted by the Greek letter alpha (α) and is also known as the level of significance. The probability of making a type II error is denoted by the Greek letter beta (β).

In statistical hypothesis testing, it is important to try to minimize the chances of both type I and type II errors. However, it is generally not possible to completely eliminate the risk of either type of error, and a trade-off must often be made between the two.

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

Answer: In statistics, the normal distribution is a very common continuous probability distribution. It is also known as the Gaussian distribution and the bell curve. It is a symmetric distribution with a single peak, and the mean, median, and mode of the distribution are all equal.

The normal distribution is defined by its mean, which is the average value of the distribution, and its standard deviation, which is a measure of the spread of the

distribution. The normal distribution is completely determined by these two parameters.

The normal distribution is often used to model random variables that take on a large number of values, such as measurements of physical quantities. It is widely used in many fields, such as economics, engineering, and the natural and social sciences, because it is a good approximation for many real-world phenomena.

5. What is correlation and covariance in statistics?

Answer: Correlation and covariance are measures of the relationship between two or more variables.

Correlation measures the strength and direction of a linear relationship between two variables. A positive correlation means that as the value of one variable increases, the value of the other variable also increases. A negative correlation means that as the value of one variable increases, the value of the other variable decreases. The strength of the correlation is measured using a coefficient, called the Pearson correlation coefficient, which ranges from -1 to 1. A value of 0 indicates that there is no correlation between the variables.

Covariance is a measure of the degree to which two variables vary together. It is similar to correlation, but it is not limited to a linear relationship between the variables. Like correlation, a positive covariance means that the variables tend to increase or decrease together, while a negative covariance means that the variables tend to move in opposite directions. The covariance of two variables is calculated as the product of the standard deviations of the two variables and the correlation between them.

Both correlation and covariance can be used to quantify the relationship between two variables, but they differ in their interpretation and the types of relationships they can measure. Correlation is generally easier to interpret and is more widely used, but covariance is a more general measure and can be useful in certain situations.

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

Answer: Univariate analysis is the analysis of a single variable. It is used to describe the characteristics of a single variable, such as its mean, median, mode, range, and standard deviation. Univariate analysis is used to summarize and understand the distribution of a single variable.

Bivariate analysis is the analysis of two variables. It is used to investigate the relationship between two variables, such as whether there is a correlation between the variables or whether one variable is a predictor of the other. Bivariate analysis can be used to explore the relationship between two variables and to understand how changes in one variable are related to changes in the other.

Multivariate analysis is the analysis of three or more variables. It is used to investigate the relationships among multiple variables and to understand how these variables are related to one another. Multivariate analysis is a powerful tool for understanding complex data sets and can be used to identify patterns and relationships that may not be apparent in univariate or bivariate analysis.

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

Answer: Sensitivity is a measure of the ability of a test or model to correctly identify true positive cases. It is defined as the number of true positive cases divided by the total number of positive cases. Mathematically, it can be expressed as follows:

Sensitivity = True Positives / (True Positives + False Negatives)

For example, suppose a test for a certain disease is administered to 100 people, and the results are as follows:

True Positives: 30

False Positives: 10

True Negatives: 50

False Negatives: 10

The sensitivity of the test can be calculated as follows:

Sensitivity = 30 / (30 + 10) = 0.75

This means that the test correctly identified 75% of the people who had the disease.

It is important to note that sensitivity is not a measure of the overall accuracy of the test. It only measures the ability of the test to correctly identify positive cases. To measure the overall accuracy of the test, you would need to consider other measures such as specificity, which is the ability of the test to correctly identify negative cases, and positive and negative predictive value, which are measures of the overall probability that a positive or negative test result is actually true.

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

Answer: Hypothesis testing is a statistical method used to test a hypothesis about a population parameter. It involves specifying a hypothesis (called the null hypothesis, or H0) and an alternative hypothesis (called H1), and then collecting data to determine whether the data support the null hypothesis or the alternative hypothesis.

The null hypothesis is a statement that assumes that there is no relationship between the variables being studied, or that there is no difference between the groups being compared. The null hypothesis is usually denoted as H0.

The alternative hypothesis is a statement that contradicts the null hypothesis. It asserts that there is a relationship between the variables or a difference between the groups. The alternative hypothesis is usually denoted as H1.

9. What is quantitative data and qualitative data?

Answer: Quantitative data is data that is numerical and can be measured or counted. It is usually collected using instruments such as surveys, experiments, or observational studies. Quantitative data is often used to test hypotheses, make predictions, and draw conclusions.

Examples of quantitative data include:

- Age
- Height
- Weight
- Income
- Test scores
- Time elapsed

Qualitative data is data that describes qualities or characteristics. It is often collected using methods such as interviews, focus groups, or open-ended surveys. Qualitative data is used to understand people's experiences, attitudes, behaviors, and motivations.

Examples of qualitative data include:

- Opinions
- Attitudes
- Beliefs
- Motivations
- Perceptions
- Experiences

10. How to calculate range and interquartile range?

Answer: Range is a measure of dispersion that indicates the spread of a data set. It is calculated by subtracting the minimum value from the maximum value in the data set. For example, if the data set is {1, 2, 3, 4, 5}, the minimum value is 1 and the maximum value is 5, so the range is 5 - 1 = 4.

Interquartile range (IQR) is another measure of dispersion that is used to describe the spread of a data set. It is calculated as the difference between the upper quartile (Q3) and the lower quartile (Q1). The upper quartile is the value that separates the top 25% of the data from the bottom 75%, and the lower quartile is the value that separates the bottom 25% from the top 75%.

To calculate the IQR, you first need to calculate the quartiles of the data set. To do this, you will need to rank the data set from smallest to largest and then find the median value (Q2). The lower quartile (Q1) is the median of the lower half of the data set, and the upper quartile (Q3) is the median of the upper half of the data set.

Once you have calculated the quartiles, you can then calculate the IQR by subtracting the lower quartile (Q1) from the upper quartile (Q3). For example, if the lower quartile is 10 and the upper quartile is 20, the IQR is 20 - 10 = 10.

The IQR is a more robust measure of dispersion than the range because it is not influenced by outliers (extremely high or low values) in the data set. It is a useful tool for identifying the central tendency and spread of a data set and for comparing data sets to one another.

11. What do you understand by bell curve distribution?

Answer: The bell curve, also known as the normal distribution, is a continuous probability distribution that is symmetric and bell-shaped. It is defined by its mean (μ) and standard deviation (σ). The mean is the average value of the distribution, and the standard deviation is a measure of the spread of the distribution.

The bell curve is commonly used to model random variables that take on a large number of values, such as measurements of physical quantities. It is a good approximation for many real-world phenomena because it is a continuous distribution that is symmetrical and has a single peak.

The shape of the bell curve is determined by the mean and standard deviation of the distribution. If the mean is large and the standard deviation is small, the bell curve will be tall and narrow. If the mean is small and the standard deviation is large, the bell curve will be short and wide.

The bell curve is used in many fields, including economics, engineering, and the natural and social sciences, because it is a good model for many real-world phenomena. It is also used in statistical analysis to test hypotheses, make predictions, and draw conclusions.

12. Mention one method to find outliers.

Answer: One common method for finding outliers in a data set is to use the interquartile range (IQR). The IQR is a measure of dispersion that is calculated as the difference between the upper quartile (Q3) and the lower quartile (Q1). It is a robust measure of dispersion that is not influenced by extreme values (outliers) in the data set.

To use the IQR to identify outliers, you can calculate the following thresholds:

Lower threshold = Q1 - (1.5 * IQR) Upper threshold = Q3 + (1.5 * IQR)

Any values in the data set that are below the lower threshold or above the upper threshold are considered to be outliers.

For example, suppose you have the following data set: {2, 3, 4, 5, 6, 7, 8, 9, 10}

To find the IQR, you would first need to calculate the quartiles of the data set. The lower quartile (Q1) is the median of the lower half of the data set, which is 4. The upper quartile (Q3) is the median of the upper half of the data set, which is 8. The IQR is then calculated as Q3 - Q1 = 8 - 4 = 4.

Using the IQR to identify outliers, the lower threshold is Q1 - (1.5 * IQR) = 4 - (1.5 * 4) = -1, and the upper threshold is Q3 + (1.5 * IQR) = 8 + (1.5 * 4) = 13. There are no values in the data set that are below the lower threshold or above the upper threshold, so there are no outliers in this data set.

13. What is p-value in hypothesis testing?

Answer: In hypothesis testing, the p-value is the probability of obtaining a test statistic at least as extreme as the one observed, assuming that the null hypothesis is true. The p-value is used to evaluate the evidence against the null hypothesis. If the p-value is small, it means that the observed data are unlikely to have occurred by chance if the null hypothesis is true, and therefore the null hypothesis is rejected in favor of the alternative hypothesis.

The p-value is calculated based on the sampling distribution of the test statistic and the observed value of the test statistic. If the observed value of the test statistic is

unusual or extreme relative to the sampling distribution, it means that it is unlikely to have occurred by chance, and the p-value will be small. On the other hand, if the observed value of the test statistic is not unusual or extreme relative to the sampling distribution, it means that it is likely to have occurred by chance, and the p-value will be large.

The p-value is usually compared to a pre-specified significance level, such as 0.05 or 0.01, to determine whether the null hypothesis should be rejected. If the p-value is less than the significance level, it means that the evidence against the null hypothesis is strong enough to reject it in favor of the alternative hypothesis. If the p-value is greater than the significance level, it means that the evidence against the null hypothesis is not strong enough to reject it, and the null hypothesis is accepted.

14. What is the Binomial Probability Formula?

Answer: The binomial probability formula is a statistical formula used to calculate the probability of a specific outcome in a binomial experiment. A binomial experiment is a statistical experiment that has the following characteristics:

- It consists of n independent trials.
- Each trial has only two possible outcomes: success or failure.
- The probability of success is the same for each trial.

The binomial probability formula is used to calculate the probability of a specific number of successes in a binomial experiment. It is defined as the probability of x successes in n trials, given a probability p of success on each trial. Mathematically, it can be expressed as follows:

P(x successes in n trials) = (n! / (x! * (n - x)!) * p^x * (1 - p)^(n - x))

Where:

- n is the total number of trials.
- x is the number of successes.
- p is the probability of success on each trial.
- ! denotes the factorial operator (e.g. 5! = 5 * 4 * 3 * 2 * 1).

15. Explain ANOVA and it's applications.

Answer: ANOVA (Analysis of Variance) is a statistical technique used to test the difference between the means of two or more groups. It is used to determine whether there is a significant difference between the means of the groups, or whether the differences between the means are due to random chance.

ANOVA is used to compare the means of three or more independent groups. It is a parametric test, which means that it assumes that the data are normally distributed and that the variances of the groups are equal.

There are three main types of ANOVA:

7. One-way ANOVA: This is used to compare the means of two or more independent groups.
8. Two-way ANOVA: This is used to compare the means of two or more groups that have been divided into subgroups.
9. Three-way ANOVA: This is used to compare the means of three or more groups that have been divided into subgroups.

ANOVA is used in many fields, including psychology, economics, biology, and engineering. It is a widely used tool for analyzing data and making inferences about populations based on sample data.