Statistics Worksheet-4

Q1to Q15 are descriptive types. Answer in brief.

Q1) What is central limit theorem and why is it important?

Statement of Central limit Theorem:

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.

We can calculate the mean of the sample means for the random samples we choose from the population:

$\mu\bar{X}=\mu$

As well as the standard deviation of sample means:

$\sigma\bar{X}=\sigma n$

According to the central limit theorem, the form of the sampling distribution will approach normalcy as the sample size is sufficiently large (usually n>30). regardless of the population distribution.

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.

Q2) What is sampling? How many sampling methods do you know?

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.

1) Simple Random Sampling
2) Systematic Sampling
3) Cluster sampling
4) Stratified Random Sampling

Q3) What is the difference between type-I and type-II error?

- Type I error is an error that takes place when the outcome is a rejection of null hypothesis which is, in fact, true. Type II error occurs when the sample results in the acceptance of null hypothesis, which is actually false.
- Type I error or otherwise known as false positives, in essence, the positive result is equivalent to the refusal of the null hypothesis. In contrast, Type II error is also known as false negatives, i.e. negative result, leads to the acceptance of the null hypothesis.
- When the null hypothesis is true but mistakenly rejected, it is type I error. As against this, when the null hypothesis is false but erroneously accepted, it is type II error
- Type I error tends to assert something that is not really present, i.e. it is a false hit. On the contrary, type II error fails in identifying something, that is present, i.e. it is a miss.
- The probability of committing type I error is the sample as the level of significance. Conversely, the likelihood of committing type II error is same as the power of the test.
- Greek letter 'α' indicates type I error. Unlike, type II error which is denoted by Greek letter 'β'.

Q4) What do you understand by the term Normal distribution?

The normal distribution is a continuous probability distribution that is symmetrical on both sides of the mean, so the right side of the center is a mirror image of the left side.

The area under the normal distribution curve represents probability and the total area under the curve sums to one.

Most of the continuous data values in a normal distribution tend to cluster around the mean, and the further a value is from the mean, the less likely it is to occur. The tails are asymptotic, which means that they approach but never quite meet the horizon (i.e. x-axis).

Q5) What is correlation and covariance in statistics?

Covariance and correlation are two mathematical concepts used in statistics. Both terms are used to describe how two variables relate to each other. Covariance is a measure of how two variables change together. The terms covariance vs correlation is very similar to each other in probability theory and statistics.

Covariance signifies the direction of the linear relationship between the two variables. By direction we mean if the *variables* are directly proportional or inversely proportional to each other. (Increasing the value of one variable might have a positive or a negative impact on the value of the other variable)

Correlation analysis is a method of statistical evaluation used to study the strength of a relationship between two, numerically measured, continuous variables.

It not only shows the kind of relation (in terms of direction) but also how strong the relationship is. Thus, we can say the correlation values have standardized notions, whereas

the covariance values are not standardized and cannot be used to compare how strong or weak the relationship is because the magnitude has no direct significance. It can assume values from -1 to +1.

Q6) Differentiate between univariate, Bivariate and multivariate analysis.

Univariate analysis is the simplest of the three analyses where the data you are analysing is only one variable. There are many different ways people use univariate analysis. The most common univariate analysis is checking the central tendency (mean, median and mode), the range, the maximum and minimum values, and standard deviation of a variable.

Bivariate analysis is where you are comparing two variables to study their relationships. These variables could be dependent or independent to each other. In Bivariate analysis is that there is always a Y-value for each X-value.

Multivariate analysis is similar to Bivariate analysis but you are comparing more than two variables. For three variables, you can create a 3-D model to study the relationship (also known as Trivariate Analysis). However, since we cannot visualize anything above the third dimension, we often rely on other software and techniques for us to be able to grasp the relationship in the data.

Q7) What do you understand by sensitivity and how would you calculate it?

The term sensitivity was introduced by Yerushalmy in the 1940s as a statistical index of diagnostic accuracy.

It is also called the true positive rate, the recall, or probability of detection.

It has been defined as the ability of a test to identify correctly all those who have the disease, which is “true-positive”.

A 90 percent sensitivity means that 90 percent of the diseased people screened by the test will give a “true-positive” result and the remaining 10 percent a “false-negative” result.

Thus, a highly sensitive test rarely overlooks an actual positive (for example, showing “nothing bad” despite something bad existing).

The sensitivity of a diagnostic test is expressed as the probability (as a percentage) that a sample tests positive given that the patient has the disease.

The following equation is used to calculate a test’s sensitivity

Sensitivity = (No of true positive)/(No. of True positives + No. of false negatives)

Q8) What is hypothesis testing? What is H0 and H1? What is H0 and H1 for two-tail test?

Hypothesis testing is a statistical interpretation that examines a sample to determine whether the results stand true for the population.

Hypothesis testing uses sample data to validate the research. Researchers speculate on relationships between various factors. They then collect data to test those relationships. Based on the data, researchers draw conclusions. In **statistics**, it is very important to eliminate randomness. The data should not have been caused by chance or a random factor. Hypothesis testing eliminates such uncertainties.

$\mathbf{H_0}$ = null hypothesis

$\mathbf{H_1}$ = alternate hypothesis

Q9) What is quantitative data and qualitative data?

Quantitative data: The data collected on the grounds of the numerical variables are quantitative data. Quantitative data are more objective and conclusive in nature. It measures the values and is expressed in numbers. The data collection is based on "how much" is the quantity. The data in quantitative analysis is expressed in numbers so it can be counted or measured. The data is extracted from experiments, surveys, market reports, matrices, etc.

Qualitative data: The data collected on grounds of categorical variables are qualitative data. Qualitative data are more descriptive and conceptual in nature. It measures the data on basis of the type of data, collection, or category. The data collection is based on what type of quality is given. Qualitative data is categorized into different groups based on characteristics. The data obtained from these kinds of analysis or research is used in theorization, perceptions, and developing hypothetical theories. These data are collected from texts, documents, transcripts, audio and video recordings, etc.

Q10) How to calculate range and interquartile range?

Statistics, the **range** is the smallest of all the measures of dispersion. It is the difference between the two extreme conclusions of the distribution. In other words, the range is the difference between the maximum and the minimum observation of the distribution.

It is defined by

Range = $X_{max} - X_{min}$

Where X_{max} is the largest observation and X_{min} is the smallest observation of the variable values.

Interquartile Range Definition

The interquartile range defines the difference between the third and the first quartile. Quartiles are the partitioned values that divide the whole series into 4 equal parts. So, there are 3 quartiles. First Quartile is denoted by Q_1 known as the lower quartile, the second

Quartile is denoted by Q_2 and the third Quartile is denoted by Q_3 known as the upper quartile. Therefore, the interquartile range is equal to the upper quartile minus lower quartile.

Interquartile Range Formula

The difference between the upper and lower quartile is known as the interquartile range. The formula for the interquartile range is given below

Interquartile range = Upper Quartile – Lower Quartile = $Q_3 – Q_1$

where Q_1 is the first quartile and Q_3 is the third quartile of the series.

Q11) What do you understand by bell curve distribution?

The term bell curve is used to describe the mathematical concept called **normal distribution**, sometimes referred to as Gaussian distribution. "Bell curve" refers to the bell shape that is created when a line is plotted using the data points for an item that meets the criteria of normal distribution.

Q12) Mention one method to find outliers.

We can use software to **visualize** your data with a box plot, or a box-and-whisker plot, and can see the data distribution at a glance. This type of chart highlights minimum and maximum values (the range), the median, and the interquartile range for data.

Q13) What is p-value in hypothesis testing?

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.

Q14) What is the Binomial Probability Formula?

In 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

Q15) Explain ANOVA and its applications

Analysis of variance (ANOVA) is an analysis tool used in statistics that splits an observed aggregate variability found inside a data set into two parts: systematic factors and random factors. The systematic factors have a statistical influence on the given data set, while the random factors do not. Analysts use the ANOVA test to determine the influence that independent variables have on the dependent variable in a regression study.

The t- and z-test methods developed in the 20th century were used for statistical analysis until 1918, when Ronald Fisher created the analysis of variance method.12 ANOVA is also called the Fisher analysis of variance, and it is the extension of the t- and z-tests.