STATISTICS WORKSHEET SOLUTIONS:

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

The Central Limit Theorem tells us that as sample sizes get larger, the sampling distribution of the mean will become normally distributed, even if the data within each sample are not normally distributed.

The Central Limit Theorem is important for statistics because 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.

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

The sample is the specific group of individuals that you will collect data from. The sampling frame is the actual list of individuals that the sample will be drawn from. Ideally, it should include the entire target population (and nobody who is not part of that population).

There are two types of sampling methods:

- Probability sampling involves random selection, allowing you to make strong statistical inferences about the whole group.
- Non-probability sampling involves non-random selection based on convenience or other criteria, allowing you to easily collect data.

Q3) What is the difference between type1 and type II error?

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.

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

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.

Q5) What is correlation and covariance in statistics?

Covariance and Correlation are two mathematical concepts which are commonly used in the field of probability and statistics. Both concepts describe the relationship between two variables.

Covariance –

1. It is the relationship between a pair of random variables where change in one variable causes change in another variable.
2. It can take any value between -infinity to +infinity, where the negative value represents the negative relationship whereas a positive value represents the positive relationship.

Correlation –

1. It show whether and how strongly pairs of variables are related to each other.
2. Correlation takes values between -1 to +1, wherein values close to +1 represents strong positive correlation and values close to -1 represents strong negative correlation.
3. In this variable are indirectly related to each other.
4. It gives the direction and strength of relationship between variables.

Q6) Differentiate between univariate , Bivariate and multivariate analysis.

- Univariate statistics summarize only one variable at a time. The main purpose of univariate analysis is to describe the data and find patterns that exist within it.
- Bivariate statistics compare two variables.
- Multivariate statistics compare more than two variables.

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

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.

The sensitivity is calculated by dividing the percentage change in output by the percentage change in input.

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

Hypothesis testing is an act in statistics whereby an analyst tests an assumption regarding a population parameter. Hypothesis testing is used to assess the plausibility of a hypothesis by using sample data. Such data may come from a larger population, or from a data-generating process.

All analysts use a random population sample to test two different hypotheses: the null hypothesis and the alternative hypothesis.

The null hypothesis would be represented as Ho: P = 0.5. The alternative hypothesis would be denoted as "H1" and be identical to the null hypothesis.

For a two tailed test, the null hypothesis (H0) should be rejected when the test value is in either of two critical regions on either side of the distribution of the test value and vice versa for alternate hypothesis.

Q9) What is quantitative data and qualitative data?

Quantitative data is information about quantities, and therefore numbers, examples are length, mass, temperature, and time whereas qualitative data is descriptive, and regards phenomenon which can be observed but not measured, such as language.

Q10) How to calculate range and interquartile range?

The Range is the difference between the lowest and highest values. Example: In {10, 6, 9, 3, 2} the lowest value is 2, and the highest is 10. So the range is 10 − 2 = 8.

We can find the interquartile range or IQR in four simple steps:

1. Order the data from least to greatest
2. Find the median
3. Calculate the median of both the lower and upper half of the data
4. The IQR is the difference between the upper and lower medians

Q11) What do you understand by bell curve distribution?

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. The mean, in turn, refers to the average of all data points in the data set or sequence and will be found at the highest point on the bell curve.

Q12) Mention one method to find outliers.

One of the methods to Calculate the Outlier is by using the Interquartile Range

1. Take your IQR and multiply it by 1.5 and 3. We'll use these values to obtain the inner and outer fences. ...
2. Calculate the inner and outer lower fences. Take the Q1 value and subtract the two values from step 1. ...
3. Calculate the inner and outer upper fences.

Q13) What is p-value in hypothesis testing?

The p-value, or probability value, tells you how likely it is that your data could have occurred under the null hypothesis.The p-value is a proportion: if your p-value is 0.05, that means that

5% of the time you would see a test statistic at least as extreme as the one you found if the null hypothesis was true.

Q14) What is the Binomial Probability Formula?

The binomial distribution formula is:

$b(x; n, P) = {}_nC_x * P^x * (1 - P)^{n-x}$
Where:
b = binomial probability
x = total number of "successes" (pass or fail, heads or tails etc.)
P = probability of a success on an individual trial
n = number of trials

Q15) Explain ANOVA and it's applications.

Analysis of variance (ANOVA) is a statistical technique that is used to check if the means of two or more groups are significantly different from each other. ANOVA checks the impact of one or more factors by comparing the means of different samples.

There are many industries that can use the ANOVA test to identify issues or variances between samples. The ANOVA is a good statistical technique for testing. Businesses that might consider the use of the ANOVA include manufacturing, healthcare, service, food, and more.