Statistics Worksheet-4

Q1-What is central limit theorem and why is it important?
Answer- Central Limit Theorem s a statistical theory that states that - if you take a sufficiently large sample size from a population with a finite level of variance, the mean of all samples from that population will be roughly equal to the population mean.

Normal distribution - Distribution is always normal irrespective of sample size
Non-Normal distribution - If sample size is adequate (appr > 30 sample), distribution starts looking normal.
$\mu\bar{x} = \mu$ (Mean of sample mean = Population mean)
$\sigma\bar{x} = \sigma/\sqrt{n}$ (Population std / sqrt of sample size)
Where,
μ = Population mean
σ = Population standard deviation
μx = Sample mean
σx= Sample standard deviation
n = Sample size
$\bar{x}$ = Sample mean

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.**

Q2- What is sampling? How many sampling methods do you know?
Answer-Sampling- A sample is a subset of individuals from a larger population. Sampling means selecting the group that you will actually collect data from in your research. For example, if you are researching the opinions of students in your university, you could survey a sample of 100 students.

In statistics, sampling allows you to test a hypothesis about the characteristics of a population.
There are two primary types of sampling methods that you can use in your research:

1) Probability sampling involves random selection, allowing you to make strong statistical inferences about the whole group. means that every member of the population has a chance of being selected. It is mainly used in quantitative research. If you want to produce results that are representative of the whole population, probability sampling techniques are the most valid choice.
There are four main types of probability sample.

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

2) Non-probability sampling involves non-random selection based on convenience or other criteria, allowing you to easily collect data. This type of sample is easier and cheaper to access, but it has a higher risk of sampling bias. That means the inferences you can make about the population are weaker than with probability samples, and your conclusions may be more limited. If you use a non-probability sample, you should still aim to make it as representative of the population as possible.

- Convenience sampling
- Voluntary response sampling
- Purposive sampling
- Snowball sampling

Q3-What is the difference between type1 and typeII error?

Answer- **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.

Q4- What do you understand by the term Normal distribution?
Answer- **A normal distribution is a type of continuous probability distribution in which most data points cluster toward the middle of the range, while the rest taper off symmetrically toward either extreme. The middle of the range is also known as the mean of the distribution**.
Graphically, a normal distribution is a bell curve because of its flared shape. Regardless of its exact shape, a normal distribution bell curve is always symmetrical about the mean.

Q5- What is correlation and covariance in statistics?
Answer5- **Covariance** is an indicator of the extent to which 2 random variables are dependent on each other. A higher number denotes higher dependency. The value of covariance lies in the range of -∞ and +∞.

Correlation is a statistical measure that indicates how strongly two variables are related. Correlation is limited to values between the range -1 and +1

Q7- Differentiate between univariate ,Biavariate,and multivariate analysis?
Answer-
Univariate statistics summarize only one variable at a time. The analysis of univariate data is thus the simplest form of analysis since the information deals with only one quantity that changes. 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 example of a univariate data can be height.
Bivariate statistics compare two variables. 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.Example of bivariate data can be temperature and ice cream sales in summer season.

Multivariate statistics compare more than two variables. it is categorized under multivariate. Example of this type of data is suppose an advertiser wants to compare the popularity of four advertisements on a website, then their click rates could be measured for both men and women and relationships between variables can then be examined.

Q7- What do you understand by sensitivity and how would you calculate it?
Answer- **The technique used to determine how independent variable values will impact a particular dependent variable under a given set of assumptions is defined as sensitive analysis.** It's usage will depend on one or more input variables within the specific boundaries, such as the effect that changes in interest rates will have on a bond's price.
It is also known as the what – if analysis. Sensitivity analysis can be used for any activity or system. All from planning a family vacation with the variables in mind to the decisions at corporate levels can be done through sensitivity analysis.
Below are mentioned the steps used to conduct sensitivity analysis:

1. Firstly the base case output is defined; say the NPV at a particular base case input value (V1) for which the sensitivity is to be measured. All the other inputs of the model are kept constant.
2. Then the value of the output at a new value of the input (V2) while keeping other inputs constant is calculated.
3. Find the percentage change in the output and the percentage change in the input.
4. 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?
Answer**- The procedure to decide whether the sample data are agreeable or consistent with the null hypothesis is called statistical hypothesis testing or simply hypothesis testing or test of significance.**

- **Null Hypothesis(H0):** A default or unverified statementabout the population parameter is called null hypothesis.

- **Alternative Hypothesis(H1)**: A claim or statement about the population parameter that contradicts or against the null hypothesis. A useful notation is H1.

Q9- What is quantitative data and qualitative data?
Answer: Quantitative data are measures of values or counts and are expressed as numbers.
Quantitative data are data about numeric variables (e.g. how many; how much; or how often).

Quantitative = Quantity

Qualitative data are measures of 'types' and may be represented by a name, symbol, or a number code.
Qualitative data are data about categorical variables (e.g. what type)

Qualitative = Quality

Q10- How to calculate range and interquartile range?
Answer: **Range**- is calculated by subtracting the lowest value from the highest value.
The formula to calculate the range is:

R = H-L

- *R* = range
- *H* = highest value
- *L* = lowest value

Interquartile Range :To find the interquartile range (IQR), **first find the median (middle value) of the lower and upper half of the data**. These values are quartile 1 (Q1) and quartile 3 (Q3). The IQR is the difference between Q3 and Q1.

IQR= Q3-Q1

- IQR- **Interquartile Range**
- Q1- quartile 1 (25% of data)
- Q3- quartile 3 (75% of data)

Q11- What do you understand by bell curve distribution ?
Answer- **A bell curve is a common type of distribution for a variable, also known as the normal distribution**. 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 top of the curve shows the mean, mode, and median of the data collected.
- Its standard deviation depicts the bell curve's relative width around the mean.
- Bell curves (normal distributions) are used commonly in statistics, including in analyzing economic and financial data.

Q12- Mention one method to find outliers.
Answer-Outliers are extreme values that differ from most other data points in a dataset. They can have a big impact on your statistical analyses and skew the results of any hypothesis tests.

<u>Interquartile Range Method</u>

1. Sort your data from low to high
2. Identify the first quartile (Q1), the median, and the third quartile (Q3).
3. Calculate your IQR = Q3 – Q1
4. Calculate your upper fence = Q3 + (1.5 * IQR)
5. Calculate your lower fence = Q1 – (1.5 * IQR)
6. Use your fences to highlight any outliers, all values that fall outside your fences.

Your outliers are any values greater than your upper fence or less than your lower fence.

Q13- 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.

Q14- What is the Binomial Probability Formula?
P_{x} = {n \ x} p^{x} q^{n-x}

Where,
P=binomial probability
x=number of times for a specific outcome within n trials
{n \ x}   =number of combinations
p=probability of success on a single trial
q=probability of failure on a single trial
n=number of trials

Q15 -Explain ANOVA and it's applications.
Answer- **Analysis of Variance (ANOVA)** is a statistical formula used to compare variances across the means (or average) of different groups.
ANOVA is **helpful for testing three or more variables**. It is similar to multiple two-sample t-tests. However, it results in fewer type I errors and is appropriate for a range of issues. ANOVA groups differences by comparing the means of each group and includes spreading out the variance into diverse sources.