STATISTICS WORKSHEET-4

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

Answer: The Central Limit Theorem states that the distribution of a sample mean that approximates the normal distribution, as the sample size becomes larger, assuming that all the samples are similar, and no matter what the shape of the population distribution is. The central limit theorem holds for the sample of size greater than or equal to 30. This theorem is very important for testing hypotheses in statistical analysis.

****
****

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

Answer: Sampling is the process of studying the population by gathering information and analysing that data.

There are 2 methods of Sampling: 1) Probability Sampling and Non-Probability Sampling.

Probability Sampling is a sampling technique in which samples from a larger population are chosen using a method based on the theory of probability.

Non-probability sampling is a sampling technique in which the researcher selects samples based on the researcher's subjective judgment rather than random selection.

In this 2 methods there are various types of sampling involved.

****
****

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

Answer: Type I error: If Null Hypothesis is True and we reject.

Type 2 error: If Null Hypothesis is False and you fail to reject

****
****

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

Answer: Normal Distribution data are symmetrically distributed with no skew. A proper bell shaped curve. In a normal distribution the mean is 0 and the standard deviation is 1.

****
****

5) What is correlation and covariance in statistics?

Answer: **Correlation**: measure which determines the change in one variable due to change in other variable. Correlation are of 2 types--- +ve and -ve.. It can take any value between -1 to +1 wherein values close to +1 represents strong +ve correlation and values close to -1 is an indicator of strong -ve correlation.

There are 4 measures of correlation:

i)Scatter Diagram

ii)Product-moement correlation coefficient

iii)Rank correlation coefficient

iv) Coefficient of concurrent deviation

Covariance: Relationship between a random variable. If one changes other will also change. e.g Age and Exp. Between -infinity to +infinity. +ve number--- +ve relationship and -ve number ---- -ve relationship. If x and y are same then covariance will be zero. If x and y unit changes--- There is no change in the strength of relationship.

****
****

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

Answer: **Univariate** statistics summarize only one variable at a time**. Bivariate** statistics compare two variables and **multivariate** statistics compare more than two variables.

****
****

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

Answer: Sensitivity analysis is an analysis technique that works on the basis of what-if analysis like how independent factors can affect the dependent factor and is used to predict the outcome when analysis is performed under certain conditions.

It is commonly used by investors who takes into consideration the conditions that affect their potential investment to test, predict and evaluate result. The sensitivity is **calculated** by dividing the percentage change in output by the percentage change in input.

****
****

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 or a population probability distribution. First, a tentative assumption is made about the parameter or distribution.

#**Null Hypothesis(HO)**- Decision will never change

Alternate Hypothesis (H1)- Decision will change

Our null hypothesis for two tailed test is that the mean is equal to x. A two-tailed test will test both if the mean is significantly greater than x and if the mean significantly less than x.

In a two-tailed test, the generic null and alternative hypotheses are the following: Null: The effect equals zero. Alternative: The effect does not equal zero.

****
****

9) What is quantitative data and qualitative data?

Answer: Quantitative data is data expressing a certain quantity, amount or range. Usually, there are measurement units associated with the data and are expressed as numbers. They are all about numeric values.

Qualitative data are measures of 'types' and may be represented by a name, symbol, or a number code.

****
****

10) How to calculate range and interquartile range?

Answer: The range is calculated by subtracting the lowest value from the highest value. While a large range means high variability, a small range means low variability in a distribution.

```
# Find the IQR(inter quantile range) to identify outliers

 # 1st Quantile
q1=data.quantile(0.25)

 # 3rd Quantile
q3=data.quantile(0.75)

 #IQR
iqr=q3-q1
```

****
****

11) What do you understand by bell curve distribution ?

Answer: Bell curve means the data distribution is normal. The data is distributed normally on both the sides forming a perfect bell shaped curve. There is no skewness in the data. In a normal distribution the mean is 0 and the STD is 1.

****
****

12) Mention one method to find outliers.

Answer: We can find outliers through IQR. Outlier Detection Formula.

Higher Side= Q3+(1.5*iqr) and Lower Side= q1-(1.5*iqr)

IQR detection formula:

```
# 1st Quantile
q1=data.quantile(0.25)
# 3rd Quantile
q3=data.quantile(0.75)
#IQR
iqr=q3-q1
```

****
****

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.

```
#Important Points
# 1) standard alpha value= 0.05
# 2) If P value is <0.05 then we reject the null hypothesis (HO)
# 3) If p value is >0.05 then we fail to reject the HO. That means we accept the null value
```

****
****

14) What is the Binomial Probability Formula?

Answer: The binomial distribution formula helps to check the probability of getting "x" successes in "n" independent trials of a binomial experiment. To recall, the binomial distribution is a type of probability distribution in statistics that has two possible outcomes. In probability theory, the binomial distribution comes with two parameters n and p.

Formula: $P(x) = {}^{n}C_{x} \cdot p^{x} (1 - p)^{n-x}$
Where,

- n = Total number of events
- r (or) x = Total number of successful events.
- p = Probability of success on a single trial.
- 1 – p = Probability of failure.

****
****

15) Explain Annova and its application?

Answer: Anova is used to compare differences of means among more than 2 groups. It does this by looking at variation in the data and where that variation is found.
Hypothesis Construction
The null hypothesis for Anova says- average of dependent variables are same for all the given groups.
The alternative hypothesis says- mean of dependent variable are not same for the given groups
HO=MuA=MuB=MuC
Ha= not all are equal
# **Application**
Step 1 Calculate the means of each group
Step 2 calculate the grand mean
Step 3 variation between,within
STEP 4 calculate the mean squared variance of between and within
Step 5 calculate F statistics and corresponding p value
****
****