STATISTICS WORKSHEET-1

Q1. Ans- True

Q2. Ans- Central Limit Theorem

Q3. Ans- Modelling bounded count data

Q4. Ans- All of the mentioned above

Q5. Ans-Poisson

Q6. Ans-False

Q7. Ans- Hypothesis

Q8. Ans- 0

Q9. Ans-Outliers cannot conform to the regression relationship

Q.10 Normal Distribution:-

The most important distribution of probability in statistics is **normal distribution**, also called **Gaussian or Gauss distribution** since it suits many natural phenomena.

Any distribution is known as normal distribution if it has the following characteristics:

The mean, median and mode of the distribution coincide

The curve of the distribution is bell-shaped and symmetrical about the line x=μ

The total area under the curve is 1

Exactly half of the values are to the left of the centre and the other half to the right.

The general form of its density of probability function is

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

The boundary μ is the mean and the boundary σ is its standard deviation. The fluctuation of the dispersion is σ2.

Normal Curve:

The shape of the normal curve is bell shaped. The curve has a single peak thus it is unimodel. The mean of a normally distributed population lies at the centre of its curve. Because of the symmetry, the median (Positional average) a mode (concentration of frequency) also lies at the centre. Thus in normal distribution mean, median and mode all coincide. The two tails of the curve extend indefinitely and never touches the X –axis.

Properties of Parameters:-

There are only two parameters of normal distribution i.e. mean (μ) and variance (σ2). The mean determines the location of the curve and variance determines the shape of the curve. Given variance, a change in mean will shift the curve as a whole along the x-axis. The following figure depicts the three normal curves, having same variance but different means.

Standard Normal Distribution

In this sub-section, we will study how a particular normal distribution is transformed into a standard normal distribution. This special distribution has a mean 0 and a standard deviation 1 and is written as N (0, 1).

Standard normal distribution is the distribution of another normal variable called Z-scores, which is defined as,

$z = (X-\mu)/\sigma$

The Z – score is the difference of an observation X from the mean (μ) expressed in term of standard deviation (σ). It is called Z-scores because random variable take on many different units of measures. Since we use only one table which should be standard unit and is represented by 7 this table gives the value or area of probability between variable and mean.

Q11. Handling Missing Data

Types of Missing Data:-

Understanding the nature of missing data is critical in determining what treatments can be applied to overcome the lack of data. Data can be missing in the following ways:

Missing Completely At Random (MCAR): When missing values are randomly distributed across all observations, then we consider the data to be missing completely at random. A quick check for this is to compare two parts of data – one with missing observations and the other

without missing observations. On a t-test, if we do not find any difference in means between the two samples of data, we can assume the data to be MCAR.

Missing At Random (MAR): The key difference between MCAR and MAR is that under MAR the data is not missing randomly across all observations, but is missing randomly only within sub-samples of data. For example, if high school GPA data is missing randomly across all schools in a district, that data will be considered MCAR. However, if data is randomly missing for students in specific schools of the district, then the data is MAR.

Not Missing At Random (NMAR): When the missing data has a structure to it, we cannot treat it as missing at random. In the above example, if the data was missing for all students from specific schools, then the data cannot be treated as MAR.

Common Methods:-

1. Mean or Median Imputation

When data is missing at random, we can use list-wise or pair-wise deletion of the missing observations. However, there can be multiple reasons why this may not be the most feasible option:

- There may not be enough observations with non-missing data to produce a reliable analysis
- In predictive analytics, missing data can prevent the predictions for those observations which have missing data
- External factors may require specific observations to be part of the analysis

In such cases, we impute values for missing data. A common technique is to use the mean or median of the non-missing observations. This can be useful in cases where the number of missing observations is low. However, for large number of missing values, using mean or median can result in loss of variation in data and it is better to use imputations. Depending upon the nature of the missing data, we use different techniques to impute data that have been described below.

2. Multivariate Imputation by Chained Equations (MICE)

MICE assumes that the missing data are Missing at Random (MAR). It imputes data on a variable-by-variable basis by specifying an imputation model per variable. MICE uses predictive mean matching (PMM) for continuous variables, logistic regressions for binary variables, bayesian polytomous regressions for factor variables, and proportional odds model for ordered variables to impute missing data.

3. Random Forest

Random forest is a non-parametric imputation method applicable to various variable types that works well with both data missing at random and not missing at random. Random forest uses multiple decision trees to estimate missing values and outputs OOB (out of bag) imputation error estimates.

Q12. A/B Testing

A/B testing is a basic randomized control experiment. It is a way to compare the two versions of a variable to find out which performs better in a controlled environment.

For instance, let's say a company wants to increase the sales of its product. Here, either he can use random experiments, or one can apply scientific and statistical methods. A/B testing is one of the most prominent and widely used statistical tools.

Q13- Ans:

The process of replacing null values in a data collection with the data's mean is known as mean imputation.

Let's have a look at a very simple example to visualize the problem. The following table have 3 variables: Age, Gender and Fitness Score. It shows a Fitness Score results (0–10) performed by people of different age and gender.

	Age	Gender	Fitness_Score
0	20	M	8
1	25	F	7
2	30	M	7
3	35	M	7
4	36	F	6
5	42	F	5
6	49	M	6
7	50	F	4
8	55	M	4
9	60	F	5
10	66	M	4
11	70	F	3
12	75	M	3
13	78	F	2

Table with correct, non-missing data

Now let's assume that some of the data in Fitness Score is actually missing, so that after using a mean imputation we can compare results using both tables.

	Age	Gender	Fitness_Score
0	20	M	NaN
1	25	F	7.0
2	30	M	NaN
3	35	M	7.0
4	36	F	6.0
5	42	F	5.0
6	49	M	6.0
7	50	F	4.0
8	55	M	4.0
9	60	F	5.0
10	66	M	4.0
11	70	F	NaN
12	75	M	3.0
13	78	F	NaN

	Age	Gender	Fitness_Score
0	20	M	5.1
1	25	F	7.0
2	30	M	5.1
3	35	M	7.0
4	36	F	6.0
5	42	F	5.0
6	49	M	6.0
7	50	F	4.0
8	55	M	4.0
9	60	F	5.0
10	66	M	4.0
11	70	F	5.1
12	75	M	3.0
13	78	F	5.1

Imputed values don't really make sense — in fact, they can have a negative effect on accuracy when training our ML model. For example, 78 year old women now has a Fitness Score of 5.1, which is typical for people aged between 42 and 60 years old. Mean imputation doesn't take into account a fact that Fitness Score is correlated to Age and Gender features. It only inserts 5.1, a mean of the Fitness Score, while ignoring potential feature correlations.

Q14. Linear Regression

Regression is the measure of average relationship between two or more variable in terms of the original units of the data

The term 'regression analysis' refers to the methods by which estimates are made of the values of a variable from a knowledge of the values of one or more other variables and to the measurement of the errors involved in this estimation process

The variable which is used to predict the variable of interest is called the independent variable or explanatory variable and the variable we are trying to predict is called the dependent variable or "explained" variable. The independent variable is denoted by X and the dependent variable by Y. The analysis used is called the simple linear regression analysis – simple because there

is only one predictor or independent variable, and linear because of the assumed linear relationship between the dependent and the independent variables.

The term “linear’’ means that an equation of a straight line of the form Y = a+bX where a and b are constants, is used to describe the average relationship when change in the independent variable (say X) by one unit leads to constant absolute change in the dependent variable (Y). When two variables have linear relationship the regression lines can be used to find out the values of dependent variable. When we plot two variables (say X and Y) on a scatter diagram and draw two lines of best fit which pass through the plotted points, these lines are called regression lines.

In linear regression, these lines are straight ones. These regression lines are based on two equations called regression equations which give best estimate of one variable when the other is exactly known or given.

Q15. Branches of Statistics:-

The two branches of statistics are descriptive statistics and inferential statistics. All these branches of statistics follow a specific scientific approach which makes them equally essential.

Descriptive Statistics

Descriptive statistics is considered as the first part of statistical analysis which deals with collection and presentation of data. Scientifically, descriptive statistics can be defined as brief explanatory coefficients that are used by statisticians to summarize a given data set. Generally, a data set can either represent a sample of a population or the entire populations. Descriptive statistics can be categorized into

- Measures of central tendency
- Measures of variability

Measures of Central Tendency

Measures of central tendency specifically help the statisticians to estimate the center of values distribution. These measures of tendency are:

Mean

This is the conventional method used in describing central tendency. Usually, to compute an average of values, you add up all the values and then divide them with the number of values available.

Median

This is the score found at the middle of a set of values. A simple way to calculate a median is to arrange the scores in numerical orders and then locate the score which is at the center of the arranged sample.

Mode

This is the frequently occurring value in a given set of scores.

Measures of Variability

The measure of variability help statisticians to analyze the distribution spread out of a given set of data. Some of the examples of measures of variability include quartiles, range, variance and standard deviation.

Inferential Statistics

Inferential statistics are techniques that enable statisticians to use the gathered information from a sample to make inferences, decisions or predictions about a given population. Inferential statistics often talks in probability terms by using descriptive statistics. These techniques are majorly used by statisticians to analyze data, make estimates and draw conclusions from the limited information which is obtained by sampling and testing how reliable the estimates are.

The different types of calculation of inferential statistics include:

- Regression analysis
- Analysis of variance (ANOVA)
- Analysis of covariance (ANCOVA)
- Statistical significance (t-test)
- Correlation analysis