STATISTICS WORKSHEET 1

Q1] a) True

Q2] a) Central Limit Theorem

Q3] b) Modeling bounded count data

Q4] d) All of the mentioned

Q5] c) Poisson

Q6] b) False

Q7] b) Hypothesis

Q8] a) 0

Q9] c) Outliers cannot conform to the regression relationship

Q10] Normal distribution

A normal distribution is an asymmetric arrangement of most of the values around the mean, which forms curve ooks like a bell. It has two key parameters: the mean and the standard deviation The possible outcomes of the function are given in terms of whole real numbers lying between -∞ to +∞.

Empirical rule:

A) 68% of all observations fall within +/- one standard deviation.
B)About 95% of all observations fall within +/- two standard deviations .
C)Nearly 99.7% of all observations fall within +/- three standard deviations

Bell-shaped Curve: Most of the values lie at the center, and fewer values lie at the tail extremities. This results in a bell-shaped curve.

Mean and Standard Deviation: This data representation is shaped by mean and standard deviation.

Equal Central Tendencies: The mean, median, and mode of this data are equal.

Symmetric: The normal distribution curve is centrally symmetric. Therefore, half of the values are to the left of the center, and the remaining values appear on the right.

Skewness and Kurtosis: Skewness is the the symmetry. The skewness for a normal distribution is zero. Kurtosis studies the tail of the represented data. For a normal distribution, the kurtosis is 3.

Total Area = 1: The total value of the standard deviation, i.e., the complete area of the curve under this probability function, is one. Also, the entire mean is zero.

Q11] There are two methods of handling missing data to avoid error .i.e Imputation and removal data.

The imputation method involves guesses for missing data. It's most useful when the percentage of missing data is low.

If the portion of missing data is too high, other option is to remove data . Removing data may not be the best option if there are not enough observations to result in a reliable analysis. In some situations, observation of specific events or factors may be required.

There are two primary methods for deleting data when dealing with missing data: listwise andpairwise, dropping variables.

a) In this method, all data for an observation that has one or more missing values are deleted. The analysis is run only on observations that have a complete set of data. If the data set is small, Deleting the instances with missing observations can result in biased parameters and estimates and reduce the statistical power of the analysis.b) Pairwise deletion assumes data are missing completely at random (MCAR), but all the cases with data, even those with missing data, are used in the analysis. Pairwise deletion allows data scientists to use more of the data. c)If data is missing for more than 60% of the observations, it may be wise to discard it if the variable is insignificant.

Imputation :

Mean, Median and Mode: This is one of the most common methods of imputing values when dealing with missing data. In cases where there are a small number of missing observations, data scientists can calculate the mean or median of the existing observations. However, when there are many missing variables, mean or median results can result in a loss of variation in the data.

This mean median mode can be impute with pandas techniques like filina() with .mean(), .mode() or directly can be import SimpleImputer from sklearn.impute

KNN Imputer : In this method, imputer choose a distance measure for k neighbors, and the average is used to impute an estimate. We must select the number of nearest neighbors and the distance metric. KNN can identify the most frequent value among the neighbors and the mean among the nearest neighbors. It can be imported from sklearn.impute as KnnImputer

Iterative imputer : In this method ,Immputer is going to fit on all provided variables by treating missing value variable as label and other as feature nad get trained by iteration along available variables. This forms linear equation to predict missing values . it can be imported from sklearn.experimental library.

Q13] Mean Imputation technique

advantages of the method:

1. Missing values in your data do not reduce your sample size, as it would be the case with listwise deletion . Since mean imputation replaces all missing values, you can keep your whole database.
2. Mean imputation is very simple to understand and to apply. You can explain the imputation method easily understandable.
3. If the response mechanism is MCAR, the sample mean of your variable is not biased. Mean substitution might be a valid approach, in case that the univariate average of your variables is the only metric your are interested in.

Drawbacks of mean imputation:

1. Mean substitution leads to bias in multivariate estimates such as correlation or regression coefficients. Values that are imputed by a variable's mean may have a correlation of zero with other variables. Relationships between variables are therefore biased toward zero.
2. Standard errors and variance of imputed variables are biased. For instance, let's assume that we would like to calculate the standard error of a mean estimation of an imputed variable. Since all imputed values are exactly the mean of our variable, we would be too sure about the correctness of our mean estimate. In other words, the confidence interval around the point estimation of our mean would be too narrow.
3. If the response mechanism is MAR or MNAR, even the sample mean of your variable is biased . Assume that you want to estimate the mean of a population's income and people with high income are less likely to respond; Your estimate of the mean income would be biased downwards.

It is not good practice to impute missing valus with mean.I do not recommend to apply .

Q14] Linear Regression

Linear regression is a basic and commonly used type of predictive analysis. The overall idea of regression is to examine two things: (1) does a set of predictor variables do a good job in predicting an outcome (dependent) variable? (2) Which variables in particular are significant predictors of the outcome variable.

These regression estimates are used to explain the relationship between one dependent variable and one or more independent variables.

The simplest form of the regression equation with one dependent and one independent variable is defined by the formula y = c + b*x, where y = estimated dependent variable score, c = constant, b = regression coefficient, and x = score on the independent variable.

Three major uses for regression analysis are (1) determining the strength of predictors, (2) forecasting an effect, and (3) trend forecasting.

Linear Regression is an analysis that assesses whether one or more predictor variables explain the dependent (criterion) variable. The regression has five key assumptions:

Linear relationship

Multivariate normality

No or little multicollinearity.

No auto correlation

Homoscedasticity

Linear regression involves finding best fit of line by iterating on dataset , to minimize the loss function. The loss function is summation difference of actual value of response and predicted value of response. This loss function is key role in finding score R2 in linear regression .

Error for each observation is given distance between actual value and predicted value of label or target variable. E_i = actual value – predicted value

Residual sum of squares are calculated with adding the squares of difference between actual and predicted label for each observation .

Coefficient of determination (R2)is evaluated by 1-(RSS/TSS)

Performance of linear regression is estimated with R2 and mean squared error, mean absolute error, and root mean squared error.

Q15] There are two main branches of statistics

Descriptive statistics : Descriptive statistics is the first part of statistics that deals with the collection of data. The statisticians need to be aware of the design and experiments. They also need to select the correct focus group and keep away from biases. On the contrary, Descriptive statistics are used to do various kinds of analysis on different studies.

Descriptive statistics is divided into two parts:

1) Central tendency measures : (Mean,Median,Mode)
2) Variability measure : (Quartiles,Ranges,Variance,Standard deviation).

Inferential Statistics

Inference statistics are techniques that enable statisticians to use the information collected from the sample to conclude, bring decisions, or predict a defined population.Inference statistics often speak in terms of probability by using descriptive statistics. Besides, a statistician uses these techniques for data analysis, drafting, and making conclusions from limited information. That is obtained by taking samples and testing how reliable they are.

Different types of inferential statustics involves Regression Analysis, Analysis of Variance(ANOVA), Analysis of Covariance(ANCOVA), Statistical signifiacnce(t_test),. correlation Analysis

- **Regression analysis:** It is a set of statistical methods used to estimate relationships between a dependent variable and one or more independent variables. It includes several variations, like linear, multiple linear, and nonlinear. The most well-known models are simple linear and multiple linear.
- **Analysis of variance (ANOVA):** ANOVA is a statistical method that distributes observed variance data into various components. A one-way ANOVA is applied for three or more data groups to gain information about the relationship between the dependent and independent variables.
- **Analysis of covariance (ANCOVA):** It is used to test categorical variables' main and interaction effects on constant dependent variables and keep control for the impact of selected other constant variables. The control variables are known as covariates.
- **Statistical significance (t-test):** It is used to determine a significant difference between the means of two groups related to particular features. A t-test studies the t-statistic, the t-distribution values, and the degree of freedom to learn the statistical significance.
- **Correlation analysis:** It is a statistical method that is used to find the relationship between two variables or datasets and discover how strong the relationship may be.

Q12] A/B Testing

A/B testing (also known as bucket testing or split-run testing) is a user experience research methodology. A/B tests consist of a randomized experiment with two variants, A and B. It includes application of statistical hypothesis testing or "two-sample hypothesis testing" as used in the field of statistics. A/B testing is a way to compare two versions of a single variable, typically by testing a subject's response to variant A against variant B, and determining which of the two variants is more effective. A/B tests are widely considered the simplest form of controlled experiment. However, by adding more variants to the test, its complexity grows.

A/B tests are useful for understanding user engagement and satisfaction of online features like a new feature or product.[7] Large social media sites like LinkedIn, Facebook, and Instagram use A/B testing to make user experiences more successful and as a way to streamline their services.[

A/B tests, also known as split tests, allow you to compare 2 versions of something to learn which is more effective. Simply put, do your users like version A or version B?

The concept is similar to the scientific method. If you want to find out what happens when you change one thing, you have to create a situation where only that one thing changes.

A/B testing is not only cost effective, it's time efficient. You test 2 or 3 elements and get your answer. From there, it's easy to decide whether to implement a change or not. If real-life data doesn't hold up to your test results, it's always possible to revert back to an older version.