Student Name: - Kiran Digambar Yadav.

Assignment Name: - STATISTICS WORKSHEET-1

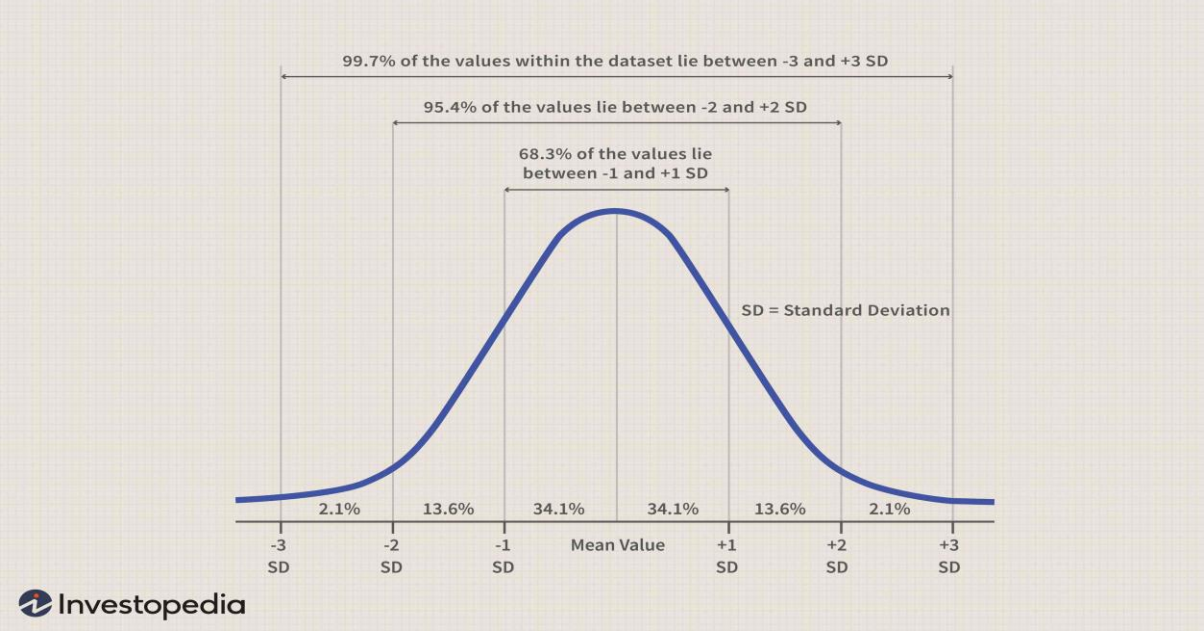
Internship: - 35

STATISTICS WORKSHEET-1

Question No	Answers
1	a) True
2	a) Central Limit Theorem
3	b) Modeling bounded count data.
4	d) All of the mentioned.
5	c) Poisson.
6	b) False.
7	b) Hypothesis.
8	a) 0
9	c) Outliers cannot conform to the regression relationship.

Q10 Answer: -

Normal distribution also known as the gaussian distribution. It 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. The normal distribution is the proper term for a probability bell curve. In the normal distribution the mean is zero, the standard deviation is 1, it has zero skew and a kurtosis of 3. For all normal distribution 68.2% of the observations will appear within the plus or minus one standard deviation of the mean. 95.4% of the observations will fall within plus or minus two standard deviation and 99.7% within plus or minus three standard deviations. The normal distribution is symmetric and has a skewness of zero.



Q11 Answer: -

Missing data can reduce the accuracy of the model. While doing preprocessing, the visualization that we get for a particular feature can be misleading because of the presence of null values. The model created at the end can be biased. These null values can create problems in real life.

How to handle missing values in a dataset: -

I. Removal or deletion of missing value.
II. Impute missing value with (Mean/Median/Mode).
III. Prediction model (Regression and classification)
IV. Use sklearn impute model (SimpleImputer, IterativeImputer, KNNImputer).
V. Imputation using deep learning library.
VI. Use python fiilna function.
VII. Use interpolation function.
VIII. Use python replace function.

Imputation techniques that we recommended.

I. Mean imputation: - Simply calculate the mean of the observed values for that variable for all individuals who are non-missing.
II. Substitution: - Impute the value from a new individual who was not selected to be in the sample.
III. Hot deck imputation: - Find all the sample subjects who are similar on other variables, then randomly choose one of their values on the missing variable.
IV. Cold deck imputation: - A systematically chosen value from an individual who has similar values on other variables. This is like hot deck in most ways, but removes the random variation.
V. Regression imputation: - The predicted value obtained by regressing the missing variable on other variables. So instead of just taking the mean you are taking the predicted value based on the other variables. This preserves relationship among variables involved in the imputation model, but not variability around predicted values.
VI. Stochastic regression imputation: - The predicted value from a regression plus a random residual value. This has all the advantage of regression imputation but adds in the advantages of the random component. Most multiple imputation is based off some form of stochastic regression imputation.

VII. Interpolation and extrapolation: - An estimated value from other observations from the same individual. It usually works in longitudinal data. Extrapolation means you are estimating beyond the actual range of the data and that requires making more assumptions that you should.

VIII. Single or multiple imputation: - Single refers to the fact that you come up with a single estimates of the missing value using one of the seven methods listed above. Multiple imputation can be used in cases where the data are missing completely at random, missing at random and even when the data are missing not at random.

Q12 answer: -

A/B testing is a basic randomized control experiment. It is way to compare the two versions of a variable to find out which performs better in a controlled environment. For instances, let us say you own a company and want to increase the sales of your product. Here either you can use random experiments or you can apply scientific and statistical methods. A/B testing is one of the most prominent and widely used statistical tools. In the above scenario, you may divide the products into two parts i.e. A & B. here A will remain unchanged while you make significant changes in B. now based on the responses from customer groups who used A & B respectively, you try to decide which is performing better. It is a hypothetical testing methodology for making decisions that estimate population parameters based on sample statistics. The population refers to all the customers buying your product, while the sample refers to the number of customers that participated in the test.

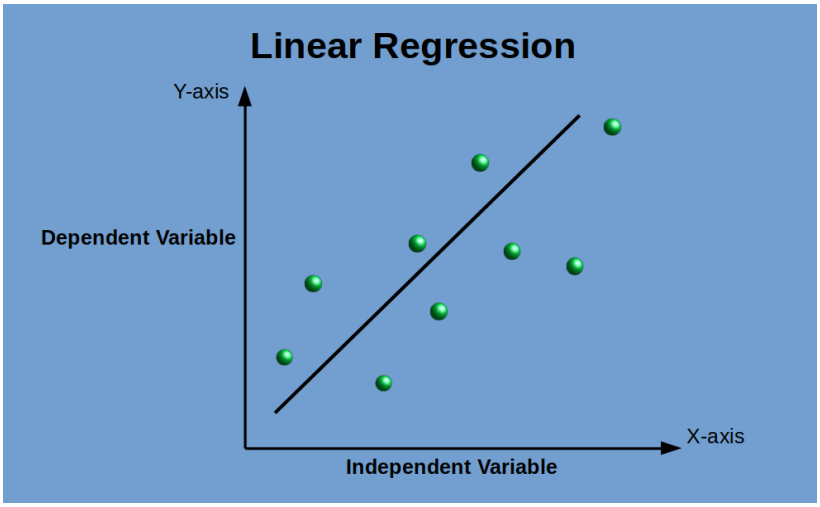
Q13 Answer: -

The process of replacing null values in a data collection with the data's mean is known as mean imputation. Mean imputation is typically considers as bad practice since it ignores feature correlation. Second thing is mean imputation decrease the variance of our data while increasing bias, as a result of the reduced variance the model is less accurate and the confidence level interval is less accurate.

Q14 Answer: -

Linear regression attempts to model the relationship between two variables by fitting a linear equation to observed data. One variable is called as explanatory variable and the other is consider as dependent variable. A scatter plot can be helpful tool in determining the strength of the relationship between two variables. If there appears to be no association between the proposed explanatory and

dependent variables (i.e., the scatter plot does not indicate any increasing or decreasing trends.) a valuable numerical variable measure of association between two variables is the correlation coefficient, which is the value between -1 and 1 indicating the strength of the association of the observed data for the two variables. A linear regression line has an equation of the form Y = a + bX, where X is the explanatory variable and Y is the dependent variable. The slope of the line is b, and a is intercept.



Q15 Answer: -

There are generally two main branches of Statistics, one is Descriptive statistics and another one is inferential statistics. Let us understand each in detail.

I. Descriptive Statistics: - It deals with the 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 represents a sample of a population or the entire populations. Descriptive statistics can be categorized into two, 1) Measure of central tendency and 2) Measure of variability.

II. 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. Different types of calculation of inferential statistics include, 1) Regression analysis, 2) Analysis of variance, 3) Analysis of covariance, 4) Statistical significance or t-test, 5) correlation analysis.