**STATISTICS - WORKSHEET 4**

**(Solutions)**

**ANS 1)** a) True

**ANS 2)** a) Central Limit Theorem

**ANS 3)** b) Modeling bounded count data

**ANS 4)** d) All of the mentioned

**ANS 5)** c) Poisson

**ANS 6)** a) True

**ANS 7)** b) Hypothesis

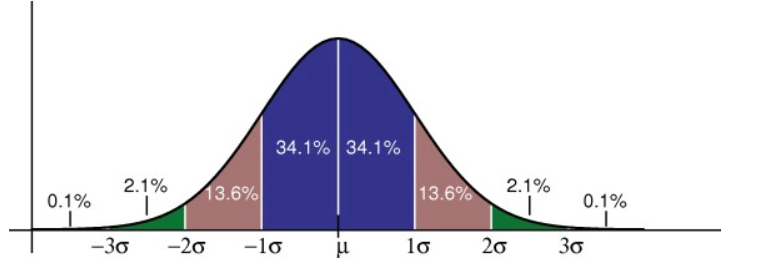
**ANS 8)** a) 0

**ANS 9)** c) Outliers cannot conform to the regression relationship

**ANS 10)** Normal distribution, also known as the Gaussian distribution, 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. In graph form, normal distribution will appear as a bell curve.

Percentage of your data falls within a certain number of [standard deviations](https://www.statisticshowto.com/probability-and-statistics/standard-deviation/) from the [mean](https://www.statisticshowto.com/mean):

* 68% of the data falls within one [standard deviation](https://www.statisticshowto.com/probability-and-statistics/standard-deviation/) of the [mean](https://www.statisticshowto.com/mean).
* 95% of the data falls within two [standard deviations](https://www.statisticshowto.com/probability-and-statistics/standard-deviation/) of the [mean](https://www.statisticshowto.com/mean).
* 99.7% of the data falls within three [standard deviations](https://www.statisticshowto.com/probability-and-statistics/standard-deviation/) of the [mean](https://www.statisticshowto.com/mean).

****

**ANS 11**) Handling missing values is very crucial step for interpretation of better results.

Most of the time the common approach to handle missing value is to delete the null values from the entire dataset ,but that results in loss of data, so a more logical approach is imputation, which simply means replacing nan values,for a continuous series or numerical data,we can replace the values with mean ,like fillna(mean of that column) ,for categorical variables the values our replace with mode .

There are two types of imputer, One type of imputation algorithm is univariate, which imputes values in the i-th feature dimension using only non-missing values in that feature dimension (e.g. impute.SimpleImputer). By contrast, the other multivariate imputation algorithms use the entire set of available feature dimensions to estimate the missing values (e.g. impute.IterativeImputer).

**ANS 12)** 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 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 – A and B. Here A will remain unchanged while you make significant changes in B’s packaging. Now, on the basis of the response from customer groups who used A and 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.

**ANS 13)**

* In this method missing values in the dataset are replaced by the mean values.
* This method is applicable for continuous variables only.
* It is very fast method, but the disadvantage of the mean imputation reduces variance in the dataset.

**ANS 14)**

**Simple linear regression** is a statistical method that allows us to summarize and study relationships between two continuous (quantitative) variables:

* One variable, denoted *x*, is regarded as the **predictor**, **explanatory**, or **independent** variable.
* The other variable, denoted *y*, is regarded as the **response**, **outcome**, or **dependent** variable.

**ANS 15)**

The two main branches of statistics are [descriptive statistics](https://explorable.com/descriptive-statistics) and [inferential statistics](https://explorable.com/inferential-statistics).

**Descriptive Statistics:**

Descriptive statistics is the first part of statistics that deals with the collection of data. People seem it too easy, but it is not that easy. The statisticians need to be aware of the designing and experiments. They also need to choose the right focus group and avoid biases. In contrast, Descriptive statistics are used in use to do various kinds of analysis on different studies.

**Descriptive statistics have two parts**

* Central tendency measures
  + Mean, Median and Mode
* Variability measures

**Inferential Statistics:**

The 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, these techniques are used primarily by a statistician 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 statistics include:**

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