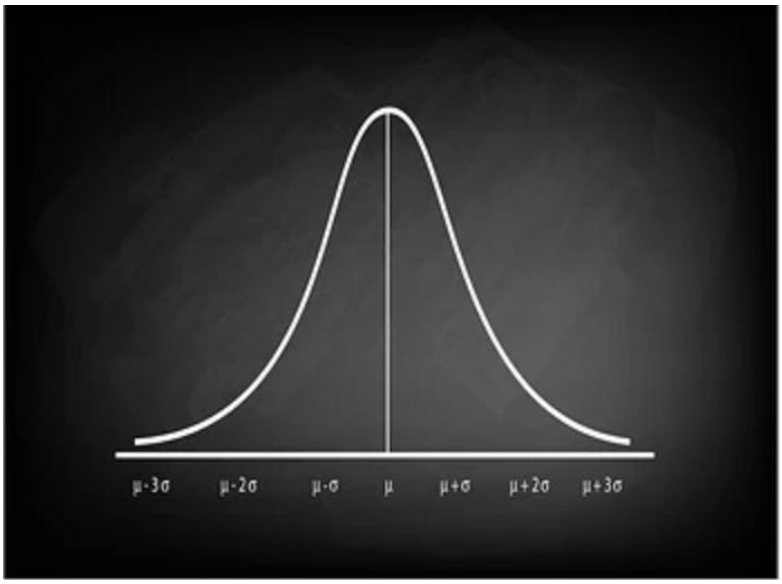
STATISTICS WORKSHEET 1

Q.no	Solution
1.	(A) True
2.	(A) Central limit theorem
3.	(B) Modeling bounded count data
4.	(D) All are correct
5.	(C) Poisson
6.	(B) False
7.	(B) Hypothesis
8.	(A) 0
9.	(C) Outliers cannot conform regression relationship

Q:- What do you understand by the term Normal Distribution?

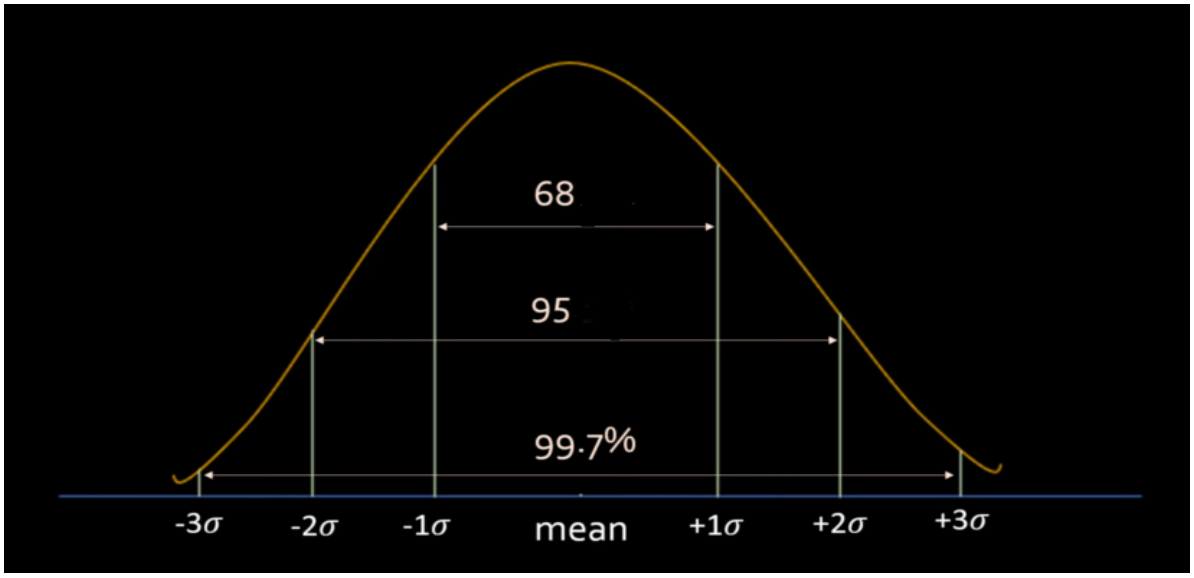
Ans:- It is a **classic bell shaped curve** distribution and is majorly determined by two parameters i.e **Mean and standard deviation**. In this the mean, median and mode are lined up in such a way that the center of the distribution is the mean and half of the result fall to the either side of the mean. This makes the distribution **symmetric.** The mean indicates where the bell is centered and standard deviation depicts wideness. It is also referred to as **Gaussian Distribution**.

A common representation of a normal curve:-



A normal distribution with mean=0 and std=1 is called a standard normal distribution.

In a normal distribution 68% of the data lies within one standard deviation, 95% of data lies within two standard deviation and 99.7% of data lies within three standard deviation which is also called **three sigma rule/Empirical rule**. If any points/data is distant from the given values in the dataset are called **outliers** and are detected through **z-score**.



Q:- How do you handle missing data? What imputation techniques do you recommend?

Ans:- The various methods through which one can handle missing data:

- Drop the variables/data entry
- Replace it with mean value
- Replace it with the mode value.
- Replace it with the median.
- Replace the missing values with a constant using fillna()

Drop the variables/Data entry: In this we can drop either the particular column if major of the values in the column are missing or we can drop the rows with the missing values.

Replace with Mean:- In this we can replace the missing values with the mean value of that particular column.

Replace it with Mode:-This method is majorly used for dealing missing values in categorial data where we will be replacing the missing values with the mode of the variable i.e the most frequently occurred data in the particular variable.

Replace it with Median:- we can replace the missing values with the median of the particular varuable/column.

Replace the missing values with a constant using fillna():- we can also replace the missing values with a constant value by using a method called fillna().

Techniques	Pros	Cons
Drop the variables	Deleting a row/column with no information gives more accurate results	Sometimes in large datasets important information is lost
Replace with mean value	Prevents loss of data.	In large datasets may leads to bias results
Replace with median value	Prevents loss of data.	In large datasets may leads to bias results
Replace with mode value	Prevents loss of data.	In large datasets may leads to bias results
Replace with constant using fillna()	Definitely information is retained	The replaced value sometimes is out of context and may leads to wrong results.

The Imputation technique that I would recommend to replace missing variables in a dataset is **replacing the missing values with mean for numeric/continuous data and when data distribution is symmetric in case if the data is skewed one can use either median or mode imputation and for categorial data I would recommend to replace missing values with the mode value i.e the most frequently occurring value** for a particular variable. As with the help of these methods major of the information is retained and the best outcomes are achieved.

Q:- What is A/B testing?

Ans:-A/B is an experiment with two groups to establish which of the two products, procedures or the practice is superior. In A/B testing, "A" refers to the **control** or the original testing variable whereas "B" refers to the **variation** or the new version of the original testing variable. This A/B testing is typically based on hypothesis testing only.

Some of the examples of A/B testing:-

1. Testing two prices which yields more net profit.
2. Testing two web ads to determine which generates more conversion.
3. Testing two headlines which generates more clicks and increase TRP's.

Let us consider an e-commerce company ABC who wants to generate more conversions for there product through advertisement. They already had one adv. and named as A and created another adv. with some minute changes and named it as B.

Out of population a random sample size of 500 people was created and out of which 250 were for control group(A) who were shown adv. A and 250 people were for variation/Treatment group(B) and the best conversion rates were recorded. This how a A/B testing is performed and best results are recorded.

Q:- What is linear Regression in Statistics?

Ans:- Simple linear regression is used to depict a relation between two continuous variables. It estimates how much one variable will change if the changes are made in the other variable. Considering two variables Y and X the linear equation can be represented as:

$$Y=A+BX+e$$

Where,

Y=Dependent variable/output/label/predicted

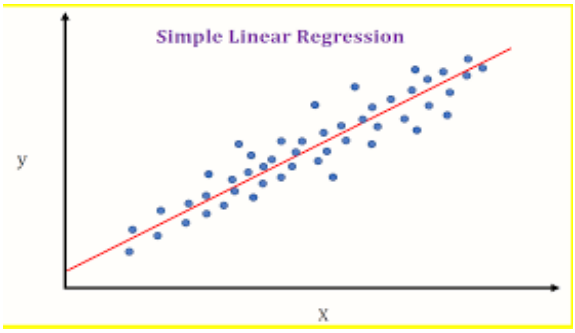
A= Intercept

B=coefficients.

X= independent variable/predictors.

e= residual error.

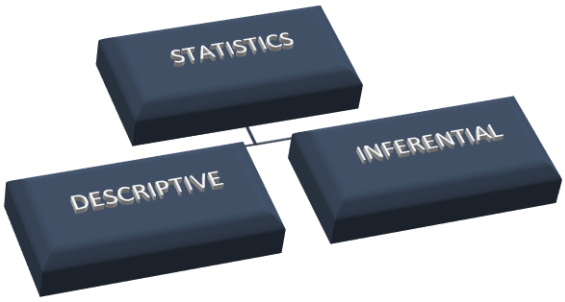
Let us consider an example, X will represent the number of experience of working persons have and Y will represent the salary the different persons are getting as of there experience. As the no. of years of experience will increase the salary will increase this is what will portrayed through Linear Regression. In the below scatter plot representation X-axis represent the number of experience person has and Y-axis will represent salary and the best fit line is drawn.



We usually use linear regression when we want to understand the relationship between the two variables i.e either they are positively/negatively co-related or is there is no relation between them.When we have multiple predictors, i.e multiple input variables the equation can be extended are represented as:-

$$Y=A+B^1X^1+B^2X^2.....B^NX^N+e$$

Q:- What are the various branches of statistics?



Statistics may be defined as the branch of science that deals with the collection, analysis, interpretation of numerical data.

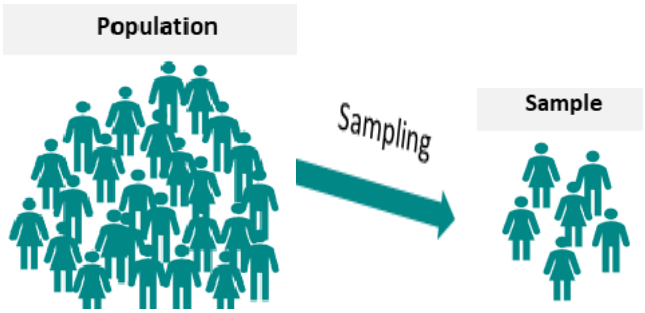
It focuses on four major areas:

- Data collection
- Data presentation
- Data analysis
- Interpretation of Data

Statistics is further divided into two main categories:

Descriptive Statistics:- It describes the important properties of data using **measure of central tendency** like mean/median/mode and **measure of dispersion** like standard deviation, variance, range, skewness, percentile.

Inferential Statistics:- when the data is too large and it becomes difficult to explain the data we will be taking out samples from large population and will be making inferences based on the samples selected. The goal of the inferential statistics is to draw conclusions from the samples and generalize them to population. Methods for the above is hypothesis testing, ANOVA etc.



Q: Is mean imputation of missing data is acceptable practice?

Ans:- Major advantage of mean Imputation of missing variables is is the whole of the dataset is preserved no information is lost and the results obtained is much accurate.

Replacing the missing values with the mean of the particular variable is simple to understand and easy to implement simply calculating the mean and replacing the NaN values with mean of that variable.

Definitely it will also be having some cons and is considered as one of the last resort to get rid of the missing values by major of the data scientists.

As replacing the missing values with mean value will reduce the variance of the variable and it distorts the relationships among different variables.

The standard deviations of the variables with the imputed mean is also affected as it reduces the standard deviation data and will leads to less errors will generate biased results.

Hence, we can conclude that the mean imputation is one of the popular techniques but practitioners usually do not go will this process because of the reduced variance distorted relations are depicted bias results will be generated.