Q.1) Ans:- *A normal distribution is a type of continuous probability distribution in which most data points cluster toward the middle of the range, while the rest taper off symmetrically toward either extreme.*

The middle of the range is also known as the mean of the distribution.

The normal distribution is an important probability distribution in math and statistics because many continuous data in nature and psychology display this bell-shaped curve when compiled and graphed.

Example of Normally Distributed Data. Heights

Height data are normally distributed. The distribution in this example fits real data that I collected from 14-year-old girls during a study. The distribution is symmetric. The number of girls shorter than average equals the number of girls taller than average. In both tails of the distribution, extremely short girls occur as infrequently as extremely tall girls.

Parameters of the Normal Distribution

As with any probability distribution, the parameters for the normal distribution define its shape and probabilities entirely. The normal distribution has two

parameters, the mean and standard deviation. The Gaussian distribution does not have just one form. Instead, the shape changes based on the parameter values .

Q2Ans :- *Missing data can be dealt with in a variety of ways. I believe the most common reaction is to ignore it. Choosing to make no decision, on the other hand, indicates that your statistical program will make the decision for you.*

Your application will remove things in a list wise sequence most of the time. Depending on why and how much data is gone, list wise deletion may or may not be a good idea.

Another common strategy among those who pay attention is imputation. Who pay attention is imputation. Imputation is the process of substitution an estimate for missing values and analyzing the entire data set as if the imputed values were the true observed values.

Following are some of the most prevalent methods

Mean imputation : Calculate the mean of the observed values for that variable for all non-missing people. It has the advantage of maintaining the same mean and

sample size, but it also has a slew of drawbacks. Almost all of the methods described below are superior to mean imputation.

Substitution : Assume the value from a new person who was not included in the sample. To put it another way, pick a new subject and employ their worth instead.

Regression Imputation : The result of regressing the missing variable on other factors to get a predicted value. As a result, instead of utilizing the mean, you're relying on the anticipated value, which is influenced by other factors. This keeps the associations between the variables in the imputation model, but not the variability around the anticipated values.

Q3Ans :- *A/B testing, also known as split testing, refers to a randomized experimentation process wherein two or more versions of a variable(web page, page element, etc.) are shown to different segments of website visitors at the same time to determine which version leaves at the same time to determine which version leaves the maximum impact and drives business metrics.*

In a typical A/B test, traffic is randomly assigned to each page variant based upon a predetermined

weighting. For example , if you are running a test with two page variants, you might split the traffic 50/50 or 60/40.

A/B testing is essentially an experiment where two or more variants of a page are shown to users at random, and statistical analysis is used to determine which variation performs better for a given conversion goal.

***Q4.Ans** :- The process of replacing null values in a data collection with the data's mean is knows as mean imputation.*

Mean imputation is typically considered terrible practice since it ignores feature correlation. Consider the following scenario: we have a table with age and fitness scores, and an eight-year-old has missing fitness score. If we average the fitness scores of people between the ages of 15 and 80, the eighty-year-old will appear to have a significantly greater fitness level than he actually does.

Second, mean imputation decreases the variance of our data while increasing bias. As a result of the reduced variance, the model is less accurate and the confidence interval in narrower.

We all know the pain when the dataset we want to use for Machine Learning contains missing data. The quick and easy workaround is to substitute a mean for numerical features and use a mode for categorical ones. Even better, someone might just insert 0's or discard the data and proceed to the training of the model. In the following article, I will explain why using a mean or mode can significantly reduce the model's accuracy and bias the results. I will also points you to few alternative imputation algorithms which have their respective Python libraries that you can use out of the box.

Q5.Ans:- *Linear regression analysis is used to predict the value of a variable based on the value of another variable. The variable you want to predict is called the dependent variable. The variable you are using to predict the other variable's value is called the independent variable.*

We could use the equation to predict weight if we knew an individual's height. In this example, if an individual was 70 inches tall, we would predict his weight to be: weight = 80 + 2X(70) = 220 lbs. In this simple linear regression , we are examining the impact of one independent variable on the outcome.

When to use regression

We are often interested in understanding the relationship among several variables. Scatterplots and scatterplot matrices can be used to explore potential relationships between parts of variables. Correlation provides a measure of the linear association between pairs of variables, but it doesn't tell us about more complex relationships. For example, if the relationship is curvilinear, the correlation might be near zero. You can use regression to develop a more formal understanding of relationships between variables. In regression , and in statistical modeling in general, we want to model the relationship between an output variable, or a response, and one or more input variables, or factors. Depending on the context, output variables might also be referred to as dependent variables, outcome, or simply Y variables , and input variables might be referred to as explanatory variables, effects, predictors or X variables.

Q6.Ans:- *There are three real branches of statistics :*

Data collection

Descriptive statistics

Inferential statistics.

The two main branches of statistics are ***Descriptive Statistics and Inferential Statistics.*** *Both of these are employed in scientific analysis of data and both are equally important for student of statistics .*

Descriptive Statistics :- *Descriptive statistics deals with the presentation and collection of data. This is usually the first part of statistical analysis. It is usually not as simple, as it sounds, and the statistician needs to be aware of designing experiments, choosing the right focus group and avoid biases that are so easy to creep into the experiment*

Inferential Statistics :- *Inferential statistics, as the name suggests, involve drawing the right conclusions from the statistical analysis that has been performed using descriptive statistics. In the end, it is the inferences that make studies important and this aspect is dealt with in inferential statistics.*

Both descriptive and inferential statistics go hand in hand and one cannot exist without the other. Good scientific methodology needs to be followed in both these steps of statistical analysis and both these branches of statistics are equally important for a researcher.