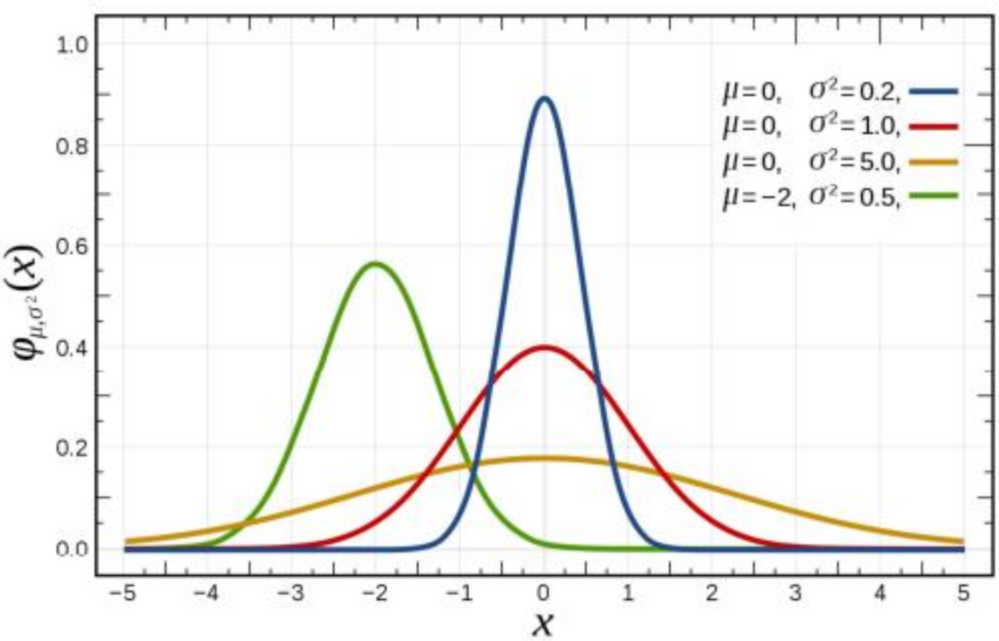
Question Number	**Answers**
1	A
2	A
3	B
4	D
5	C
6	B
7	B
8	A
9	C

10. What do you understand by the term Normal Distribution?

Data is usually distributed in different ways with a bias to the left or to the right or it can all be jumbled up. However, there are chances that data is distributed around a central value without any bias to the left or right and reaches normal distribution in the form of a bell-shaped curve.



The random variables are distributed in the form of a symmetrical, bell-shaped curve. Properties of Normal Distribution are as follows:

1. Unimodal (Only one mode)

2. Symmetrical (left and right halves are mirror images)

3. Bell-shaped (maximum height (mode) at the mean)

4. Mean, Mode, and Median are all located in the center

5. Asymptotic

11. How do you handle missing data? What imputation techniques do you recommend?

One of the easiest ways to handle missing or corrupted data is to drop those rows or columns or replace them entirely with some other value.

There are two useful methods in Pandas:

● IsNull() and dropna() will help to find the columns/rows with missing data and drop them

● Fillna() will replace the wrong values with a placeholder value.

Imputation technique for handling missing data during prediction time

- Validate input data before feeding into ML model; Discard data instances with missing values
- Predicted value imputation
- Distribution-based imputation
- Unique value imputation
- Reduced feature models
-

Input Data Validation- Discard Data Instance with Missing Data

Most trivial of all the missing data imputation techniques is discarding the data instances which do not have values present for all the features. In other words, before sending the data to the model, the consumer/caller program validates if data for all the features are present. If the data for all of the features are not present, the caller program do not invoke the model at all and takes on some value or show exceptions. For beginners, this could be a technique to start with. If this technique is used during training model training/testing phase, it could result in model bias.

Predicated Value Imputation

In this technique, one of the following methods is followed to impute missing data and invoke the model appropriately to get the predictions:

- Impute with mean, median or mode value: In place of missing value, mean, median or mode value is taken appropriately for continuous and categorical data respectively. Recall that the mean, median and mode are the central tendency measures of any given data set. The goal is to find out which is a better measure of central tendency of data and use that value for replacing missing values appropriately.
    - When to use mean: If the data is symmetrically distributed, one can make use of mean for replacing missing value. One can use box plot or distribution plot to find out about the data distribution.
    - When to use median: If the data is skewed or if the data consists of outliers, one may want to use median.
    - When to use mode: If the data is skewed, one may want to use mode.

- Impute with predicted value: Another technique is understanding/learning the relationship between missing data and other features value in other test instances where data were found for feature representing missing data, and appropriately predict the missing data based on the value of other features for the instances where data is found to be missing. One could, however, argue that if a feature value can be estimated using other feature values, isn't it the case of correlates and thus the feature could be imputed. One needs to watch out for feature imputability scenarios.

Distribution Based Imputation

In this technique, for the estimated distribution over the values of features for which data is missing, one may estimate the expected distribution of the target variable. The final predication will be mean or mode value of all predication. This strategy is common for applying classification trees in AI research and practice. This technique is fundamentally different from predicated value imputation because it combines the classification across the distribution of feature's possible value rather than merely making the classification based on its most likely value.

Unique Value Imputation

In this technique, a unique value is imputed in place of missing values. This technique is recommended when it can be determined if the data is generally found to be missing for a particular label/class value and, this dependence is found during model training/testing phase. One of the techniques used for imputing missing data with unique value is randomly selecting the similar records. This is also termed as hot deck cold deck imputation technique. The random selection for missing data imputation could be instances such as selection of last observation (also termed Last observation carried forward – LOCF**).**

Reduced Feature Models

In this technique, different models are built with the different set of features with the idea that appropriate models with only those set of features are used for making predictions for which the data is available. This is against applying imputation to missing data using one of the above techniques. For example, let's say that a model is built with feature A, B, AB, C, D. As part of analysis it is found out that most of the time, data related to feature D would be missing. Thus, using the reduced feature modelling technique, another model using features A, B, AB, and C is built. In production, both the models get deployed and in case the data is found to be missing data for feature D, the model trained with features A, B, AB and C is used or else, the model with all features including feature D is used.

It has been experimentally found that reduced feature modelling is a superior technique from performance perspective out of all the other techniques mentioned above. However, reduced feature modelling is an expensive one at the same time from different perspectives such as resource intensive, maintenance etc.

12. What is A/B testing?

It is a hypothesis testing for a randomized experiment with two variables A and B. The goal of A/B Testing is to identify any changes to the web page to maximize or increase the outcome of interest. A/B testing is a fantastic method for figuring out the best online promotional and marketing strategies for your business. It can be used to test everything from website copy to sales emails to search ads. An example of this could be identifying the click-through rate for a banner ad.

13. Is mean imputation of missing data acceptable practice?

Mean imputation is the practice of replacing null values in a data set with the mean of the data.

Mean imputation is generally bad practice because it doesn't take into account feature correlation. For example, imagine we have a table showing age and fitness score and imagine that an eighty-year-old has a missing fitness score. If we took the average fitness score from an age range of 15 to 80, then the eighty-year-old will appear to have a much higher fitness score that he actually should. Second, mean imputation reduces the variance of the data and increases bias in our data. This leads to a less accurate model and a narrower confidence interval due to a smaller variance.

14. What is linear regression in statistics?

In statistics, linear regression is a linear approach for modelling the relationship between a scalar response and one or more explanatory variables (also known as dependent and independent variables). The case of one explanatory variable is called simple linear regression; for more than one, the process is called multiple linear regressions.

Assumptions in linear regression:

1. There is a linear relationship between the dependent variables and the regressors, meaning the model you are creating actually fits the data
2. The errors or residuals of the data are normally distributed and independent from each other
3. There is minimal multicollinearity between explanatory variables
4. Homoscedasticity. This means the variance around the regression line is the same for all values of the predictor variable.

Algebraic Method

Algebraic method develops two regression equations of X on Y, and Yon X.

Regression equation of Y on X

$Y=a + bX$

Where −

- Y = Dependent variable
- X = Independent variable
- a = Constant showing Y-intercept
- b = Constant showing slope of line

Values of a and b is obtained by the following normal equations:

$\sum Y= Na + b\sum X$

$\sum XY = a\sum X + b\sum X^2$

Where −

- N= Number of observations

Regression equation of X on Y

$X = a + bY$

Where −

- X = Dependent variable
- Y = Independent variable
- a = Constant showing Y-intercept
- b = Constant showing slope of line

Values of a and b is obtained by the following normal equations:

$\sum X = Na + b\sum Y$

$\sum XY = a\sum Y + b\sum Y^2$

Where −

- N = Number of observations

15. What are the various branches of statistics?

Statistics have majorly categorised into two types:

1. Descriptive statistics
2. Inferential statistics

Descriptive Statistics

In this type of statistics, the data is summarised through the given observations. The summarisation is one from a sample of population using parameters such as the mean or standard deviation.

Descriptive statistics is a way to organise, represent and describe a collection of data using tables, graphs, and summary measures. For example, the collection of people in a city using the internet or using Television.

Descriptive statistics are also categorised into four different categories:

- Measure of frequency
- Measure of dispersion
- Measure of central tendency
- Measure of position

The frequency measurement displays the number of times a particular data occurs. Range, Variance, Standard Deviation are measures of dispersion. It identifies the spread of data. Central tendencies are the mean, median and mode of the data. And the measure of position describes the percentile and quartile ranks.

Inferential Statistics

This type of statistics is used to interpret the meaning of Descriptive statistics. That means once the data has been collected, analysed and summarised then we use these stats to describe the meaning of the collected data. Or we can say, it is used to draw conclusions from the data that depends on random variations such as observational errors, sampling variation, etc.

Inferential Statistics is a method that allows us to use information collected from a sample to make decisions, predictions or inferences from a population. It grants us permission to give statements that goes beyond the available data or information. For example, deriving estimates from hypothetical research.