Q.1. What do you understand by the term Normal Distribution?

Ans. 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.

Normal Distribution Formula:

The probability density function of normal or gaussian distribution is given by;

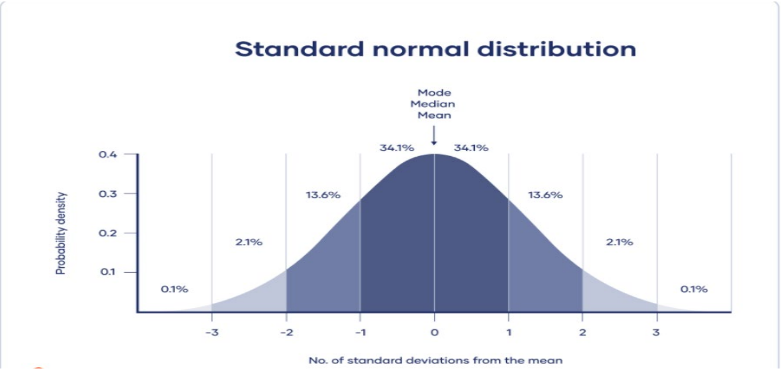
$$f(x, \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

Where,

- x is the variable
- μ is the mean
- σ is the standard deviation

Normal distributions have key characteristics that are easy to spot in graphs:

- The Mean, Median and Mode are exactly the same.
- The distribution is symmetric about the mean—half the values fall below the mean and half above the mean.
- The distribution can be described by two values: the mean and the standard deviation.
- The mean is the location parameter while the standard deviation is the scale parameter.



Q.2. How do you handle missing data? What imputation techniques do you recommend?

Ans. Missing data is a huge problem for data analysis because it distorts findings. The best option available to data scientists is to work with powerful, processing tools that can make the data capturing and analysis process significantly easier. It is the best way to handle missing data.

- Data scientists can use data imputation techniques:

Data scientists use two data imputation techniques to handle missing data : Average imputation and common-point imputation.

Average imputation uses the average value of the responses from other data entries to fill out missing values. However, a word of caution when using this method – it can artificially reduce the variability of the dataset.

Common-point imputation, on the other hand, is when the data scientists utilise the middle point or the most commonly chosen value.

For example, on a five-point scale, the substitute value will be 3. Something to keep in mind when utilising this method is the three types of middle values: mean, median and mode, which is valid for numerical data (it should be noted that for non-numerical data only the median and mean are relevant).

Q.3. What is A/B testing?

Ans. A/B testing (also known as split testing or bucket testing) is a method of comparing two versions of a webpage or app against each other to determine which one performs better.

Essentially, A/B testing eliminates all the guesswork out of website optimization and enables experience optimizers to make data-backed decisions. In A/B testing, A refers to ‘control’ or the original testing variable. Whereas B refers to ‘variation’ or a new version of the original testing variable.

The version that moves your business metric(s) in the positive direction is known as the ‘winner.’ Implementing the changes of this winning variation on your tested page(s) / element(s) can help optimize your website and increase business ROI.

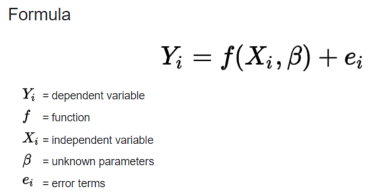
Q.4. Is mean imputation of missing data acceptable practice?
Ans. No. The process of replacing null values in a data collection with the data’s mean is known as mean imputation.

Outliers data points will have a significant impact on the mean and hence, in such cases, it is not recommended to use the mean for replacing the missing values. Using mean values for replacing missing values may not create a great model.

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 a 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 not.

Q.5. What is linear regression in statistics?
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

Linear regression is a kind of statistical analysis that attempts to show a relationship between two variables. Linear regression looks at various data points and plots a trend line.



Formula

$$Y_i = f(X_i, \beta) + e_i$$

Y_i = dependent variable
f = function
X_i = independent variable
β = unknown parameters
e_i = error terms

Types of Linear Regression

Simple linear regression
Multiple linear regression
Logistic regression
Ordinal regression
Multinomial regression
Discriminant analysis

Q.6. What are the various branches of statistics?
Ans.

Descriptive Statistics

Descriptive statistics is the first part of statistics that deals with the collection of data. People think it is too easy, but it is not that easy. The statisticians need to be aware of the design and experiments. They also need to select the correct focus group and keep away from biases. On the contrary, Descriptive statistics are used to do various kinds of analysis on different studies.

Example of Descriptive Statistics: The average score of the college students in the math test. The average age of the people who voted for the winning candidate in the last election. The average length of the statistics book.

Inferential Statistics

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, a statistician uses these techniques for data analysis, drafting, and making conclusions from limited information. That is obtained by taking samples and testing how reliable they are.

Example of Inferential Statistics:
Suppose you want to get an idea about the percentage of the people who love shopping at FILA. We take the sample of the population and find the proportion of individuals who love the FILA brand. With the assistance of probability, this sample proportion allows us to make a few assumptions about the population proportion. This study belongs to inferential statistics.