STATISTICS WORKSHEET-1

Q1 to Q9 have only one correct answer. Choose the correct option to answer your question.

1. Bernoulli random variables take (only) the values 1 and 0.

Ans(a) True

2. Which of the following theorem states that the distribution of averages of iid variables, properly normalized, becomes that of a standard normal as the sample size increases?

Ans(a) Central Limit Theorem

3. Which of the following is incorrect with respect to use of Poisson distribution?

a) Modeling event/time data
b) Modeling bounded count data
c) Modeling contingency tables
d) All of the mentioned

Ans(B)

4. Point out the correct statement.

a) The exponent of a normally distributed random variables follows what is called the log- normal distribution
b) Sums of normally distributed random variables are again normally distributed even if the variables are dependent
c) The square of a standard normal random variable follows what is called chi-squared distribution
d) All of the mentioned

Ans(D)

5. ______ random variables are used to model rates.

Ans(c) Poisson

6. 10. Usually replacing the standard error by its estimated value does change the CLT.

Ans(b) False

7. 1. Which of the following testing is concerned with making decisions using data?

Ans(b) Hypothesis

8. 4. Normalized data are centered at______and have units equal to standard deviations of the original data.

Ans(a) 0

9. Which of the following statement is incorrect with respect to outliers?

a) Outliers can have varying degrees of influence
b) Outliers can be the result of spurious or real processes
c) Outliers cannot conform to the regression relationship
d) None of the mentioned

Ans(c)

Q10and Q15 are subjective answer type questions, Answer them in your own words briefly.

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

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 also known as a *Gaussian distribution* or *probability bell curve*. It is symmetric about the mean and indicates that values near the mean occur more frequently than the values that are farther away from the mean.

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

Ans Analyze each column with missing values carefully to understand the reasons behind the missing of those values, as this information is crucial to choose the strategy for handling the missing values.

There are 2 primary ways of handling missing values:

1. Deleting the Missing values
2. Imputing the Missing Values

Data imputation is a method for retaining the majority of the dataset's data and information by substituting missing data with a different value. These methods are employed because it would be impractical to remove data from a dataset each time. Additionally, doing so would substantially reduce the dataset's size, raising questions about bias and impairing analysis.

These are some of the data imputation techniques-

- Next or Previous Value
- K Nearest Neighbors
- Maximum or Minimum Value
- Missing Value Prediction
- Most Frequent Value
- Average or Linear Interpolation
- (Rounded) Mean or Moving Average or Median Value
- Fixed Value

1. Next or Previous Value

For time-series data or ordered data, there are specific imputation techniques. These techniques take into consideration the dataset's sorted structure, wherein nearby values are likely more comparable than far-off ones. The next or previous value inside the time series is typically substituted for the missing value as part of a common method for imputed incomplete data in the time series. This strategy is effective for both nominal and numerical values.

2. K Nearest Neighbors

The objective is to find the k nearest examples in the data where the value in the relevant feature is not absent and then substitute the value of the feature that occurs most frequently in the group.

3. Maximum or Minimum Value

You can use the minimum or maximum of the range as the replacement cost for missing values if you are aware that the data must fit within a specific range [minimum, maximum] and if you are aware from the process of data collection that the measurement instrument stops recording and the message saturates further than one of such boundaries. For instance, if a price cap has been reached in a financial exchange and the exchange procedure has indeed been halted, the missing price can be substituted with the exchange boundary's minimum value.

4. Missing Value Prediction

Using a machine learning model to determine the final imputation value for characteristic x based on other features is another popular method for single imputation. The model is trained using the values in the remaining columns, and the rows in feature x without missing values are utilized as the training set.

Depending on the type of feature, we can employ any regression or classification model in this situation. In resistance training, the algorithm is used to forecast the most likely value of each missing value in all samples.

A basic imputation approach, such as the mean value, is used to temporarily impute all missing values when there is missing data in more than a feature field. Then, one column's values are restored to missing. After training, the model is used to complete the missing variables. In this manner, an is trained for every feature that has a missing value up until a model can impute all of the missing values.

5. Most Frequent Value

The most frequent value in the column is used to replace the missing values in another popular technique that is effective for both nominal and numerical features.

6. Average or Linear Interpolation

The average or linear interpolation, which calculates between the previous and next accessible value and substitutes the missing value, is similar to the previous/next value imputation but only applicable to numerical data. Of course, as with other operations on ordered data, it is crucial to accurately sort the data in advance, for example, in the case of time series data, according to a timestamp.

7. (Rounded) Mean or Moving Average or Median Value

Median, Mean, or rounded mean are further popular imputation techniques for numerical features. The technique, in this instance, replaces the null values with mean, rounded mean, or median values determined for that feature across the whole dataset. It is advised to utilize the median rather than the mean when your dataset has a significant number of outliers.

8. Fixed Value

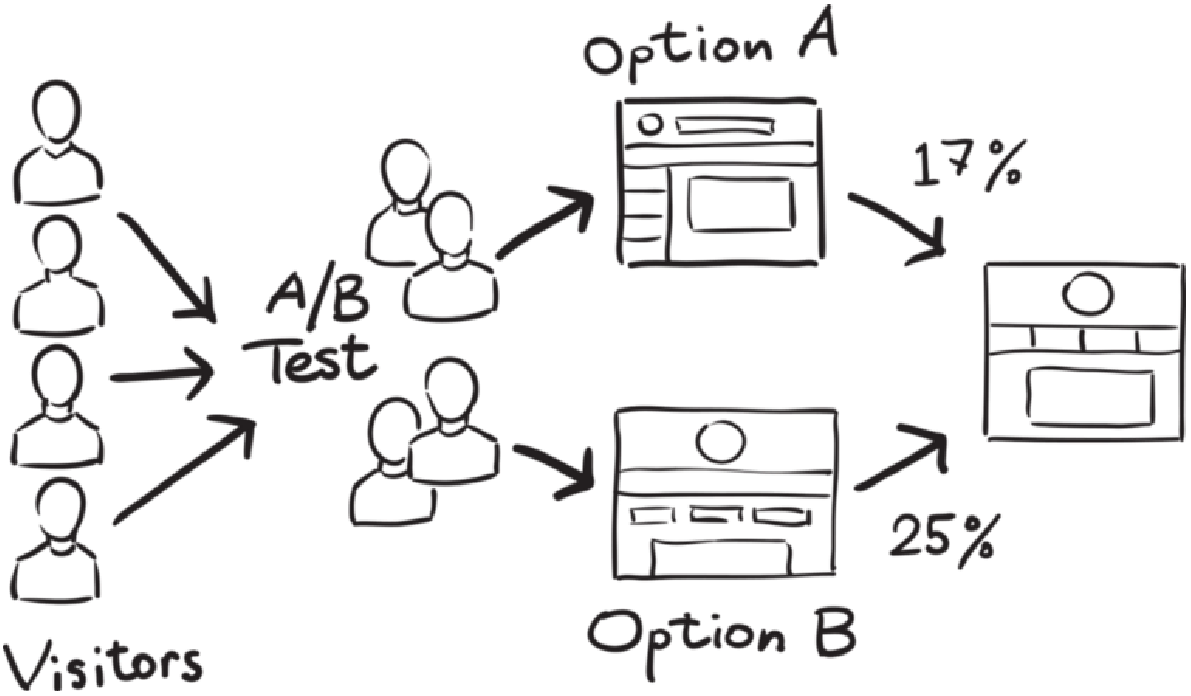
Fixed value imputation is a universal technique that replaces the null data with a fixed value and is applicable to all data types. You can impute the null values in a survey using "not answered" as an example of using fixed imputation on nominal features.

12. What is A/B testing?

Ans 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.

13. Is mean imputation of missing data acceptable practice?

Ans The process of replacing null values in a data collection with the data's mean is known 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 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.

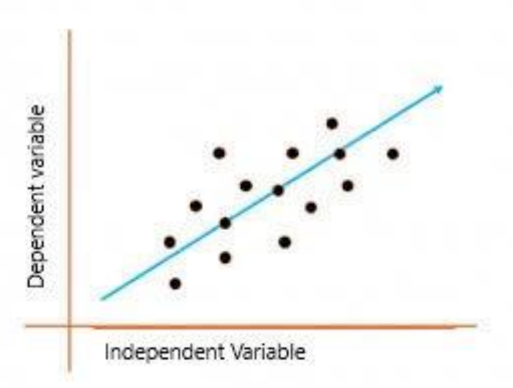
Second, mean imputation decreases the variance of our data while increasing bias. As a result of the reduced variance, the model is less accurate and the confidence interval is narrower.

14. 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.

This form of analysis estimates the coefficients of the linear equation, involving one or more independent variables that best predict the value of the dependent variable. Linear regression fits a straight line or surface that minimizes the discrepancies between predicted and actual output values. There are simple linear regression calculators that use a "least squares" method to discover the best-fit line for a set of paired data. You then estimate the value of X (dependent variable) from Y (independent variable).

The simplest form of the regression equation with one dependent and one independent variable is defined by the formula y = c + b*x, where y = estimated dependent variable score, c = constant, b = regression coefficient, and x = score on the independent variable.



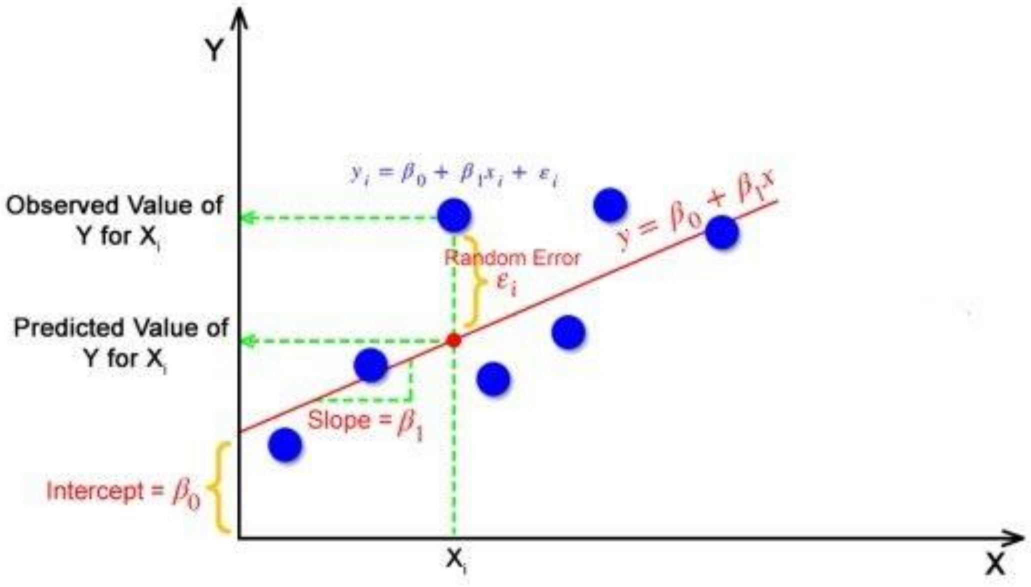
The above graph presents the linear relationship between the output(y) variable and predictor(X) variables. The blue line is referred to as the *best fit* straight line. Based on the given data points, we attempt to plot a line that fits the points the best.

*T*o calculate best-fit line linear regression uses a traditional slope-intercept form which is given below,

$Y_i = \beta_0 + \beta_1 X_i$

where Y_i = Dependent variable, β_0 = constant/Intercept, β_1 = Slope/Intercept, X_i = Independent variable.

This algorithm explains the linear relationship between the dependent(output) variable y and the independent(predictor) variable X using a straight line $Y= B_0 + B_1 X$.


Y
$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$
Observed Value of Y for X_i
Random Error ε_i
$y = \beta_0 + \beta_1 x$
Predicted Value of Y for X_i
Slope = β_1
Intercept = β_0
X_i
X

But how the linear regression finds out which is the best fit line?

The goal of the linear regression algorithm is to get the **best values for B_0 and B_1** to find the best fit line. The best fit line is a line that has the least error which means the error between predicted values and actual values should be minimum.

15. What are the various branches of statistics?

Ans The two main branches of statistics are descriptive statistics and inferential statistics.

(1) Descriptive Statistics

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.

(Or)

Descriptive statistics deals with the collection of data, its presentation in various forms, such as tables, graphs and diagrams and finding averages and other measures which would describe the data.

For example: Industrial statistics, population statistics, trade statistics, etc. Businessmen make use of descriptive statistics in presenting their annual reports, final accounts, and bank statements.

(2) 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

For example: Suppose we want to have an idea about the percentage of the illiterate population of our country. We take a sample from the population and find the proportion of illiterate individuals in the sample. With the help of probability, this sample proportion enables us to make some inferences about the population proportion. This study belongs to inferential statistics.