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.
a) True
b) False

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?
a) Central Limit Theorem
b) Central Mean Theorem
c) Centroid Limit Theorem
d) All of the mentioned

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) Modeling bounded count data

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) All of the mentioned

5. _____ random variables are used to model rates.
a) Empirical
b) Binomial
c) Poisson
d) All of the mentioned

Ans: c) Poisson

6. Usually replacing the standard error by its estimated value does change the CLT.
a) True
b) False

Ans: b) False

7. Which of the following testing is concerned with making decisions using data?
a) Probability
b) Hypothesis
c) Causal
d) None of the mentioned

Ans: b) Hypothesis

8. Normalized data are centered at_____ and have units equal to standard deviations of the original data.
a) 0
b) 5
c) 1
d) 10

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) Outliers cannot conform to the regression relationship

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 the proper term for a probability bell curve. In a normal distribution, the mean is zero and standard deviation is 1.

The normal distribution is also called the Gaussian Distribution. The normal distribution is the most commonly seen continuous distribution in nature. Just as the binomial dustribution, every event is independent from one another. In the normal distribution, the mean, median, and mode all line up such that the center of the distribution is the mean. Because of this, exactly half of the results fall to either side of the mean. The normal distribution is identifiable by its bell shape and may sometimes be referred to as a bell curve.

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

Ans: There are 2 primary ways of handling missing values:

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

1. Deleting the Missing values

Generally, this approach is not recommended. It is one of the quick and dirty techniques one can use to deal with missing values.

If the missing value is of the type Missing Not At Random (MNAR), then it should not be deleted.

If the missing value is of type Missing At Random (MAR) or Missing Completely At Random (MCAR) then it can be deleted.

The disadvantage of this method is one might end up deleting some useful data from the dataset.

There are 2 ways one can delete the missing values:

Deleting the entire row

If a row has many missing values then you can choose to drop the entire row.

If every row has some (column) value missing then you might end up deleting the whole data.

Deleting the entire column

If a certain column has many missing values then you can choose to drop the entire column.

2. **Imputing the Missing Values**

Imputation is a technique used for replacing the missing data with some substitute value to retain most of the data/information of the dataset.

There are different ways of replacing the missing values.

Replacing With Mean

This is the most common method of imputing missing values of numeric columns. If there are outliers then the mean will not be appropriate. In such cases, outliers need to be treated first.

Replacing With Mode

Mode is the most frequently occurring value. It is used in the case of categorical features.

Replacing With Median

Median is the middlemost value. It's better to use the median value for imputation in the case of outliers.

Replacing with previous value – Forward fill

In some cases, imputing the values with the previous value instead of mean, mode or median is more appropriate. This is called forward fill. It is mostly used in time series data.

Replacing with next value – Backward fill

In backward fill, the missing value is imputed using the next value.

Interpolation

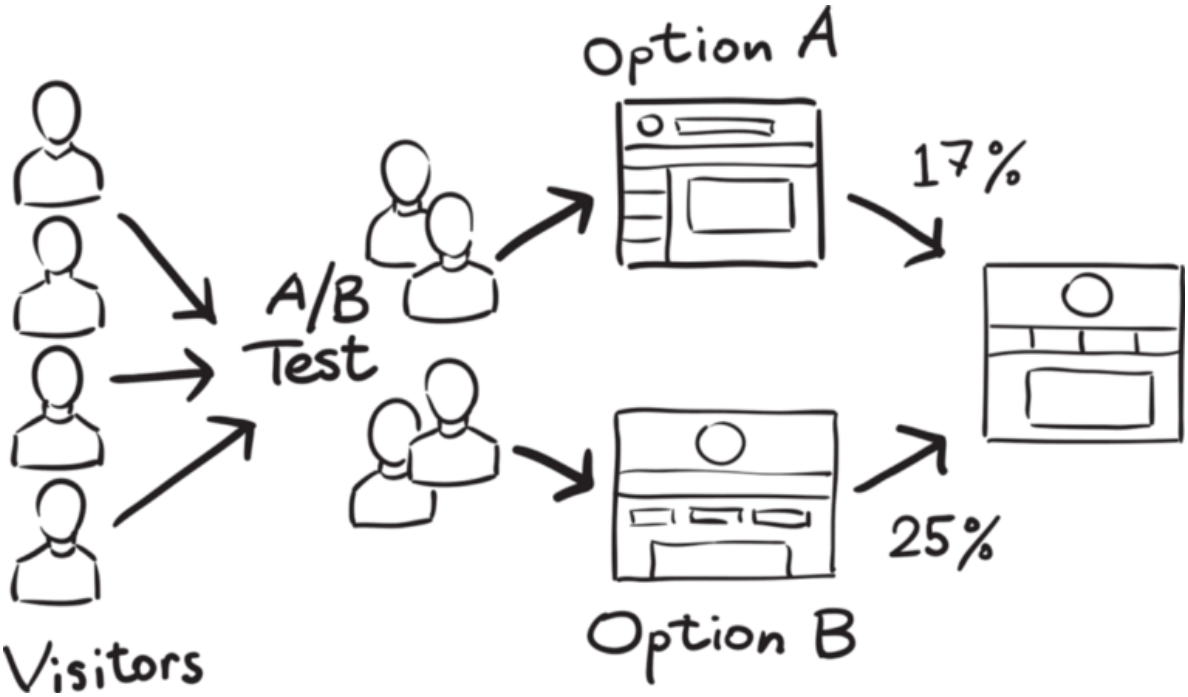
Missing values can also be imputed using interpolation. Pandas interpolate method can be used to replace the missing values with different interpolation methods like 'polynomial', 'linear', 'quadratic'. Default method is 'linear'.

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?

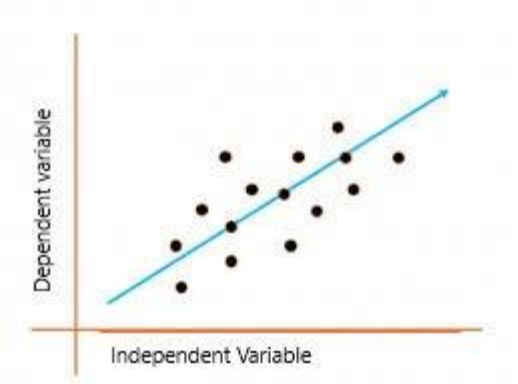
Ans: The process of replacing null values in a data collection with the data's mean is known as mean imputation.

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 is a quiet and the simplest statistical regression method used for predictive analysis in machine learning. Linear regression shows the linear relationship between the independent (predictor) variable i.e. X-axis and the dependent (output) variable i.e. Y-axis, called linear regression. If there is a single input variable X (dependent variable), such linear regression is called *simple linear regression*.



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,

$$\mathbf{Y_i = \beta_0 + \beta_1 X_i}$$

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 $Y= B_0 + B_1 X$.

15. What are the various branches of statistics?

Ans: The various branches of statistics are:

1. **Descriptive Statistics -**Descriptive statistics is a concept that allows us to analyze and summarize data and organize the same in the form of numbers graph, bar plots, histogram, pie chart, etc. Descriptive statistics is simply a process to describe our existing data. It transforms the raw observations into some meaningful data that can be further interpreted and used. Concepts like standard deviation, central tendency are widely used around the world when it comes to learning descriptive statistics.

2. **Inferential Statistics –** Inferential statistics on the other hand is an important concept that deals with drawing conclusions based on small samples collected from the entire

population. For example, during an election poll, people will often want to predict the exit poll results so they will conduct a survey in various parts of state or country and record their opinion. Based on the information they have collected they tend to draw conclusions and make inferences to predict results for the entire population.