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 - Correct
b) False

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 - Correct
b) Central Mean Theorem
c) Centroid Limit Theorem
d) All of the mentioned

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

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

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 - Correct

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

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

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

a) Probability
b) Hypothesis - Correct
c) Causal
d) None of the mentioned

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

a) 0 - Correct
b) 5
c) 1
d) 10

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 - Correct
d) None of the mentioned

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

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

The normal distribution, also known as the Gaussian distribution, is the most important probability distribution in statistics for independent, random variables. Most people recognize its familiar bell-shaped curve in statistical reports.The normal distribution is a continuous probability distribution that is symmetrical around its mean, most of the observations cluster around the central peak, and the probabilities for values further away from the mean taper off equally in both directions. Extreme values in both tails of the distribution are similarly unlikely. While the normal distribution is symmetrical, not all symmetrical distributions are normal**.**

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

Missing data appear when no value is available in one or more variables of an individual. Due to Missing data, the statistical power of the analysis can reduce, which can impact the validity of the results.
First we have to find the missing values and what type of data are missing, and then we have to check if there is any relationship between the values which is missing and present in data. When we collect all the necessary information then we start imputing the data as per our requirements.

Following are the few techniques of imputation :

List wise deletion:

List wise deletion is preferred when there is a Missing Completely at Random case. In List wise deletion entire rows (which hold the missing values) are deleted. It is also known as complete-case analysis as it removes all data that have one or more missing values.

In python we use dropna() function for List wise deletion.

List wise deletion is not preferred if the size of the dataset is small as it removes entire rows if we eliminate rows with missing data then the dataset becomes very short and the machine learning model will not give good outcomes on a small dataset.

Pair wise Deletion:

Pair wise Deletion is used if missingness is missing completely at random i.e MCAR.

Pair wise deletion is preferred to reduce the loss that happens in List wise deletion. It is also called an available-case analysis as it removes only null observation, not the entire row.

Dropping complete columns

If a column holds a lot of missing values, say more than 80%, and the feature is not meaningful, that time we can drop the entire column.

Imputation techniques:

The imputation technique replaces missing values with substituted values. The missing values can be imputed in many ways depending upon the nature of the data and its problem. Imputation techniques can be broadly they can be classified as follows:

Imputation with constant value:

It replaces the missing values with either zero or any constant value.

We will use the SimpleImputer class from sklearn.

Imputation using Statistics:

The syntax is the same as imputation with constant only the SimpleImputer strategy will change. It can be "Mean" or "Median" or "Most_Frequent".

"Mean" will replace missing values using the mean in each column. It is preferred if data is numeric and not skewed.

"Median" will replace missing values using the median in each column. It is preferred if data is numeric and skewed.

"Most_frequent" will replace missing values using the most_frequent in each column. It is preferred if data is a string(object) or numeric.

Before using any strategy, the foremost step is to check the type of data and distribution of features(if numeric).

Advanced Imputation Technique:

Unlike the previous techniques, Advanced imputation techniques adopt machine learning algorithms to impute the missing values in a dataset. Followings are the machine learning algorithms that help to impute missing values.

K_Nearest Neighbor Imputation:

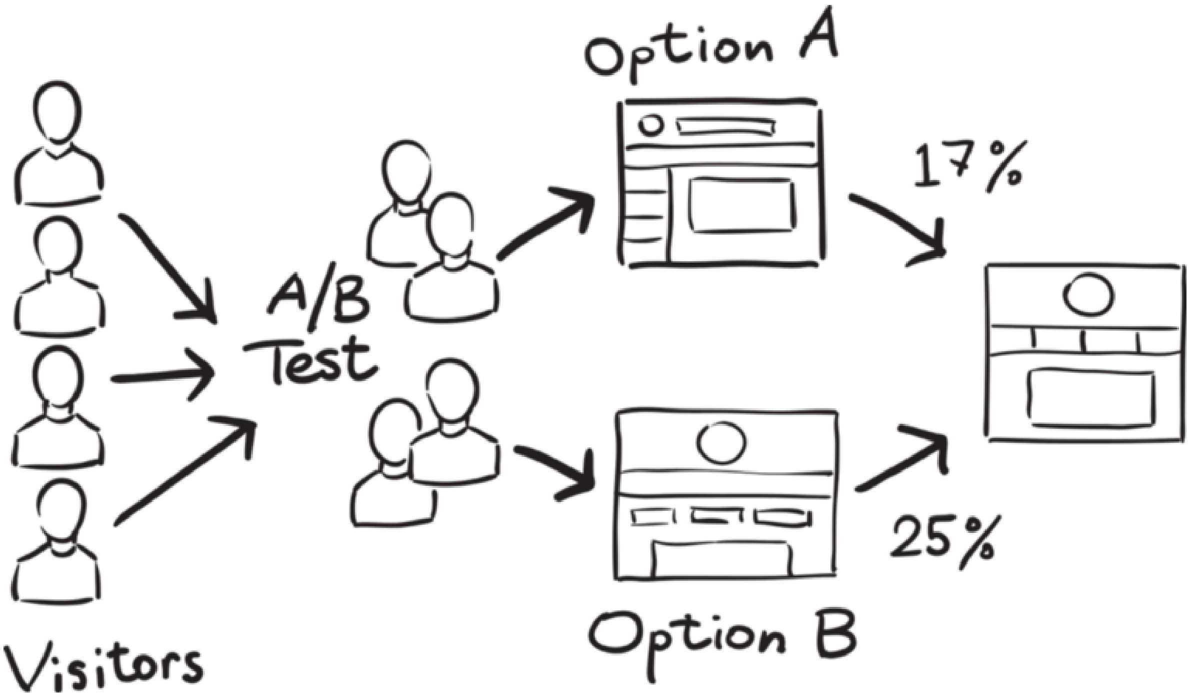
The KNN algorithm helps to impute missing data by finding the closest neighbors using the Euclidean distance metric to the observation with missing data and imputing them based on the non-missing values in the neighbors. The fundamental weakness of KNN doesn't work on categorical features. We need to convert them into numeric using any encoding method. It requires normalizing data as KNN Imputer is a distance-based imputation method and different scales of data generate biased replacements for the missing values.

12. What is A/B testing?

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.



Source

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?

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?

Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. Different regression models differ based on – the kind of relationship between dependent and independent variables they are considering, and the number of independent variables getting used.

Linear Regression is one of the most fundamental and widely known Machine Learning Algorithms which people start with. Building blocks of aLinear Regression Model are:

- Discreet/continuous independent variables
- A best-fit regression line

- Continuous dependent variable. i.e., A Linear Regression model predicts the dependent variable using a regression line based on the independent variables. The equation of the Linear Regression is:

$$Y=a+b*X + e$$

Where, a is the intercept, b is the slope of the line, and e is the error term. The equation above is used to predict the value of the target variable based on the given predictor variable(s).

15. What are the various branches of statistics?

There are **two main branches** of statistics:
- Inferential Statistic.
- Descriptive Statistic.

Inferential Statistics:
Inferential statistics used to make inference and describe about the population. These stats are more useful when its not easy or possible to examine each member of the population.

Descriptive Statistics:
Descriptive statistics are use to get a brief summary of data. You can have the summary of data in numerical or graphical form.