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

7. 1. 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. 4. 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

10. What do you understand by the term Normal Distribution?
Ans. Normal distribution, also known as Gaussian or general distribution, is the distribution of opportunities that cover all your values equally, and many results are found near the definition of opportunities. Values may be equally sorted above or below the definition. A common distribution is the distribution of the probability of (approximately) defining many common data sets in the real world. It is the most common type of distribution, and occurs naturally in mathematics using random sampling techniques. Nowadays, it often appears as a "lifetime" model of a product, such as a lamp, or the result of a standard test, such as IQ. Biological measurements, such as height or weight, are usually measured by standard distribution.

11. How do you handle missing data? What imputation techniques do you recommend?
Ans: 1. Mean or Median Imputation

When data is missing at random, we can use list-wise or pair-wise deletion of the missing observations. However, there can be multiple reasons why this may not be the most feasible option:

- There may not be enough observations with non-missing data to produce a reliable analysis
- In predictive analytics, missing data can prevent the predictions for those observations which have missing data
- External factors may require specific observations to be part of the analysis

In such cases, we impute values for missing data. A common technique is to use the mean or median of the non-missing observations. This can be useful in cases where the number of missing observations is low. However, for large number of missing values, using mean or median can result in loss of variation in data and it is better to use imputations. Depending upon the nature of the missing data, we use different techniques to impute data that have been described below.

2. Multivariate Imputation by Chained Equations (MICE)

MICE assumes that the missing data are Missing at Random (MAR). It imputes data on a variable-by-variable basis by specifying an imputation model per variable. MICE uses predictive mean matching (PMM) for continuous variables, logistic regressions for binary variables, bayesian polytomous regressions for factor variables, and proportional odds model for ordered variables to impute missing data.

To set up the data for MICE, it is important to note that the algorithm uses all the variables in the data for predictions. In this case, variables that may not be useful for predictions, like the ID variable, should be removed before implementing this algorithm.

3. Random Forest

Random forest is a non-parametric imputation method applicable to various variable types that works well with both data missing at random and not missing at random. Random forest uses multiple decision trees to estimate missing values and outputs OOB (out of bag) imputation error estimates.

One caveat is that random forest works best with large datasets and using random forest on small datasets runs the risk of overfitting. The extent of overfitting leading to inaccurate imputations will depend upon how closely the distribution for predictor variables for non-missing data resembles the distribution of predictor variables for missing data. For example, if the distribution of race/ethnicity for non-missing data is similar to the distribution of race/ethnicity for missing data, overfitting is not likely to throw off results. However, if the two distributions differ, the accuracy of imputations will suffer.

The MICE library in R also allows imputations by random forest by setting the method to "rf". The authors of the MICE library have provided an example on how to implement the random forest method .

To sum up data imputations is tricky and should be done with care. It is important to understand the nature of the data that is missing when deciding which algorithm to use for imputations. While using the above algorithms, predictor variables should be set up carefully to avoid confusion in the methods implemented during imputation. Finally, you can test the quality of your imputations by normalized root mean square error (NRMSE) for continuous variables and proportion of falsely classified (PFC) for categorical variables.

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.

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 quantifies the relationship between one or more predictor variable(s) and one outcome variable. Linear regression is commonly used for predictive analysis and modeling. For example, it can be used to quantify the relative impacts of age, gender, and diet (the predictor variables) on height (the outcome variable). Linear regression is also known as multiple regression, multivariate regression, ordinary least squares (OLS), and regression. This post will show you examples of linear regression, including an example of simple linear regression and an example of multiple linear regression.

15. What are the various branches of statistics?

Ans: 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 graphycal form.