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.
   1. True
   2. 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?
   1. Central Limit Theorem
   2. Central Mean Theorem
   3. Centroid Limit Theorem
   4. All of the mentioned
3. Which of the following is incorrect with respect to use of Poisson distribution?
   1. Modeling event/time data
   2. Modeling bounded count data
   3. Modeling contingency tables
   4. All of the mentioned
4. Point out the correct statement.
   1. The exponent of a normally distributed random variables follows what is called the log- normal distribution
   2. Sums of normally distributed random variables are again normally distributed even if the variables are dependent
   3. The square of a standard normal random variable follows what is called chi-squared distribution
   4. All of the mentioned
5. random variables are used to model rates.
   1. Empirical
   2. Binomial
   3. Poisson
   4. All of the mentioned
6. 10. Usually replacing the standard error by its estimated value does change the CLT.
   1. True
   2. False
7. 1. Which of the following testing is concerned with making decisions using data?
   1. Probability
   2. Hypothesis
   3. Causal
   4. None of the mentioned
8. 4. Normalized data are centered at and have units equal to standard deviations of the original data.
   1. 0
   2. 5
   3. 1
   4. 10
9. Which of the following statement is incorrect with respect to outliers?
   1. Outliers can have varying degrees of influence
   2. Outliers can be the result of spurious or real processes
   3. Outliers cannot conform to the regression relationship
   4. None of the mentioned

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

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

Ans. A normal distribution is a type of continuous probability 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 used to analyze data when there is an equally likely chance of being above or below the mean for continuous data whose histogram fits a bell curve.

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

Ans. Missing data can skew anything for data scientists, from economic analysis to clinical trials. After all, any analysis is only as good as the data. A data scientist doesn’t want to produce biased estimates that lead to invalid results.

When dealing with missing data, data scientists can use two primary methods to solve the error:

Imputation or the removal of data.

The imputation method develops reasonable guesses for missing data. It’s most useful when the percentage of missing data is low. If the portion of missing data is too high, the results lack natural variation that could result in an effective model.

The other option is to remove data. When dealing with data that is missing at random, related data can be deleted to reduce bias. Removing data may not be the best option if there are not enough observations to result in a reliable analysis. In some situations, observation of specific events or factors may be required.

1. 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. A/B testing is a type of experiment in which you split your web traffic or user base into two groups and show two different versions of a web pages, app, email and so on with the goal of comparing the results to find the more successful.

There are three subtypes of A/B testing:

1. Split testing
2. Multivariate testing
3. Multi-page testing
4. 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 score 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 reduced variance, the model is less accurate and the confidence interval is narrower.

1. What is linear regression in statistics?

Ans. Linear regression is a regression model that estimates the relationship between one independent variable and one dependent variable using a straight line, both variables should be quantitative. Linear regression is commonly used for predictive analysis and modeling .

Y=mx+c or y =a+bx+e

1. What are the various branches of statistics?

Ans. There are two types of statistics known descriptive statistics describe the properties of sample and population data, and inferential statistics, which uses those properties to test hypotheses and draw conclusions.

Descriptive statistics is a way to organize, 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.

Descriptive statistics are also categorized into four different categories:

1. Measure of frequency
2. Measure of dispersion
3. Measure of central tendency
4. Measure of position

Inferential statistics is used to interpret the meaning of descriptive statistics, that means once the data has been collected, analysed and summarized then we use these states to describe the meaning of the collected data.

