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?

* Normal distribution, also known as the Gaussian distribution, is a probability distribution that is symmetric about the mean, showing that data near the mean are more frequent in occurrence than data far from the mean. In graph form, normal distribution will appear as a bell curve
* The normal distribution is the most common type of distribution assumed in technical stock market analysis and in other types of statistical analyses

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

* Handling the missing values is one of the greatest challenges faced by analysts, because making the right decision on how to handle it generates robust data models.
* Methods are;
* Deleting Rows:
  + This method commonly used to handle the null values. Here, we either delete a particular row if it has a null value for a particular feature and a particular column if it has more than 70-75% of missing values
* Replacing with Mean/Median/Mode
  + This strategy can be applied on a feature which has numeric data like the age of a person or the ticket fare. We can calculate the mean, median or mode of the feature and replace it with the missing values.
  + This method is also called as leaking the data while training. Another way is to approximate it with the deviation of neighboring values. This works better if the data is linear
* Assigning A Unique Category:
  + A categorical feature will have a definite number of possibilities, such as gender, for example. Since they have a definite number of classes, we can assign another class for the missing values
* Predicting the Missing Values:
  + Using the features which do not have missing values, we can predict the nulls with the help of a machine learning algorithm.
  + We will be using linear regression to replace the nulls in the feature ‘age’, using other available features. One can experiment with different algorithms and check which gives the best accuracy instead of sticking to a single algorithm.

1. What is A/B testing?

* A/B testing is the act of running a simultaneous experiment between two or more variants of a page to see which one performs the best.
* Imagine, for instance, that you want to test your hypothesis that one headline will generate more leads than another. Sure, you could just make the change and cross your fingers. But what if you’re wrong? Mistakes can get costly.
* By sending half your traffic to one version of the page and half to another, you can first gather evidence about which one works best before you commit to the change.
* Essentially, A/B testing lets you play scientist—and make decisions based on data about how people actually behave when they hit your page

1. Is mean imputation of missing data acceptable practice?

* It is a non-standard, but a fairly flexible imputation algorithm. It uses RandomForest at its core to predict the missing data. It can be applied to both continuous and categorical variables which makes it advantageous over other imputation algorithms.

1. What is linear regression in statistics?

* Linear regression is a basic and commonly used type of predictive analysis. The overall idea of regression is to examine two things: (1) does a set of predictor variables do a good job in predicting an outcome (dependent) variable? (2) Which variables in particular are significant predictors of the outcome variable, and in what way do they–indicated by the magnitude and sign of the beta estimates–impact the outcome 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.

1. What are the various branches of statistics?

* The two main branches of statistics are descriptive statistics and inferential statistics. Both of these are employed in scientific analysis of data and both are equally important for the student of statistics.

