WORKSHEET

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

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

3. Which of the following is incorrect with respect to use of Poisson distribution?
b) Modeling bounded count data

4. Point out the correct statement.
d) All of the mentioned

5. ______ random variables are used to model rates.
c) Poisson

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

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

8. 4. Normalized data are centered at______and have units equal to standard deviations of the original data.
a) 0

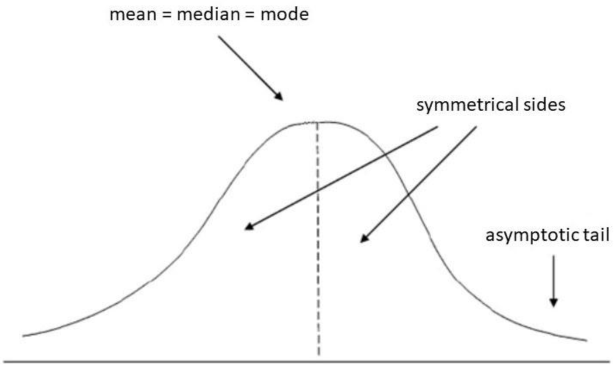
9. Which of the following statement is incorrect with respect to outliers?
c)Outliers cannot conform to the regression relationship

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. In normal distribution, data tends to like ‘Bell Curve’. Moreover, The Normal Distribution has:

- mean = median = mode
- symmetry about the center
- 50% of values less than the mean and 50% greater than the mean.



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

ANS. Generally, these are the ways to handel missing data:

- Deleting Rows
- Replacing with Mean/Median/Mode
- Assigning an Unique Category
- Predicting The Missing Values
- Using Algorithms Which Support Missing Values
- Imputation Methods

The simplest imputation method is replacing missing values with the **mean or median** values of the dataset at large, or some similar summary statistic. This has the advantage of being the simplest possible approach, and one that doesn't introduce any undue bias into the dataset. However, there are many other methods like K-Nearest Neighbors, Tree models, Linear Methods are also available.

12. What is A/B testing?

ANS. A/B testing (also known as split testing) is the process of comparing two versions variables such as a web page, email, or other marketing asset and measuring the difference in performance. You do this giving one version to one group and the other version to another group. Then you can see how each variation performs.

13. Is mean imputation of missing data acceptable practice?

ANS. Mean imputation: So simple and yet, so dangerous because Mean imputation does not preserve the relationships among variables and Leads to an Underestimate of Standard Errors so avoid this method if doesn’t suit the dataset.

14. What is linear regression in statistics?

ANS. Linear regression attempts to model the relationship between two variables by fitting linear equation to observed data. One variable is considered to be an explanatory variable, and the other is considered to be a dependent variable. For example, a modeler might want to relate the weights of individuals to their heights using a linear regression model.

Before attempting to fit a linear model to observed data, a modeler should first determine whether or not there is a relationship between the variables of interest. This does not necessarily imply that one variable causes the other (for example, higher SAT scores do not cause higher college grades), but that there is some significant association between the two variables. If there appears to be no association between the proposed explanatory and dependent variables then fitting a linear regression model to the data probably will not provide a useful model. A valuable numerical measure of association between two variables is the correlation coefficient, which is a value between -1 and 1 indicating the strength of the association of the observed data for the two variables.

A linear regression line has an equation of the form $Y = a + bX$, where X is the explanatory variable and Y is the dependent variable. The slope of the line is b, and a is the intercept (the value of y when $x = 0$)

15. What are the various branches of statistics?
ANS.

Descriptive Statistics:

Descriptive statistics deals with the presentation and collection of data. This is usually the first part of a statistical analysis. It is usually not as simple as it sounds, and the statistician needs to be aware of designing experiments, choosing the right focus group and avoid biases that are so easy to creep into the experiment.

Different areas of study require different kinds of analysis using descriptive statistics. For example, a physicist studying turbulence in the laboratory needs the average quantities that vary over small intervals of time. The nature of this problem requires that physical quantities be averaged from a host of data collected through the experiment.

Inferential Statistics:

Inferential statistics, as the name suggests, involves drawing the right conclusions from the statistical analysis that has been performed using descriptive statistics. In the end, it is the inferences that make studies important and this aspect is dealt with in inferential statistics.

Most predictions of the future and generalizations about a population by studying a smaller sample come under the purview of inferential statistics. Most social sciences experiments deal with studying a small sample population that helps determine how the population in general behaves. By designing the right experiment, the researcher is able to draw conclusions relevant to his study.

While drawing conclusions, one needs to be very careful so as not to draw the wrong or biased conclusions. Even though this appears like a science, there are ways in which one can manipulate studies and results through various means. For example, data dredging is increasingly becoming a problem as computers hold loads of information and it is easy, either intentionally or unintentionally, to use the wrong inferential methods.

Both descriptive and inferential statistics go hand in hand and one cannot exist without the other. Good scientific methodology needs to be followed in both these steps of statistical analysis and both these branches of statistics are equally important for a researcher.