STATISTICS WORKSHEET-1

Solutions:

1. (a) True

2. (a) Central Limit Theorem

3. (b) Modeling bounded count data

4. (d)

5. (c) Poisson

6. (b) False

7. (b) Hypothesis.

8. (a) 0

9. (c) Outliers cannot conform to the regression relationship

Descriptive:

10. What do you understand by the term Normal Distribution?
Ans: The normal distribution is the most important and most widely used distribution in statistics. It is sometimes called the "bell curve,". It helps describe all the possible values a random variable can take within a given range with most of the distribution area is in the middle and few are in the tails, at the extremes.

11. How do you handle missing data? What imputation techniques do you recommend?
Ans: The most common way of dealing with missing data is removing all the missing rows or columns if there are not many rows with missing data. There are some other ways to deal with missing data:

- Imputation with mean: Missing data is replaced by the mean of the column. This is a commonly used technique. However, this might not be appropriate if the data is not unimodal (for example suppose we fill missing value of weights, the mean of weights for males might be different from females and this might not be a unimodal distribution).
- Imputation with median: Missing data is replaced by the median of the column. A median is better than the mean when there are outliers, but once again, if the data is multi-model with multiple clusters, median might not work.

- Imputation with Mode: Missing data is replaced with mode of the column. This also leads to similar problems as the above two methods.

12. What is A/B testing?

Ans: A/B testing (also known as bucket testing or split-run testing) is a user experience research methodology. A/B tests consist of a randomized experiment with two variants, A and B. It includes application of statistical hypothesis testing or "two-sample hypothesis testing" as used in the field of statistics. A/B testing is a way to compare two versions of a single variable, typically by testing a subject's response to variant A against variant B. and determining which of the two variants is more effective.

13. Is mean imputation of missing data acceptable practice?

Ans:.It is a non-standard, but a fairly flexible imputation algorithm. It can be applied to both continuous and categorical variables which makes it advantageous over other imputation algorithms.

- Bad practice in general
- If just estimating means: mean imputation preserves the mean of the observed data
- Leads to an underestimate of the standard deviation

14. What is linear regression statistics?

Ans: In statistics, linear regression is a linear approach to modelling the relationship between a scalar response and one or more explanatory variables. The case of one explanatory variable is called simple linear regression; for more than one, the process is called multiple linear regression. Linear regression attempts to model the relationship between two variables by fitting a linear equation to observed data.

15. What are the various branches of statistics?

Ans: The two main branches of statistics are descriptive statistics and inferential statistics.

Descriptive statistics:

It organizes raw data into meaningful information. A house hold articles manufacturing company would like to know what people feel about their products. For that purpose, the company forms a team of people and tries to collect information from the public. The team of people formed by the company is trying to collect data from the public directly. The data which is being collected directly from the public will always not be meaning full Hence, the data which is being collected directly from the public has to be converted in to meaningful information. This is the work being done in this particular branch "descriptive-statistics". That is, it focuses on collecting, summarizing and presenting set of data.

Inferential Statistics:

It analyses sample data to draw conclusion about population. It analyses sample data to draw conclusion about population Marketing research team of a company wants to know how far the people need a particular product manufactured by the company. There are one hundred thousand populations in a particular city. It is bit difficult to go and ask all one hundred thousand people, due to time consumption and other factors. Hence, it takes a sample of 1000 people to draw conclusion for the whole population. That is making general statement from the study of particular cases or any treatment of data, which leads to prediction or inference concerning a larger group of data.