Statistics Worksheet – 1

Submission by: Sahil Kumar

Project: Internship 24

Date: 15 January 2022

1. Bernoulli random variables take (only) the values 1 and 0.

Answer: **(A) True**

2. Which of the following theorem states that the distribution of average iid variables, properly normalized, becomes that of a standard normal as the sample size increases?

Answer: **(A) Central Limit Theorem**

3. Which of the following is incorrect with respect to use of Poisson distribution?

Answer: **(B) Modeling bounded count data**

4. Point out the correct statement:

Answer: **(D) All of the mentioned**

5. _____ random variables are used to model rates.

Answer: **(C) Poisson**

6. Usually replacing the standard error by its estimated value does change the CLT.

Answer: **(B) False**

7. Which of the following testing is concerned with making decisions using data?

Answer: **(B) Hypothesis**

8. Normalized data are centered at _____ and have units equal to standard deviations of the original data.

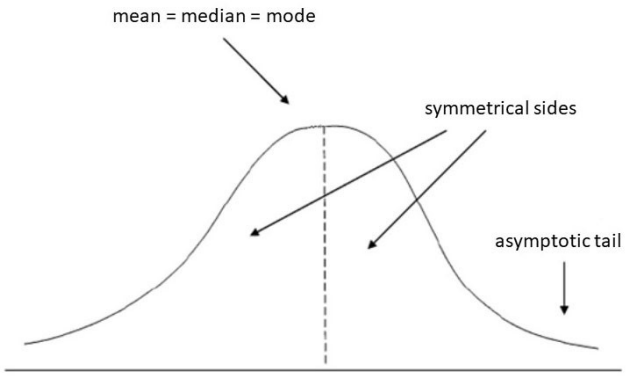
Answer: **(A) 0**

9. Which of the following statement is incorrect with respect to outliers?

Answer: **(C) Outliers cannot conform to the regression relationship**

10. **What do you understand by the term Normal Distribution?**

Answer: 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. In a normal distribution, the mean is zero, the standard deviation is 1 and it has zero skew.



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

Answer: According to SeleritySAS, Missing data is a big problem for data analysis because it has the potential to misrepresent or distort our insights and findings. Data scientists usually classify missing data into 3 types:

- Missing Completely at Random (MCAR): when data is completely missing at random across the dataset with no discernable pattern
- Missing at Random (MAR): when data is not missing randomly, but only within sub-samples of data
- Not Missing at Random (NMAR): when there is a noticeable trend in the way data is missing

The following techniques have been considered as the best to handle missing data:

(i) Deletion: this method only works for certain datasets where participants have missing fields. It means deleting any participants or data entries with missing values. This method is particularly advantageous to samples where there is a large volume of data because values can be deleted without significantly distorting readings. Alternatively, data scientists can fill out the missing values by contacting the participants in question.

(ii) Using regression analysis for systematic elimination of data: this method can be used to predict the null value using other information from the dataset. Regression methods can be successful in finding the missing data, but this largely depends on how well connected the remaining data is.

(iii) Data imputation techniques: We use two data imputation techniques to handle missing data:

   a. Average imputation: using the average value of the responses from other data entries to fill out missing values. However, it can artificially reduce the variability of the dataset. Uses mean of the data.

   b. Common-point imputation: utilising the middle point or the most commonly chosen value. Uses median or mode of the data

12. **What is A/B Testing?**

Answer: According to AnalyticsVidhya, “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.”. It is one of the most prominent and widely used statistical tools. It is a hypothetical testing methodology for making decisions that estimate population parameters based on sample statistics.

It usually follows the following process:

(i) Make a hypothesis

   a. Null hypothesis, or H_o

   b. Alternative hypothesis, or H_a

(ii) Create a control group and test group

(iii) Conduct the A/B test and collect the data

13. **Is mean imputation of missing data acceptable practice?**

Answer: According to TheAnalysisFactor, mean imputation is the replacement of a missing observation with the mean of the non-missing observations for that variable.

Imputing the mean preserves the mean of the observed data. If the data is missing completely at random, the estimate of the mean remains unbiased, however, the bias on the standard error of the data is affected. Also, by imputing the mean, we are able to keep the sample size up to the full sample size. Since most research studies are interested in the relationship among variables, mean imputation is not a good solution. Thus, in using mean imputation, our analysis will present a stronger relationship than there actually is.

Moreover, any statistic that uses the imputed data will have a standard error that is lower than what should have been. Since the imputations are themselves estimates, there is error associated with them, however, the analysis software in use will treat it as real data. This may result in lower standard errors, lower p-values, and potentially resulting in Type-I errors.

14. **What is linear regression in statistics?**

Answer: In statistics, linear regression is a linear approach for modelling the relationship between a scalar response and one or more explanatory variables (also known as dependent and independent variables). The case of one explanatory variable is called simple linear regression; for more than one, the process is called multiple linear regression. (Freedman 2009).

It is the process of predicting a label (or dependent variable) based on the features (independent variables) at hand. It is used for time series modelling and finding the causal effect relationship between the variables and forecasting.

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

Answer: Statistics can primarily be classified in to three real branches. These are data collection, descriptive statistics, and inferential statistics.

- Data collection: Data collection is all about how the actual data is collected. There are issues in the collection of the data; we need to make sure that the data has been collected fairly before we go on to deal with it, and try to present it and make conclusions. The population is the entire set of data, and a sample is a subset of the population.
- Descriptive Statistics: Descriptive statistics is the part of statistics that deals with presenting the data we have. This can take two basic forms – presenting aspects of the data either visually (via graphs, charts, etc.) or numerically (via averages and so on). The basic aim of descriptive statistics is to ‘present the data’ in an understandable way.
- Inferential statistics: Inferential statistics is the aspect that deals with making conclusions about the data. This is quite a wide area.