Statistics worksheet-1

1. Bernoulli random variables take (only) the values 1 and 0. a) True b) False

Ans. a) True

Bernoulli's random variable takes either the value 1 or 0 depending on whether a single trial of an experiment resulted in a success or a failure

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?

Ans. a) Central limit theorem

Most of the phenomenon that we see in a population when considering a continuous data is normally distributed. So, when the sample size increases, the variability seen in the data points reduces and the distribution approaches more towards a normal distribution.

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

a) Modeling event/time data

b) Modeling bounded count data

c) Modeling contingency tables

d) All of the mentioned

Ans. b) Modeling bounded count data

4. Point out the correct statement.

a) The exponent of a normally distributed random variables follows what is called the log-normal distribution

b) Sums of normally distributed random variables are again normally distributed even if the variables are dependent

c) The square of a standard normal random variable follows what is called chi-squared distribution

d) All of the mentioned

Ans. d) All of the mentioned

5. ______ random variables are used to model rates. a) Empirical b) Binomial c) Poisson d) All of the mentioned

Ans c)Poisson

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

Ans False

7. Which of the following testing is concerned with making decisions using data? a) Probability b) Hypothesis c) Causal d) None of the mentioned

Ans b) Hypothesis

8. 4. Normalized data are centered at ______and have units equal to standard deviations of the original data. a) 0 b) 5 c) 1 d) 10

Ans 0

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

a) Outliers can have varying degrees of influence

b) Outliers can be the result of spurious or real processes

c) Outliers cannot conform to the regression relationship

d) None of the mentioned

Ans c) Outliers cannot conform to regression relationship

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

Normal distribution (also called Bell Curve) is a probability distribution of continuous data of a sample/population that resembles a bell shaped curve where the mean, median and mode lines up at the centre of the distribution such that the centre of the distribution is its mean. It finds majority of its data points symmetric about the mean. As we recede away from the mean, the frequency also reduces on either side of the mean symmetrically.

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

Ans. Missing data occurs when one or more observations are not recorded for a variable and the data entry for it is left empty, represented by NaN or 0. This could be a problem as it could lead to inaccurate statistical data analysis while giving inaccurate outcome or prediction. There are various ways to handle missing values which we apply based on the situation we have at hand. If we are dealing with a large dataset and there are very few missing data the dropping the missing values is the either by eliminating the variable or removing the data entry is the best option. However, we can go for imputation which involves replacing the missing values in cases where we want to retain the entire sample size. One way of doing this is by replacing the missing values by the mean of the data points in case it is a continuous data. But if we have a missing data that falls under categorical data, we may replace it with a value that is seen with higher frequency (mode). And sometimes, it is left as a missing data

Before recommending an imputation technique, we have to first analyse why the data is missing. If data is missing completely at random, then the best option would be to drop the data entry, however if we compute missing value if it shares some relationship with the observed values, then going for mean, median or mode imputation seems like a plausible option. At times, the data gatherer may know additional information regarding the missing data,then missing value can be predicted more conclusively using more sophisticated imputation technique.

12. What is A/B testing?

Ans. A/B testing is a methodology implemented to determine the better version of a variant by the application of hypothetical testing. In hypothetical testing, a tentative assumption, which we call as a null hypothesis is made which states that there is no difference between the control and the variants. An alternative hypothesis is then defined which states opposite to the null hypothesis. A test is then conducted by collecting data from a randomly sampled population (to avoid bias) and we determine if there is enough statistical evidence to reject the null hypothesis. The statistical significance of a test is assessed by measuring the p-value, which is typically set as 0.05, though we can further lower it if the experiment allows very little error. So if we set the p-value as 0.05, and we obtain a result were less than 5% of the collected data supports the null hypothesis, then we get conclusive evidence to reject the null hypothesis.

13. Is mean imputation of missing data acceptable practice?

Ans. The technique of handling missing data differs depending on the type of missing data we are dealing with. Mean imputation of missing data cannot be implemented to variables in datasets that are categorical. Further, even when handling continuous data, replacing the missing data by its mean value will reduce the variance across the data thus distorting its distribution. Nevertheless, mean imputation can be implemented if the missing values is less than 10% of the original dataset and the correlation between the variables is low. Hence, mean imputation of missing is generally not recommended to deal with missing data due to limitations that we have discussed above.

14. What is linear regression in statistics?

Linear regression helps establish a relationship between the dependent variable and at least one independent variable. Once a linear relationship is established, it helps to predict the outcome (dependent variable) using the independent variables.

When we deal with a single independent variable, we call it a simple linear regression. The equation is given by:

$$\mathbf{y = \beta_0 + \beta_1 x}$$

Where y is the dependent variable

x is the independent variable

$\boldsymbol{\beta_0}$ **:** is a constant

$\boldsymbol{\beta_1}$: is a coefficient of X1

When there is more than one independent variable, it is called multiple linear regression where the dependent value is evaluated using the equation given below:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$$

The annotation for this equation remains the same as the above equation.

15. What are the various branches of statistics?

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

In descriptive statistics, we get an overview of the dataset and it also tells how the variables in the sample or population are related through statistical measures. Usually a data set is described through Measures of Central tendency (mean, median and mode), which helps to understand the values around which the sample/population is centred around, and Measures of Variability such as standard deviation, variance etc which describes how the data is dispersed or how spread out the values are in a dataset.

When we deal with a large population, it is not realistic to collect data from every individual from the population. Rather we sample the population and collect the data to perform statistical analysis that will then represent the population. This branch of statistics is called inferential statistics.