❖ STATISTICS WORKSHEET-1 SOLUTION

(* Answers are represented in red bold font)

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

b) 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?

a) Central Limit Theorem

b) Central Mean Theorem

c) Centroid Limit Theorem

d) All of the mentioned

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

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

5. _____ random variables are used to model rates.

a) Empirical

b) Binomial

c) Poisson

d) All of the mentioned

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

a) True

b) 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

8. Normalized data are centered at_____and have units equal to standard deviations of the original data.

a) 0

b) 5

c) 1

d) 10

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

Q10and Q15 are subjective answer type questions, Answer them in your own words briefly.

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

A normal distribution or Gaussian distribution refers to a probability distribution where the values of a random variable are distributed symmetrically. These values are equally distributed on the left and the right side of the central tendency. Thus, a bell-shaped curve is formed.

The normal distribution is described by the mean (μ) and the standard deviation (σ).

The normal distribution is often referred to as a 'bell curve' because of it's shape:

- Most of the values are around the center ()
- The median and mean are equal
- It has only one mode
- It is symmetric, meaning it decreases the same amount on the left and the right of the center

The area under the curve of the normal distribution represents probabilities for the data.

The area under the whole curve is equal to 1, or 100%

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

When data is missing, it may make sense to delete data. However, that may not be the most effective option. For example, if too much information is discarded, it may not be possible to complete a reliable analysis. Or there may be insufficient data to generate a reliable prediction for observations that have missing data.

Instead of deletion, data scientists have multiple solutions to impute the value of missing data. Depending why the data are missing, imputation methods can deliver reasonably reliable results. These are examples of single imputation methods for replacing missing data.

Mean, Median and Mode

This is one of the most common methods of imputing values when dealing with missing data. In cases where there are a small number of missing observations, data scientists can calculate the mean or median of the existing observations. However, when there are many missing variables, mean or median results can result in a loss of variation in the data. This method does not use time-series characteristics or depend on the relationship between the variables.

Time-Series Specific Methods

Another option is to use time-series specific methods when appropriate to impute data. There are four types of time-series data:

No trend or seasonality.

Trend, but no seasonality.

Seasonality, but no trend.

Both trend and seasonality.

The time series methods of imputation assume the adjacent observations will be like the missing data. These methods work well when that assumption is valid. However, these methods won't always produce reasonable results, particularly in the case of strong seasonality.

Last Observation Carried Forward (LOCF) & Next Observation Carried Backward (NOCB)

These options are used to analyze longitudinal repeated measures data, in which follow-up observations may be missing. In this method, every missing value is replaced with the last observed value. Longitudinal data track the same instance at different points along a timeline. This method is easy to understand and implement. However, this method may introduce bias when data has a visible trend. It assumes the value is unchanged by the missing data.

Linear Interpolation

Linear interpolation is often used to approximate a value of some function by using two known values of that function at other points. This formula can also be understood as a weighted average. The weights are inversely related to the distance from the end points to the unknown point. The closer point has more influence than the farther point.

When dealing with missing data, you should use this method in a time series that exhibits a trend line, but it's not appropriate for seasonal data.

Seasonal Adjustment with Linear Interpolation

When dealing with data that exhibits both trend and seasonality characteristics, use seasonal adjustment with linear interpolation. First you would perform the seasonal adjustment by computing a centered moving average or taking the average of multiple averages – say, two one-year averages – that are offset by one period relative to another. You can then complete data smoothing with linear interpolation as discussed above.

Multiple Imputation

Multiple imputation is considered a good approach for data sets with a large amount of missing data. Instead of substituting a single value for each missing data point, the missing values are exchanged for values that encompass the natural variability and uncertainty of the right values. Using the imputed data, the process is repeated to make multiple imputed data sets. Each set is then analyzed using the standard analytical procedures, and the multiple analysis results are combined to produce an overall result.

The various imputations incorporate natural variability into the missing values, which creates a valid statistical inference. Multiple imputations can produce statistically valid results even when there is a small sample size or a large amount of missing data.

K Nearest Neighbors

In this method, data scientists choose a distance measure for k neighbors, and the average is used to impute an estimate. The data scientist must select the number of nearest neighbors and the distance metric. KNN can identify the most frequent value among the neighbors and the mean among the nearest neighbors.

12. What is A/B testing?

A/B testing, at its most basic, is a way to compare two versions of something to figure out which performs better.

For instance, let's say you own a company and want to increase the sales of your product. Here, either you can use random experiments, or you can apply scientific and statistical methods. A/B testing is one of the most prominent and widely used statistical tools.

In the above scenario, you may divide the products into two parts – A and B. Here A will remain unchanged while you make significant changes in B's packaging. Now, on the basis of the response from customer groups who used A and B respectively, you try to decide which is performing better.

13. Is mean imputation of missing data acceptable practice?

Generally mean imputation of missing data is not acceptable practice due to following reasons:

Mean imputation reduces the variance of the imputed variables.

Mean imputation shrinks standard errors, which invalidates most hypothesis tests and the calculation of confidence interval.

Mean imputation does not preserve relationships between variables such as correlations.

14. What is linear regression in statistics?

Linear Regression is one of the most fundamental and widely known machine Learning Algorithm which people start with. Building blocks of a Linear Regression Model are:

Discreet/continuous independent variables

A best-fit regression line

Continuous dependent variable.

A Linear Regression model predicts the dependent variable using a regression line based on the independent variables. The equation of the Linear Regression is:

$$Y=a+b*X + e$$

Where, a is the intercept, b is the slope of the line, and e is the error term. The equation above is used to predict the value of the target variable based on the given predictor variable(s).

15. What are the various branches of statistics?

There are two branches of Statistics.

Descriptive Statistics: It is a statistics or a measure that describes the data.

Inferential Statistics : Using a random sample of data taken from a population to describe and make inferences about the population is called Inferential Statistics.

Descriptive Statistics

Descriptive Statistics is summarizing the data at hand through certain numbers like mean, median etc. so as to make the understanding of the data easier. It does not involve any generalization or inference beyond what is available. This means that the descriptive statistics are just the representation of the data (sample) available and not based on any theory of probability.

Commonly Used Measures

Measures of Central Tendency

Measures of Dispersion (or Variability)

Inferential Statistics :

In Inferential statistics, we make an inference from a sample about the population. The main aim of inferential statistics is to draw some conclusions from the sample and generalise them for the population data. E.g., we have to find the average salary of a data analyst across India. There are two options.

The first option is to consider the data of data analysts across India and ask them their salaries and take an average.

The second option is to take a sample of data analysts from the major IT cities in India and take their average and consider that for across India.

The first option is not possible as it is very difficult to collect all the data of data analysts across India. It is time-consuming as well as costly. So, to overcome this issue, we will look into the second option to collect a small sample of salaries of data analysts and take their average as India average. This is the inferential statistics where we make an inference from a sample about the population.