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
   b) False

   Answer: a

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

   Answer: a

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

   Answer: b

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

   Answer: d

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

   Answer: c

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

   Answer: b

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

c) Causal
d) None of the mentioned

Answer: b

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

Answer: a

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

Answer: c

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: The normal distribution is the most common type of distribution. Shape of normal distribution curve is similar to the shape of bell curve. It is the arrangement of data set in which most values cluster in the middle of the range and the rest taper off symmetrically toward either extreme ends of the x-axis. In a normal distribution, the mean, mode and median are all the same. The standard normal distribution is a type of normal distribution. It mostly appears when a normal random variable has a mean value equal to 0 and value of standard deviation is equal to 1

Formula:

z=(x-mu)/sigma →x=normal variable,

mu=mean of the data,

sigma=standard deviation of the data

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

Answer:

- Imputation is the process of replacing missing values with substituted data. It is done as a preprocessing step.
- If the data is numerical, we can use mean and median values to replace else if the data is categorical, we can use mode which is a frequently occurring value.
- Imputation of missing values from predictive techniques assumes that the nature of such missing observations are not observed completely at random and the variables chosen to impute such missing observations have some relationship with it, else it could yield imprecise estimates.
- Predictive model could be used to impute the missing values for Device, OS, Revenues. There are various statistical methods like regression techniques, machine learning methods like SVM and/or data mining methods to impute such missing values.

12. What is A/B testing?

- A/B Testing(also known as Split testing) is one of the best way to compare two or more versions of an application or a web page. It enables you to determine which one of them performs better and can generate better conversion rates. It is one of the easiest ways to analyze an application or a web page to create a new version that is more effective.
- A/B Testing is used to make business decisions based on the results derived from data, instead of just making predictions.
- There are two significant methods through which you can check conversion rates using A/B Testing
  - Sampling of Data
  - Confidence Intervals

13. Is mean imputation of missing data acceptable practice?

Answer:

Mean imputation is generally bad practice because it doesn’t take into account feature correlation.
It’s a popular solution to missing data, despite its drawbacks. Mainly because it’s easy. It can be really painful to lose a large part of the sample you so carefully collected, only to have little power.
But that doesn’t make it a good solution, and it may not help you find relationships with strong parameter estimates. Even if they exist in the population. On the other hand, there are many alternatives to mean imputation that provide much more accurate estimates, and if the data are missing completely at random, mean imputation will not bias your parameter estimate.

14. What is linear regression in statistics?

Linear regression is a linear approach for modeling the relationship between a dependent and independent variables. Once the degree of relationship between variables has been established using co-relation analysis, it is natural to delve into the nature of relationship. Regression analysis helps in determining the cause and effect relationship between variables.

Formula: Y=a+bX

b =Slope of the line.
a =Y-intercept of the line.
X= Values of the first data set.
Y= Values of the second data set.

15. What are the various branches of statistics?

Answer: Statistics have majorly categorized into two types: 1)Descriptive statistics
2)Inferential statistics

1)Descriptive Statistics:

In this type of statistics, the data is summarized through the given observations. The summarisation is one from a sample of population using parameters such as the mean or standard deviation.

Descriptive statistics are also categorised into four different categories:

- Measure of frequency
- Measure of dispersion
- Measure of central tendency
- Measure of position

2)Inferential statistics:

Inferential Statistics is a method that allows us to use information collected from a sample to make decisions, predictions or inferences from a population. It grants us permission to give statements that goes beyond the available data or information. For example, deriving estimates from hypothetical research. That means once the data has been collected, analyzed and summarized then we use these stats to describe the meaning of the collected data. Or we can say, it is used to draw conclusions from the data that depends on random variations such as observational errors, sampling variation, etc.