

STATISTICS WORKSHEET-1

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

Ans: a) True

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

Ans: a) Central Limit Theorem

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

Ans: b) False

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

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: a) 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 the regression relationship

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

A distribution is called to be normal when the variation between mean and the one standard deviation on both sides of the mean, we will have 34.1% of the data.

- For $\mu \pm \sigma$ distribution to be 34.1% on each side
- For $\mu \pm 2\sigma$ distribution to be 13.6% on each side
- For $\mu \pm \sigma$ distribution to be 2.1% on each side

A normal distribution is an arrangement of a data set in which most values are clustered near the mean value of the data set and the rest taper off symmetrically toward either extreme.

A homogeneous set have similar attributes with minor variations. Through normal distribution many natural phenomena can be explained.

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

Missing data can be handled in different ways. Namely:

- Predict missing values through logistic regression, discriminant regression, data imputation etc.
- Substitution mean values if the data set is homogeneous.
- Ignore the missing data sets or values.

Imputation:

Imputation is the method of filling in the estimated values in the missing data places and analyzing the whole data set as if were the filled in data values are assuming to be true.

The following imputation techniques can be recommended.

- Mean imputation
- Hot deck imputation
- Cold deck imputation
- Regression imputation
- Stochastic regression imputation
- Interpolation and extrapolation
- Single or multiple imputation

12. What is A/B testing? Ans:

A/B testing is a method of statistical testing, it is a way of comparing the two versions of a data sets to find out which performs better in a controlled environment. In A/B Testing the data is divided into two sets that is control and experimental.

This is also known as the split test. Here, the users are either shown the primary data or the changed data but not both and with the help of statistical analysis, the better version of data will be selected.

13. Is mean imputation of missing data acceptable practice?

Ans:

The process of replacing null values in a data set with the data's mean is known as mean imputation.

The mean imputation of missing data is a bad practice as, the mean imputation decreases the variance of our data while increasing bias. As a result of the reduced variance, the model is less accurate and the confidence interval is narrower.

14. What is linear regression in statistics? Ans:

Linear regression is a method of predictive analysis, to predict the value of a variable based on the value of another variable.

The variable we want to find is called as dependent variable and the other one is called as the explanatory/independent variable.

- A linear equation is fit in for estimating the required variable.
- Before attempting to fit the linear equation on the preferred data set, one has to determine if there is a relationship with the variables of interest.
- Linear-regression models are relatively simple and provide an easy-to-interpret mathematical formula that can generate predictions. Linear regression can be applied to various areas in business, life sciences, academic study etc.

15. What are the various branches of statistics? Ans:

There are two main branches of statistics

1. Inferential Statistics.
2. Descriptive Statistics. **Inferential Statistics:**
 - This is a branch of statistics which deals with techniques used for analysis of data, making estimates that lead to predictions and drawing conclusions or inferences from limited

information taken on sample basis and testing the reliability of the estimates or predictions.

- Inferential statistics is used to make comparisons or predictions about a larger group taking a sample out of it **Descriptive Statistics:**
- This is a branch of statistics which deals with methods of collection of data, its presentation and organization in various forms, such as distribution tables, graphs, diagrams and finding measures of central tendency and measures of dispersion or spread which are used in the description of data.
- Descriptive statistics is used to present the data in an understandable way, so that a meaningful description can be made.

*****End of STATISTICS WORKSHEET-1*****