

STATISTICS WORKSHEET-1

1. Bernoulli random variables take (only) the values 1 and 0.
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
3. Which of the following is incorrect with respect to use of Poisson distribution?
b) Modeling bounded count data
4. Point out the correct statement.
d) All of the mentioned
5. _____ random variables are used to model rates.
c) Poisson
6. Usually replacing the standard error by its estimated value does change the CLT.
b) False
7. 1. Which of the following testing is concerned with making decisions using data?
b) Hypothesis
8. 4. Normalized data are centered at _____ and have units equal to standard deviations of the original data.
a) 0
9. Which of the following statement is incorrect with respect to outliers?
c) Outliers cannot conform to the regression relationship

Q10 and Q15 are subjective answer type questions, Answer them in your own words briefly.

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

Ans: It is a probability distribution curve with symmetrical distribution of data resulting in bell shaped curve when plotted and whose mean and median are same with standard deviation as 1.

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

Ans: handling of missing data is carried out by:

- 1) Deleting missing data : if data is vast and missing values are almost negligible comparative to actual data then missing data is deleted. But this technique is not recommended mostly.
- 2) Filling of null values with either mean in terms of continuous data or mode in terms of categorical data.

Other than using mean and mode to impute data, other imputation techniques includes simple imputer, knn imputer and iterative imputer

12. What is A/B testing?

Ans: A/B testing is a kind of tests on sample A and B based on hypothesis that is making an assumption without any proof. There are 2 types of hypothesis i.e.

(I) Null hypothesis denoted as (H_0) - Here decision always leads to status quo that means the status or hypothesis will not change.

(II) Alternative hypothesis denoted as (H_a) -Decision leads to opposite of Null hypothesis.
The tests used are: 1) One tail test and 2) Two tail test.

13. Is mean imputation of missing data acceptable practice?

Ans: If the given data is vast and the null values are negligible then using mean of that data is favourable but doing so sometimes hampers the model accuracy. Further mean imputation should always be done on continuous data.

14. What is linear regression in statistics?

Ans: Logistic regression predicts the dependant variable using a regression line on the independent variables with equation as $y=a+b*x+e$;

where a is the intercept, b is the slope of the line, e is the error

It is an important tool for analysing and modelling data where we fit a curve/line into the data points in such a manner that the difference between the actual data points from the plotted curve line is minimum.

15. What are the various branches of statistics?

Ans: There are 2 branches of statistics:

A) Descriptive statistics- when one is able to describe things and provide opinion then it is called descriptive statistics. This is true when data is less.

B) Inferential statistics- when one is able to know the data very well or able to describe things then you apply population and sample concepts to offer final opinion.