

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
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

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?

- normal distribution also called Gaussian distribution_ the most common distribution function for independent, randomly generated variables
- A normal distribution, sometimes called the bell curve, is a distribution that occurs naturally in many situations
- The graph of the normal distribution is characterized by two parameters: the mean, or average
- normal distribution is symmetric from the peak of the curve, where the mean is. This means that most of the observed data is clustered near the mean, while the data become less frequent when farther away from the mean. The resultant graph appears as bell-shaped where the mean, median, and mode are of the same values and appear at the peak of the curve.

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

- There is lots of technics handle missing data :
 - 1) Deleting Rows with missing values.
 - 2) Impute missing values for continuous variable.
 - 3) Impute missing values for categorical variable
 - 4) Using Algorithms that support missing values.
 - 5) Prediction of missing values
- Imputation technics are :
 - 1) Incompatible with most of the Python libraries used in Machine Learning
 - 2) Distortion in Datasets

12. What is A/B testing?

- A/B testing, also known as split testing, refers to a randomized experimentation process wherein two or more versions of a variable (like web page, page element) are shown to different segments of website visitors at the same time to determine which version leaves the maximum impact and drive business metrics.

Conduct A/B testing in this ways :

- Pick one variable to test. ...
- Identify your goal. ...
- Create a control and a challenger....
- Split your sample groups equally and randomly. ...
- Determine your sample size (if applicable). ...
- Decide how significant your results need to be. ...
- Make sure you're only running one test at a time on any campaign.

13. Is mean imputation of missing data acceptable practice?

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

14. What is linear regression in statistics?

- linear regression is a regression model that estimates the relationship between one independent variable and one dependent variable using a straight line
- linear regression is divided into two types: **Multiple linear regression** and **Simple linear regression**.

1) Simple linear regression :

- In Simple L. Regression we try to find the relationship between a single independent variable and a corresponding dependent variable

2) Multiple linear regression :

- In Multiple L. Regression, we try to find the relationship between 2 or more independent variables and the corresponding dependent variable

15. What are the various branches of statistics?

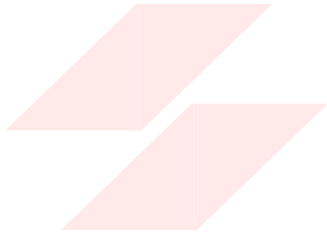
- Statistics have majorly categorised into two types:

1) Descriptive statistics :

- In the descriptive Statistics, the Data is described in a summarized way. The summarization is done from the sample of the population using different parameters like Mean or standard deviation. Descriptive Statistics are a way of using charts, graphs, and summary measures to organize, represent, and explain a set of Data.
- For example :
the collection of people in a city using the internet or using Television
- Descriptive statistics are also categorised into four different categories:
 - 1) Measure of frequency
 - 2) Measure of dispersion
 - 3) Measure of central tendency
 - 4) Measure of position

2) Inferential statistics :

In the Inferential Statistics, we try to interpret the Meaning of descriptive Statistics. After the Data has been collected, analyzed, and summarised we use Inferential Statistics to describe the Meaning of the collected Data.



FLIP ROBO