Worksheet-1

STATISTICS WORKSHEET

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

Worksheet-1

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

a) True
b) False

Answer (b)

7. 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. 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)

Worksheet-1

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?

Answer-

The normal distribution is also known as the Gaussian distribution. It is a type of continuous probability distribution in which most data points cluster towards the middle of the range while the rest taper off symmetrically towards either extreme.

Graphically normal distribution is a bell curve because of its flared sh

The skewness for normal distribution is zero and kurtosis is 3. The mean, median and mode are all the same.

In normal distribution the mean is zero and the standard deviation is 1.

Formula for the normal distribution

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$$

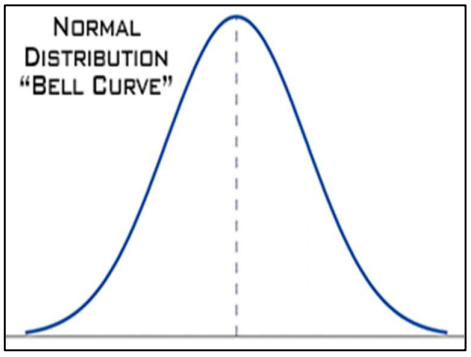
Here,

x is the value of variable

f(x) is the probability density function

μ is the mean

σ is the standard deviation



Worksheet-1

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

Answer-

There are two ways of handling missing data

a) Removal of the data
b) Imputation

Removal of the data

Missing data can be handled by deleting the rows or columns having null values. This mothed is not recommended because it might end up deleting some useful data from the dataset.

Imputation

Imputation is a technique used for replacing the missing data with some substitute value to retain most of the data of the dataset. Some imputation techniques are

- Replacing missing data with previous data
- Replacing missing data with next data
- Replacing missing data with mean/median/mode
- KNN imputation

Mean/Median/Mode:

Replacing with the mean is most common method of imputing missing values of numeric columns. Columns in the dataset which have missing values can be replaced with mean or median of remaining values in the column. Mode is used in the case of categorical features.

KNN imputation:

It operates by replacing missing data with the average mean of the neighbors nearest to it.

Worksheet-1

12. What is A/B testing?

Answer-

A/B testing is a process to compare two versions of a content to figure out which performs better. It is also known as the split testing and bucket testing.

A/B testing is one of the most popular controlled experiments used to optimize web marketing strategies. It allows us to see which variation works better for our audience based on statistical analysis.

A/B testing is a form of statistical and two-sample hypothesis testing. Statistical hypothesis testing is a method in which a sample dataset is compared against the population data. Two-sample hypothesis testing is a method in determining whether the differences between the two samples are statistically significant or not.

Steps for conducting A/B testing

- Determine the null hypothesis and alternative hypothesis.
- Create control group and variant group. There are two concepts to consider in this step, random samplings and sample size.
- Conduct the test and collect the data
- Compare the results and reject or do not reject the null hypothesis.

13. Is mean imputation of missing data acceptable practice?

Answer-

Mean imputation is a method in which missing value is replaced by the mean of remaining values in that column. This method is not considered as good practice. Using mean imputation can significantly reduce the model's accuracy and bias the results.

The drawbacks of using mean imputation

- Mean imputation ignores feature correlations.
- Mean imputation reduces a variance of the data.

Worksheet-1

14. What is linear regression in statistics?

Answer-

In statistics, linear regression is the process of predicting a dependent variable using a regression line based on the independent variables. It is most basic and commonly used predictive analysis. It shows the significant relationship between dependent variable and independent variables. The independent variable is also known as the predictor or feature and the dependent variable is known as the outcome or target or label.

Linear regression used for time series modelling. Building blocks of a linear regression model are:

- Discreet/continuous independent variables
- A best-fit regression line
- Continuous dependent variable.

Equation of linear regression

$Y = a + b*X + e$

Where, a = intercept

b = slope of the line

e = error term

15. What are the various branches of statistics?

Answer-

There are three real branches of statistics

- Data collection
- Descriptive statistics
- Inferential statistics

Descriptive statistics are used to describe the basic features of the data in a study. They provide simple summaries about the sample and the measures. Descriptive statistics can be categorized into

Worksheet-1

- Measures of central tendency
- Measures of variability

Inferential statistics are techniques that use to gathered information from a sample to make inferences, decisions or predictions about a given population. The different types of calculation of inferential statistics include:

- Regression analysis
- Analysis of variance (ANOVA)
- Analysis of covariance (ANCOVA)
- Statistical significance (t-test)
- Correlation analysis