WORKSHEET

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
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:- a) True

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

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:- A normal distribution is a type of continuous probability distribution in which most data point cluster toward the middle of the range, while the rest taper off symmetrically toward either extreme.

Example:- The weight of a new born baby.

Normal distribution :- Mean = Mode = Median = 0

Standard deviation = +/- 1

Example:- 100 student weight = 60

= +1 =61

= -1 = 59

Ideally condition is every thing is same no difference in data .

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

Ans:-The real- world data often has a lot of missing values. The cause of missing values can be data corruption or failure to record data. The handling of missing data is very important during the preprocessing of the dataset as many machine learning algorithms do not support missing value.

*Deleting Rows with missing value .

*Impute missing values for continuous variable .

*Impute missing values for categorical variable.

* Prediction of missing value.

Example:-

```
x . shape
y . shape
x_train; x_test; y_train; y_test; =train_test_split
(x,y,test_size =30 ,random_state =42)
x_train . shape
y_ train . shape
x_test . shape
y_test . shape
LM = linear regression()
LM . fit (x_train,y_train)
LM . coef_
LM . intercept
bos . columns
LM . score (x_train, y_train)
#predict the value
Pred = LM . predict (x_test)
Pred (predicted result price : " , pred)
print ("actual price" , y_test)
```

12. What is A/B testing?

Ans:- Hypothesis test = An objective method of making decisions or inferences from sample data (evidence)

There are two types of hypothesis test.

1. Null hypothesis(HO)
2. Alternative hypothesis(HA)

Example = selecte some 10 to 12 question and randomly ask to 3000 ya 5000 people
Then analysis all answer with some test (chi test, regression test) then test the data . then check all test p value(probability value). P.value give idea for which value are accepted and which value are rejected
If p. value >0.05
Then I decied null hypothesis is accepted and alternative is rejected
If p. value <0.05
Then opposite .

13. Is mean imputation of missing data acceptable practice?

Ans :-Mean imputation (MI) is one such method in which the mean of the observed values for each variable is computed and the missing values for that variable are imputed by this mean.

Example:- we have a table with age and fitness scores, and an eight-year-old has a missing fitness score. If we average the fitness scores of people between the ages of 15 and 80, the eighty-year-old will appear to have a significantly greater fitness level than he actually does., 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 **the most basic and commonly used predictive analysis**. Regression estimates are used to describe data and to explain the relationship.

Linear regression analysis is used to predict the value of a variable based on the value of another variable. The variable you want to predict is called the dependent variable. The variable you are using to predict the other variable's value is called the independent variable.

Y = a(bo) + bx

(depended variable/ Label / target / output) (intercept) (where b are slope/cofficent value and x is input or independent variable)

Y= a +bx + error

15. What are the various branches of statistics?

Ans:- There are twol branches of statistics: 1. Descriptive statistics
2. Inferential statistics

1. Descriptive statistics = i Central tendancy (mean, mode, median)
ii Disperson of data (Range, variance, standard deviation, persontial, skew)
2. Inferential statistics = z score.