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

a) True

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

---

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

---

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

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 an arrangement of a data set in which most values cluster in the middle of the range and the rest taper off symmetrically toward either extreme.

A graphical representation of a normal distribution is sometimes called a bell curve because of its flared shape. The precise shape can vary according to the distribution of the population but the peak is always in the middle and the curve is always symmetrical. In a normal distribution, the mean, mode and median are all the same.

Normal distribution curves are sometimes designed with a histogram inside the curve. The graphs are commonly used in mathematics, statistics and corporate data analytics.

For a normal distribution, 68% of the observations are within +/- one standard deviation of the mean, 95% are within +/- two standard deviations, and 99.7% are within +- three standard deviations.

---

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

Ans:First, understand that there is NO good way to deal with missing data. Every software and technique that says they deal well with missing data is lying. Do everything you can to avoid it. But sometimes the cost of patching the data up are less than throwing the data out. The exception is when the missing data can be predicted with 100% accuracy (e.g. my city is New York, and my state is missing).

We can adopt to some of these ways to handle missing data:

- Deleting Rows with missing values
- Impute missing values for continuous variable
- Impute missing values for categorical variable
- Mean or Median imputation
- Using Algorithms that support missing values
- Prediction of missing values
- Imputation using Deep Learning Library — Datawig

I also recommended

- Deleting Rows with missing values
- Mean or Median imputation
- Prediction of missing values

---

12. What is A/B testing?

Ans: A/B tests, also known as split tests, allow you to compare 2 versions of something to learn which is more effective. Simply put, do your users like version A or version B?

The concept is similar to the scientific method. If you want to find out what happens when you change one thing, you have to create a situation where only that one thing changes.

Think about the experiments you conducted in elementary school. If you put 2 seeds in 2 cups of dirt and put one in the closet and the other by the window, you'll see different results. This kind of experimental setup is A/B testing.

13. Is mean imputation of missing data acceptable practice?

Ans: The mean imputation is not considered as a good practice of data imputation for low sample size.

There are three problems with using mean-imputed variables in statistical analyses:

- Mean imputation reduces the variance of the imputed variables.
- Mean imputation shrinks standard errors, which invalidates most hypothesis tests and the calculation of confidence interval.
- Mean imputation does not preserve relationships between variables such as correlations.

14. What is linear regression in statistics?

Ans: Linear regression models use a straight line, while logistic and nonlinear regression models use a curved line. Regression allows you to estimate how a dependent variable changes as the independent variable(s) change.

linear regression is used to estimate the relationship between two quantitative variables. You can use linear regression when you want to know:

How strong the relationship is between two variables (e.g. the relationship between rainfall and soil erosion).

The value of the dependent variable at a certain value of the independent variable (e.g. the amount of soil erosion at a certain level of rainfall).

Example:

You are a social researcher interested in the relationship between income and happiness. You survey 500 people whose incomes range from $15k to $75k and ask them to rank their happiness on a scale from 1 to 10.

Your independent variable (income) and dependent variable (happiness) are both quantitative, so you can do a regression analysis to see if there is a linear relationship between them.

15. What are the various branches of statistics?

Ans: There are two main branches of statistics

Inferential Statistic.

Descriptive Statistic.

Inferential Statistics:

Inferential statistics used to make inference and describe about the population. These stats are more useful when its not easy or possible to examine each member of the population.

Descriptive Statistics:

Descriptive statistics are use to get a brief summary of data. You can have the summary of data in numerical or graphycal form.

N.B:-In MCQ question right answer marked by yellow colour.