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

   Answer : (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

   Answer: (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

   Answer: (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

   Answer: (d) All of the mentioned

FLIP ROBO

5. _____random variables are used to model rates.
   a) Empirical
   b) Binomial
   c) Poisson
   d) All of the mentioned

   Answer: (c) poisson

6. 10. Usually replacing the standard error by its estimated value does change the CLT.
   a) True
   b) False

   Answer: (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

   Answer: (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

   Answer: (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

   Answer : © 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?

Answer: The normal (or Gaussian) distribution is one particular kind of a bell shaped curve. It is unimodal (that is, there is one peak"), symmetric (that is you can flip it around its mid point) and its mean, median and mode are all equal. However, it is only *one* such distribution - others meet all those conditions and are not normal

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

- Answer: **Ignore the records with missing values.**

Many tools ignore records with missing values. When the percentage of records with missing values is small, we could ignore those records.

- **Substitute a value such as mean**.

When the percentage is large and also when it makes sense to do something to avoid bias modeling results, substituting a value (e.g. mean, median) is a commonly used way. But this method could cause bias distribution and variance. That's where the following imputation methods come in.

- **Predict missing values.**

Depending on the type of the imputed variable (i.e. continuous, ordinal, nominal) and missing data pattern (i.e. monotone, non-monotone), below are a few commonly used models. If you plan to do it in SAS, there are SAS codes that you can write to identify the missing data pattern.

  - Logistic Regression
  - Discriminant Regression
  - Markov Chain Monte Carlo (MCMC)
  - …

- **Predict missing values - Multiple Imputation**. Although there are pros & cons, MI is considered to be superior to single imputation, and it better measures the uncertainty of the missing values.

In addition, there are a few required **statistical assumptions** for multiple imputation:

1. Whether the data is missing at random (MAR).
2. Multivariate normal distribution, for some of the modeling methods mentioned above (e.g. regression, MCMC).
3. …

At last, if you have to think of **what to report** -

- The type of imputation algorithm used.
- Some justification for choosing a particular imputation method.
- The proportion of missing observations.
- The number of imputed datasets (m) created.
- The variables used in the imputation model.

12. What is A/B testing?

Answer: A/B testing is a basic randomized control experiment. It is a way to compare the two versions of a variable to find out which performs better in a controlled environment.

13. Is mean imputation of missing data acceptable practice?

Answer: Mean imputation is typically considered terrible practice since it ignores feature correlation. 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?

Answer: Linear regression attempts to model the relationship between two variables by fitting a linear equation to observed data. One variable is considered to be an explanatory variable, and the other is considered to be a dependent variable.

15. What are the various branches of statistics?

Answer: There are three real branches of statistics: data collection, descriptive statistics and inferential statistics.

FLIP ROBO