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.

Bernoulli distribution, is the discrete probability distribution of a random variable which takes the value of 1 when with the probability p and the value 0 with the probability q= 1-p.

2. Which of the following theorem states that the distribution of averages of its 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

States that whenever a random sample of size n is taken from any distribution with mean and variance, then the sample mean will be approximately normally distributed with mean and variance. The larger the value of the sample size, the better the approximation to the normal.

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 the mentioned

Answer: c) Modeling contingency tables

Poisson distribution: frequency at which an event occurs within a specific interval.

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

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

Answer: c) Poisson

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

Answer: 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

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: c) Outliers cannot conform to 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: Normal distribution is also known as Gaussian distribution. It is a probability distribution that is symmetric about the mean, showing that data near the mean are more frequent in occurrence than data far from the mean. Normal distribution will appear as a bell-curve.

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

Answer: First of all, we need to determine the pattern of missing data.

There are three kinds of missing data:

1) Missing completely at random: no pattern in the missing data. Best case scenario.
2) Missing at random: Pattern in the missing data but not on your primary dependent variables.
3) Missing not at random: There is a pattern in the missing data that affect your primary dependent variables. Worst-case scenario.

There are 7 ways to deal with missing data:

1) Listwise deletion: Delete data with missing values.
2) Recover the missing values

Imputation: replacing missing values with substitute values

3) Educated guess: infer a missing value
4) Average imputation: Use the average value
5) Common point imputation
6) Regression imputation: Use multiple regression analysis to estimate the missing values.
7) Multiple imputation: Most sophisticated and popular approach, uses correlations, software creates plausible values.

12. What is A/B testing?

Answer: A/B testing can also be known as split or bucket testing. It is a method to compare two versions of a webpage or an app against each other to determine which performs better.

A/B testing is essentially an experiment where two or more variants of a page are shown to users at random, and statistical analysis is used to determine which variation performs better for given conversion goal.

A/B testing allows to make careful changes to the user experiences while collecting data on the results. It allows them to construct hypotheses and to learn why certain elements of their experiences impact user behavior.

13. Is mean imputation of missing data acceptable practice?

Answer: Mean imputation is the replacement of a missing observation with the mean of the non-missing observations for that variable.

Issues with mean imputation practice:

1) Mean imputation does not preserve the relationships among variables.
2) Mean Imputation leads to an underestimate of standard errors.

14. What is linear regression in statistics?

Answer: Linear regression is the most widely used statistical technique. It is a way to model a relationship between two sets of variables. The result is a linear regression equation that can be used to make predictions about data.

15. What are the various branches of statistics?

Answer: There are two main branches of statistics:

1) Descriptive: deals with presentation and collection of the data. The first step of statistical analysis.
2) Inferential: deals with drawing the right conclusion from the statistical analysis performed using descriptive statistics.

Both descriptive and inferential statistics go hand in hand and one cannot exist without the other. Good scientific_methodology needs to be followed in both these steps of statistical analysis and both these branches of statistics are equally important for a researcher.