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

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

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

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

5. _____ random variables are used to model rates.
a) Empirical
b) Binomial
c) Poisson
d) All of the mentioned
Answer: C

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

Ans) B

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

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

9. Which of the following statements 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

Q10and Q15 are subjective answer type questions, Answer them in your own words briefly.

10. What do you understand by the term Normal Distribution?

A normal distribution is the continuous probability distribution with a probability density function that gives you a symmetrical bell curve. Simply put, it is a plot of the probability function of a variable that has maximum data concentrated around one point and a few points taper off symmetrically towards two opposite ends.

In this definition of a normal distribution, you will explore the following terms:

1. Continuous Probability Distribution: A probability distribution where the random variable, X, can take any given value, e.g., amount of rainfall. You can record the rainfall received at a certain time as 9 inches. But this is not an exact value. The actual value can be 9.001234 inches or an infinite amount of other numbers. There is no definitive way to plot a point in this case, and instead, you use a continuous value.

2. Probability Density Function: An expression that is used to define the range of values that a continuous random variable can take.

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

Understanding the nature of missing data is critical in determining what treatments can be applied to overcome the lack of data. Data can be missing in the following ways:

1. Missing Completely At Random (MCAR)
2. Missing At Random (MAR)
3. Not Missing At Random (NMAR)

Methods to Handle Missing Data are:-

1. Mean or Median Imputation
2. Multivariate Imputation by Chained Equations (MICE)
3. Random Forest

Imputation Techniques

Mean imputation

Calculate the mean of the observed values for that variable for all non-missing people.It has the advantage of maintaining the same mean and sample size, but it also has a slew of drawbacks. Almost all of the methods described below are superior to mean imputation.

Substitution

Assume the value from a new person who was not included in the sample.To put it another way, pick a new subject and employ their worth instead.

Hot deck imputation

A value picked at random from a sample member who has comparable values on other variables.To put it another way, select all the sample participants who are comparable on other factors, then choose one of their missing variable values at random.

One benefit is that you are limited to just feasible values. In other words, if age is only allowed to be between 5 and 10 in your research, you will always obtain a value between 5 and 10.Another

factor is the random element, which introduces some variation. For exact standard errors, this is crucial.

Cold deck imputation

A value picked deliberately from an individual with similar values on other variables.In most aspects, this is comparable to Hot Deck, but without the random variance. As an example, under the same experimental condition and block, you can always select the third individual.

Regression imputation

The result of regressing the missing variable on other factors to get a predicted value.As a result, instead of utilising the mean, you're relying on the anticipated value, which is influenced by other factors. This keeps the associations between the variables in the imputation model, but not the variability around the anticipated values.

Stochastic regression imputation

The predicted value of a regression plus a random residual value.This has all of the benefits of regression imputation plus the random component's benefits.The majority of multiple imputation is based on stochastic regression imputation.

Interpolation and extrapolation

An estimate based on other observations made by the same person. It generally only works with data that is collected over time.Proceed with caution, though. For a variable like height in children–one that cannot be reduced through time–interpolation would make more sense. Extrapolation entails estimating beyond the data's true range, which necessitates making more assumptions than is necessary.

Single or Multiple Imputation

- Single and multiple imputation are the two forms of imputation. When people say imputation, they usually mean single.
- The term "single" refers to the fact that you only use one of the seven methods to estimate the missing number outlined above.
- It's popular since it's simple to understand and generates a sample with the same number of observations as the complete data set.
- When listwise deletion eliminates a considerable amount of the data set, single imputation appears to be a tempting option. It does, however, have certain restrictions.
- Unless the data is Missing Completely at Random, certain imputation processes, such as means, correlations, and regression coefficients, result in skewed parameter estimations. The bias is frequently worse than with listwise deletion, which is most software's default.

- The level of the bias is determined by a number of factors, including the imputation technique, the missing data mechanism, the fraction of missing data, and the information in the data set.

12. What is A/B testing?

A/B testing is one of the most important concepts in data science and in the tech world in general because it is one of the most effective methods in making conclusions about any hypothesis one may have. A/B testing lets organizations quickly experiment and iterate in order to continually improve their business. In data science, A/B tests can also be used to choose between two models in production, by measuring which model performs better in the real world.

13. Is mean imputation of missing data acceptable practice?

No Mean Imputation of missing data is not acceptable practice.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. This method can lead into severely biased estimates even if data are MCAR

14. What is linear regression in statistics?

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.

This form of analysis estimates the coefficients of the linear equation, involving one or more independent variables that best predict the value of the dependent variable. Linear regression fits a straight line or surface that minimizes the discrepancies between predicted and actual output values. There are simple linear regression calculators that use a “least squares” method to discover the best-fit line for a set of paired data. You then estimate the value of X (dependent variable) from Y (independent variable).

15. What are the various branches of statistics?

There are two main branches of statistics

- Inferential Statistics.

- Descriptive Statistics.

Inferential Statistics:

Inferential statistics used to make inferences and describe the population. These stats are more useful when it's not easy or possible to examine each member of the population.

Descriptive Statistics:

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