STATISTICS WORKSHEET

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**

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**

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**

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 **ANSWER- 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 **ANSWER- 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 **ANSWER- A**

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**

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

ANSWER

The normal distribution, also known as the Gaussian or standard normal distribution, is the probability distribution that plots all of its values in a symmetrical fashion, and most of the results are situated around the probability's mean.

Values are equally likely to plot either above or below the mean.

The normal distribution is a probability distribution that (roughly) describes many common datasets in the real world. It is the most common type of distribution, and it arises naturally in statistics through random sampling techniques.

Nowadays, it is more common to show up as a model for the "lifespan" of a product, like a lightbulb, or the outcome of standardized tests, like IQ. Biological measurements, like height or weight, are often estimated with normal distributions .

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

ANSWER

Missing data is a huge problem for data analysis because it distorts findings. It's difficult to be fully confident in the insights when you know that some entries are missing values. Hence, why they must be addressed. According to data scientists, there are three types of missing data. These are Missing Completely at Random (MCAR) – when data is completely missing at random across the dataset with no discernable pattern. There is also Missing At Random (MAR) – when data is not missing randomly, but only within sub-samples of data. Finally, there is Not Missing at Random (NMAR), when there is a noticeable trend in the way data is missing.

Best techniques to handle missing data

Use deletion methods to eliminate missing data- The deletion methods only work for certain datasets where participants have missing fields. There are several deleting methods – two common ones include Listwise Deletion and Pairwise Deletion. It means deleting any participants or data entries with missing values. This method is particularly advantageous to samples where there is a large volume of data because values can be deleted without significantly distorting readings.

Use regression analysis to systematically eliminate data- Regression is useful for handling missing data because it can be used to predict the null value using other information from the dataset. There are several methods of regression analysis, like Stochastic regression. Regression methods can be successful in finding the missing data, but this largely depends on how well connected the remaining data is. Of course, the one drawback with regression analysis is that it requires significant computing power, which could be a problem if data scientists are dealing with a large dataset.

Data scientists can use data imputation techniques- Data scientists use two data imputation techniques to handle missing data: Average imputation and common-point imputation. Average imputation uses the average value of the responses from other data entries to fill out missing values. However, a word of caution when using this method – it can artificially reduce the variability of the dataset.

Keeping things under control- Missing data is a sad fact of life when it comes to data analytics. We cannot avoid situations like these entirely because there are several remedial steps data scientists need to take to make sure it doesn't adversely affect the analytics process.

12. What is A/B testing?

ANSWER

A/B testing, also known as split testing, refers to a randomized experimentation process wherein two or more versions of a variable (web page, page element, etc.) are shown to different segments of website visitors at the same time to determine which version leaves the maximum impact and drive business metrics.

Essentially, A/B testing eliminates all the guesswork out of website optimization and enables experience optimizers to make data-backed decisions. In A/B testing, A refers to 'control' or the original testing variable. Whereas B refers to 'variation' or a new version of the original testing variable.

The version that moves your business metric(s) in the positive direction is known as the 'winner.' Implementing the changes of this winning variation on your tested page(s) / element(s) can help optimize your website and increase business ROI.

The metrics for conversion are unique to each website. For instance, in the case of eCommerce, it may be the sale of the products. Meanwhile, for B2B, it may be the generation of qualified leads.

A/B testing is one of the components of the overarching process of Conversion Rate Optimization (CRO), using which you can gather both qualitative and quantitative user insights. You can further use this collected data to understand user behavior, engagement rate, pain points, and even satisfaction with website features, including new features, revamped page sections, etc. If you're not A/B testing your website, you're surely losing out on a lot of potential business revenue.

13. Is mean imputation of missing data acceptable practice?

ANSWER

The process of replacing null values in a data collection with the data's mean is known as mean imputation.

Mean imputation is typically considered terrible practice since it ignores feature correlation. Consider the following scenario: 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.

Second, 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 quantifies the relationship between one or more predictor variable(s) and one outcome variable. Linear regression is commonly used for predictive analysis and modeling.

For example, it can be used to quantify the relative impacts of age, gender, and diet (the predictor variables) on height (the outcome variable). Linear regression is also known as multiple regression, multivariate regression, ordinary least squares (OLS), and regression. This post will show you examples of linear regression, including an example of simple linear regression and an example of multiple linear regression.

15. What are the various branches of statistics?

ANSWER

Statistics is a study of presentation, analysis, collection, interpretation and organization of data

There are two main branches of statistics

- Inferential Statistic
- Descriptive Statistic
- Data collection

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.

Data collection :

Data collection is all about how the actual data is collected. For the most part, this needn't concern us too much in terms of the mathematics (we just work with what we are given), but there are significant issues to consider when actually collecting data.