Statistics Worksheet

|  |  |
| --- | --- |
|  | (A) |
|  | (A) |
|  | (B) |
|  | (D) |
|  | (C) |
|  | (B) |
|  | (B) |
|  | (A) |
|  | (D) |

Answer 10.

The standard normal distribution has a mean of 0 and a standard deviation of 1. Many distributions follow a normal distribution and hence many at times datasets are normalized to come closer to a normal distribution since it would be easier to model them.

Answer 11.

When we replace a missing value with a new value derived from the data, using summary statistics or data from other observations it is called imputation and is a common way to handle missing data. In pandas the common methods like **fillna()** is used to fill in empty values. The scikit library also has the module impute for imputing missing data. **impute.SimpleImputer** imputes values in the i-th feature dimension using only non-missing values in that feature dimension whereas **impute.IterativeImputer** uses multivariate imputation algorithms use the entire set of available feature dimensions to estimate the missing values

Answer 12.

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. A/B testing is a form of statistical and two-sample hypothesis testing.

Answer 13.

The 2 problems associated with mean imputation are

(1) results in a lower variance in the dataset which leads to a more narrow confidence interval in the probability distribution

(2) Doesn’t take into account the correlation between features and hence will produce incorrect values

Answer 14.

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. A linear regression line has an equation of the form ***Y = a + bX***, where ***X*** is the explanatory variable and ***Y*** is the dependent variable. The slope of the line is ***b***, and ***a*** is the intercept (the value of ***y*** when ***x*** = 0).

Answer 15.

The two major areas of statistics are known as **descriptive statistics**, which describes the properties of sample and population data, and inferential statistics, which uses those properties to test hypotheses and draw conclusions.