Statistics Worksheet 1 Solution

Question 1: Bernoulli random variables take (only) the values 1 and 0.
Solution: a) True

Question 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?
Solution: a) Central Limit Theorem

Question 3: Which of the following is incorrect with respect to use of Poisson distribution?
Solution: c) Modeling contingency tables

Question 4: Point out the correct statement
Solution: d) All of the mentioned

Question 5: ______ random variables are used to model rates.
Solution: c) Poisson

Question 6: Usually replacing the standard error by its estimated value does change the CLT.
Solution: b) False

Question 7: Which of the following testing is concerned with making decisions using data?
Solution: b) Hypothesis

Question 8: Normalized data are centered at______ and have units equal to standard deviations of the original data
Solution: a) 0

Question 9: Which of the following statement is incorrect with respect to outliers?
Solution: c) Outliers cannot conform to the regression relationship

Question 10: What do you understand by the term Normal Distribution?
Solution:
Normal Distribution is a probability distribution where the occurrence of the data is more towards the mean than far away from it. This is also known as Gaussian Distribution. It follows a Bell-shaped curve. The mean is zero (0) and the standard deviation is one (1). The mean, median and mode are equal. It has a skewness of zero (0).

Question 11: How do you handle missing data? What imputation techniques do you recommend?
Solution:
Missing Data can be handled in many ways depending on the observations where the data is missing.
If only a few observations are missing data in 1 or more columns, deletion technique can be used.
In case there are many missing values, we use imputation techniques.
First is Simple Imputer. Here we can replace the missing values with Mean, median, mode or a constant value
We can also use ML algorithms to implement advanced imputation.
KNNImputer can be used to handle missing values by calculating Euclidean distance.

Question 12: What is A/B testing?
Solution:
A/B testing is a randomized control experiment to compare two versions of a variable in order to find out which version performs better in a controlled environment.
It involves making a hypothesis (null and alternate), creating a control group and test group and then conduct the A/B test and collecting the data.
In this hypothesis testing there are 2 types of errors.

a) **Type – 1 error – Rejecting null hypothesis when its True**
b) **Type – 2 error – Failing to reject Null hypothesis when its false.**

A/B testing works best for incremental changes.

Question 13: Is mean imputation of missing data acceptable practice?
Solution:
It is not a good solution as this would ignore the correlation between the features.
Imputing the missing values by mean decreases the variance, thus increasing bias. Due to reduced variance the accuracy of the model is less and the confidence interval is narrow.

Question 14: What is linear regression in statistics?
Solution:
Linear Regression is a very basic type of predictive analysis. The estimates od the Linear Regression are used to showcase the relationship between a dependent variable and one or more independent variables.
It is defined as y = c + b*x
Where y is estimated dependent variable score, c is a constant and x is the score of the independent variable.

Question 15: What are the various branches of statistics?
Solution:
Statistics is a branch of Maths which involves dealing with data.
There are two branches in Statistics

a) **Descriptive statistics: This is a branch of statistics which deals with collection of data and presenting the data either visually or numerically.**
b) **Inferential Statistics: This branch involves making conclusion about the data**