Statistics worksheet 1

1. Bernoulli random variables takes (only) the variable 1 and 0.
   True
2. Which of the following theorem states that the distribution of average of iid variables, properly normalized, because that of a standard normal as the sample size increase?
   Central limit theorem
3. Which of the following is incorrect in respect with respect to use of poisson distribution?
   Modelling bounded count data is the incorrect answer
4. Point out the correct statement.
   All the mentioned are correct statement.
5. ____ random variables are used to model rates.
   Poisson
6. Usually replacing the standard error by its estimated value does change the CLT.
   False
7. Which of the following testing is concerned with making decision using data?
   Hypothesis
8. Normalized data are centered at _____and have units equal to standard deviations of the original data.
   0
9. Which of the following statement is incorrect with respect to outliers?
   Outliers cannot conform to the regression relationship

10. <u>What do understand by the term normal distribution?</u>
    Normal distribution is a graphical representation in a bell shape curve that implies the mean, median, mode to be converging at the same point. If we see any kind of skewness in left or right side, then the data at hand consist of outliers which can affect the outcome of a model built on such a data. There are various ways to make skewed data into normal distribution data but there are times when an approximate bell curve is formed instead of a perfect one on real time data. The peak of a normal distribution consists of the maximum data points while the bottom part reduces in terms of frequency of data.

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

Missing data or missing values is defined as the data value that is not stored for a variable in the observation of interest. The problematic of missing data is relatively mutual in almost all research and can have a noteworthy effect on the conclusion that can be drawn from the data. The best possible method of handling the missing data is to avert the problem by well planning the study and collecting the data wisely.
The study design should limit the collection of data to those who are contributing to the study.
These are missing completely at random (MCAR) – when data is completely missing at random across the dataset with no noticeable pattern. There is also missing (MAR)- when data is not missing there is a noticeable movement in the way data is missing.
Imputation techniques to handle missing data

- Mean imputation
- Substitution
- Hot deck imputation
- Regression imputation
- Stochastic regression imputation
- Interpolation and extrapolation

12. What is A/B testing?

A/B testing is a test also known as split test, is an experiment for determining which of the different variation of an offline experience performs better by presenting each version to users at random and analyzing the results.

13. Is mean imputation of missing data acceptable practice?

True, imputing the mean preserve the mean of the observed data. So, if the data are missing completely at random, the estimate of the mean remain unbiased. That's a good thing. Plus, by imputing the mean, you can keep your sample size up to the full sample size.

14. What is linear regression in statistics?

    Linear regression is a supervised machine learning algorithm that can be used to predict continuous data. The underlying equation used by linear regression model is y=mx*c where m is the slope and c are the intercept of the best fit line. There are 2 types of linear regression, and they are simple linear regression and multiple linear regression techniques. linear regression is majorly used to predict the labels with the help of one or more feature variable ensuring that the features selected in the equation provide the necessary input to obtain the desired label without creating an overfitting or underfitting model that provides a reasonable accuracy on future prediction.

15. What are the various branches of statistics?

    The two main branches of statistics are descriptive statistics and inferential statistics. Both are employed in scientific analysis of data, and both are equally important part of statistics.