

STATISTICS WORKSHEET-1

1. Answer is (a) – True
2. Answer is (a) – Central Limit Theorem
3. Answer is (b) – Modeling bounded count data.
4. Answer is (d) – All of the mentioned
5. Answer is (c) – Poisson
6. Answer is (b) – False
7. Answer is (b) – Hypothesis
8. Answer is (a) – 0
9. Answer is (c) - Outliers cannot conform to the regression relationship
10. **Normal distribution** is also known as Gaussian distribution. It is a probability distribution that is symmetric about the mean, showing that data near the mean are more frequent in occurrence than data far from the mean. In graphical form, the normal distribution appears as a "Bell curve".
11. **Missing data** can be dealt with in a variety of ways. Imputation is the process of substituting an estimate for missing values and analyzing the entire data set as if the imputed values were the true observed values. To choose that estimate, the following are some of the most prevalent methods-
 - (a) **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.
 - (b) **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.
 - (c) **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.

(d) **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.

(e) **Regression imputation** - The result of regressing the missing variable on other factors to get a predicted value. As a result, instead of utilizing 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.

12. **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 drives business metrics.
13. **Mean imputation of missing data is acceptable** when the missing value proportion is not large enough. But, when the missing values are large enough and you impute them with the mean, the standard errors will be lesser than what they actually would have been.
14. **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.
15. The two main branches of statistics are
 - (A) **Descriptive statistics** - It deals with the presentation and collection of data. This is usually the first part of a statistical analysis.
 - (B) **Inferential statistics** – It involves drawing the right conclusions from the statistical analysis that has been performed using descriptive statistics.