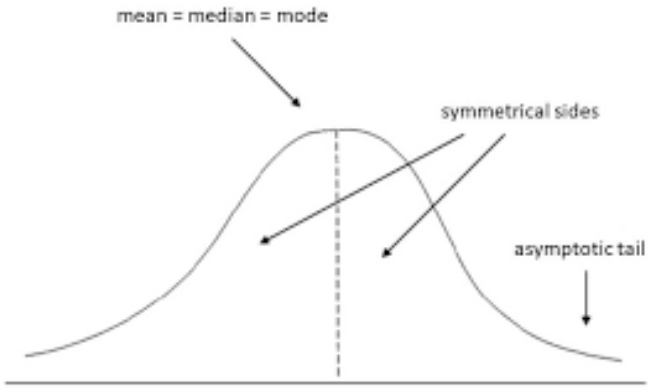
STATISTICS WORKSHEET-1

1. a) True
2. a) Central Limit Theorem
3. b) Modeling bounded count data
4. d) All of the mentioned
5. c) Poisson
6. b) False
7. b) Hypothesis
8. a) 0
9. c) Outliers cannot conform to the regression relationship

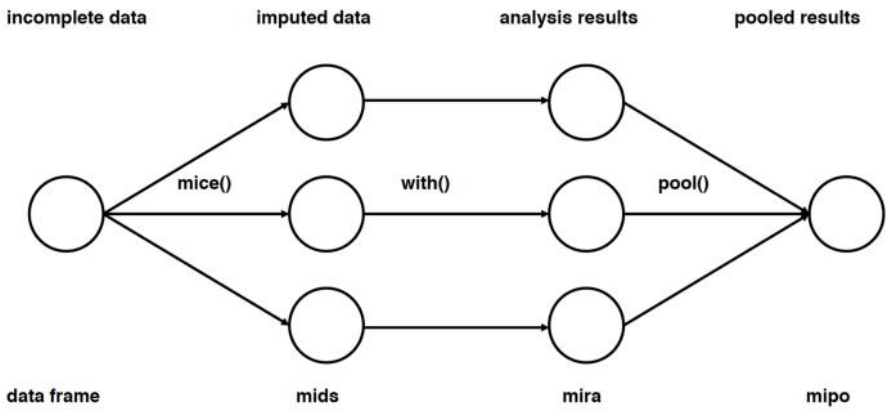
10. The normal distribution, also known as the Gaussian distribution, is a symmetric probability distribution centred on the mean, indicating that data around the mean occur more frequently than data far from it. The normal distribution will show as a bell curve on a graph.



11. When dealing with missing data, we have two options for resolving the problem: imputation or data removal.
    - For missing data, the imputation process generates credible predictions. When the percentage of missing data is low, it's the most beneficial. If the percentage of missing data is too high, the results will be devoid of natural variation, which will make it difficult to build an effective model.

- The other option is to delete information. To eliminate bias when dealing with data that is absent at random, relevant data can be erased. If there aren't enough observations to make a reliable analysis, removing data may not be the best solution. Observation of specific events or factors may be essential in some circumstances.

  Recommended method of imputation for missing data is Multivariate Imputation by Chained Equation (MICE). This method of imputation works by repeatedly filling in the missing data. Multiple Imputations (MIs) are far superior to single imputations in terms of measuring the uncertainty of missing values. The chained equations method is also particularly adaptable, as it can handle a wide range of variables and data types.



12. A basic randomised control experiment is A/B testing. It's a method of comparing two versions of a variable in a controlled environment to see which performs better.
    A/B testing, in essence, eliminates all guessing and allows optimizers to make data-driven judgments. The 'control' or original testing variable is referred to as A in A/B testing. 'Variation,' or a new version of the original testing variable, is denoted by B.

The 'winner' is the variant that advances the metric(s) in a positive direction. Implementing the winning variation's adjustments on your tested page(s) / data(s) then help improve your results.

13. Utilizing mean-imputed variables in statistical analysis has three drawbacks:

- The variance of the imputed variables is reduced using mean imputation.
- Mean imputation reduces standard errors, rendering most hypothesis tests and confidence interval calculations invalid.
- Relationships between variables, such as correlations, are not preserved by mean imputation.

14. By fitting a linear equation to observed data, linear regression seeks to model the relationship between two variables. One variable is regarded as an explanatory variable, while the other is regarded as a dependent variable.

The simplest equation for a linear regression line is $Y = a + bX$, with X as the explanatory variable and Y as the dependent variable. The intercept (the value of y when $x = 0$) is a, while the slope of the line is b.

Three major uses for regression analysis are
(1) determining the strength of predictors,
(2) forecasting an effect, and
(3) trend forecasting

15. The two main branches of statistics are descriptive statistics and inferential statistics.

- Descriptive statistics is concerned with gathering data, presenting it in various formats such as tables, graphs, and diagrams, and calculating averages and other measures to characterize the data.

- Inferential statistics is concerned with approaches for analyzing data, creating estimates, and drawing inferences from restricted data gained through sampling, as well as assessing the estimates' reliability.