STATISTICS WORKSHEET-1

1) Bernoulli random variables take (only) the values 1 and 0

Ans: - **a) True (Because it arises only Binary Outcome)**

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?

Ans: - **a) Central Limit Theorem**

3) Which of the following is incorrect with respect to use of Poisson distribution?

Ans: - **b) Modeling bounded count data (since it used for Unbounded count data)**

4) Point out the correct statement.

Ans: - d**) All of the mentioned**

5) _____ random variables are used to model rates

Ans: - **c) Poisson**

6) Usually replacing the standard error by its estimated value does change the CLT.

Ans: **- b) False**

7) Which of the following testing is concerned with making decisions using data?

Ans: **- b) Hypothesis**

8) Normalized data are centered at_____and have units equal to standard deviations of the original data.

Ans: - **a) 0**

9) Which of the following statement is incorrect with respect to outliers?

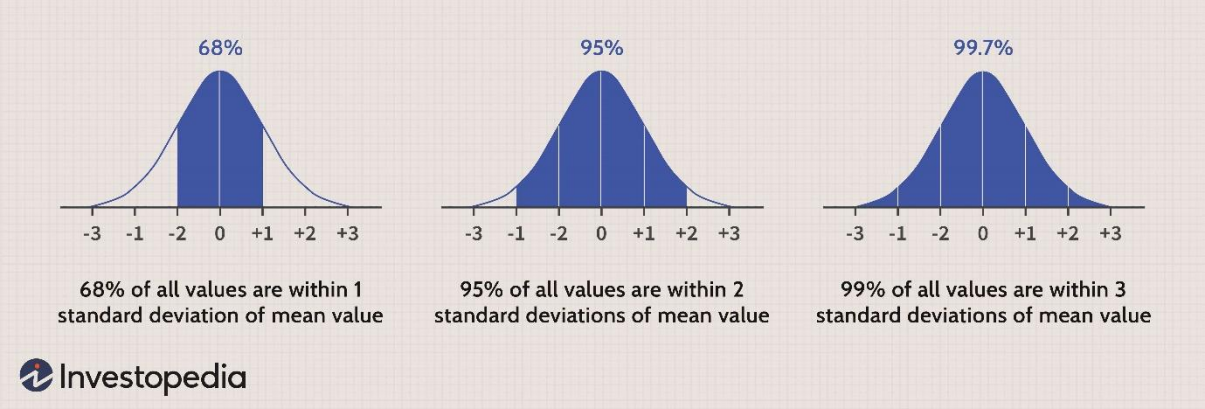
Ans: - **c) Outliers cannot conform to the regression relationship**

10) What do you understand by the term Normal Distribution?

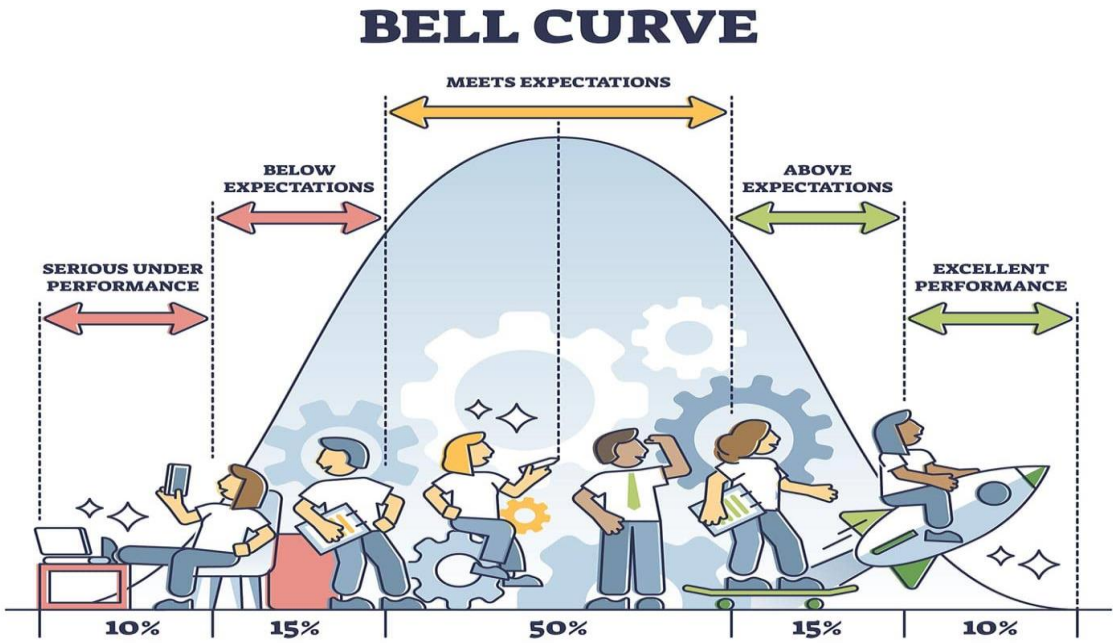
Ans: **- Normal or Gaussian Distribution is called as 'Belled Curve'**. It describes the tendency for data to cluster around the central Value. This central value is population Mean **('μ')** is located in the middle of the curved.

To draw a Normal Distribution, we need to know: -

- Average Measurement (tells us the centred of curve goes)
- Standard Deviation (tells us how wide the curve should be)
- Width of Curve (Wider curve = Shorter Height
  Narrower Curve = Taller Height)



The distribution of data is normal when it is symmetric around the mean, when +-68% of the data lies within one standard deviation of the mean, 95% within two standard deviation of the mean and 99.8% of it within three standard deviations of the mean.



Normal Distribution are always centred at the average value. The width of the curve is defined by the standard deviation. Knowing the standard deviation is helpful because the curve is drawn in such a way that 95% of the measurements fall between +/- 2 standard deviation around the mean.

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

Ans: -Before approaching to handle missing data by imputation technique, first understand it is missing. There are four types of missing data: -

a. **MCAR (Missing Completely at Random): -** In this the data is randomly missing and we have no clue why is it missing, we cannot figure out why is it missing. There is no pattern or logic behind it, this is less commonly scenario seen in industries. Even if we try impute those missing values, we are sure how it will work.

b. **MAR (Missing at Random):** - In this it is missing at random not completely but with some pattern or logic behind it. If one columns value is missing then we can understand the logic behind it and by imputation technique we can guess the value. We can at least try something to find those missing data. Here we are not completely clueless.
c. **MNAR (Missing Not at Random):** - If the character of the data does not meet those of MCAR and MAR then it falls under the MNAR. Here the values are intentionally missing. These data are problematic.
d. **SM (Structured Missing):** - Here we know the exact reason behind those missing data. Compare to above three missing data in this we have more accuracy for the reason behind it. More understanding we have then better imputation we can do for it.

The missing data are Univariate (here only single column has random missing data) and Multivariate (here more than one column has random missing data).

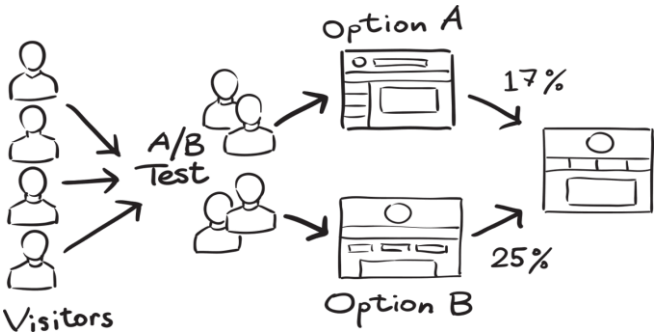
Imputation techniques: -

1. **Listwise Deletion**
2. **Pairwise Deletion**
3. **Mean Substitution**
4. **Regression Imputation**
5. **Last observation carried forward (LOCF)**
6. **Maximum Likelihood**
7. **Expectation-Maximization**
8. **Multiple Imputation**
9. **Sensitivity Analysis**

- The targeted data cannot eliminate the potential bias, more attention should be there towards missing data in the design and performance.
- Best solution is to maximize the data collection. Statistical analysis technique should be only after maximum efforts are applied to reduce missing data.
- Single imputation or LOCF are not optimal for final analysis it can cause bias and led to invalid conclusion.
- It is difficult to know whether Multiple imputation or full maximum likelihood is best, both are superior approaches.
- Multiple Imputation is good to approach generally.

12) What is A/B testing?

Ans: -



A/B testing is also known as **split testing**. It is way to compare two or more variable to find out which performs and leaves maximum impact in business metrics. Here **A** refers to **'Control'** or the original testing and **B** refers to the **'variation'** of a new version of original testing variable.

It is hypothetical testing for making decisions that estimate population parameter based on sample statistics. It is one of the components using which we can gather the qualitative and quantitative user insight. We can further use this collected data to understand user's behaviour, pain points, engagement rate and even satisfaction.

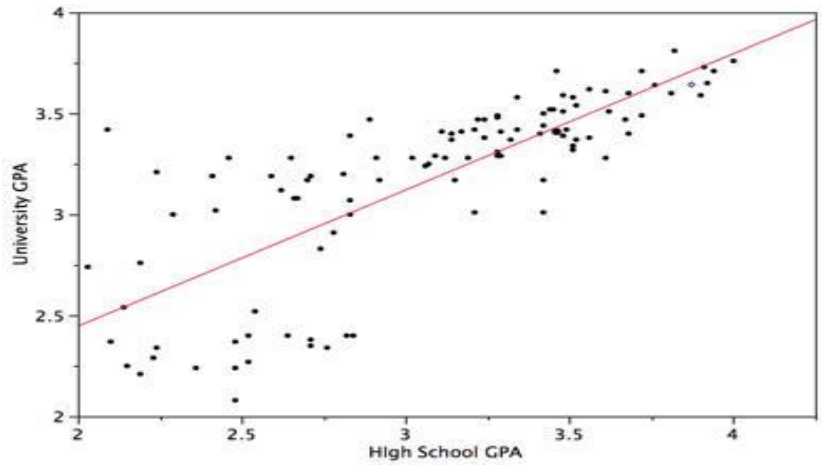
13) Is mean imputation of missing data acceptable practice?

Ans: - The replacing of null values in data collection with the data' mean is Mean Imputation. It is basically considered the terrible practise since it ignores the feature correlation. It decreases the variance of our data while increasing bias, as a result model is less accurate and confidence is narrower.

- Mean imputation does not preserve the relationships among variables.
- Mean Imputation Leads to An Underestimate of Standard Errors.

14) What is linear regression in statistics?

Ans: - It is a regression model that estimates the relationship between one independent variable and one dependent variable using a straight line. Below one is just an example of linear regression graph.



We can use simple linear regression when you want to know: -

- Find how strong is the relationship is between two variables.
- The value of the dependent variable at a certain value of the independent variable
- The size of the error in our prediction doesn't change significantly across the values of the independent variable.
- The observations in the dataset were collected using statistically valid sampling methods, and there are no hidden relationships among observations.
- The data follows a normal distribution.
- The relationship between the independent and dependent variable is **linear.**

15) What are the various branches of statistics?

Ans: -There are two main branches of statistics
Inferential Statistic.
Descriptive Statistic.

Inferential Statistics:
Inferential statistics used to make inference and describe about the population. These stats are more useful when it's not easy or possible to examine each member of the population. It involves drawing conclusions or generalizations or making predictions from the gathered data. The types of it as: -

- Regression analysis
- Analysis of variance (ANOVA)
- Analysis of covariance (ANCOVA)
- Statistical significance (t-test)
- Correlation analysis

Descriptive Statistics:
Descriptive statistics are used to get a brief summary of data. You can have the summary of data in numerical or graphical form. It is involved with gathering and organizing data so that it can be analyzed and presented.