# Worksheet-1

## STATISTICS ASSIGNMENT

Q1. a) True

Q2. a) Central Limit Theorem

Q3. b) Modeling bounded count data

Q4. d) All of the mentioned

Q5. c) Poisson

Q6. b) False

Q7. b) Hypothesis

Q8. a) 0

Q9. c) Outliers cannot conform to the regression relationship

Q10. What do you understand by the term Normal Distribution?

Ans. Normal Distribution is a bell-shaped frequency distribution curve which helps describe all the possible values a random variable can take within a given range with most of the distribution area is in the middle and few are in the tails, at the extremes. This distribution has two key parameters: the mean (μ) and the standard deviation (σ) which plays key role in assets return calculation and in risk management strategy.

Q11. How do you handle missing data? What imputation techniques do you recommend?

Ans. Many real-world datasets may contain missing values for various reasons. They are often encoded as NaNs, blanks or any other placeholders. Training a model with a dataset that has a lot of missing values can drastically impact the machine learning model's quality. Some algorithms such as *scikit-learn estimators* assume that all values are numerical and have and hold meaningful value.

One way to handle this problem is to get rid of the observations that have missing data. However, you will risk losing data points with valuable information. A better strategy would be to impute the missing values. In other words, we need to infer those missing values from the existing part of the data. There are three main types of missing data:

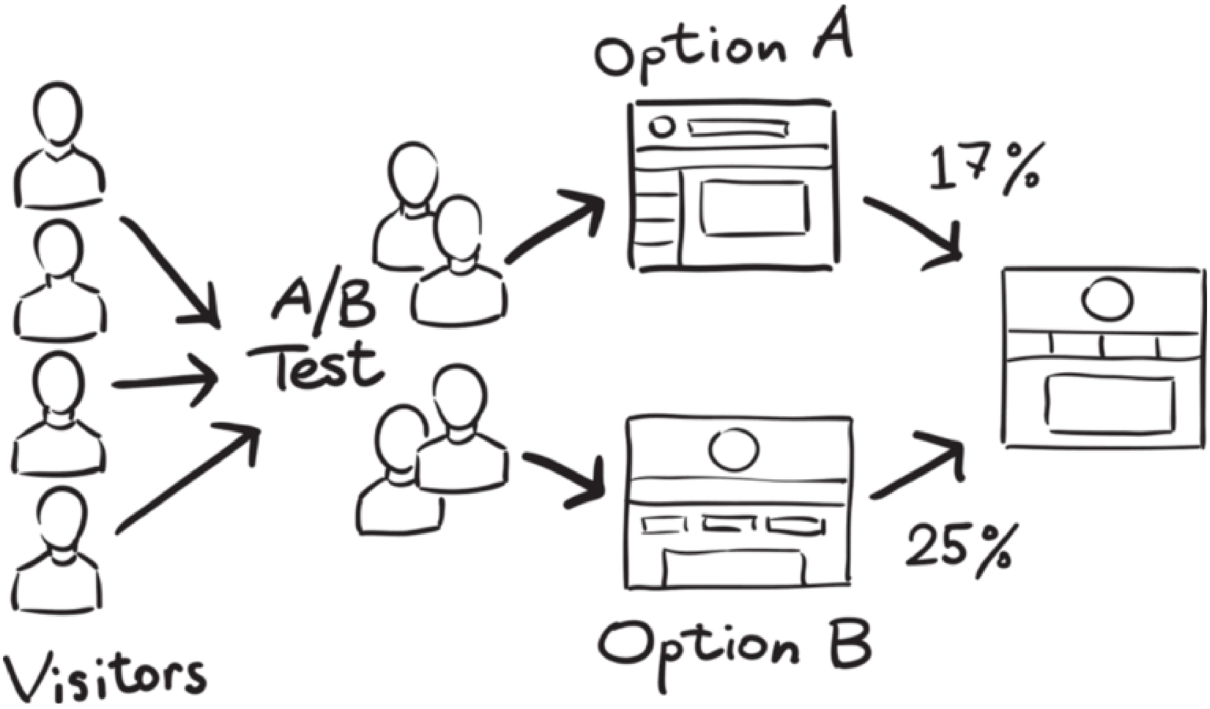
- Missing completely at random (MCAR)
- Missing at random (MAR)
- Not missing at random (NMAR)

The popular ways for data imputation for datasets:

1. Do nothing.
2. Imputation using mean/median values.
3. Imputation using most frequent or zero/constant values.
4. Imputation using KNN.
5. Imputation using Multivariate Imputation by Chained Equations (MICE).
6. Imputation using Deep Learning.

Q12. What is A/B testing?

Ans. A/B testing is a basic randomized control experiment. It is a way to compare the two versions of a variable to find out which performs better in a controlled environment.
For instance, let's say you own a company and want to increase the sales of your product. Here, either you can use random experiments, or you can apply scientific and statistical methods. A/B testing is one of the most prominent and widely used statistical tools.
In the above scenario, you may divide the products into two parts – A and B. Here A will remain unchanged while you make significant changes in B's packaging. Now, on the basis of the response from customer groups who used A and B respectively, you try to decide which is performing better.



Q13. Is mean imputation of missing data acceptable practice?

Ans.

- Bad practice in general
- If just estimating means: mean imputation preserves the mean of the observed data
- Leads to an underestimate of the standard deviation
- Distorts relationships between variables by “pulling” estimates of the correlation toward zero

Q14. What is linear regression in statistics?

Ans. Regression analysis is a statistical method to model the relationship between a dependent (target) and independent (predictor) variables with one or more independent variables. More specifically, Regression analysis helps us to understand how the value of the dependent variable is changing corresponding to an independent variable when other independent variables are held fixed. It predicts continuous/real values such as temperature, age, salary, price, etc.

We can understand the concept of regression analysis using the below example:

Example: Suppose there is a marketing company A, who does various advertisement every year and get sales on that. The below list shows the advertisement made by the company in the last 5 years and the corresponding sales:

| Advertisements | Sales |
|---|---|
| $90 | $1000 |
| $120 | $1300 |
| $150 | $1800 |
| $100 | $1200 |
| $130 | $1380 |
| $200 | ?? |

Now, the company wants to do the advertisement of $200 in the year 2019 and wants to know the prediction about the sales for this year. So, to solve such type of prediction problems in machine learning, we need regression analysis.

Regression is a supervised learning technique which helps in finding the correlation between variables and enables us to predict the continuous output variable based on the one or more predictor variables. It is mainly used for prediction, forecasting, time series modelling, and determining the causal-effect relationship between variables.

In Regression, we plot a graph between the variables which best fits the given datapoints, using this plot, the machine learning model can make predictions about the data. In simple words, "Regression shows a line or curve that passes through all the datapoints on target-predictor graph in such a way that the vertical distance between the datapoints and the regression line is minimum." The distance between datapoints and line tells whether a model has captured a strong relationship or not.

Q15. What are the various branches of statistics?

Ans. There are mainly two branches of the statistics:

- Descriptive Statistics
- Inferential Statistics

Descriptive statistics is divided into two parts:

a. Central Tendency Measures
b. Variability Measures

The central tendency measures are:

1. Mean
2. Median
3. Mode

The variability measures are:

1. Quartiles,
2. Ranges,
3. Variances, and
4. Standard Deviation.

Inferential statistics include:

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