Q.1 ANS - a) True

Q.2 ANS - a) Central Limit Theorem

Q.3 ANS - b) Modeling bounded count data

Q.4 ANS - d) All of the mentioned

Q.5 ANS - c) Poisson

Q.6 ANS - b) False

Q.7 ANS- b) Hypothesis

Q.8 ANS- a) 0

Q.9 ANS- c) Outliers cannot conform to the regression relationship

**Q.10. What do you understand by the term Normal Distribution?**

ANS – A normal distribution of data has bell shape curve where mean and median are equal , also for perfect normal distribution mean is 0 and each standard deviation is by 1 unit on both left and right side of the mean. In normal distribution 68% of data falls within 1st standard deviation, 95% of data falls within 2nd standard deviation and 99.7% of data falls within 3rd standard deviation of mean.

Less the range of distribution more it is normalized so, Normal distribution has less range as compared to Non-normal distribution, also unlike Non-normal distribution where there is right or left skew ness, Normal distribution is symmetric on both sides of mean.

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

ANS- I handle missing data in following ways-

1. **Dropping columns or variables** –

This is not the best practice as we may lose some valuable information but I do it in following cases.

* More than 90-95% data is missing in column.
* I check multi-correlation and output correlation of that column and if one of these is less then I drop that column.
* In case of column with large no of missing values , I also check variance of the column which reflect whether that column is showing any significant trend or not and if I found it nearby 0 I opt to drop it.

1. **Simple Imputer –**

Simple imputer using mean, median or mode is very basic imputation method. It is very fast and sometimes gives good result but the disadvantages outnumber the advantages. The first main disadvantage is that it does not take any advantage of the time series characteristics or relationship between the variables or columns, also the mean imputation reduces variance in the dataset.

There are numbers of Imputation techniques present like Average imputation, Common-Point imputation, **Regression substitution**, **multiple imputations** and **KNN (K Nearest Neighbors)** Imputation which have their own advantages and disadvantages.

Among above imputation techniques Multiple Imputations and KNN are the widely used and preferred techniques and I also recommend these two.

**Multiple Imputation**

1. Imputation: Impute the missing entries of the incomplete data sets m times (m=3 in the figure). Note that imputed values are drawn from a distribution. Simulating random draws doesn’t include uncertainty in model parameters. Better approach is to use Markov Chain Monte Carlo (MCMC) simulation. This step results in m complete data sets.
2. Analysis: Analyze each of the m completed data sets.
3. Pooling: Integrate the m analysis results into a final result

**KNN Imputation**

In this method, k neighbors are chosen based on some distance measure and their average is used as an imputation estimate. The method requires the selection of the number of nearest neighbors, and a distance metric. KNN can predict both discrete attributes (the most frequent value among the k nearest neighbors) and continuous attributes (the mean among the k nearest neighbors)  
The distance metric varies according to the type of data:  
1. Continuous Data: The commonly used distance metrics for continuous data are Euclidean, Manhattan and Cosine  
2. Categorical Data: Hamming distance is generally used in this case. It takes all the categorical attributes and for each, count one if the value is not the same between two points. The Hamming distance is then equal to the number of attributes for which the value was different.  
One of the most attractive features of the KNN algorithm is that it is simple to understand and easy to implement. The non-parametric nature of KNN gives it an edge in certain settings where the data may be highly “unusual”.  
One of the obvious drawbacks of the KNN algorithm is that it becomes time-consuming when analyzing large datasets because it searches for similar instances through the entire dataset. Furthermore, the accuracy of KNN can be severely degraded with high-dimensional data because there is little difference between the nearest and farthest neighbor.

**12. What is A/B testing?**

ANS - A/B testing (also known as split testing) is a process or technique of website optimization.

The more optimized the website will be the more visitors will take action on it for the business purpose and which in turn will result into the more ‘conversion rate’.

One of the most important ways to optimize website’s funnel in digital marketing is A/B testing.

A/B testing (also known as split testing) is a process of showing two variants of the same web page to different segments of website visitors at the same time and comparing which variant drives more conversions.

**Key Benefits –**

1. Reduce bounce rates

2. Solve visitor pain points

3. More conversions by investing less, etc.

**13. Is mean imputation of missing data acceptable practice?**

ANS – It depends on how much data we are missing to decide whether it is good practice or not to apply mean imputation.

Mean imputation is the most popular solution to missing data perhaps because it is easy, also there are some other benefits as listed below-

Imputing the mean preserves the mean of the observed data.  So, if the data are [missing completely at random](https://www.theanalysisfactor.com/causes-of-missing-data/), the estimate of the mean remains unbiased.

Plus, by imputing the mean, we are able to keep sample size up to the full sample size.

If the missing data is very less compared to non-missing data then we can opt for mean imputation but in general mean imputation is not considered a good practice because there are many [alternatives to mean imputation](https://www.theanalysisfactor.com/missing-data-two-recommended-solutions/) that provide much more accurate estimates and standard errors, for example multiple imputation.

The drawbacks out weights the benefits, some of the drawbacks of mean imputation are listed as below –

### Mean imputation does not preserve the relationships among variables.

### Mean Imputation Leads to An Underestimate of Standard Errors.

If the data are missing completely at random, mean imputation will not bias our parameter estimate but it will still bias our standard error.

**14. What is linear regression in statistics?**

ANS – Linear regression is the simplest and most extensively used statistical technique for predictive modelling analysis. It is a way to explain the relationship between a dependent variable (target) and one or more explanatory variables (predictors) using a straight line. There are two types of linear regression – Simple and Multiple

**15. What are the various branches of statistics?**

ANS – There are two branches of statistics as below –

1. **Descriptive Statistics** - Descriptive [statistics](https://www.investopedia.com/terms/s/statistics.asp) are brief descriptive coefficients that summarize a given data set, which can be either a representation of the entire or a sample of a population.

Descriptive statistics cannot be used to make inference or prediction. While these descriptive are helpful in understand the attributes of data, inferential statistical techniques - a separate branch of statistics - are required to understand how variables interact with one another in a data set.

* Measure of central tendency – Mean, Median and Mode
* Measure of variability (Spread) – Standard deviation, Variance, Skew ness etc.

1. **Inferential Statistics** - Inferential statistics allows us to make predictions (“inferences”) from the data. With inferential statistics, we take data from [samples](https://www.statisticshowto.com/sample/)and make generalizations about a [population](https://www.statisticshowto.com/what-is-a-population/).

* Z-Score
* T-test
* Chi- Square test
* ANOVA
* Other hypothesis tests

Descriptive statistics describes data (for example, a chart or graph) and inferential statistics allows us to make predictions (“inferences”) from that data.  With inferential statistics, we take data from [samples](https://www.statisticshowto.com/sample/)and make generalizations about a [population](https://www.statisticshowto.com/what-is-a-population/).