10. What do you understand by the term normal distribution?

Normal distribution is the statistical tool used to find the actual data collected value falls into a prospective category. In statistics, z-test is being subject to analysis for the N number of sampling data selected from the population. It is usually involves, the average value of data that usually fall in.

For example:

The PURCHASE of organic vegetables is usually done by 1.BPL 2.APL 3.MIDDLE INCOME 4.UPPER INCOME

The answer is clear and foreseen is UPPER INCOME. Hence most data fall in upper income; rather some values may fall in middle income who are conscious about the health; very few who themselves grow on their own can fall in APL category.

The UPPER INCOME will be the MEAN, MEDIAN and MODE at the centre of the distribution called MEAN. While MIDDLE INCOME and APL will fall either side of the UPPER INCOME group, so called STANDARD DEVIATIONS.

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

The Python libraries represent missing numbers as **nan** which is short for "not a number".

Drop Columns with Missing Values

Imputation

An extension to imputation

Most libraries (including scikit-learn) will give you an error if you try to build a model using data with missing values. So you'll need to choose one of the strategies below.

**Imputation Using (Mean/Median) Values**

**Imputation Using (Most Frequent) or (Zero/Constant) Values**

**Imputation Using k-NN**

**Imputation Using Multivariate Imputation by Chained Equation (MICE)**

**Imputation Using Deep Learning ([Datawig](https://github.com/awslabs/datawig))**

**Stochastic regression imputation**

**Extrapolation and Interpolation**

**Hot-Deck imputation**

Cold-deck imputation

12. what is A/B testing?

A/B testing is the randomized controlled experiment. It is tool in decision making towards which direction the results have to be in best outcomes.

Eg: A Factory needs to buy one of the parts from vendor or to have their own production if the raw material is available in the factory. Their cost can be calculated in the different aspects and can be subjected to either scientific or statistical tool for the conclusion to be made by for the cost efficiency.

It also involved in finding out the hypothesis drafted for the definite problems with unified solutions or multiple variation that is subjected to do find out the best solution for the definite problem.

A/B testing works best when testing incremental changes, such as UX changes, new features, ranking, and page load times. Here you may compare pre and post-modification results to decide whether the changes are working as desired or not.

A/B testing doesn’t work well when testing major changes, like new products, new branding, or completely new user experiences. In these cases, there may be effects that drive higher than normal engagement or emotional responses that may cause users to behave in a different manner.

13. Is mean imputation of missing data acceptable practice?

**advantages of the method**:

1. Missing values in your data **do not reduce your sample size**, as it would be the case with [listwise deletion](https://statisticsglobe.com/listwise-deletion-missing-data/) (the default of many statistical software packages, e.g. R, Stata, SAS or SPSS). Since mean imputation replaces all missing values, you can keep your whole database.
2. Mean imputation is very **simple to understand and to apply** (more on that later in the R and SPSS examples). You can explain the imputation method easily to your audience and everybody with basic knowledge in statistics will get what you’ve done.
3. If the [response mechanism is MCAR](https://statisticsglobe.com/missing-data/#response-mechanisms), the **sample mean of your variable is not biased**. Mean substitution might be a valid approach, in case that the univariate average of your variables is the only metric your are interested in.

**drawbacks of mean imputation**:

1. Mean substitution leads to **bias in multivariate estimates** such as correlation or regression coefficients. Values that are imputed by a variable’s mean have, in general, a correlation of zero with other variables. Relationships between variables are therefore biased toward zero.
2. **Standard errors and variance** of imputed variables are biased. For instance, let’s assume that we would like to calculate the standard error of a mean estimation of an imputed variable. Since all imputed values are exactly the mean of our variable, we would be too sure about the correctness of our mean estimate. In other words, the [confidence interval](https://www.mathsisfun.com/data/confidence-interval.html) around the point estimation of our mean would be too narrow.
3. If the [response mechanism is MAR or MNAR](https://statisticsglobe.com/missing-data/#response-mechanisms), even the **sample mean of your variable is biased** (compare that with point 3 above). Assume that you want to estimate the mean of a population’s income and people with high income are less likely to respond; Your estimate of the mean income would be biased downwards.

There are a few advantages, but many serious drawbacks. On top of that, we can also benefit from the advantages with more advanced [imputation methods](https://statisticsglobe.com/missing-data-imputation-statistics/) (e.g. [predictive mean matching](https://statisticsglobe.com/predictive-mean-matching-imputation-method/) or [stochastic regression imputation](https://statisticsglobe.com/regression-imputation-stochastic-vs-deterministic/)). To make it short, there is basically no excuse for using mean imputation.

14. What is linear regression in statistics?

Linear regression is a basic and commonly used type of predictive analysis.  The overall idea of regression is to examine two things: (1) does a set of predictor variables do a good job in predicting an outcome (dependent) variable?  (2) Which variables in particular are significant predictors of the outcome variable, and in what way do they–indicated by the magnitude and sign of the beta estimates–impact the outcome variable?  These regression estimates are used to explain the relationship between one dependent variable and one or more independent variables.  The simplest form of the regression equation with one dependent and one independent variable is defined by the formula y = c + b\*x, where y = estimated dependent variable score, c = constant, b = regression coefficient, and x = score on the independent variable

Naming the Variables.  There are many names for a regression’s dependent variable.  It may be called an outcome variable, criterion variable, endogenous variable, or regressand.  The independent variables can be called exogenous variables, predictor variables, or regressors.

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

15. What are various branches of statistics?

The statistics are divided into two branches namely

1.descriptive statistics - as name given, involves collection, summarizing and description of the data.

1. measure of frequency

2. measure of central tendency

3. measure of dispersion or variation

4. measures of position

2. inferential statistics – drawing conclusion from the given data and reporting.

1.correlation

2. regression

3.analysis of variance

4.hypothesis testing