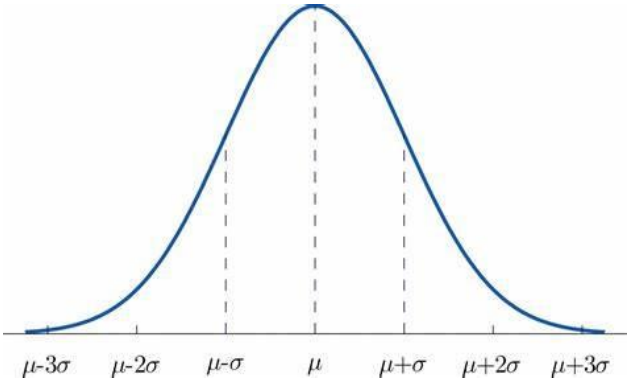
Statistics (Answer)

1. a
2. a
3. b
4. d
5. c
6. b
7. b
8. a
9. c

From 10-15 subjective questions: -

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

Answer – It is also called as bell-shaped curve because it is a symmetrical curve that looks like a bell.



Data near the mean are more frequent in occurrence than the data far from the mean.

In normal distribution –

1) The mean is zero

2) Standard deviation is 1.

The normal distribution model is motivated by the CENTRAL LIMIT THEOREM, this theory states that averages calculated from independent, identically distributed random variables have approximately normal distributions, regardless of the type of distribution from which the variables are sampled provided it has finite variance.

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

Answer – handling missing data is done by imputation technique. Handling missing data is one of important things for data so it should be done in a way that has a best possible value for the place we are filling it.

Missed data is shown by data.isnull().sum() - we can get how much data is missing is columns.

There is some Advanced technique which is used for imputation of missing data,

1)**Iterative imputer** -

This method treats other column which are not having null values it takes it as feature columns and which are having null value take it as label column and train on them.

Finally, it will predict the nan data & impute. It's just like a regression problem

Here, null column is label & non-null column as features.

2)**KNN imputer**-

KNN imputer will try to find the relation with the other columns and impute the data according the relation with other columns.

(n-neighbour: means nearest how many data point to be selected)

3)**Simple Imputer-**

Simple imputer will take mean of a column and replace it with the missing value.

Mean is default of strategy we can change it to mode or median as per our data classification.

Basic method is fillna, which is not good technique for more number of rows.

13. What is A/B testing?

ANSWER- A/B testing is also known as **Statistical hypothesis testing**

Hypothesis - making a guess (not a wild guess) based on assumption without scientific proof or explaining the situation based on reasonable assumptions.

Hypothesis is breakdown into

- Null hypothesis (Ho) - Decision always leads to status quo. Current status/assumption does not change
- Alternative hypothesis (Ha) - Decision lead to opposite of Ho

Statistical tools -

1) ttest

2) ANOVA (analysis of variance)

3) chi-square test (chi2 test)

14) Is mean imputation of missing data acceptable practice?

Answer – No, it's not a good technique because missing value can be dependent on other columns and can be categorical data so it's not a good impute value using mean. Best approaches can be iterative imputer and knn imputer.

15) What is linear regression in statistics?

Answer - In statistics, linear regression is a linear approach for modelling the relationship between a scalar response and one or more explanatory variables (also known as dependent and independent variables).

When you choose to analyze your data using linear regression, part of the process involves checking to make sure that the data you want to analyze can actually be analyzed using linear regression. You need to do this because it is only

appropriate to use linear regression if your data "passes" six assumptions that are required for linear regression to give you a valid result.

Assumption1) Your two variables should be measured at the continuous level that is they are either interval or ratio variables

Assumption 2) There needs to be a linear relationship between the two variables

Assumption 3) There should be no significant outliers

Assumption 4) You should have independence of observations, which you can easily check using the Durbin-Watson statistic

Durbin-Watson statistics - The Durbin Watson statistic is a test statistic used in statistics to detect autocorrelation in the residuals from a regression analysis

Assumption 5) Your data needs to show homoscedasticity, which is where the variances along the line of best fit remain similar as you move along the line

Assumption 6) Finally, you need to check that the residuals (errors) of the regression line are approximately normally distributed.

16) What are the various branches of statistics?

Answer – branches of statistics are:

1) **Descriptive statistics** - it is the part of statistics that deals with presenting the data we have. This can take two basic forms – presenting aspects of the data either visually (via graphs, charts, etc.) or numerically (via averages and so on)

2) **Inferential statistics** - Inferential statistics is the aspect that deals with making conclusions about the data.

3) **Data collection** - Data collection is all about how the actual data is collected.