Parth Makwana (DS0722)

STATISTICS

Q1 to Q9

1. Which of the following is the correct formula for total variation?
b) Total Variation = Residual Variation + Regression Variation

2. Collection of exchangeable binary outcomes for the same covariate data are called outcomes.
c) binomial

3. How many outcomes are possible with Bernoulli trial?
a) 2

4. If Ho is true and we reject it is called
a) Type-I error

5. Level of significance is also called:
b) Size of the test

6. The chance of rejecting a true hypothesis decreases when sample size is:
b) Increase

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

8. What is the purpose of multiple testing in statistical inference?
d) All of the mentioned

9. Normalized data are centred at and have units equal to standard deviations of the original data
a) 0

Q10 and Q15

10. What Is Bayes' Theorem?
Bayes' theorem describes the probability of occurrence of an event related to any condition. It is also considered for the case of conditional probability. Bayes theorem is also known as the formula for the probability of "causes". For example: if we have to calculate the probability of taking a blue ball from the second bag out of three different bags of balls, where each bag contains three different colour balls viz. red, blue, black. In this case, the probability of occurrence of an event is calculated depending on other conditions is known as conditional probability.

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$

Where:

P(A|B) – the probability of event A occurring, given event B has occurred
P(B|A) – the probability of event B occurring, given event A has occurred
P(A) – the probability of event A
P(B) – the probability of event B

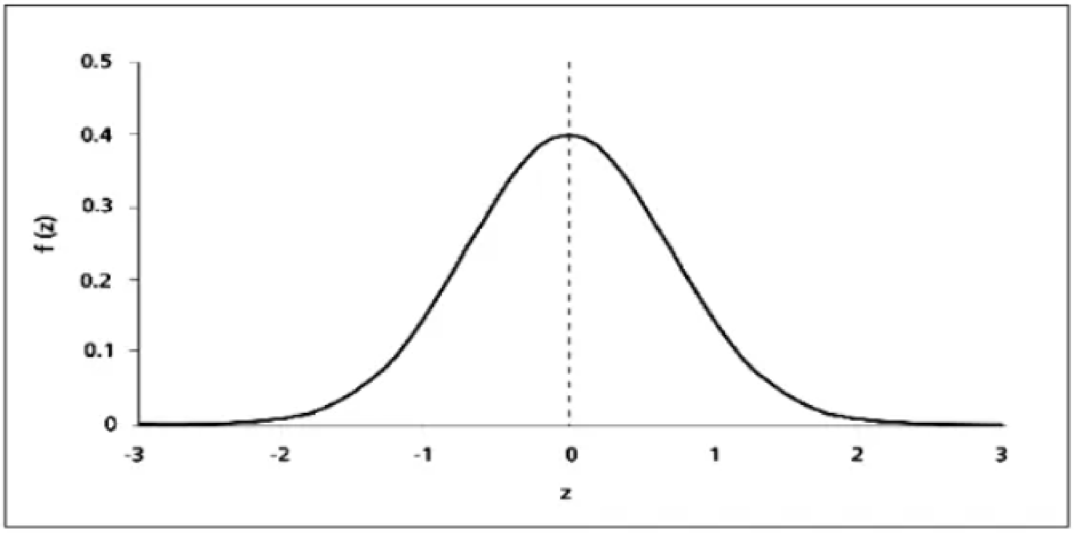
Besides statistics, the Bayes' theorem is also used in various disciplines, with medicine and pharmacology as the most notable examples. In addition, the theorem is commonly employed in different fields of finance.

Some of the applications include but are not limited to, modeling the risk of lending money to borrowers or forecasting the probability of the success of an investment.

11. What is z-score?

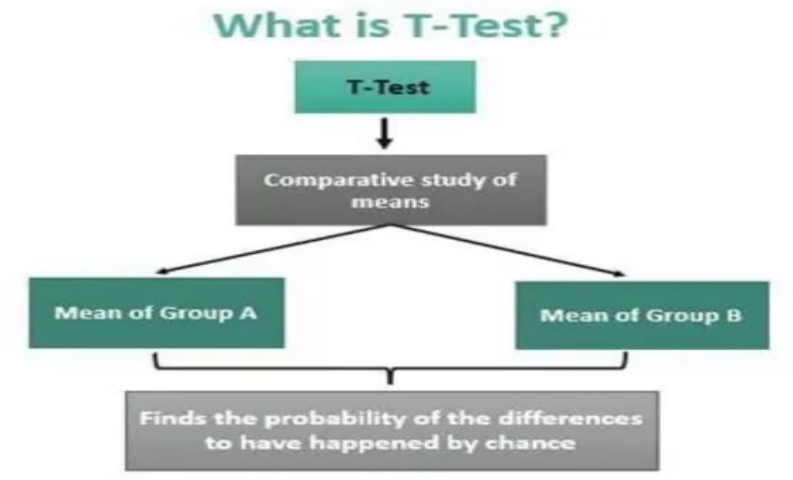
A z-score describes the position of a raw score in terms of its distance from the mean, when measured in standard deviation units. The z-score is positive if the value lies above the mean, and negative if it lies below the mean.

It is also known as a standard score, because it allows comparison of scores on different kinds of variables by standardizing the distribution. A standard normal distribution (SND) is a normally shaped distribution with a mean of 0 and a standard deviation (SD) of 1



12. What is t-test?

A T-test is the final statistical measure for determining differences between two means that may or may not be related. The testing uses randomly selected samples from the two categories or groups. It is a statistical method in which samples are chosen randomly, and there is no perfect normal distribution


What is T-Test?
T-Test
Comparative study of means
Mean of Group A
Mean of Group B
Finds the probability of the differences to have happened by chance

The type of T-test to be conducted is decided by whether the samples to be analyzed are from the same category or distinct categories. The inference obtained in the process indicates the probability of the mean differences to have happened by chance. The test is useful when comparing population age, length of crops from two different species, student grades, etc.

13. What is percentile?

A percentile (or a centile) is a measure used in statistics indicating the value below which a given percentage of observations in a group of observations fall. For example, the 20th percentile is the value (or score) below which 20% of the observations may be found.

The term percentile and the related term percentile rank are often used in the reporting of scores from norm-referenced tests.

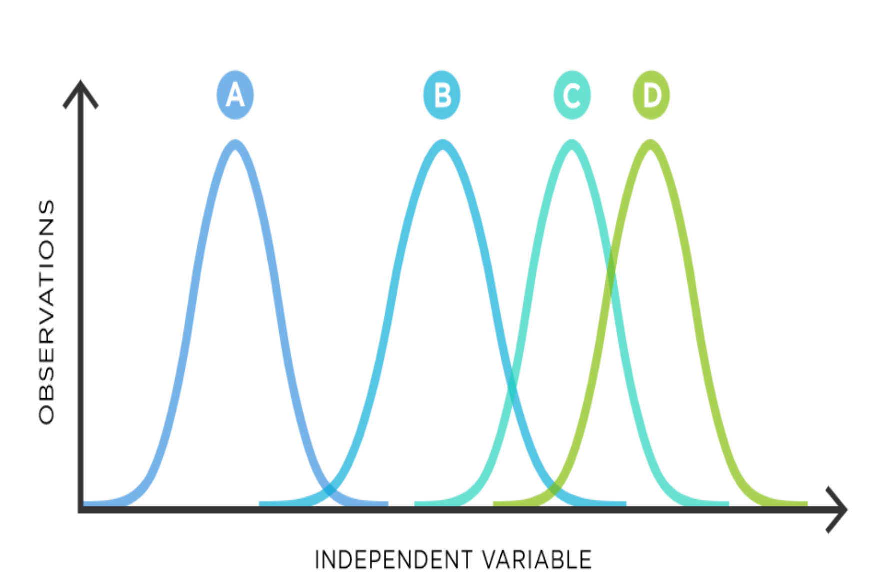
In general, percentiles and quartiles are specific types of quantiles.

It is used with the median value to report data that are markedly non-normally distributed.

Parth Makwana (DS0722)

14. What is ANOVA?

Analysis of Variance (ANOVA) is a statistical formula used to compare variances across the means (or average) of different groups. A range of scenarios use it to determine if there is any difference between the means of different groups.



For example, to study the effectiveness of different diabetes medications, scientists design and experiment to explore the relationship between the type of medicine and the resulting blood sugar level. The sample population is a set of people. We divide the sample population into multiple groups, and each group receives a particular medicine for a trial period. At the end of the trial period, blood sugar levels are measured for each of the individual participants. Then for each group, the mean blood sugar level is calculated.

ANOVA helps to compare these group means to find out if they are statistically different or if they are similar.

The outcome of ANOVA is the 'F statistic'. This ratio shows the difference between the within group variance and the between group variance, which ultimately produces a figure which allows a conclusion that the null hypothesis is supported or rejected. If there is a significant difference between the groups, the null hypothesis is not supported, and the F-ratio will be larger.

ANOVA Terminology

- **Dependent variable:** This is the item being measured that is theorized to be affected by the independent variables.
- **Independent variable/s:** These are the items being measured that may have an effect on the dependent variable.
- **A null hypothesis (H0):** This is when there is no difference between the groups or means. Depending on the result of the ANOVA test, the null hypothesis will either be accepted or rejected.
- **An alternative hypothesis (H1):** When it is theorized that there is a difference between groups and means.
- **Factors and levels:** In ANOVA terminology, an independent variable is called a factor which affects the dependent variable. Level denotes the different values of the independent variable that are used in an experiment.
- **Fixed-factor model: Some** experiments use only a discrete set of levels for factors. For example, a fixed-factor test would be testing three different dosages of a drug and not looking at any other dosages.
- **Random-factor model:** This model draws a random value of level from all the possible values of the independent variable.

15. How can ANOVA help?

One of the biggest challenges in machine learning is the selection of the most reliable and useful features that are used in order to train a model. ANOVA helps in selecting the best features to train a model. ANOVA minimizes the number of input variables to reduce the complexity of the model. ANOVA helps to determine if an independent variable is influencing a target variable.

Parth Makwana (DS0722)

An example of ANOVA use in data science is in email spam detection. Because of the massive number of emails and email features, it has become very difficult and resource-intensive to identify and reject all spam emails. ANOVA and f-tests are deployed to identify features that were important to correctly identify which emails were spam and which were no