STATISTICS WORKSHEET-3

Q1 to Q9 have only one correct answer. Choose the correct option to answer your question.

1. Which of the following is the correct formula for total variation?

Answer: b) Total Variation = Residual Variation + Regression Variation

2. Collection of exchangeable binary outcomes for the same covariate data are called ________ Outcomes

Answer: c) binomial

3. How many outcomes are possible with Bernoulli trial?

Answer: a) 2

4. If Ho is true and we reject it is called

Answer: a) Type-I error

5. Level of significance is also called:

Answer: a) Power of the test

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

Answer: b) Increase

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

Answer: b) Hypothesis

8. What is the purpose of multiple testing in statistical inference?

Answer: d) All of the mentioned

9. Normalized data are centered at and have units equal to standard deviations of the Original data

Answer: a) 0

Q10 to Q15 are subjective answer type questions, Answer them in your own Words briefly.

STATISTICS WORKSHEET-3

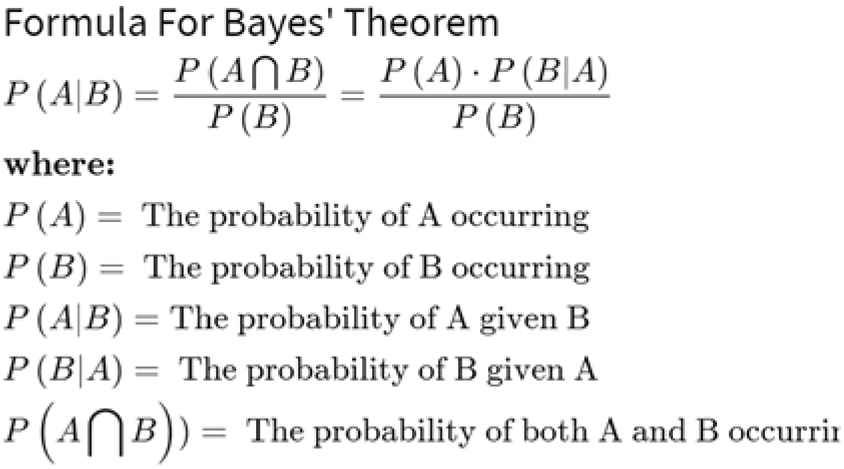
10. What Is Bayes' Theorem?

Answer:

A theorem describing how the conditional probability of each of a set of possible causes for a given observed outcome can be computed from knowledge of the probability of each cause and the conditional probability of the outcome of each cause.

Bayes' Theorem allows you to update the predicted probabilities of an event by incorporating new information.
Bayes' Theorem was named after 18th-century mathematician Thomas Bayes.
It is often employed in finance in calculating or updating risk evaluation.
The theorem has become a useful element in the implementation of machine learning.
The theorem was unused for two centuries because of the high volume of calculation capacity required to execute its transactions.

In Bayesian statistical inference, is the probability of an event occurring before new data is collected. In other words, it represents the best rational assessment of the probability of a particular outcome based on current knowledge before an experiment is performed.
Posterior probability is the revised probability of an event occurring after taking into consideration the new information. Posterior probability is calculated by updating the prior probability using Bayes' theorem. In statistical terms, the posterior probability is the probability of event (A) occurring given that event (B) has occurred.



Formula For Bayes' Theorem

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

where:

$P(A) =$ The probability of A occurring

$P(B) =$ The probability of B occurring

$P(A|B) =$ The probability of A given B

$P(B|A) =$ The probability of B given A

$P\left(A \bigcap B\right)) =$ The probability of both A and B occurri

11. What is z-score?

Answer:

Z-score is also known as standard score gives us an idea of how far a data point is from the mean. It indicates how many standard deviations an element is from the mean. Hence, Z-Score is measured in terms of standard deviation from the mean. For example, a standard deviation of 2 indicates the value is 2 standard deviations away from the mean. In order to use a z-score, we need to know the population mean (μ) and also the population standard deviation (σ).

The Formula for Z-Score

A z-score can be calculated using the following formula.

$z = (X – \mu) / \sigma$

Where,
z = Z-Score
X = the value of the element
μ = the population mean
σ = the population standard deviation

Usually, the population mean ((μ), the population standard deviation (σ), and the observed value (x) are provided in the problem statement, and substituting the same in the above Z-score equation yields us the Z-Score value. Depending upon whether the given Z-Score is positive or negative.

The Z-score, by contrast, is the number of standard deviations a given data point lies from the mean.

For data points that are below the mean, the Z score is negative. In most large data set, 99% of values have a z score between -3 and 3, meaning they lie within three standard deviation above and below the mean.

12. What is t-test?

Answer:
A t-test is a statistical test that is used to compare the means of two groups. It is often used in hypothesis testing to determine whether a process or treatment actually has an effect on the population of interest, or whether two groups are different from one another.

The t-test is a parametric test of difference, meaning that it makes the same assumptions about your data as other parametric tests. The t-test assumes your data:

- Independent.
- Normally distributed.
- Have a similar amount of variance within each group being compared.

One-sample, two-sample, or paired t-test:

- If the groups come from a single population perform a paired t-test.
- If the groups come from two different populations perform a two-sample t-test.
- If there is one group being compared against a standard value perform a one-sample t-test.

One-tailed or two-tailed t-test:

- If you only care whether the two populations are different from one another, perform a two-tailed t-test.
- If you want to know whether one population mean is greater than or less than the other, perform a one-tailed t-test.

T-test formula:

The formula for the two-sample t-test

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{(s^2(\frac{1}{n_1} + \frac{1}{n_2}))}}$$

In this formula, t is the t-value, x1 and x2 are the means of the two groups being compared, s2 is the pooled standard error of the two groups, and n1 and n2 are the number of observations in each of the groups.

A larger t-value shows that the difference between group means is greater than the pooled standard error, indicating a more significant difference between the groups.

13. What is percentile?

Answer:

Percentile is used to indicate the value below which the group of percentage of data falls below.

Example:

STATISTICS WORKSHEET-3

Consider if your score is 75th percentile, which you scored for better than 75% of people who took part in the test.

It is most commonly applicable in indicating the scores from the norm referenced tests such as SAT, GRE, and LSAT.

PERCENTILE EXAMPLE:

How to calculate percentile for the given example:

There are 25 test scores such as:

72, 54, 56, 61, 62, 66, 68, 43, 69, 69, 70, 71, 77, 78, 79, 85, 87, 88, 89, 93, 95, 96, 98, 99, 99.

Find the 60th percentile?

Solution:

Step 1: Arrange the data in the ascending order.

43, 54, 56, 61, 62, 66, 68, 69, 69, 70, 71, 72, 77, 78, 79, 85, 87, 88, 89, 93, 95, 96, 98, 99, 99.

Step 2: Find Rank, Rank = percentile/100 = 60/100 = 0.60

Step 3: Find 60th percentile = 0.60 * 25(Test Score) = 15

Step 4: Count the value in the given data set from left to right until you reach the number 15 value in data set.

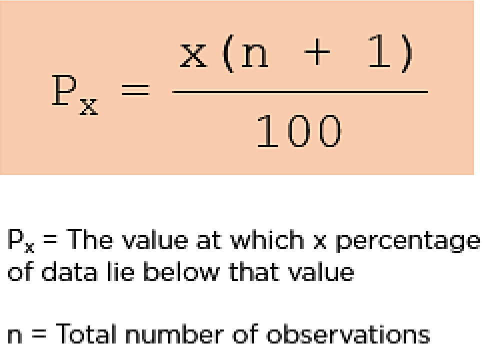
From the given data set, 15th number is 79.

Now, take the 15th number and 16th number and find the average:
79 + 85/2 = 164/2 = 82

Hence, 60th percentile of given data set is 82.

STATISTICS WORKSHEET-3

A percentile is a comparison score between a particular score and the scores of the rest of a group. It shows the percentage of scores that a particular score surpassed.



$$P_x = \frac{x(n + 1)}{100}$$

P_x = The value at which x percentage of data lie below that value

n = Total number of observations

14. What is ANOVA?

Answer:
ANOVA ---> Analysis Of Variance

Analysis of variance (ANOVA) is an analysis tool used in statistics that splits an observed aggregate variability found inside a data set into two parts: systematic factors and random factors. The systematic factors have a statistical influence on the given data set, while the random factors do not. Analysts use the ANOVA test to determine the influence that independent variables have on the dependent variable in a regression study.

The t- and z-test methods developed in the 20th century were used for statistical analysis until 1918, when Ronald Fisher created the analysis of variance method.

ANOVA is also called the Fisher analysis of variance, and it is the extension of the t- and z-tests. The term became well-known in 1925, after appearing in Fisher's book, "Statistical Methods for Research Workers."

- Analysis of variance, or ANOVA, is a statistical method that separates observed variance data into different components to use for additional tests.
- A one-way ANOVA is used for three or more groups of data, to gain information about the relationship between the dependent and independent variables.
- If no true variance exists between the groups, the ANOVA's F-ratio should equal close to 1.

STATISTICS WORKSHEET-3

The ANOVA test is the initial step in analyzing factors that affect a given data set. Once the test is finished, an analyst performs additional testing on the methodical factors that measurably contribute to the data set's inconsistency. The analyst utilizes the ANOVA test results in an f-test to generate additional data that aligns with the proposed regression models.

The ANOVA test allows a comparison of more than two groups at the same time to determine whether a relationship exists between them. The result of the ANOVA formula, the F statistic (also called the F-ratio), allows for the analysis of multiple groups of data to determine the variability between samples and within samples.

If no real difference exists between the tested groups, which is called the null hypothesis, the result of the ANOVA's F-ratio statistic will be close to 1. The distribution of all possible values of the F statistic is the F-distribution. This is actually a group of distribution functions, with two characteristic numbers, called the numerator degrees of freedom and the denominator degrees of freedom.

15. How can ANOVA help?

Answer:

Analysis of variance (ANOVA) is a statistical technique that is used to check if the means of two or more groups are significantly different from each other. ANOVA checks the impact of one or more factors by comparing the means of different samples.
ANOVA is helpful for testing three or more variables. It is similar to multiple two-sample t-tests. However, it results in fewer type I errors and is appropriate for a range of issues.

There are two main types of ANOVA:

- One-way ANOVA
- Two-way ANOVA

There also variations of ANOVA. For example, MANOVA (multivariate ANOVA) differs from ANOVA as the former tests for multiple dependent variables simultaneously while the latter assesses only one dependent variable at a time. One-way or two-way refers to the number of independent variables in your analysis of variance test. A one-way ANOVA evaluates the impact of a sole factor on a sole response variable. It determines whether all the samples are the same. The one-way ANOVA is used to determine whether there are any statistically significant differences between the means of three or more independent (unrelated) groups.

STATISTICS WORKSHEET-3

A two-way ANOVA is an extension of the one-way ANOVA. With a one-way, you have one independent variable affecting a dependent variable. With a two-way ANOVA, there are two independents. For example, a two-way ANOVA allows a company to compare worker productivity based on two independent variables, such as salary and skill set. It is utilized to observe the interaction between the two factors and tests the effect of two factors at the same time.

ANOVA HYPOTHESIS:

- **Null Hypothesis**: Groups means are equal (no variation in means of groups)
  - HO=μ1=μ2=""μp
- **Alternative Hypothesis:** At least, one group mean is different from other groups.
  - H1= All μ are not equal.

HOW ANOVA WORKS

- Check sample size: equal number of observation in each group
- Calculate Mean square for each group (MS) (SS of group/Level-1); level-1 is A degree of freedom (DF) for a group.
- Calculate Mean square error (MSE) (SS error/df of residuals)
- Calculate F value (MS of group/MSE)