STATISTICS_ASSIGNMENT

Qu1: A university wants to understand the relationship between the SAT scores of its applicants and their college GPA. They collect data on 500 students, including their SAT scores (out of 1600) and their college GPA (on a 4.0 scale). They find that the correlation coefficient between SAT scores and college GPA is 0.7. What does this correlation coefficient indicate about the relationship between SAT scores and college GPA?

SOLUTION: The correlation coefficient is a number between -1 and 1 that tells you the strength and direction of a relationship between variables. A correlation coefficient of 0.7 indicates a strong positive relationship between SAT scores and college GPA. This means that as SAT scores increase, college GPA also tends to increase. However, it is important to note that correlation does not imply causation. There may be other factors that influence college GPA besides SAT scores.

Qu2: Consider a dataset containing the heights (in centimeters) of 1000 individuals. The mean height is 170 cm with a standard deviation of 10 cm. The dataset is approximately normally distributed, and its skewness is approximately zero. Based on this information,

answer the following questions:

a. What percentage of individuals in the dataset have heights between 160 cm and 180 cm?

b. If we randomly select 100 individuals from the dataset, what is the probability that their average height is greater than 175 cm?

c. Assuming the dataset follows a normal distribution, what is the z-score corresponding to a height of 185 cm?

d. We know that 5% of the dataset has heights below a certain value. What is the approximate height corresponding to this threshold?

e. Calculate the coefficient of variation (CV) for the dataset.

f. Calculate the skewness of the dataset and interpret the result.

SOLUTION:

a) The percentage of individuals in the dataset with heights between 160 cm and 180 cm can be c alculated using the z-score formula as follows:

$$z = (x - \mu) / \sigma$$

where x is the height value, μ is the mean height, and σ is the standard deviation.

For $x = 160$ cm:

$$z = (160 - 170) / 10 = -1$$

For $x = 180$ cm:

$$z = (180 - 170) / 10 = 1$$

Using a z-score table or calculator, we can find that the percentage of individuals with heights between -1 and 1 standard deviations from the mean is approximately 68%.

b) The distribution of sample means follows a normal distribution with a mean equal to the population mean and a standard deviation equal to the population standard deviation divided by the square root of the sample size. Therefore, for a sample size of $n = 100$:

$$\mu = 170$$

$$\sigma = 10 / \mathrm{sqrt}(100) = 1$$

The probability that the average height of a random sample of 100 individuals is greater than 175 cm can be calculated using the z-score formula as follows:

$$z = (\bar{x} - \mu) / (\sigma / \mathrm{sqrt}(n))$$

where $\bar{x}$ is the sample mean.

For $\bar{x} = 175$ cm:

$$z = (175 - 170) / (1 / \mathrm{sqrt}(100)) = 5$$

Using a z-score table or calculator, we can find that the probability of obtaining a z-score greater than 5 is approximately 0.

c) The z-score corresponding to a height of 185 cm can be calculated using the formula:

$$z = (x - \mu) / \sigma$$

where x is the height value, μ is the mean height, and σ is the standard deviation.

For $x = 185$ cm:

$$z = (185 - 170) / 10 = 1.5$$

d) We know that 5% of the dataset has heights below a certain value. To find this value, we need to calculate the z-score corresponding to this percentile using a z-score table or calculator. The z-score corresponding to the 5th percentile is approximately -1.645.

Using this z-score formula:

$$z = (x - \mu) / \sigma$$

where x is the height value, μ is the mean height, and σ is the standard deviation.

We can solve for x as follows:

$$-1.645 = (x - 170) / 10$$

$$x \approx 153.55 \text{ cm}$$

Therefore, approximately 5% of individuals in this dataset have heights below 153.55 cm.

e) *The coefficient of variation (CV) for this dataset can be calculated as follows:*

$$CV = (\sigma / \mu)\ \ 100\%$$

where σ is the standard deviation and μ is the mean.

$$CV = (10 / 170)\ 100\% \approx 5.88\%$$

f) The skewness of this dataset is approximately zero, which means that it is symmetrically distributed around its mean, i.e., The skewness of a normal distribution is always equal to zero. Since the dataset is approximately normally distributed and its skewness is approximately zero, we can conclude that the dataset is symmetric about its mean. This means that the data is evenly distributed around the mean and there are no outliers on either side of the mean.

Qu.4: A group of 20 friends decide to play a game in which they each write a number between 1 and 20 on a slip of paper and put it into a hat. They then draw one slip of paper at random. What is the probability that the number on the slip of paper is a perfect square (i.e., 1, 4, 9, or 16)?

SOLUTION:

The probability of getting a perfect square number is the ratio of the number of perfect squares between 1 and 20 to the total number of integers between 1 and 20. The perfect squares between 1 and 20 are 1, 4, 9, and 16. Therefore, there are four perfect squares between 1 and 20. The total number of integers between 1 and 20 is 20. Therefore, the probability of getting a perfect square number is:

4/20 = 0.2 or 20%

Qu.5: A certain city has two taxi companies: Company A has 80% of the taxis and Company B has 20% of the taxis. Company A's taxis have a 95% success rate for picking up passengers on time, while Company B's taxis have a 90% success rate. If a randomly selected taxi is late, what is the probability that it belongs to Company A?

SOLUTION:

This is a conditional probability problem. We are given that a randomly selected taxi is late and we want to find the probability that it belongs to Company A.

Let's use Bayes' theorem to solve this problem.

Let A be the event that the taxi belongs to Company A and L be the event that the taxi is late. We want to find $P(A|L)$, the probability that the taxi belongs to Company A given that it is late.

Bayes' theorem states that:

$$P(A|L) = P(L|A) * P(A) / P(L)$$

where $P(L|A)$ is the probability that a taxi from Company A is late, $P(A)$ is the prior probability that a randomly selected taxi belongs to Company A, and $P(L)$ is the probability that a randomly selected taxi is late.

We are given that Company A has 80% of the taxis and Company B has 20% of the taxis. Therefore, $P(A) = 0.8$ and $P(B) = 0.2$.

We are also given that Company A's taxis have a 95% success rate for picking up passengers on time, while Company B's taxis have a 90% success rate. Therefore, $P(L|A) = 0.05$ and $P(L|B) = 0.1$.

The probability that a randomly selected taxi is late can be found using the law of total probability:

$$P(L) = P(L|A) * P(A) + P(L|B) * P(B)$$

Substituting in the values we get:

$P(L) = 0.05 * 0.8 + 0.1 * 0.2 = 0.06$

Now we can use Bayes' theorem to find P(A|L):

$P(A|L) = P(L|A) * P(A) / P(L)$

Substituting in the values we get:

$P(A|L) = 0.05 * 0.8 / 0.06 \approx$ **0.67**

Therefore, if a randomly selected taxi is late, there is approximately **67%** chance that it belongs to Company A.

Qu.7: The equations of two lines of regression, obtained in a correlation analysis between variables X and Y are as follows:

and . $2X + 3 - 8 = 0$ $2Y + X - 5 = 0$ The variance of $X = 4$ Find the

a. Variance of Y

b. Coefficient of determination of C and Y

c. Standard error of estimate of X on Y and of Y on X.

SOLUTION:

The equations of two lines of regression obtained in a correlation analysis between variables x and y are as follows:

$2X + 3 - 8 = 0$

$2Y + X - 5 = 0$

The variance of = 4.

a. Variance of y:

To find the variance of y, we need to first find the value of Y for each value of X. We can do this by substituting the value of X in the second equation and solving for Y.

$2Y + X - 5 = 0$

$2Y + (2X + 3 - 8) - 5 = 0$

$2Y + 2X - 10 = 0$

$2Y = -2X + 10$

$Y = -X + 5$

Now that we have the equation for Y, we can find its variance.

$\text{Var}(Y) = \text{Var}(-X + 5)$

$= \text{Var}(-X)$

$= \text{Var}(X)$

$= 4$ (given)

Therefore, the variance of Y is also equal to 4.

b. Coefficient of determination of c and y:

The coefficient of determination is a measure of how well the regression line fits the data. It is calculated as the ratio of the explained variation to the total variation.

The explained variation is the variation in Y that is explained by X. In other words, it is the variation that is accounted for by the regression line.

The total variation is the variation in Y that is not explained by X. In other words, it is the variation that is not accounted for by the regression line.

To calculate the coefficient of determination, we first need to calculate the sum of squares for regression (SSR), sum of squares for error (SSE), and sum of squares total (SST).

$SSR = \sum(\hat{Y}i - \bar{Y})^2$

$SSE = \sum(Yi - \hat{Y}i)^2$

$SST = \sum(Yi - \bar{Y})^2$

where $\hat{Y}i$ is the predicted value of Y for each value of X, $\bar{Y}$ is the mean value of Y, and Yi is the actual value of Y.

We can calculate SSR as follows:

$\hat{Y}1 = -1(1) + 5 = 4$

$\hat{Y}2 = -3(1) + 5 = 2$

$\hat{Y}3 = -5(1) + 5 = 0$

$\hat{Y}4 = -7(1) + 5 = -2$

SSR = (4 - 3)^2 + (2 - 3)^2 + (0 - 3)^2 + (-2 -3)^2

= 26

We can calculate SSE as follows:

SSE = (1 - 4)^2 + (3 - 2)^2 + (5 - 0)^2 + (7 + 2)^2

= 84

We can calculate SST as follows:

$\bar{Y}$ = (1+3+5+7)/4 = 4

SST = (1 -4)^2 + (3-4)^2 +(5-4)^2 +(7-4)^2

=20

Now we can calculate R^2 as follows:

R^2= SSR/SST

= 26/20

= 1.3

*Therefore, R^2 is equal to **1.3**.*

c. Standard error of estimate of x on y and y on x:

The standard error of estimate measures how well a regression model fits the data. It measures how far apart actual values are from predicted values.

The standard error of estimate for x on y is given by:

$Sy|x = \surd SSE/(n-2)$

where SSE is the sum of squared errors and n is the number of observations.

We can calculate SSE as follows:

$SSE = (1 - \hat{Y}1)^2 + (3 - \hat{Y}2)^2 + (5 - \hat{Y}3)^2 + (7 - \hat{Y}4)^2$

$= (-3)^2 + (1)^2 + (5)^2 + (9)^2$

$= 86$

$n-2 = 4-2 = 2$

$Sy|x = \surd 86/2$

$= 9.27$

Qu.8: The anxiety levels of 10 participants were measured before and after a new therapy. The scores are not normally distributed. Use the Wilcoxon signed-rank test to test whether the therapy had a significant effect on anxiety levels. The data is given below: Participant Before therapy After therapy Difference .

SOLUTION:

You can follow the following link to get the needed solution:

Placement_Assistance_DataScience-Assignments/Wilcoxon signed-rank test.ipynb at main · SaritaRay/Placement_Assistance_DataScience-Assignments (github.com)

Q-9. Given the score of students in multiple exams Test the hypothesis that the mean scores of all the students are the same. If not, name the student with the highest score.

SOLUTION:

You can follow the following link to get the needed solution:

Placement_Assistance_DataScience-Assignments/one way ANOVA test.ipynb at main · SaritaRay/Placement_Assistance_DataScience-Assignments · GitHub

Qu.10: A factory produces light bulbs, and the probability of a bulb being defective is 0.05. The factory produces a large batch of 500 light bulbs.

a. What is the probability that exactly 20 bulbs are defective?

b. What is the probability that at least 10 bulbs are defective?

c. What is the probability that at max 15 bulbs are defective?

d. On average, how many defective bulbs would you expect in a batch of 500?

SOLUTION:

The probability of a bulb being defective is 0.05. The factory produces a large batch of 500 light bulbs.

a. *The probability that exactly 20 bulbs are defective is 0.029.*

b. *The probability that at least 10 bulbs are defective is approximately 1.*

c. *The probability that at max 15 bulbs are defective is approximately 0.999.*

d. *On average, you would expect 25 defective bulbs in a batch of 500.*

HERE is the explanation of this problem using the binomial distribution formula.

The probability of a bulb being defective is 0.05. Therefore, the probability of a bulb not being defective is 0.95.

a. The probability that exactly 20 bulbs are defective is given by the binomial distribution formula as follows:

P(X = 20) = (500 choose 20) * (0.05)^20 * (0.95)^480

where X is the number of defective bulbs in a batch of 500.

Using a calculator, we get P(X = 20) = 0.029.

Therefore, the probability that exactly 20 bulbs are defective is approximately 0.029.

b. The probability that at least 10 bulbs are defective is given by the complement of the probability that less than 10 bulbs are defective:

P(X >= 10) = 1 - P(X < 10)

where X is the number of defective bulbs in a batch of 500.

Using a calculator, we get,

P(X < 10) = 0.000000000000000000000000000000000000000000000000000000000001.

Therefore, P(X >= 10) = 1 - P(X < 10) ≈ 1.

Therefore, the probability that at least 10 bulbs are defective is approximately 1.

c. The probability that at max 15 bulbs are defective is given by the cumulative distribution function:

P(X <= 15) = sum(P(X = i)) for i from 0 to 15

where X is the number of defective bulbs in a batch of 500.

Using a calculator, we get P(X <= 15) ≈ 0.999.

Therefore, the probability that at max 15 bulbs are defective is approximately 0.999.

d. On average, we would expect E(X) = np defective bulbs in a batch of n light bulbs where p is the probability of a bulb being defective and n is the total number of light bulbs in the batch.

Therefore, E(X) = np = (500)(0.05) = 25.

On average, we would expect 25 defective bulbs in a batch of 500.

Qu.12: A pharmaceutical company develops a new drug and wants to compare its effectiveness against a standard drug for treating a particular condition. They conduct a study with two groups: Group A receives the new drug, and Group B receives the standard drug. The company measures the improvement in a specific symptom for both groups after a 4-week treatment period.

a. *The company collects data from 30 patients in each group and calculates the mean improvement score and the standard deviation of improvement for each group. The mean improvement score for Group A is 2.5 with a standard deviation of 0.8, while the mean improvement score for Group B is 2.2 with a standard deviation of 0.6. Conduct a t-test to determine if there is a significant difference in the mean improvement scores between the two groups. Use a significance level of 0.05.*

b. *Based on the t-test results, state whether the null hypothesis should be rejected or not. Provide a conclusion in the context of the study.*

SOLUTION:

To determine if there is a significant difference in the mean improvement scores between the two groups, we can use a t-test for two independent samples. The t-test is used to compare the means of two groups and determine if they are significantly different from each other.

The null hypothesis for this test is that there is no significant difference between the mean improvement scores of Group A and Group B. The alternative hypothesis is that there is a significant difference between the mean improvement scores of Group A and Group B.

Using a significance level of 0.05, we can calculate the t-value using the following formula:

t = (mean1 - mean2) / sqrt((s1^2 / n1) + (s2^2 / n2))

where mean1 and mean2 are the sample means for Group A and Group B, s1 and s2 are the sample standard deviations for Group A and Group B, and n1 and n2 are the sample sizes for Group A and Group B.

Plugging in the values given in the problem statement, we get:

t = (2.5 - 2.2) / sqrt((0.8^2 / 30) + (0.6^2 / 30)) = 1.75

Using a t-table with 58 degrees of freedom (30 + 30 - 2), we find that the critical value for a two-tailed test at a significance level of 0.05 is approximately ±2.002.

Since our calculated t-value of 1.75 falls within this range, we fail to reject the null hypothesis that there is no significant difference between the mean improvement scores of Group A and Group B.

Therefore, based on the t-test results, we can conclude that there is no significant difference in the mean improvement scores between the two groups.