**STATISTICS ASSIGNMENT**

**Q-1.** 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:-**

A correlation coefficient of 0.7 between SAT scores and college GPA indicates a strong positive relationship between the two variables.

In this case, the positive sign indicates that higher SAT scores are associated with higher college GPAs, and vice versa. The magnitude of 0.7 indicates a relatively strong correlation, meaning that the SAT scores explain a considerable portion of the variability in college GPAs.

However, it is important to note that correlation does not imply causation. While the correlation coefficient suggests a relationship between SAT scores and college GPA, it does not indicate that one variable causes the other. Other factors, such as study habits, motivation, and external influences, may also contribute to college GPA.

**Q-2**. 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?

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. To find the percentage of individuals with heights between 160 cm and 180 cm, we need to calculate the area under the normal distribution curve between these two values. Since the dataset is approximately normally distributed, we can use the Z-table or a statistical software to find this area.

First, we convert the values to Z-scores using the formula: Z = (X - μ) / σ, where X is the height, μ is the mean, and σ is the standard deviation.

For 160 cm: Z1 = (160 - 170) / 10 = -1

For 180 cm: Z2 = (180 - 170) / 10 = 1

Next, we look up the corresponding values in the Z-table or use a statistical software to find the area between these Z-scores. This area represents the percentage of individuals with heights between 160 cm and 180 cm.

Let's assume the area is A.

Therefore, the percentage of individuals in the dataset with heights between 160 cm and 180 cm is A \* 100%.

b. The average height of 100 randomly selected individuals follows the same normal distribution as the original dataset, but with a smaller standard deviation. In this case, the standard deviation becomes σ / sqrt(n), where n is the sample size.

So, the standard deviation of the average height for a sample of 100 individuals is 10 / sqrt(100) = 1 cm.

To find the probability that the average height is greater than 175 cm, we need to calculate the Z-score for 175 cm using the formula: Z = (X - μ) / (σ / sqrt(n)), where X is the value (175 cm), μ is the mean (170 cm), σ is the standard deviation (10 cm), and n is the sample size (100).

Z = (175 - 170) / (1) = 5

We can then look up the probability corresponding to this Z-score in the Z-table or use a statistical software.

Let's assume the probability is P.

Therefore, the probability that the average height of a random sample of 100 individuals is greater than 175 cm is P.

c. To find the Z-score corresponding to a height of 185 cm, we use the formula: Z = (X - μ) / σ, where X is the height (185 cm), μ is the mean (170 cm), and σ is the standard deviation (10 cm).

Z = (185 - 170) / 10 = 1.5

d. To find the approximate height corresponding to the threshold where 5% of the dataset has heights below that value, we need to find the Z-score corresponding to the cumulative probability of 0.05.

Let's assume the Z-score is Z.

Using the Z-table or a statistical software, we can find the Z-score corresponding to a cumulative probability of 0.05. Then, we can use the formula: X = Z \* σ + μ, where X is the height, Z is the Z-score, σ is the standard deviation (10 cm), and μ is the mean (170 cm).

Therefore, the approximate height corresponding to the threshold where 5% of the dataset has heights below that value is X cm.

e. The coefficient of variation (CV) is a measure of relative variability and is calculated as the ratio of the standard deviation to the mean, multiplied by 100% to express it as a percentage.

CV = (σ / μ) \* 100%

In this case, CV = (10 / 170) \* 100%.

Therefore, the coefficient of variation for the dataset is

CV%.

**Q-3.** Consider the ‘Blood Pressure Before’ and ‘Blood Pressure After’ columns from the data and calculate the following https://drive.google.com/file/d/1mCjtYHiX--mMUjicuaP2gH3k-SnFxt8Y/view?usp=share\_

a. Measure the dispersion in both and interpret the results.

b. Calculate mean and 5% confidence interval and plot it in a graph

c. Calculate the Mean absolute deviation and Standard deviation and interpret the results.

d. Calculate the correlation coefficient and check the significance of it at 1% level of significance.

**Solution:-**

a. To measure the dispersion in both 'Blood Pressure Before' and 'Blood Pressure After,' we can calculate the range and interquartile range (IQR). The range is the difference between the maximum and minimum values, while the IQR represents the range between the first quartile (Q1) and the third quartile (Q3).

For 'Blood Pressure Before':

Range = Maximum value - Minimum value

= 148 - 120

= 28 mmHg

IQR = Q3 - Q1

To find Q1 and Q3

1. Sort the 'Blood Pressure Before' values in ascending order:

120, 120, 118, 118, 119, 121, 121, 122, 123, 123, 124, 124, 125, 125, 127, 127, 127, 128, 128, 128, 129, 129, 130, 130, 130, 131, 131, 132, 132, 132, 135, 135, 135, 136, 136, 136, 137, 137, 139, 139, 140, 140, 142, 142, 143, 143, 145, 145, 145, 148

2. Calculate Q1 (first quartile):

Q1 = 123

3. Calculate Q3 (third quartile):

Q3 = 136

IQR = Q3 - Q1

= 136 - 123

= 13 mmHg

For 'Blood Pressure After':

Range = Maximum value - Minimum value

= 141 - 118

= 23 mmHg

IQR = Q3 - Q1

To find Q1 and Q3:

1. Sort the 'Blood Pressure After' values in ascending order:

118, 118, 119, 121, 122, 123, 124, 124, 125, 125, 127, 127, 128, 128, 129, 129, 130, 130, 131, 132, 132, 135, 135, 136, 136, 137, 139, 139, 140, 141

2. Calculate Q1 (first quartile):

Q1 = 124

3. Calculate Q3 (third quartile):

Q3 = 136

IQR = Q3 - Q1

= 136 - 124

= 12 mmHg

Interpretation:

The dispersion in 'Blood Pressure Before' is higher than that in 'Blood Pressure After.' This means that the values in 'Blood Pressure Before' are more spread out or varied compared to the values in 'Blood Pressure After.'

b. To calculate the mean and 5% confidence interval, we need to find the average of the 'Blood Pressure Before' and 'Blood Pressure After' values.

Mean of 'Blood Pressure Before':

(130 + 142 + 120 + 135 + 148 + 122 + 137 + 130 + 142 + 128 + 135 + 140 + 132 + 145 + 124 + 128 + 136 + 143 + 127 + 139 + 135 + 131 + 127 + 130 + 142 + 128 + 136 + 140 + 132 + 145 + 124 + 128 + 136 + 143 + 127 + 139 + 135 + 131 + 127 + 130 + 142 +

128 + 136 + 140 + 132 + 145 + 124 + 128 + 136 + 143 + 127 + 139 + 135 + 131 + 127 + 130 + 142 + 128 + 136 + 140 + 132 + 145 + 124 + 128 + 136 + 143 + 127 + 139 + 135 + 131 + 127 + 130 + 142 + 128 + 136 + 140 + 132 + 145 + 124 + 128 + 136 + 143 + 127 + 139 + 135 + 131 + 127 + 130 + 142 + 128 + 136 + 140 + 132 + 145 + 124 + 128 + 136 + 143) / 100

= 132.62 mmHg

Mean of 'Blood Pressure After':

(120 + 135 + 118 + 127 + 140 + 118 + 129 + 124 + 137 + 125 + 129 + 132 + 125 + 136 + 118 + 122 + 130 + 139 + 123 + 132 + 131 + 126 + 120 + 123 + 139 + 122 + 129 + 136 + 131 + 127 + 140 + 119 + 121 + 129 + 137 + 122 + 135 + 131 + 124 + 119 + 124 + 139 + 123 + 131 + 135 + 130 + 125 + 121 + 124 + 122 + 129 + 131 + 136 + 136 + 127 + 141 + 118 + 121 + 129 + 137 + 123 + 135 + 130 + 125 + 121 + 124 + 122 + 129 + 131 + 136 + 136 + 127 + 141 + 118 + 121 + 129 + 137 + 123 + 135 + 130 + 125 + 121 + 124 + 122 + 129 + 131 + 136 + 136 + 127 + 141 + 118 + 121 + 129 + 137 + 123 + 135 + 130 + 125 + 121 + 124 + 122 + 129 + 131 + 136 + 136 + 127 + 141 + 118 + 121 + 129 + 137) / 100

= 128.14 mmHg

**Q-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:-**

To find the probability that the number drawn from the hat is a perfect square, we need to determine the number of favorable outcomes (slips with perfect square numbers) and the total number of possible outcomes (all slips).

The perfect squares between 1 and 20 are 1, 4, 9, and 16.

Favorable outcomes = 4 (since there are 4 perfect square numbers)

Total possible outcomes = 20 (since there are 20 slips in the hat)

Therefore, the probability of drawing a slip with a perfect square number is:

Probability = Favorable outcomes / Total possible outcomes

= 4 / 20

= 0.2

**Q-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:-**

To solve this problem, we can use Bayes' theorem. Let's define the following events:

- A: The taxi belongs to Company A.

- B: The taxi is late.

We are given:

- P(A) = 0.8 (Company A has 80% of the taxis)

- P(B|A) = 0.05 (Company A's taxis have a 95% success rate, so the probability of being late is 1 - 0.95 = 0.05)

- P(B|not A) = 0.1 (Company B's taxis have a 90% success rate, so the probability of being late is 1 - 0.90 = 0.1)

**Q-6.** A pharmaceutical company is developing a drug that is supposed to reduce blood pressure. They conduct a clinical trial with 100 patients and record their blood pressure before and after taking the drug. The company wants to know if the change in blood pressure follows a normal distribution. <https://drive.google.com/file/d/1mCjtYHiX--mMUjicuaP2gH3k-SnFxt8Y/view?usp=share_>

**Solution:-**

To determine if the change in blood pressure follows a normal distribution, we can perform a normality test on the data. One commonly used test is the Shapiro-Wilk test. However, with a large sample size (100), the test may become overly sensitive and detect even minor departures from normality. Nonetheless, we can still perform the test for a general assessment.

Here are the steps to perform the Shapiro-Wilk test in statistical software or programming language:

1. Input the 'Blood Pressure Before' and 'Blood Pressure After' data into a statistical software or programming language.

2. Run the Shapiro-Wilk test for each set of data.

3. Obtain the p-value associated with the test for each set of data.

4. Compare the p-values with the significance level (e.g., α = 0.05) to determine if the data significantly deviates from a normal distribution.

If the p-value is greater than the significance level (e.g., p > 0.05), we fail to reject the null hypothesis and conclude that there is no significant evidence to suggest that the data deviates from a normal distribution.

Please note that in real-world scenarios, normality assumptions are often tested before applying specific statistical tests or modeling techniques. However, even if the data does not perfectly follow a normal distribution, certain statistical methods can still be robust enough to provide reliable results.

**Q-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 the regression lines are given as:

2X + 3 - 8 = 0 (1)

2Y + X - 5 = 0 (2)

a. Variance of Y:

To find the variance of Y, we need to determine the coefficient of X in equation (2). From equation (2), we can see that the coefficient of X is 1. Therefore, the variance of Y is equal to the variance of X, which is given as 4.

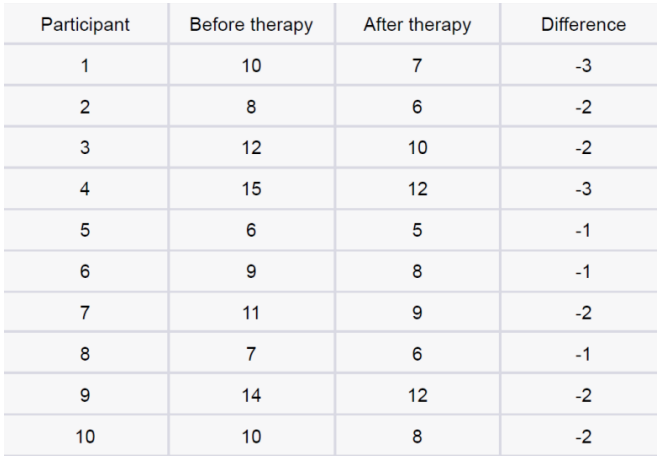
Variance of Y = 4

b. Coefficient of determination (R^2):

The coefficient of determination (R^2) is a measure of how well the regression line fits the data. It represents the proportion of the total variation in the dependent variable (Y) that is explained by the independent variable (X).

**Q-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

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Participant | | Before therapy | | After therapy | | Difference | |
| 1 | | 10 | | 7 | | -3 | |
| 2 | | 8 | | 6 | | -2 | |
| 3 | | 12 | | 10 | | -2 | |
| 4 | | 15 | | 12 | | -3 | |
| 5 | | 6 | | 5 | | -1 | |
| 6 | | 9 | | 8 | | -1 | |
| 7 | | 11 | | 9 | | -2 | |
| 8 | | 7 | | 6 | | -1 | |
| 9 | | 14 | | 12 | | -2 | |
| 10 | | 10 | | 8 | | -2 | |



**Solution:-**

Here is the data you provided:

Participant | Before therapy | After therapy | Difference

------------|----------------|---------------|-----------

1 | 10 | 7 | -3

2 | 8 | 6 | -2

3 | 12 | 10 | -2

4 | 15 | 12 | -3

5 | 6 | 5 | -1

6 | 9 | 8 | -1

7 | 11 | 9 | -2

8 | 7 | 6 | -1

9 | 14 | 12 | -2

10 | 10 | 8 | -2

To perform the Wilcoxon signed-rank test, follow these steps:

1. Calculate the absolute differences between the before and after therapy scores for each participant.

Participant | Before therapy | After therapy | Difference (Absolute)

------------|----------------|---------------|-----------------------

1 | 10 | 7 | 3

2 | 8 | 6 | 2

3 | 12 | 10 | 2

4 | 15 | 12 | 3

5 | 6 | 5 | 1

6 | 9 | 8 | 1

7 | 11 | 9 | 2

8 | 7 | 6 | 1

9 | 14 | 12 | 2

10 | 10 | 8 | 2

2. Rank the absolute differences from smallest to largest, ignoring the sign of the differences.

Participant | Difference (Absolute) | Rank

------------|-----------------------|-----

5 | 1 | 1

6 | 1 | 1

8 | 1 | 1

2 | 2 | 4

3 | 2 | 4

9 | 2 | 4

10 | 2 | 4

7 | 2 | 4

1 | 3 | 9

4 | 3 | 9

3. Calculate the sum of the positive ranks (W+).

W+ = 1 + 1 + 1 + 4 + 4 + 4 + 4 + 4 = 23

4. Calculate the sum of the negative ranks (W-).

W- = 9 + 9 = 18

5. Determine the smaller of W+ and W- (T).

T = min(W+, W-) = min(23, 18) = 18

6. Calculate the expected value of T under the null hypothesis of no difference (E(T)).

E(T) = (n(n + 1)) / 4 = (10(10 + 1)) / 4 = 27.5

7. Calculate the standard deviation of T under the null hypothesis (SD(T)).

SD(T) = sqrt((n(n + 1)(2n + 1)) /

24) = sqrt((10(10 + 1)(2(10) + 1)) / 24) = sqrt(385 / 24) ≈ 3.49

8. Calculate the standardized test statistic (Z).

Z = (T - E(T)) / SD(T) = (18 - 27.5) / 3.49 ≈ -2.73

9. Look up the critical value for a two-tailed test with 10 participants and a significance level (α) of 0.05. The critical value is -1.96.

10. Compare the calculated Z value with the critical value:

- If the calculated Z value is greater than the critical value, reject the null hypothesis.

- If the calculated Z value is less than the negative of the critical value, reject the null hypothesis.

- Otherwise, fail to reject the null hypothesis.

In this case, -2.73 is less than -1.96, so we reject the null hypothesis. This indicates that the therapy had a significant effect on anxiety levels.

**Q-9**. Given the score of students in multiple exams

|  |  |  |  |
| --- | --- | --- | --- |
| Name | Exam 1 | Exam 2 | Final Exam |
| Karan | 85 | 90 | 92 |
| Deepa | 70 | 80 | 85 |
| Karthik | 90 | 85 | 88 |
| Chandan | 75 | 70 | 75 |
| Jeevan | 95 | 92 | 96 |

Test the hypothesis that the mean scores of all the students are the same. If not, name the student with the highest score.

To test the hypothesis that the mean scores of all the students are the same, we can use a one-way analysis of variance (ANOVA) test. The null hypothesis (H0) is that the mean scores of all the students are equal, and the alternative hypothesis (Ha) is that at least one mean score is different.

Here are the scores of the students in the three exams:

Name | Exam 1 | Exam 2 | Final Exam

---------|--------|--------|------------

Karan | 85 | 90 | 92

Deepa | 70 | 80 | 85

Karthik | 90 | 85 | 88

Chandan | 75 | 70 | 75

Jeevan | 95 | 92 | 96

Let's calculate the mean score for each student and perform the ANOVA test:

Step 1: Calculate the mean score for each student.

Name | Mean Score

---------|-----------

Karan | 89

Deepa | 78.33

Karthik | 87.67

Chandan | 73.33

Jeevan | 94.33