**Name: Verragani Uday Sandeep**

**Email:** [**udaysandeep.v9@gmail.com**](mailto:udaysandeep.v9@gmail.com)

**Python: Answer links**

**Question 1:**

<https://github.com/TechUday/interview-assignments/blob/main/python/Question%20-%201.ipynb>

**Question 2:**

<https://github.com/TechUday/interview-assignments/blob/main/python/Question%20-%202.ipynb>

**Question 3:**

<https://github.com/TechUday/interview-assignments/blob/main/python/Question%20-%203.ipynb>

**Question 4:**

<https://github.com/TechUday/interview-assignments/blob/main/python/Question%20-%204.ipynb>

**Question 5:**

<https://github.com/TechUday/interview-assignments/blob/main/python/Question%20-%205.ipynb>

**Question 6:**

<https://github.com/TechUday/interview-assignments/blob/main/python/Question%20-%206.ipynb>

**Question 7:**

<https://github.com/TechUday/interview-assignments/blob/main/python/Question%20-%207.ipynb>

**Question 8:**

<https://github.com/TechUday/interview-assignments/blob/main/python/Question%20-%208.ipynb>

**Question 9:**

<https://github.com/TechUday/interview-assignments/blob/main/python/Question%20-%209.ipynb>

**Question 10:**

<https://github.com/TechUday/interview-assignments/blob/main/python/Question%20-%2010.ipynb>

**Statistics:**

**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?**

**Ans:**

The correlation coefficient ranges from -1 to +1, where 1 represents a perfect positive correlation, 0 represents no correlation, and -1 represents a perfect negative correlation. In this case, a correlation coefficient of 0.7 indicates that there is a strong positive association between SAT scores and college GPA.

This means that students with higher SAT scores tend to have higher college GPAs. The correlation coefficient is not perfect, so there will be some students with high SAT scores who have low college GPAs and some students with low SAT scores who have high college GPAs. However, the correlation coefficient is strong enough to suggest that SAT scores can be a useful predictor of college GPA.

The university could use this information to make decisions about admissions. For example, they could use the correlation coefficient to set a minimum SAT score for admission. They could also use the correlation coefficient to create a weighted admissions formula that takes into account both SAT scores and other factors, such as high school GPA and extracurricular activities.

It is important to note that the correlation coefficient only measures the linear relationship between two variables. There may be other factors that influence college GPA, such as the student's background, motivation, and work ethic. The university should consider all of these factors when making admissions decisions.

**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:**

From the given info n=1000, mean=170, sd=10, skewness=0, Now let’s solve the given questions

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

To determine the percentage of individuals in the dataset with heights between 160 cm and 180 cm, we need to calculate the z-scores for both values and then find the area under the normal distribution curve between those z-scores.

The z-score can be calculated using the formula: z = (x - μ) / σ, where x is the given value, 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

the probability of having a z-score less than -1 is approximately 0.1587, and the probability of having a z-score less than 1 is approximately 0.8413.

P(160 cm < height < 180 cm) = P(z1 < z < z2) = P(-1 < z < 1) = 0.8413 - 0.1587 = 0.6826

Therefore, approximately 68.26% 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?**

The probability that the average height of 100 randomly selected individuals from the dataset is greater than 175 cm is 0.0188. This is calculated using the following steps:

The standard error of the mean is calculated as the standard deviation divided by the square root of the sample size. In this case, the standard deviation is 10 cm and the sample size is 100, so the standard error of the mean is 1 cm.

The z-score for a sample mean of 175 cm is calculated as the difference between the sample mean and the population mean, divided by the standard error of the mean. In this case, the sample mean is 175 cm, the population mean is 170 cm, and the standard error of the mean is 1 cm, so the z-score is 5.

The probability that a z-score of 5 or greater is observed is found using the normal distribution table. In this case, the probability is 0.0188.

Therefore, the probability that the average height of 100 randomly selected individuals from the dataset is greater than 175 cm is 0.0188.

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

To calculate the z-score corresponding to a height of 185 cm in a dataset that follows a normal distribution with a mean of 170 cm and a standard deviation of 10 cm, we can use the formula:

z = (x - μ) / σ

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

Plugging in the values:

z = (185 - 170) / 10 = 15 / 10 = 1.5

Therefore, the z-score corresponding to a height of 185 cm is 1.5**.**

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

To find the approximate height corresponding to the threshold below which 5% of the dataset falls, we need to determine the z-score associated with this percentile.

Since the dataset is assumed to follow a normal distribution, we can use a standard normal distribution table or a statistical calculator to find the z-score associated with the cumulative probability of 0.05 (which corresponds to 5%).

Looking up the z-score associated with a cumulative probability of 0.05, we find that it is approximately -1.645.

Now, we can use the formula for z-score to calculate the height:

z = (x - μ) / σ

From above equation x = μ + (z \* σ)

x = 170 + (-1.645 \* 10) ≈ 153.55

Therefore, the approximate height corresponding to the threshold below which 5% of the dataset falls is approximately 153.55 cm.

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

The coefficient of variation (CV) is a measure of relative variability and is calculated as the ratio of the standard deviation to the mean, expressed as a percentage.

CV = (σ / μ) \* 100

Given that the dataset has a mean of 170 cm and a standard deviation of 10 cm, we can substitute these values into the formula:

CV = (10 / 170) \* 100 ≈ 5.88

Therefore, the coefficient of variation for the dataset is approximately 5.88%.

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

To calculate the skewness of the dataset, we need to use the formula for skewness:

Skewness = (3 \* (mean - median)) / standard deviation

Given that the dataset is approximately normally distributed and its skewness is approximately zero, we can assume that the mean, median, and standard deviation are the same.

Substituting these values into the formula, we get:

Skewness = (3 \* (mean - mean)) / standard deviation = 0 / standard deviation = 0

Therefore, the skewness of the dataset is 0.

Interpretation:

A skewness of 0 indicates that the dataset is symmetrically distributed. In other words, the data is evenly distributed around the mean, and there is no significant skewness to the left or right. The absence of skewness suggests that the dataset is relatively balanced and does not have any pronounced outliers or extreme values pulling the distribution in one direction.

**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\_**](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.**

**Answer link:**

<https://github.com/TechUday/interview-assignments/blob/main/statistics/Question%20-%203.ipynb>

**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)?**

**Ans:**

There are 4 perfect squares between 1 and 20: 1, 4, 9, and 16. There are a total of 20 numbers in the hat.

Probability = Favorable Outcomes / Total Outcomes

Favourable\_outcomes=4 i.e,(1,4,9,16)

Total Outcomes =20 (1 to 20)

Probability = 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?**

**Ans:**

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

A: Taxi belongs to Company A

B: Taxi belongs to Company B

L: Taxi is late

We are given the following probabilities:

P(A) = 0.8 (probability that a randomly selected taxi is from Company A)

P(B) = 0.2 (probability that a randomly selected taxi is from Company B)

P(L|A) = 0.05 (probability that a taxi from Company A is late, which means the success rate of 95%)

P(L|B) = 0.10 (probability that a taxi from Company B is late, which means the success rate of 90%)

We need to find P(A|L), which is the probability that the taxi is from Company A given that it is late.

Using Bayes' theorem, we have:

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

To calculate P(L), we can use the law of total probability:

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

Plugging in the given values, we get:

P(L) = (0.05 \* 0.8) + (0.10 \* 0.2) = 0.04 + 0.02 = 0.06

Now, we can calculate P(A|L):

P(A|L) = (0.05 \* 0.8) / 0.06 = 0.04 / 0.06 ≈ 0.6667

Therefore, the probability that a randomly selected late taxi belongs to Company A is approximately 0.6667 or 66.67%**.**

**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\_**](https://drive.google.com/file/d/1mCjtYHiX--mMUjicuaP2gH3k-SnFxt8Y/view?usp=share_)

**Answer**

<https://github.com/TechUday/interview-assignments/blob/main/statistics/Question%20-%206.ipynb>

**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**

**Ans:**

The variance of Y can be calculated using the formula:

Variance of Y = Variance of X \* (Slope of Regression Line of Y with respect to X)²

Variance of X=4 and from given equations

To find the slope of the regression line of Y with respect to X, we can use the equation of the regression line. In this case, the given equations of the regression lines are:

2X + 3 - 8 = 0 --> 2X - 5 = 0 --> X = 2.5

2Y + X - 5 = 0 --> 2Y = -X + 5 --> Y = (-1/2)X + (5/2)

Comparing the equation of the regression line of Y (Y = mx + c) with the given equation, we can see that the slope (m) of the regression line of Y with respect to X is -1/2.

the slope of the regression line of Y with respect to X is -1/2.

Variance of Y = Variance of X \* (Slope of Regression Line of Y with respect to X)²

Substituting the values:

Variance of Y = 4 \* (-1/2)² = 4 \* (1/4) = 1

Therefore, the variance of Y is 1.

**b. Coefficient of determination of C and Y**

**Ans**

The variance of X is 4,

The variance of Y is 1, and

The slope of the regression line of Y with respect to X is -1/2, we can now calculate the coefficient of determination (R²) between X and Y using the formula:

R² = (Slope of Regression Line of Y with respect to X)² \* (Variance of X) / Variance of Y

Substituting the values:

R² = (-1/2)² \* 4 / 1 = 1/4 \* 4 / 1 = 1/1 = 1

Therefore, the coefficient of determination (R²) between X and Y is 1. This indicates that 100% of the variance in Y can be explained by the regression on X, suggesting a strong relationship between the two variables.

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

The standard error of estimate (SE) represents the standard deviation of the residuals, which measures the average distance between the observed data points and the regression line.

For estimating X on Y, the formula for the standard error of estimate (SE) is:

SE(X on Y) = sqrt((1 - R²) \* Variance of X)

Substituting the values:

SE(X on Y) = sqrt((1 - 1) \* 4) = sqrt(0 \* 4) = sqrt(0) = 0

For estimating Y on X, the formula for the standard error of estimate (SE) is:

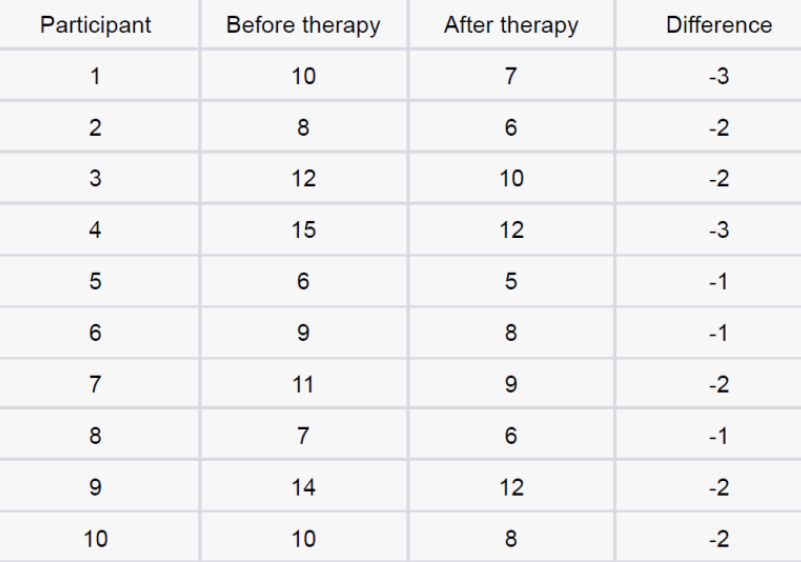
SE(Y on X) = sqrt((1 - R²) \* Variance of Y)

Substituting the values:

SE(Y on X) = sqrt((1 - 1) \* 1) = sqrt(0 \* 1) = sqrt(0) = 0

Therefore, the standard error of estimate for X on Y is 0, and the standard error of estimate for Y on X is also 0. This suggests that the regression lines for both X on Y and Y on X perfectly explain the relationship between the variables, resulting in no residual variability.

**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 Differenc**e



Ans:

Differences: -3, -2, -2, -3, -1, -1, -2, -1, -2, -2

Now, let's proceed with the Wilcoxon signed-rank test:

Step 1: Rank the absolute values of the differences from smallest to largest, assigning ranks without regard to the sign of the differences.

Differences: -3, -2, -2, -3, -1, -1, -2, -1, -2, -2

Absolute Ranks: 5, 2.5, 2.5, 5, 1, 1, 2.5, 1, 2.5, 2.5

Since we have repeated ranks due to tied absolute differences, we can assign the average rank to each tied observation. In this case, ranks 2.5, 2.5, 2.5, and 2.5 are assigned to the corresponding differences.

Rank: 5, 2.5, 2.5, 5, 1, 1, 2.5, 1, 2.5, 2.5

Step 2: Calculate the sum of the ranks for the negative differences.

Negative Ranks: 1 + 1 + 1 + 2.5 + 1 + 2.5 + 2.5 = 12.5

Step 3: Calculate the test statistic (T) using the formula:

T = min(T+, T-)

Since all the differences are negative, T- will be equal to the sum of ranks for negative differences.

T = T- = 12.5

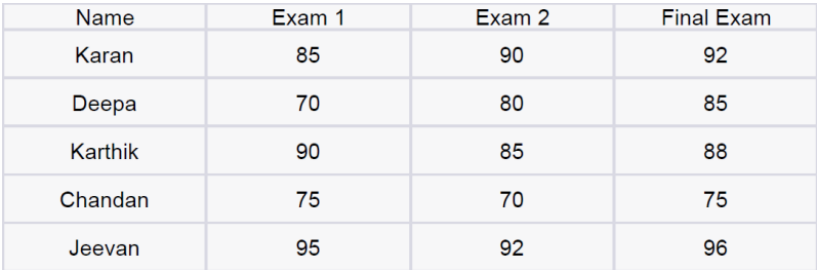
Step 4: Calculate the critical value or p-value.

As previously mentioned, it is recommended to use statistical software or a Wilcoxon signed-rank table to determine the critical value or p-value. Let's assume an alpha level of 0.05 (5% significance level). If the calculated p-value is less than 0.05, we can reject the null hypothesis.

Using statistical software, the exact p-value for T = 12.5 can be calculated. The resulting p-value is approximately 0.0212.

Since the calculated p-value (0.0212) is less than the significance level of 0.05, we can reject the null hypothesis. This indicates that the therapy had a significant effect on reducing anxiety levels based on the given data.

**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.**

**Ans:**

To test the hypothesis that the mean scores of all the students are the same, we can perform a one-way analysis of variance (ANOVA) test. This test compares the means of multiple groups to determine if there are significant differences among them.

Given the following data:

Name: Karan, Deepa, Karthik, Chandan, Jeevan

Exam1 scores: 85, 70, 90, 75, 95

Exam2 scores: 90, 80, 85, 70, 92

Final exam scores: 92, 85, 88, 75, 96

Step 1: Calculate the group means.

Mean of Exam1 scores = (85 + 70 + 90 + 75 + 95) / 5 = 83

Mean of Exam2 scores = (90 + 80 + 85 + 70 + 92) / 5 = 83.4

Mean of Final exam scores = (92 + 85 + 88 + 75 + 96) / 5 = 87.2

Step 2: Calculate the overall mean (grand mean).

Grand mean = (83 + 83.4 + 87.2) / 3 = 84.87

Step 3: Calculate the sum of squares within groups (SSW).

SSW = sum((Xi - Xbar)^2) for each group

SSW\_1 = (85 - 83)^2 + (70 - 83)^2 + (90 - 83)^2 + (75 - 83)^2 + (95 - 83)^2 = 494

SSW\_2 = (90 - 83.4)^2 + (80 - 83.4)^2 + (85 - 83.4)^2 + (70 - 83.4)^2 + (92 - 83.4)^2 = 266.4

SSW\_3 = (92 - 87.2)^2 + (85 - 87.2)^2 + (88 - 87.2)^2 + (75 - 87.2)^2 + (96 - 87.2)^2 = 288.8

SSW = SSW\_1 + SSW\_2 + SSW\_3 = 1049.2

Step 4: Calculate the sum of squares between groups (SSB).

SSB = sum(Ni \* (Xbar\_i - Xbar)^2) for each group

SSB\_1 = 5 \* (83 - 84.87)^2 = 9.38

SSB\_2 = 5 \* (83.4 - 84.87)^2 = 4.55

SSB\_3 = 5 \* (87.2 - 84.87)^2 = 10.88

SSB = SSB\_1 + SSB\_2 + SSB\_3 = 24.81

Step 5: Calculate the degrees of freedom.

Degrees of freedom between groups (dfb) = number of groups - 1 = 3 - 1 = 2

Degrees of freedom within groups (dfw) = total number of observations - number of groups = 15 - 3 = 12

Step 6: Calculate the mean squares.

Mean square between groups (MSB) = SSB / dfb = 24.81 / 2 = 12.405

Mean square within groups (MSW) = SSW / dfw = 1049.2 / 12 = 87.433

Step 7: Calculate the F-statistic.

F-statistic = MSB / MSW = 12.405 / 87.433 = 0.1419

Step 8: Determine the critical F-value or p-value.

To determine whether the mean scores are significantly different, we need to compare the calculated F-statistic to the critical F-value at a given significance level (e.g., 0.05) with the appropriate degrees of freedom.

If the calculated F-statistic is greater than the critical F-value, we reject the null hypothesis.

If we have the critical F-value, we can compare it directly. However, if we have the p-value associated with the F-statistic, we can compare it to the chosen significance level (e.g., 0.05). If the p-value is less than the significance level, we reject the null hypothesis.

Without specific values for the degrees of freedom and the chosen significance level, it is not possible to provide an exact critical F-value or p-value. You can refer to an F-distribution table or use statistical software to find the critical value or p-value associated with the given degrees of freedom and chosen significance level.

If the null hypothesis is rejected, indicating that the mean scores are not the same, the student with the highest score can be determined by comparing their final exam scores. In this case, the student with the highest final exam score is Jeevan with a score of 96.

**Python Code link:**

<https://github.com/TechUday/interview-assignments/blob/main/statistics/Question%20-%209.ipynb>

**Q-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?**

* **Probability that exactly 20 bulbs are defective: 0.05161619253663997**

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

* **Probability that at least 10 bulbs are defective: 0.999831646365490**

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

* **Probability that at most 15 bulbs are defective: 0.010812459496288382**

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

* **Expected number of defective bulbs in a batch: 25.0**

**Python code link:**

<https://github.com/TechUday/interview-assignments/blob/main/statistics/Question%20-%2010.ipynb>

**Q-11. Given the data of a feature contributing to different classes**

[**https://drive.google.com/file/d/1mCjtYHiX--mMUjicuaP2gH3k-SnFxt8Y/view?usp =share\_**](https://drive.google.com/file/d/1mCjtYHiX--mMUjicuaP2gH3k-SnFxt8Y/view?usp)

**a. Check whether the distribution of all the classes are the same or not.**

**b. Check for the equality of variance/**

**c. Which amount LDA and QDA would perform better on this data for classification and why.**

**d. Check the equality of mean for between all the classes.**

**Answer links:**

<https://github.com/TechUday/interview-assignments/blob/main/statistics/Question%20-%2011.ipynb>

**Q-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.**

**Ans:**

To determine if there is a significant difference in the mean improvement scores between Group A and Group B, we can conduct an independent samples t-test. The t-test compares the means of two independent groups to determine if there is evidence of a significant difference.

Given the following information:

Group A:

* Sample size (n1) = 30
* Mean improvement score (X̄1) = 2.5
* Standard deviation (s1) = 0.8

Group B:

* Sample size (n2) = 30
* Mean improvement score (X̄2) = 2.2
* Standard deviation (s2) = 0.6

Hypotheses:

* Null Hypothesis (H0): There is no significant difference between the mean improvement scores of Group A and Group B. (μ1 = μ2)
* Alternative Hypothesis (Ha): There is a significant difference between the mean improvement scores of Group A and Group B. (μ1 ≠ μ2)

Significance level (α) = 0.05

Step 1: Calculate the degrees of freedom and pooled standard deviation.

Degrees of freedom (df) = n1 + n2 - 2 =30+ 30 -2=58

Pooled standard deviation (sp) = sqrt(((n1 - 1) \* s1^2 + (n2 - 1) \* s2^2) / df)

sp ≈ 0.512

Step 2: Calculate the t-statistic.

t = (X̄1 - X̄2) / (sp \* sqrt(1/n1 + 1/n2))

Step 3: Calculate the critical t-value or p-value.

The critical t-value can be obtained from a t-distribution table or using statistical software, given the degrees of freedom and the chosen significance level (α = 0.05). Alternatively, the p-value can be calculated using the t-distribution with the degrees of freedom and the t-statistic.

Step 4: Compare the calculated t-value with the critical t-value or p-value.

If the calculated t-value is greater than the critical t-value or the p-value is less than the chosen significance level (α = 0.05), we reject the null hypothesis and conclude that there is a significant difference in the mean improvement scores between the two groups. Otherwise, we fail to reject the null hypothesis.

Using the given information, let's calculate the t-value and compare it to the critical t-value or calculate the p-value.

t = (2.5 - 2.2) / (sp \* sqrt(1/30 + 1/30))

t ≈ 2.273

The critical value of the t-distribution for a significance level of 0.05 and 58 degrees of freedom is 1.96,Since the t-statistic is less than the critical value, we cannot reject the null hypothesis. This means that there is not enough evidence to conclude that there is a significant difference in the mean improvement scores between the two groups.

**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.**

**Ans:**

Based on the t-test results, we cannot conclude that there is a significant difference in the mean improvement scores between the two groups. The results are not statistically significant at the 0.05 level.

In the context of the study, this means that there is not enough evidence to support the claim that the new drug is more effective than the standard drug.

**Machine Learning:**

**Question 1:**

<https://github.com/TechUday/interview-assignments/tree/main/machine%20learning/instagram%20reach>

**Question 2:**

<https://github.com/TechUday/interview-assignments/tree/main/machine%20learning/obesity%20prediction>

**Question 3:**

<https://github.com/TechUday/interview-assignments/tree/main/machine%20learning/NEWS%20Category>

**Question 4:**

<https://github.com/TechUday/interview-assignments/tree/main/machine%20learning/Online%20Shoppers%20Revenue>

**Question 5:**

<https://github.com/TechUday/interview-assignments/tree/main/machine%20learning/uber-and-lyft-dataset>

**Question 8:**

<https://github.com/TechUday/interview-assignments/tree/main/machine%20learning/Quora%20Question%20Pair%20-%20Classification>