**Python**

Total Marks: 100

Each question 10 marks

**Question 1: -**

Write a program that takes a string as input, and counts the frequency of each word in the string, there might be repeated characters in the string. Your task is to find the highest frequency and returns the length of the highest-frequency word.

Note - You have to write at least 2 additional test cases in which your program will run successfully and provide an explanation for the same.

Example input - string = “write write write all the number from from from 1 to 100”

Example output - 5

Explanation - From the given string we can note that the most frequent words are “write” and “from” and the maximum value of both the values is “write” and its corresponding length is 5

**Ans**: - <https://github.com/SatishPatri10/iNeuron-Data-Science-Assignment/blob/main/Python/Placement%20Assignment_Satish_Kumar_Patri/Python/1.%20Placement%20Assignment_Satish.ipynb>

**Question 2: -**

Consider a string to be valid if all characters of the string appear the same number of times. It is also valid if he can remove just one character at the index in the string, and the remaining characters will occur the same number of times. Given a string, determine if it is valid. If so, return YES , otherwise return NO .

Note - You have to write at least 2 additional test cases in which your program will run successfully and provide an explanation for the same.

Example input 1 - s = “abc”. This is a valid string because frequencies are { “a”: 1, “b”: 1, “c”: 1 } Example output 1- YES

Example input 2 - s “abcc”. This string is not valid as we can remove only 1 occurrence of “c”. That leaves character frequencies of { “a”: 1, “b”: 1 , “c”: 2 }

Example output 2 - NO

**Ans: -** <https://github.com/SatishPatri10/iNeuron-Data-Science-Assignment/blob/main/Python/Placement%20Assignment_Satish_Kumar_Patri/Python/2.%20Placement%20Assignment_Satish.ipynb>

**Question 3: -**

Write a program, which would download the data from the provided link, and then read the data and convert that into properly structured data and return it in Excel format.

Note - Write comments wherever necessary explaining the code written.

Link - https://raw.githubusercontent.com/Biuni/PokemonGO-Pokedex/master/pokedex.json

Data Attributes - id: Identification Number - int num: Number of the

* Pokémon in the official Pokédex - int name: Pokémon name –
* string img: URL to an image of this Pokémon - string type:
* Pokémon type -string height: Pokémon height – float
* weight: Pokémon weight - float candy: type of candy used to evolve Pokémon or given
* when transferred - string candy\_count: the amount of candies required to evolve – int
* egg: Number of kilometers to travel to hatch the egg - float spawn\_chance:
* Percentage of spawn chance (NEW) - float avg\_spawns: Number of this pokemon on 10.000 spawns (NEW) – int
* spawn\_time: Spawns most active at the time on this field. Spawn times are the same for all time zones and are expressed in local time. (NEW) - “minutes: seconds” multipliers: Multiplier of Combat Power (CP) for calculating the CP after evolution See below - list of int weakness: Types of
* Pokémon this Pokémon is weak to - list of strings next\_evolution: Number and Name of successive evolutions of Pokémon - list of dict prev\_evolution: Number and Name of previous evolutions of Pokémon - - list of dict

**Ans:** [**-** https://github.com/SatishPatri10/iNeuron-Data-Science-Assignment/blob/main/Python/Placement%20Assignment\_Satish\_Kumar\_Patri/Python/3.%20Placement%20Assignment\_Satish.ipynb](-%20https:/github.com/SatishPatri10/iNeuron-Data-Science-Assignment/blob/main/Python/Placement%20Assignment_Satish_Kumar_Patri/Python/3.%20Placement%20Assignment_Satish.ipynb)

**Question 4 -**

Write a program to download the data from the link given below and then read the data and convert the into the proper structure and return it as a CSV file.

Link - https://data.nasa.gov/resource/y77d-th95.json

Note - Write code comments wherever needed for code understanding.

Sample Data –

Excepted Output Data Attributes

* Name of Earth Meteorite - string id - ID of Earth
* Meteorite - int nametype - string recclass – string
* mass - Mass of Earth Meteorite - float year - Year at which Earth
* Meteorite was hit - datetime format reclat - float recclong – float
* point coordinates - list of int

**Ans:** [- https://github.com/SatishPatri10/iNeuron-Data-Science-Assignment/blob/main/Python/Placement%20Assignment\_Satish\_Kumar\_Patri/Python/4.%20Placement%20Assignment\_Satish.ipynb](-%20https:/github.com/SatishPatri10/iNeuron-Data-Science-Assignment/blob/main/Python/Placement%20Assignment_Satish_Kumar_Patri/Python/4.%20Placement%20Assignment_Satish.ipynb)

**Question 5 -**

Write a program to download the data from the given API link and then extract the following data with proper formatting

Link - http://api.tvmaze.com/singlesearch/shows?q=westworld&embed=episodes

Note - Write proper code comments wherever needed for the code understanding

Sample Data -

Excepted Output Data Attributes -

* id - int url – string
* name - string season
* int number – int
* type - string airdate –
* date format airtime –
* 12-hour time format
* runtime – float
* average rating – float
* summary – string
* without html tags
* medium image link – string
* Original image link - string

**Ans:** - <https://github.com/SatishPatri10/iNeuron-Data-Science-Assignment/blob/main/Python/Placement%20Assignment_Satish_Kumar_Patri/Python/5.%20Placement%20Assignment_Satish.ipynb>

**Question 6 -**

Using the data from Question 3, write code to analyse the data and answer the following questions Note

1. Draw plots to demonstrate the analysis for the following questions for better visualizations.
2. Write code comments wherever required for code understanding

Insights to be drawn -

* Get all Pokemons whose spawn rate is less than 5%
* Get all Pokemons that have less than 4 weaknesses
* Get all Pokemons that have no multipliers at all
* Get all Pokemons that do not have more than 2 evolutions
* Get all Pokemons whose spawn time is less than 300 seconds.

Note - spawn time format is "05:32”, so assume “minute: second” format and perform the analysis.

* Get all Pokemon who have more than two types of capabilities

**Ans:** - <https://github.com/SatishPatri10/iNeuron-Data-Science-Assignment/blob/main/Python/Placement%20Assignment_Satish_Kumar_Patri/Python/6.%20Placement%20Assignment_Satish.ipynb>

**Question 7 -**

Using the data from Question 4, write code to analyze the data and answer the following questions Note -

1. Draw plots to demonstrate the analysis for the following questions for better visualizations
2. Write code comments wherever required for code understanding

Insights to be drawn -

* Get all the Earth meteorites that fell before the year 2000
* Get all the earth meteorites co-ordinates who fell before the year 1970
* Assuming that the mass of the earth meteorites was in kg, get all those whose mass was more than 10000kg

**Ans:** - <https://github.com/SatishPatri10/iNeuron-Data-Science-Assignment/blob/main/Python/Placement%20Assignment_Satish_Kumar_Patri/Python/7.%20Placement%20Assignment_Satish.ipynb>

**Question 8 -**

Using the data from Question 5, write code the analyze the data and answer the following questions Note -

1. Draw plots to demonstrate the analysis for the following questions and better visualizations
2. Write code comments wherever required for code understanding

Insights to be drawn -

* Get all the overall ratings for each season and using plots compare the ratings for all the seasons, like season 1 ratings, season 2, and so on.
* Get all the episode names, whose average rating is more than 8 for every season
* Get all the episode names that aired before May 2019
* Get the episode name from each season with the highest and lowest rating
* Get the summary for the most popular (ratings) episode in every season

**Ans:** - <https://github.com/SatishPatri10/iNeuron-Data-Science-Assignment/blob/main/Python/Placement%20Assignment_Satish_Kumar_Patri/Python/8.%20Placement%20Assignment_Satish.ipynb>

**Question 9 -**

Write a program to read the data from the following link, perform data analysis and answer the following questions

Note -

1. Write code comments wherever required for code understanding

Link - <https://data.wa.gov/api/views/f6w7-q2d2/rows.csv?accessType=DOWNLOAD>

Insights to be drawn -

* Get all the cars and their types that do not qualify for clean alternative fuel vehicle
* Get all TESLA cars with the model year, and model type made in Bothell City.
* Get all the cars that have an electric range of more than 100, and were made after 2015
* Draw plots to show the distribution between city and electric vehicle type

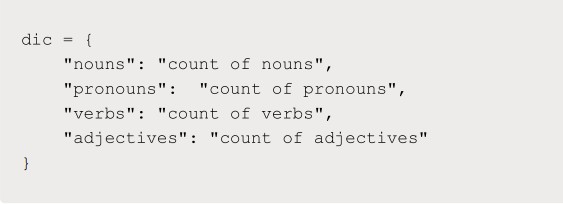
**Ans:** - <https://github.com/SatishPatri10/iNeuron-Data-Science-Assignment/blob/main/Python/Placement%20Assignment_Satish_Kumar_Patri/Python/9.%20Placement%20Assignment_Satish.ipynb>

**Question 10** -

Write a program to count the number of verbs, nouns, pronouns, and adjectives in a given particular phrase or paragraph, and return their respective count as a dictionary.

Note -

1. Write code comments wherever required for code
2. You have to write at least 2 additional test cases in which your program will run successfully and provide an explanation for the same

Example Output -

**Ans:** - <https://github.com/SatishPatri10/iNeuron-Data-Science-Assignment/blob/main/Python/Placement%20Assignment_Satish_Kumar_Patri/Python/10.%20Placement%20Assignment_Satish.ipynb>

**Statistics**

Total Marks: 120

Each question 10 marks

**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?

**Answer: -** The correlation coefficient ranges from -1 to +1, with 0 indicating no correlation, -1 indicating a strong negative relationship, and +1 indicating a strong positive relationship. In this case, a correlation coefficient of 0.7 suggests that as SAT scores increase, college GPAs tend to increase as well.

**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:

1. What percentage of individuals in the dataset have heights between 160 cm and 180 cm?
2. If we randomly select 100 individuals from the dataset, what is the probability that their average height is greater than 175 cm?
3. Assuming the dataset follows a normal distribution, what is the z-score corresponding to a height of 185 cm?
4. We know that 5% of the dataset has heights below a certain value. What is the approximate height corresponding to this threshold?
5. Calculate the coefficient of variation (CV) for the dataset.
6. Calculate the skewness of the dataset and interpret the result.

**Answer: -**

1. To find the percentage of individuals with heights between 160 cm and 180 cm, we can use the properties of the normal distribution. Since the dataset is approximately normally distributed, we can calculate the percentage using the Z-score.

The Z-score for 160 cm can be calculated as:

Z1 = (160 - 170) / 10 = -1

The Z-score for 180 cm can be calculated as:

Z2 = (180 - 170) / 10 = 1

Using a standard normal distribution table or a Z-table, we can find the corresponding probabilities. The probability of individuals having heights between 160 cm and 180 cm is:

P(-1 < Z < 1) ≈ P(Z < 1) - P(Z < -1)

Using the Z-table, the probability of Z < 1 is approximately 0.8413, and the probability of Z < -1 is approximately 0.1587. Therefore:

P(-1 < Z < 1) ≈ 0.8413 - 0.1587 = 0.6826

So, approximately 68.26% of individuals in the dataset have heights between 160 cm and 180 cm.

1. The average height of a random sample of 100 individuals can be considered to follow a normal distribution as well. In this case, the mean height of the sample would still be 170 cm, but the standard deviation of the sample mean, also known as the standard error of the mean, can be calculated as:

Standard Error = Standard Deviation / sqrt(sample size)

= 10 / sqrt(100)

= 10 / 10

= 1

We need to find the probability that the average height is greater than 175 cm. We can calculate the Z-score for 175 cm as:

Z = (175 - 170) / 1 = 5

Using the Z-table, the probability of Z > 5 is essentially zero. Therefore, the probability that the average height of a random sample of 100 individuals is greater than 175 cm is approximately 0.

1. To calculate the z-score corresponding to a height of 185 cm, we can use the formula:

Z = (X - Mean) / Standard Deviation

Z = (185 - 170) / 10 = 1.5

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

1. We know that 5% of the dataset has heights below a certain value. To find the height corresponding to this threshold, we need to find the Z-score that corresponds to a cumulative probability of 5% in the standard normal distribution.

Using the Z-table, we can find the Z-score that corresponds to a cumulative probability of 5% (or 0.05). This Z-score is approximately -1.645. We can then use the formula for the Z-score to find the corresponding height:

Z = (X - Mean) / Standard Deviation

-1.645 = (X - 170) / 10

Solving for X, we get:

X - 170 = -1.645 \* 10

X - 170 = -16.45

X = 170 - 16.45

X ≈ 153.55

So, the approximate height corresponding to the threshold of 5% is 153.55 cm.

1. The coefficient of variation (CV) is a measure of relative variability and is calculated by dividing the standard deviation by the mean, then multiplying by 100 to express it as a percentage.

CV = (Standard Deviation / Mean) \* 100

= (10 / 170)

**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\_

1. Measure the dispersion in both and interpret the results.
2. Calculate mean and 5% confidence interval and plot it in a graph
3. Calculate the Mean absolute deviation and Standard deviation and interpret the results.
4. Calculate the correlation coefficient and check the significance of it at 1% level of significance.

**Ans: -** <https://github.com/SatishPatri10/iNeuron-Data-Science-Assignment/blob/main/Python/Placement%20Assignment_Satish_Kumar_Patri/Statistic/Q3.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)?

**Answer: -**

There are 20 slips of paper in the hat, and you want to find the probability of drawing a perfect square number (1, 4, 9, or 16). The total number of possible outcomes is 20 since there are 20 slips of paper. The number of favorable outcomes (perfect square numbers) is 4 (1, 4, 9, and 16).

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

Probability = (Number of Favorable Outcomes) / (Total Number of Outcomes)

= 4 / 20

= 1 / 5

= 0.2

So, the probability of drawing a perfect square number from the hat is 0.2 or 20%

**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?

**Answer:** - To solve this problem, we can use Bayes' theorem. Let us define the events:

A: The taxi belongs to Company A

B: The taxi is late

We are looking for the conditional probability P(A|B), which is the probability that the taxi belongs to Company A given that it is late.

According to the problem statement:

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

P(B|A) = 1 - 0.95 = 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|B) = 1 - 0.90 = 0.10 (Company B's taxis have a 90% success rate, so the probability of being late is 1 - 0.90 = 0.10)

Using Bayes' theorem:

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

Plugging in the values:

P(A|B) = (0.05 \* 0.8) / (0.05 \* 0.8 + 0.10 \* 0.2)

= 0.04 / (0.04 + 0.02)

= 0.04 / 0.06

= 2/3

≈ 0.6667

Therefore, the probability that a randomly selected taxi, which is late, 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\_

**Answer -** <https://github.com/SatishPatri10/iNeuron-Data-Science-Assignment/blob/main/Python/Placement%20Assignment_Satish_Kumar_Patri/Statistic/Q6.ipynb>

Q-7. 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 𝑋 = 4 Find the

1. Variance of Y
2. Coefficient of determination of C and Y
3. Standard error of estimate of X on Y and of Y on X.

Answer: -

Given equations of the regression lines:

2𝑋 + 3 − 8 = 0

2𝑌 + 𝑋 − 5 = 0

1. Variance of Y: The equation of the regression line for Y is 2𝑌 + 𝑋 − 5 = 0. Rearranging the equation, we have: 2𝑌 = -𝑋 + 5 𝑌 = (-1/2)𝑋 + 5/2

The coefficient of determination (R²) is the proportion of the variance in Y that can be explained by the regression model. Since the variance of X is given as 4, we can calculate the variance of Y using the formula:

Variance of Y = R² \* Variance of X

Substituting the values: Variance of Y = R² \* 4

1. Coefficient of determination of X and Y: The coefficient of determination (R²) is the square of the correlation coefficient (r). The correlation coefficient can be found by comparing the slopes of the regression lines.

Slope of the regression line for X: 2 Slope of the regression line for Y: (-1/2)

The correlation coefficient (r) is the ratio of the two slopes: r = (-1/2) / 2 = -1/4

Coefficient of determination (R²) = r²

1. Standard error of estimate: The standard error of estimate measures the average distance between the predicted values from the regression line and the actual values.

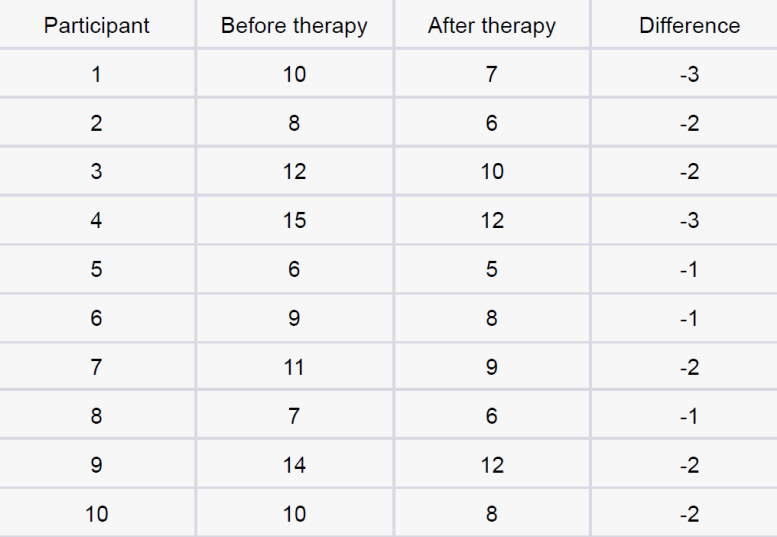
Standard error of estimate of X on Y: To calculate this, we use the formula: Standard error = √((1 - R²) \* Variance of X)

Substituting the values: Standard error of estimate of X on Y = √((1 - R²) \* 4)

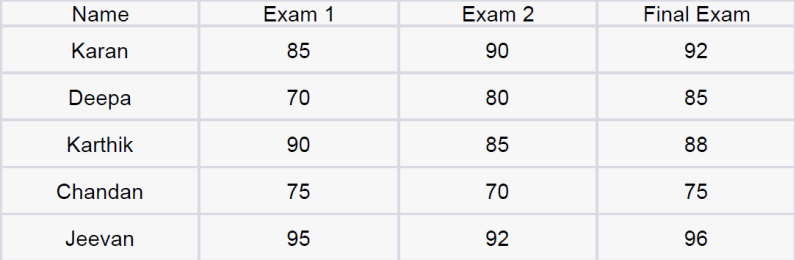
Standard error of estimate of Y on X: To calculate this, we use the formula: Standard error = √((1 - R²) \* Variance of Y)

Substituting the values: Standard error of estimate of Y on X = √((1 - R²) \* Variance of Y)

**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



**Answer:** - <https://github.com/SatishPatri10/iNeuron-Data-Science-Assignment/blob/main/Python/Placement%20Assignment_Satish_Kumar_Patri/Statistic/Q8.ipynb>

**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.

**Answer** - <https://github.com/SatishPatri10/iNeuron-Data-Science-Assignment/blob/main/Python/Placement%20Assignment_Satish_Kumar_Patri/Statistic/Q9.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.

1. What is the probability that exactly 20 bulbs are defective?
2. What is the probability that at least 10 bulbs are defective?
3. What is the probability that at max 15 bulbs are defective?
4. On average, how many defective bulbs would you expect in a batch of 500?

**Answer: -**

1. The probability of a bulb being defective is 0.05, and the factory produces 500 bulbs. We can use the binomial probability formula to calculate this:

P(X = k) = C(n, k) \* p^k \* (1 - p)^(n - k)

Where:

- P(X = k) is the probability of exactly k defective bulbs.

- C(n, k) is the number of ways to choose k items out of n, also known as the binomial coefficient.

- p is the probability of a bulb being defective.

- n is the total number of bulbs.

Using these values, we can calculate the probability:

P(X = 20) = C(500, 20) \* (0.05)^20 \* (1 - 0.05)^(500 - 20)

1. To find the probability that at least 10 bulbs are defective, we need to calculate the probability of having 10, 11, 12, ..., up to 500 defective bulbs and sum them all up:

P(X >= 10) = P(X = 10) + P(X = 11) + ... + P(X = 500)

We can use the same binomial probability formula to calculate each individual probability and then sum them up.

1. To find the probability that at most 15 bulbs are defective, we need to calculate the probability of having 0, 1, 2, ..., up to 15 defective bulbs and sum them all up:

P(X <= 15) = P(X = 0) + P(X = 1) + ... + P(X = 15)

Again, we can use the binomial probability formula to calculate each individual probability and then sum them up.

1. The average number of defective bulbs in a batch of 500 can be calculated using the expected value formula:

E(X) = n \* p

Where:

- E(X) is the expected value or mean of the binomial distribution.

- n is the total number of bulbs.

- p is the probability of a bulb being defective.

Using these values, we can calculate the expected number of defective bulbs:

E(X) = 500 \* 0.05

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

https://drive.google.com/file/d/1mCjtYHiX--mMUjicuaP2gH3k-SnFxt8Y/view?usp

=share\_

1. Check whether the distribution of all the classes are the same or not.
2. Check for the equality of variance
3. Which amount LDA and QDA would perform better on this data for classification and why.
4. Check the equality of mean for between all the classes.

**Answer -** <https://github.com/SatishPatri10/iNeuron-Data-Science-Assignment/blob/main/Python/Placement%20Assignment_Satish_Kumar_Patri/Statistic/Q11.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.

1. 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.
2. 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.

**Answer –** To determine if there is a significant difference in the mean improvement scores between Group A and Group B, we can conduct a t-test. The null hypothesis (H0) states that there is no significant difference between the two groups, while the alternative hypothesis (Ha) states that there is a significant difference.

1. Let's perform the t-test using the provided data:

Group A:

Sample size (na) = 30

Mean improvement score (x̄a) = 2.5

Standard deviation (sa) = 0.8

Group B:

Sample size (nb) = 30

Mean improvement score (x̄b) = 2.2

Standard deviation (sb) = 0.6

Using a two-sample independent t-test, we can calculate the t-value:

t = (x̄a - x̄b) / sqrt((sa^2 / na) + (sb^2 / nb))

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

t = 0.3 / sqrt((0.064 / 30) + (0.036 / 30))

t ≈ 0.3 / sqrt(0.002133 + 0.0012)

t ≈ 0.3 / sqrt(0.003333)

t ≈ 0.3 / 0.057732

t ≈ 5.19

1. Now, let's compare the calculated t-value to the critical t-value at a significance level of 0.05. Since we have equal sample sizes and can assume approximately equal variances, we will use the pooled standard deviation to calculate the degrees of freedom.

Degrees of freedom (df) = na + nb - 2 = 30 + 30 - 2 = 58

The critical t-value with df = 58 and a two-tailed test at a significance level of 0.05 is approximately ±2.001.

Since the calculated t-value (5.19) is greater than the critical t-value (2.001), we reject the null hypothesis (H0) in favor of the alternative hypothesis (Ha). This means there is a significant difference in the mean improvement scores between Group A and Group B.

Conclusion: The new drug has shown a significantly greater mean improvement in the specific symptom compared to the standard drug.