

10 MARCH ASSIGNMENT

Q1: What is Estimation Statistics? Explain point estimate and interval estimate.

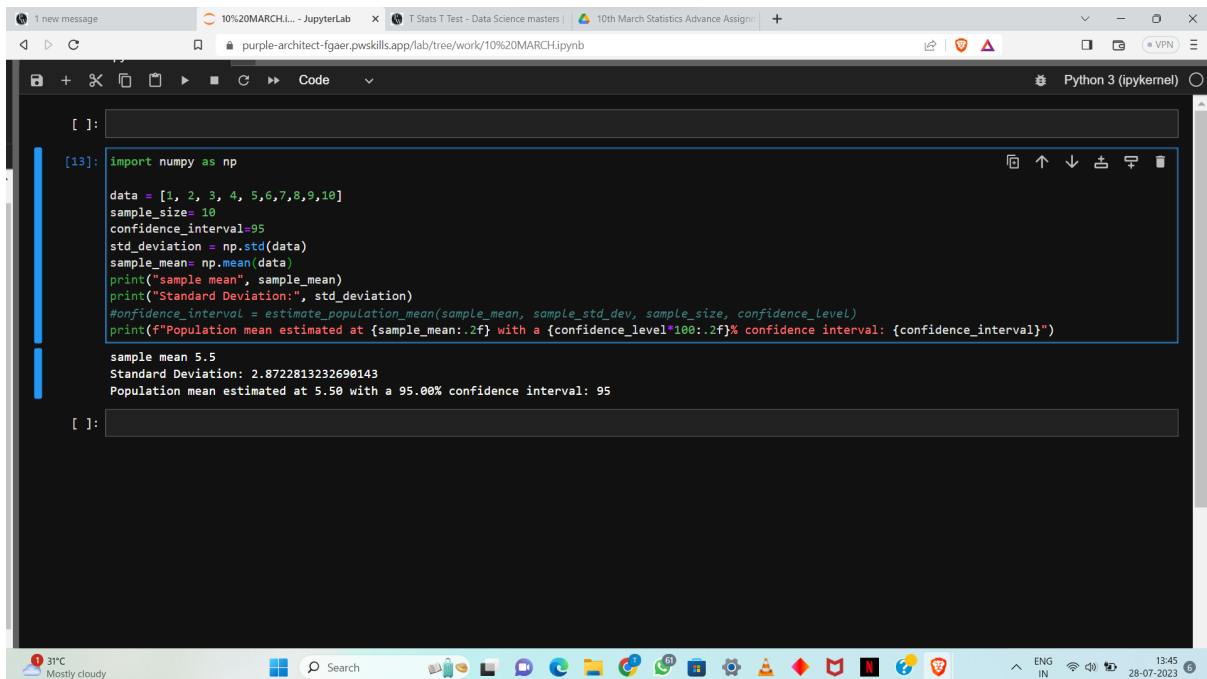
Estimation statistics is a branch of inferential statistics that focuses on estimating unknown population parameters based on sample data. In inferential statistics, we use sample data to make inferences about a larger population. Estimation plays a crucial role in this process because we usually cannot access or collect data from the entire population.

There are two main types of estimates in estimation statistics:

Point Estimate: A point estimate is a single value that serves as the best guess or approximation of the population parameter. It is obtained by calculating a statistic from the sample data that corresponds to the parameter of interest. For example, if we want to estimate the population mean, the sample mean would be the point estimate. Similarly, if we want to estimate the population proportion, the sample proportion would be the point estimate. Point estimates are easy to compute and provide a straightforward representation of the parameter's value.

Interval Estimate: An interval estimate, also known as a confidence interval, is a range of values within which the true population parameter is likely to lie, along with a degree of confidence. Instead of providing a single point estimate, interval estimates offer a range of plausible values, accounting for the uncertainty inherent in estimating population parameters from sample data. The confidence level associated with the interval indicates the percentage of times that the interval would contain the true population parameter if the same sampling and estimation process were repeated.

Q2. Write a Python function to estimate the population mean using a sample mean and standard deviation.



```
[ ]:
[13]: import numpy as np

data = [1, 2, 3, 4, 5, 6, 7, 8, 9, 10]
sample_size= 10
confidence_interval=95
std_deviation = np.std(data)
sample_mean= np.mean(data)
print("sample mean", sample_mean)
print("Standard Deviation:", std_deviation)
#confidence_interval = estimate_population_mean(sample_mean, sample_std_dev, sample_size, confidence_level)
print(f"Population mean estimated at {sample_mean:.2f} with a {confidence_level*100:.2f}% confidence interval: {confidence_interval}")

sample mean 5.5
Standard Deviation: 2.8722813232690143
Population mean estimated at 5.50 with a 95.00% confidence interval: 95

[ ]:
```

Q3: What is Hypothesis testing? Why is it used? State the importance of Hypothesis testing.

Hypothesis testing is a statistical method used to make decisions about a population based on sample data. It involves formulating two competing hypotheses, the null hypothesis (H_0) and the alternative hypothesis (H_a), and then using sample data to determine which hypothesis is more likely to be true.

Here's a breakdown of the key components of hypothesis testing:

1. **Null Hypothesis (H_0):** The null hypothesis is the default assumption or claim that there is no significant difference or effect. It represents the status quo or the idea that any observed difference or effect is due to random chance.
2. **Alternative Hypothesis (H_a):** The alternative hypothesis is the claim or assertion that contradicts the null hypothesis. It represents the researcher's interest in finding evidence of a specific effect, difference, or relationship.
3. **Test Statistic:** The test statistic is a numerical value calculated from the sample data that is used to assess how well the data supports the null hypothesis or the alternative hypothesis.

4. P-value: The p-value is the probability of obtaining results as extreme or more extreme than the observed data, assuming the null hypothesis is true. A low p-value indicates that the data is unlikely to occur if the null hypothesis is true, leading to a rejection of the null hypothesis in favor of the alternative hypothesis.

5. Significance Level (Alpha): The significance level (usually denoted by alpha, α) is a threshold chosen by the researcher to determine the level of evidence required to reject the null hypothesis. Commonly used significance levels are 0.05 (5%) and 0.01 (1%).

The importance of hypothesis testing lies in its role as a formal and systematic approach to drawing conclusions from data. Here are some reasons why hypothesis testing is essential:

1. Decision-Making: Hypothesis testing provides a structured way to make informed decisions about whether to accept or reject a claim or hypothesis based on evidence from the data.

2. Scientific Validity: In scientific research, hypothesis testing helps researchers determine the validity of their findings and whether they can confidently draw conclusions about the population under study.

3. Objectivity: Hypothesis testing introduces objectivity into the decision-making process by relying on statistical evidence rather than subjective opinions or beliefs.

4. Inference to Population: With hypothesis testing, researchers can make inferences about the population based on the results observed in the sample, helping to generalize findings to a larger group.

5. Practical Applications: Hypothesis testing is widely used in various fields, including medicine, social sciences, engineering, economics, and more, to test hypotheses and support evidence-based decision-making.

6. Identifying Relationships: Hypothesis testing allows researchers to explore and identify relationships between variables and make predictions about cause-and-effect relationships.

Q4. Create a hypothesis that states whether the average weight of male college students is greater than the average weight of female college students.

Hypothesis:

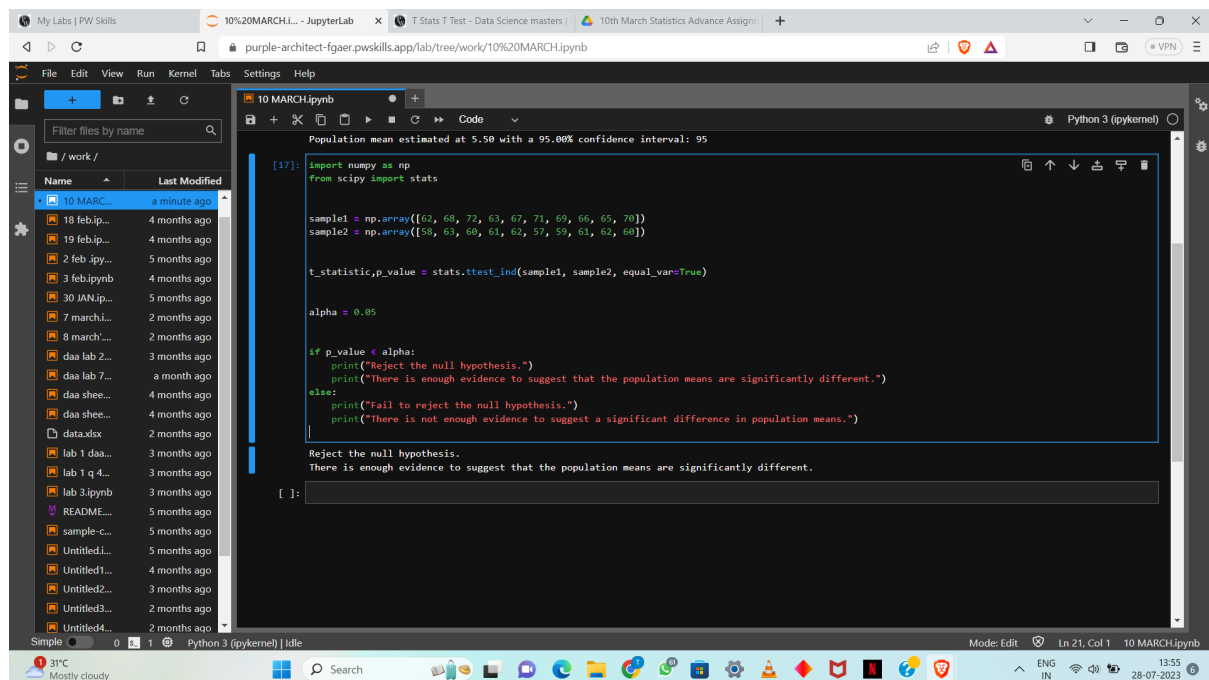
Null Hypothesis (H_0): The average weight of male college students is equal to or less than the average weight of female college students.

Alternative Hypothesis (H_a): The average weight of male college students is greater than the average weight of female college students.

In symbols:

$H_0: \mu_{\text{male}} \leq \mu_{\text{female}}$ $H_a: \mu_{\text{male}} > \mu_{\text{female}}$

Q5. Write a Python script to conduct a hypothesis test on the difference between two population means, given a sample from each population.



```
Population mean estimated at 5.50 with a 95.00% confidence interval: 95

[17]: import numpy as np
      from scipy import stats

      sample1 = np.array([62, 68, 72, 63, 67, 71, 69, 66, 65, 70])
      sample2 = np.array([58, 63, 60, 61, 62, 57, 59, 61, 62, 60])

      t_statistic, p_value = stats.ttest_ind(sample1, sample2, equal_var=True)

      alpha = 0.05

      if p_value < alpha:
          print("Reject the null hypothesis.")
          print("There is enough evidence to suggest that the population means are significantly different.")
      else:
          print("Fail to reject the null hypothesis.")
          print("There is not enough evidence to suggest a significant difference in population means.")

      Reject the null hypothesis.
      There is enough evidence to suggest that the population means are significantly different.

[ ]:
```

Q6: What is a null and alternative hypothesis? Give some examples.

In statistical hypothesis testing, the null hypothesis (H_0) and the alternative hypothesis (H_a) are two competing statements or claims about a population parameter. The hypothesis testing process involves gathering sample data and using it to determine whether there is enough evidence to reject the null hypothesis in favor of the alternative hypothesis.

1. Null Hypothesis (H_0):

The null hypothesis is the default assumption or claim that there is no significant difference, effect, or relationship in the population. It suggests that any observed difference or effect in the sample data is due to random variation or chance. In hypothesis testing, we assume the null hypothesis to be true unless there is enough evidence to suggest otherwise.

Example 1 (Coin Tossing):

H0: The probability of getting heads on a fair coin toss is 0.5.

Example 2 (Medical Treatment):

H0: A new drug has no effect on patients' recovery time compared to the standard treatment.

Example 3 (Survey):

H0: There is no difference in the average salary between male and female employees in a company.

2. Alternative Hypothesis (Ha):

The alternative hypothesis is the claim or assertion that contradicts the null hypothesis. It represents the researcher's interest in finding evidence of a specific effect, difference, or relationship in the population. The alternative hypothesis is what the researcher hopes to support with the sample data.

Example 1 (Coin Tossing):

Ha: The probability of getting heads on a fair coin toss is not 0.5 (indicating a biased coin).

Example 2 (Medical Treatment):

Ha: The new drug leads to faster patient recovery time compared to the standard treatment.

Example 3 (Survey):

Ha: There is a significant difference in the average salary between male and female employees in a company.

In hypothesis testing, we collect sample data and use statistical tests to determine whether the evidence supports the null hypothesis or favors the alternative hypothesis. If the evidence is strong enough, we reject the null hypothesis and accept the alternative hypothesis. However, if the evidence is not strong enough, we fail to reject the null hypothesis, meaning we do not have sufficient evidence to support the alternative hypothesis, and we stick with the default assumption of no significant effect or difference in the population.

Q7: Write down the steps involved in hypothesis testing.

Hypothesis testing is a structured and systematic process used to make decisions about a population based on sample data. The steps involved in hypothesis testing are as follows:

Step 1: Formulate the Hypothesis

- State the null hypothesis (H0): The default assumption that there is no significant difference, effect, or relationship in the population.
- State the alternative hypothesis (Ha): The claim or assertion that contradicts the null hypothesis and represents the researcher's interest.

Step 2: Set the Significance Level (Alpha)

- Choose a significance level (alpha, denoted by α), which represents the threshold for determining whether the evidence is strong enough to reject the null hypothesis. Commonly used values are 0.05 (5%) and 0.01 (1%).

Step 3: Collect and Prepare Data

- Gather sample data from the population of interest. Ensure the data is representative and unbiased.

Step 4: Choose a Statistical Test

- Select an appropriate statistical test based on the nature of the data and the hypotheses being tested. Common tests include t-tests, z-tests, chi-square tests, ANOVA, etc.

Step 5: Calculate the Test Statistic

- Compute the test statistic using the sample data. The test statistic quantifies the distance between the sample data and the null hypothesis.

Step 6: Determine the P-value

- Calculate the p-value, which is the probability of obtaining results as extreme or more extreme than the observed data, assuming the null hypothesis is true.

Step 7: Compare P-value with Alpha

- Compare the calculated p-value with the chosen significance level (alpha).
- If $p\text{-value} < \alpha$, reject the null hypothesis in favor of the alternative hypothesis. There is enough evidence to suggest a significant difference or effect in the population.
- If $p\text{-value} \geq \alpha$, fail to reject the null hypothesis. There is not enough evidence to suggest a significant difference or effect in the population, and we stick with the default assumption.

Step 8: Draw Conclusions

- Based on the comparison of the p-value with alpha, make a decision about whether to reject or fail to reject the null hypothesis.
- State the conclusions and interpret the results in the context of the problem.

Step 9: Report the Findings

- Communicate the results, including the hypothesis statements, the test used, the test statistic, the p-value, and the final decision, in a clear and concise manner.

It is important to note that the conclusion reached through hypothesis testing is not an absolute truth but rather a probabilistic statement based on the available sample data. The results may indicate the likelihood of an effect or difference in the population, but they do not prove causation or provide certainty about the entire population.

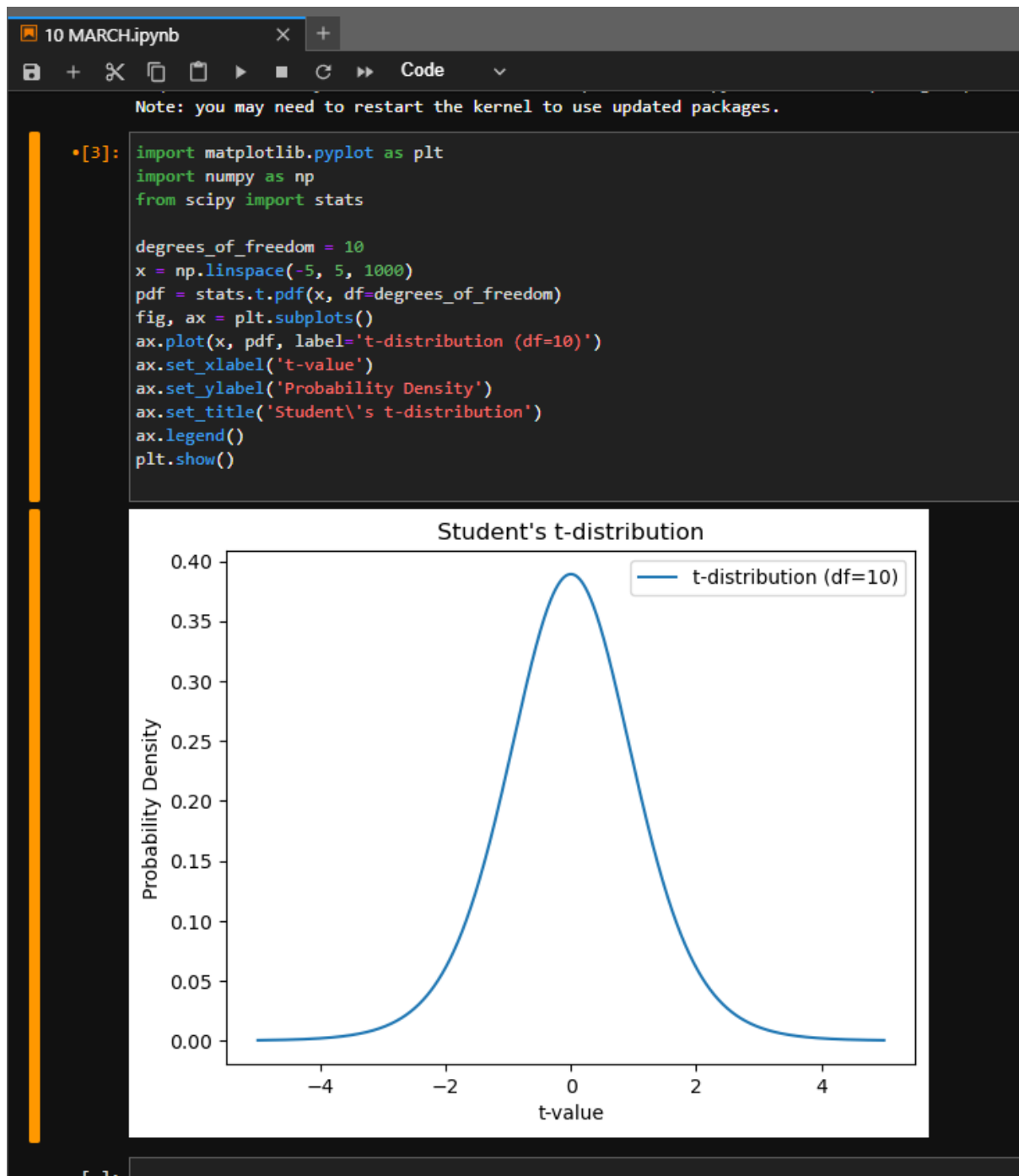
Q8. Define p-value and explain its significance in hypothesis testing.

The p-value (probability value) is a crucial concept in hypothesis testing. It represents the probability of obtaining results as extreme or more extreme than the observed data, assuming the null hypothesis is true. In other words, the p-value quantifies the strength of evidence against the null hypothesis based on the sample data.

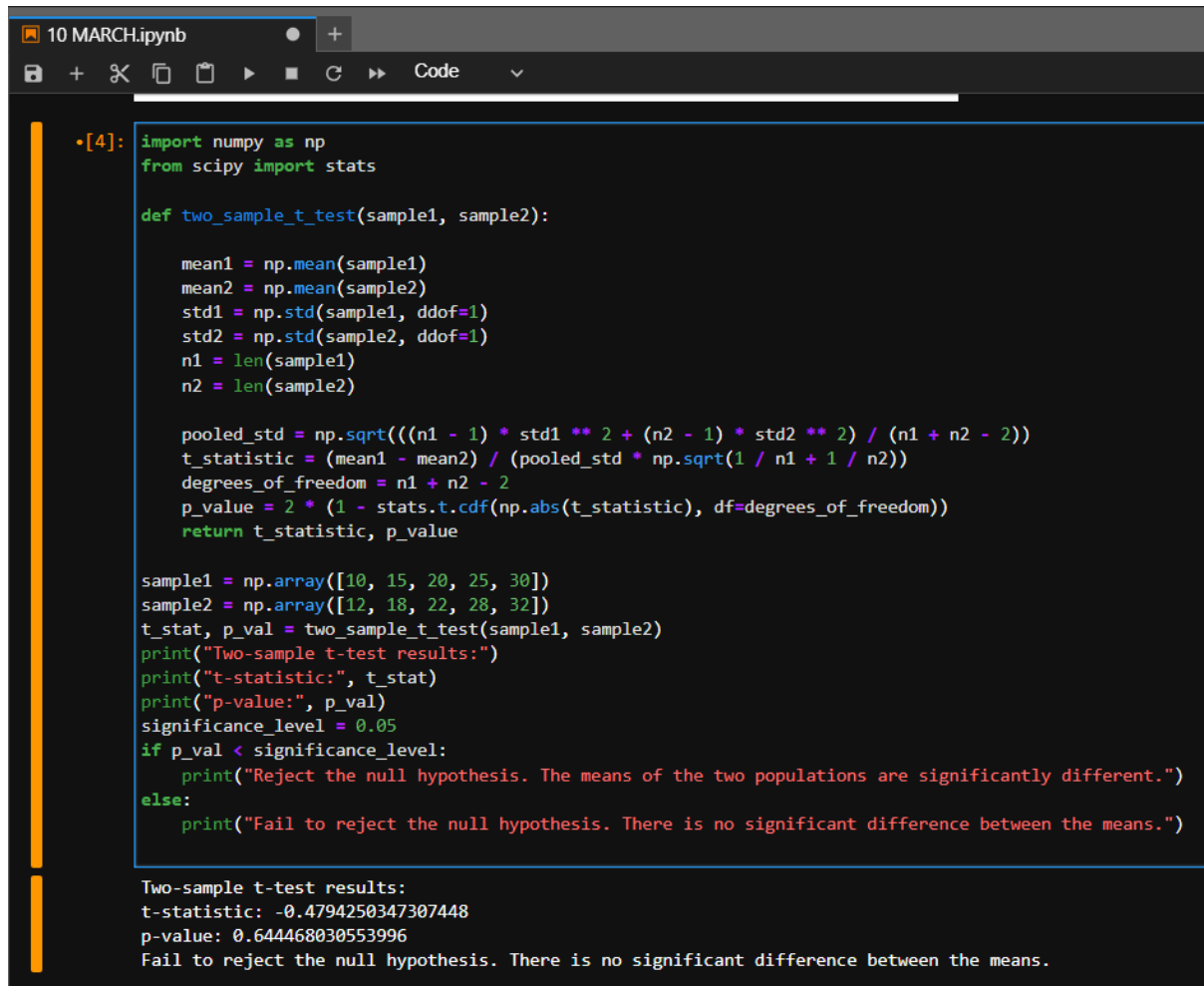
Significance of p-value in Hypothesis Testing:

1. **Testing Hypotheses:** The p-value is used to make a decision about whether to reject or fail to reject the null hypothesis. By comparing the p-value with the chosen significance level (α), which represents the threshold for statistical significance, you can draw conclusions about the hypotheses being tested.
2. **Decision Rule:** If the p-value is less than or equal to the significance level ($p \leq \alpha$), it suggests that the observed data is unlikely to occur if the null hypothesis is true. In this case, you reject the null hypothesis in favor of the alternative hypothesis. This means you have enough evidence to support the claim made in the alternative hypothesis.
3. **Probabilistic Statement:** The p-value provides a probabilistic statement about the likelihood of the observed results under the null hypothesis. It helps researchers avoid making binary decisions and acknowledges the uncertainty inherent in hypothesis testing.
4. **Level of Confidence:** A lower p-value indicates stronger evidence against the null hypothesis. A very small p-value (e.g., $p < 0.01$) suggests highly significant results, providing more confidence in rejecting the null hypothesis. Conversely, a higher p-value (e.g., $p > 0.05$) indicates weaker evidence and a failure to reject the null hypothesis.
5. **Not Proving the Null Hypothesis:** It's important to note that a non-significant p-value (e.g., $p > \alpha$) does not prove the null hypothesis to be true. Instead, it indicates that there is not enough evidence in the sample data to reject the null hypothesis. Failing to reject the null hypothesis does not imply that the null hypothesis is true; it means you do not have sufficient evidence to support the alternative hypothesis.
6. **Reproducibility:** The p-value allows other researchers to evaluate the study's results and conclusions. By reporting the p-value, the research becomes more transparent, and others can attempt to replicate the findings.

Q9. Generate a Student's t-distribution plot using Python's matplotlib library, with the degrees of freedom parameter set to 10.



Q10. Write a Python program to calculate the two-sample t-test for independent samples, given two random samples of equal size and a null hypothesis that the population means are equal.



```
10 MARCH.ipynb
+
Code
•[4]: import numpy as np
      from scipy import stats

      def two_sample_t_test(sample1, sample2):

          mean1 = np.mean(sample1)
          mean2 = np.mean(sample2)
          std1 = np.std(sample1, ddof=1)
          std2 = np.std(sample2, ddof=1)
          n1 = len(sample1)
          n2 = len(sample2)

          pooled_std = np.sqrt(((n1 - 1) * std1 ** 2 + (n2 - 1) * std2 ** 2) / (n1 + n2 - 2))
          t_statistic = (mean1 - mean2) / (pooled_std * np.sqrt(1 / n1 + 1 / n2))
          degrees_of_freedom = n1 + n2 - 2
          p_value = 2 * (1 - stats.t.cdf(np.abs(t_statistic), df=degrees_of_freedom))
          return t_statistic, p_value

      sample1 = np.array([10, 15, 20, 25, 30])
      sample2 = np.array([12, 18, 22, 28, 32])
      t_stat, p_val = two_sample_t_test(sample1, sample2)
      print("Two-sample t-test results:")
      print("t-statistic:", t_stat)
      print("p-value:", p_val)
      significance_level = 0.05
      if p_val < significance_level:
          print("Reject the null hypothesis. The means of the two populations are significantly different.")
      else:
          print("Fail to reject the null hypothesis. There is no significant difference between the means.")

      Two-sample t-test results:
      t-statistic: -0.4794250347307448
      p-value: 0.644468030553996
      Fail to reject the null hypothesis. There is no significant difference between the means.
```

Q11: What is Student's t distribution? When to use the t-Distribution.

Student's t-distribution, often referred to simply as the t-distribution, is a probability distribution that is used in hypothesis testing and confidence interval estimation when the sample size is small and the population standard deviation is unknown. It is similar to the standard normal (Z) distribution but has heavier tails, which accounts for the increased uncertainty that arises from using a sample standard deviation to estimate the population standard deviation.

Characteristics of the t-distribution:

1. Shape: The t-distribution has a bell-shaped curve, like the standard normal distribution, but with slightly fatter tails.
2. Parameter: The shape of the t-distribution depends on the degrees of freedom (df), which is related to the sample size. As the sample size increases, the t-distribution approaches the standard normal distribution.

When to use the t-Distribution:

The t-distribution is used in situations where the following conditions are met:

1. Small Sample Size: The sample size is relatively small (typically less than 30). For larger sample sizes, the t-distribution becomes very close to the standard normal distribution, making the use of Z-tests more appropriate.
2. Unknown Population Standard Deviation: The population standard deviation (σ) is unknown, so you need to use the sample standard deviation (s) to estimate it.

Common scenarios where the t-distribution is used include:

1. Two-sample t-test: When comparing the means of two independent samples.
2. Paired t-test: When comparing the means of two related samples (e.g., before and after measurements).
3. One-sample t-test: When comparing the mean of a single sample to a known or hypothesised value.
4. Confidence interval estimation: When calculating confidence intervals for population parameters (e.g., population mean) based on sample data.

Q12: What is t-statistic? State the formula for t-statistic.

The t-statistic is a value calculated from sample data and is used in hypothesis testing to determine whether there is a significant difference between the sample mean and a hypothesised population mean. It quantifies the difference between the sample mean and the population mean in terms of the standard error.

The formula for the t-statistic is as follows:

$$t = (\bar{x} - \mu) / (s / \sqrt{n})$$

Where:

t = t-statistic

\bar{x} = Sample mean

μ = Population mean (hypothesised value under the null hypothesis)

s = Sample standard deviation

n = Sample size

In this formula, the numerator ($\bar{x} - \mu$) represents the difference between the sample mean and the hypothesised population mean. The denominator (s / \sqrt{n}) is the standard error, which measures the variability of the sample mean around the population mean. The standard error accounts for the uncertainty introduced by using the sample mean to estimate the population mean.

The t-statistic is a measure of how many standard errors the sample mean is away from the hypothesised population mean. A larger t-statistic indicates a larger difference between the sample mean and the population mean, which may provide stronger evidence to reject the null hypothesis in favour of the alternative hypothesis.

In hypothesis testing, the t-statistic is compared with critical values from the t-distribution or used to calculate the p-value to determine whether the difference between the sample mean and the hypothesised population mean is statistically significant. If the t-statistic is far from zero, it suggests that the sample mean is significantly different from the population mean, leading to a rejection of the null hypothesis in favour of the alternative hypothesis.

Q13. A coffee shop owner wants to estimate the average daily revenue for their shop. They take a random sample of 50 days and find the sample mean revenue to be \$500 with a standard deviation of \$50. Estimate the population mean revenue with a 95% confidence interval.

75.82

10 March

Assignment

PAGE NO	
DATE	

Que 13

A coffee shop owner want to estimate the average daily revenue for their shop. They take a random sample of 50 days and find the sample mean revenue to be ₹ 500 with a standard deviation of ₹ 50. Estimate the population mean revenue with a 95% confidence interval.

Ans

$$\bar{x} = 500$$

$$n = 50$$

$$\sigma = 50$$

$$C.I = 95\% = 0.95$$

$$\alpha = 0.05$$

Standard Error (SE) of Sample Mean

$$SE = s/\sqrt{n}$$

$$= 50/\sqrt{50} = 7.07$$

$$ME = \text{critical value} * SE$$

$$= 2.0096 * 7.07$$

$$= 14.20$$

The population mean lies within the interval ± 14.20 of the sample mean $\bar{x} = 500$.
 i.e. $[500 - 14.20, 500 + 14.20]$
 i.e. $[485.80, 514.20]$

Q14. A researcher hypothesizes that a new drug will decrease blood pressure by 10 mmHg. They conduct a clinical trial with 100 patients and find that the sample mean decrease in blood pressure is 8 mmHg with a standard deviation of 3 mmHg. Test the hypothesis with a significance level of 0.05.

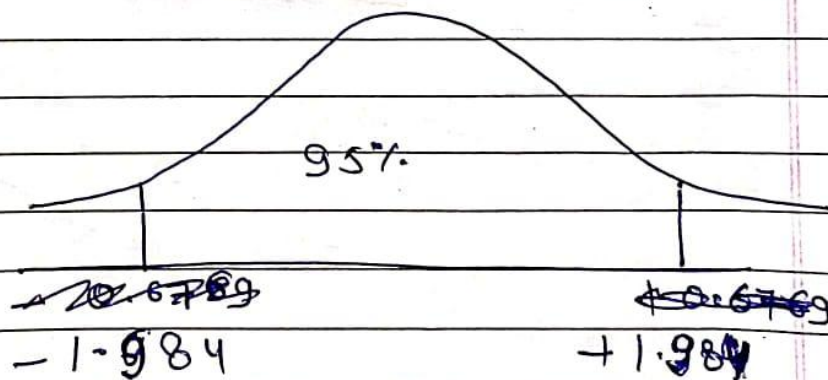
Q.14 A Researcher hypothesizes that a new drug will decrease blood pressure by 10 mmHg. They conduct a clinical trial within 100 patients and find that the sample mean decrease in blood pressure is 8 mmHg with a standard deviation of 3 mmHg. Test the Hypothesis with a significance level of 0.05.

Ans: 14. $\mu = 10 \text{ mmHg}$
 $\bar{x} = 8 \text{ mmHg}$
 $S(\sigma) = 3 \text{ mmHg}$
 $n = 100$

Null Hypothesis, $H_0: \mu = 10$
 Alternative " $H_1: \mu \neq 10$

Degree of freedom = 99

→ Decision Rule



$$T\text{-test} = \frac{\bar{x} - \mu}{S/\sqrt{n}}$$

$$= \frac{8 - 10}{3/\sqrt{100}} = -2 \times \frac{10}{3}$$

$$= -6.667$$

~~$$-6.667 < -0.6769$$~~

$$-6.667 < -0.6769 \quad \therefore \text{we accept}$$

~~we accept Null Hypothesis~~

We Reject the Null

Hypothesis

Q15. An electronics company produces a certain type of product with a mean weight of 5 pounds and a standard deviation of 0.5 pounds. A random sample of 25 products is taken, and the sample mean weight is found to be 4.8 pounds. Test the hypothesis that the true mean weight of the products is less than 5 pounds with a significance level of 0.01.

Ques An electronic company produces a certain type of product with a mean weight of 5 pounds and std 0.5 pounds. A Random Sample of 25 product and Sample Mean weight 4.8. Test the Hypothesis that the true mean weight of the product is less than 5 pounds with a Significance level of 0.01.

Ans

$$\mu = 5$$

$$\sigma = 0.5$$

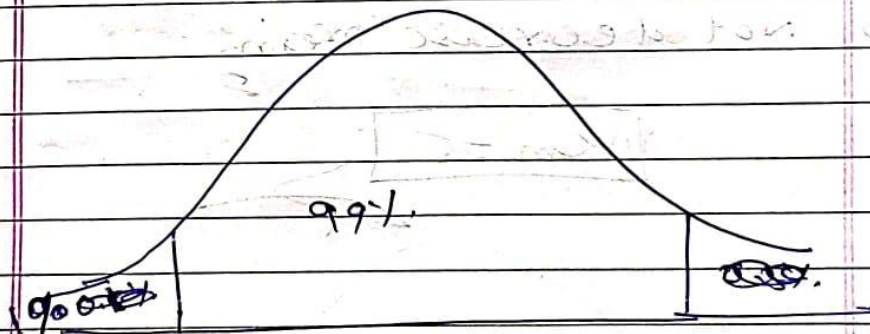
$$n = 25$$

$$\bar{x} = 4.8$$

$$\alpha = 0.01$$

Null Hypothesis: $H_0: \mu = 5$

Alternative: $H_1: \mu < 5$



$$Z\text{-test} = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

$$= \frac{4.8 - 5}{0.5 / \sqrt{25}}$$

$$= -2$$

$$p = 0.0455$$

$p < \text{significance}$

$$0.0455 < 0.01$$

We accept the Null

Hypothesis

→ Not decrease mean.

$$\mu_{\text{mean}} = 5$$

$$= 0.10 - 1$$

Q16. Two groups of students are given different study materials to prepare for a test. The first group ($n_1 = 30$) has a mean score of 80 with a standard deviation of 10, and the second group ($n_2 = 40$) has a mean score of 75 with a standard deviation of 8. Test the hypothesis that the population means for the two groups are equal with a significance level of 0.01.

Ques 16 First Group ($n_1 = 30$) mean 80
std. = 10, Second group ($n_2 = 40$)
mean = 75, std = 8. Test the
Hypothesis that the population
mean for the two groups are equal
with a significance level of 0.01

Ans 16

Null Hypothesis:- The mean of
both groups are same

Alternative Hypothesis:- The mean of
both groups are different

$$\alpha = 0.01$$

$$10.0 > 10.0$$

$$T_{\text{test}} = \frac{(80 - 75)}{\sqrt{\frac{30}{10} + \frac{40}{8}}}$$

$$= \frac{5}{\sqrt{3 + 5}} = \frac{5}{\sqrt{8}}$$

$$= \frac{5}{2.828} = 1.767$$

$$= 1.767$$

$$P = 0.0340$$

$$T\text{-test} = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

$$= \frac{(80 - 75)}{\sqrt{\frac{10^2}{30} + \frac{8^2}{40}}} = \frac{5}{\sqrt{\frac{100}{30} + \frac{64}{40}}} = \frac{5}{\sqrt{3.33 + 1.6}} = \frac{5}{\sqrt{4.93}} = \frac{5}{2.22} = 2.25$$

$$= \frac{5}{\sqrt{\frac{40 \times 100 + 30 \times 64}{30 \times 40}}} = \frac{5}{\sqrt{\frac{4000 + 1920}{1200}}} = \frac{5}{\sqrt{\frac{5920}{1200}}} = \frac{5}{\sqrt{4.93}} = \frac{5}{2.22} = 2.25$$

$$= \frac{5}{\sqrt{\frac{5920}{1200}}} = \frac{5}{\sqrt{4.93}} = \frac{5}{2.22} = 2.25$$

$$+ - \text{test} = 2.25$$

$$df = 70 - 2 = 68$$

$$p = 0.005$$

$$p < 0.01 \quad -10.0 = 0$$

$$0.005 < 0.01$$

$$(\text{test} = 0.05) = 17.20$$

$$\text{We Reject the Null Hypothesis}$$

$$H_0: \mu = 0$$

$$H_1: \mu \neq 0$$

$$H_0: \mu = 0$$

$$H_1: \mu \neq 0$$

$$H_0: \mu = 0$$

$$H_1: \mu \neq 0$$

$$H_0: \mu = 0$$

$$H_1: \mu \neq 0$$