**State the Hypotheses**

* **Null Hypothesis (H₀): There is no significant difference in performance (measured by CGPA) between students who take handwritten notes and those who rely on digital notes.**

**H0:μhandwritten=μdigitalH\_0: \mu\_{handwritten} = \mu\_{digital}H0​:μhandwritten​=μdigital​**

* **Alternative Hypothesis (H₁): Students who take handwritten notes perform better than those who rely on digital notes.**

**H1:μhandwritten>μdigitalH\_1: \mu\_{handwritten} > \mu\_{digital}H1​:μhandwritten​>μdigital​**

**2. Set the Significance Level**

* **α=0.05\alpha = 0.05α=0.05 (5%)**

**Why T-Test Instead of Z-Test?**

1. **Small Sample Size**:
   * The **Z-test** is typically used when the sample size is **large (n > 30)**, as it relies on the Central Limit Theorem to approximate the sampling distribution as normal.
   * The **T-test** is designed for smaller sample sizes (n ≤ 30) and accounts for increased variability in the data by using the **t-distribution**, which has heavier tails than the normal distribution.
2. **Unknown Population Standard Deviation**:
   * The **Z-test** assumes that the population standard deviation (σ\sigmaσ) is known.
   * In our case, we only have the sample standard deviations, so we use the **T-test**, which adjusts for this uncertainty.

**Why Use a Parametric Test Instead of a Nonparametric Test?**

1. **Data Characteristics**:
   * Parametric tests, like the T-test, assume the data follows a **normal distribution**. In this case, CGPA data is continuous and expected to follow a roughly normal distribution.
   * Nonparametric tests (e.g., Mann-Whitney U test) are used when data violates these assumptions or is ordinal/categorical.
2. **Efficiency**:
   * Parametric tests are more **statistically powerful** when their assumptions are met. This means they are more likely to detect true differences between groups.
3. **CGPA is a Quantitative Metric**:
   * Since CGPA is numerical and measured on an interval scale, a parametric test is appropriate.

**Threshold Value (α\alphaα):**

* The threshold value (α\alphaα) is **0.05** in our test.
  + This means we are willing to accept a **5% chance of incorrectly rejecting the null hypothesis (Type I error)**.
* **Why 0.05?**
  + It is a widely accepted standard in statistical analysis to balance between Type I and Type II errors.

**Explanation of the Code**

**1. Data Loading**

python

Copy code

data\_file\_path = r"D:\Bracu Course\CSE437\Project\Note-Taking-Method.xlsx"

df = pd.read\_excel(data\_file\_path)

* **What it does:** Loads the Excel file containing the survey data into a Pandas DataFrame.
* **Why:** To process and analyze the data programmatically.

**2. Remove Outliers**

python

Copy code

handwritten\_cgpa\_clean = handwritten\_cgpa[handwritten\_cgpa <= 4.0]

digital\_cgpa\_clean = digital\_cgpa[digital\_cgpa <= 4.0]

* **What it does:** Removes any CGPA values greater than 4.0 (which is the maximum valid value for CGPA).
* **Why:** Ensures data integrity by excluding invalid or erroneous entries.

**3. Descriptive Statistics**

python

Copy code

descriptive\_stats\_handwritten = handwritten\_cgpa\_clean.describe()

descriptive\_stats\_digital = digital\_cgpa\_clean.describe()

* **What it does:** Calculates summary statistics (mean, median, standard deviation, etc.) for both groups.
* **Why:** Provides an overview of the data and helps in comparing the two groups.

**4. Hypothesis Testing**

python

Copy code

t\_stat, p\_value = ttest\_ind(handwritten\_cgpa\_clean, digital\_cgpa\_clean, equal\_var=False)

* **What it does:** Performs an independent two-sample T-test to compare the mean CGPA of handwritten note-takers and digital note-takers.
* **Why:** Tests whether the difference in means between the two groups is statistically significant.

**5. Visualization**

python

Copy code

plt.boxplot([handwritten\_cgpa\_clean, digital\_cgpa\_clean],

labels=["Handwritten Notes", "Digital Notes"], patch\_artist=True,

boxprops=dict(facecolor='lightblue', color='blue'),

medianprops=dict(color='red', linewidth=2),

whiskerprops=dict(color='blue', linewidth=1.5),

capprops=dict(color='blue', linewidth=1.5))

* **What it does:** Creates a boxplot to compare the distributions of CGPA for both groups.
* **Why Use a Boxplot:**
  + **Shows Median and Spread:** Highlights the median CGPA and variability within each group.
  + **Detects Outliers:** Visualizes any data points that fall outside the whiskers.
  + **Comparative View:** Makes it easy to compare the two groups' central tendency and spread.

**6. Results Summary**

python

Copy code

results\_summary = {

"Descriptive Statistics (Handwritten Notes)": descriptive\_stats\_handwritten.to\_dict(),

"Descriptive Statistics (Digital Notes)": descriptive\_stats\_digital.to\_dict(),

"T-Test Results": {

"T-Statistic": t\_stat,

"P-Value": p\_value

},

"Conclusion": "Reject Null Hypothesis" if p\_value < 0.05 else "Fail to Reject Null Hypothesis"

}

* **What it does:** Compiles descriptive statistics, T-test results, and a conclusion based on the p-value.
* **Why:** Provides a clear and interpretable summary of the findings.

**Interpretation of the Boxplot**

1. **Boxes**:
   * Represent the interquartile range (IQR), which contains the middle 50% of CGPA values.
2. **Whiskers**:
   * Extend to the smallest and largest values within 1.5 times the IQR.
3. **Median Line**:
   * The red line indicates the median CGPA for each group.
4. **Colors**:
   * Blue and light blue are used to differentiate the groups visually.

**Conclusion**

The T-test and boxplot together provide a comprehensive analysis:

* The **T-test** quantifies whether the difference between the groups is statistically significant.
* The **boxplot** visually highlights the central tendency, spread, and potential outliers in CGPA for both note-taking methods.

**How the Boxplot Helps Understand the Hypothesis**

The hypothesis being tested is: **"Students who take handwritten notes perform better on exams than those who rely solely on digital notes."**

The boxplot visually aids in this analysis by showing how the distributions of CGPA differ between the two groups (handwritten vs. digital notes). Here's how:

**1. Comparing Central Tendency (Median)**

* The **red line** inside each box represents the **median** CGPA for the respective group.
* **Relevance to Hypothesis**:
  + If the median for handwritten notes is higher than that of digital notes, it supports the hypothesis.
  + In the boxplot:
    - If the **handwritten notes group** has a visibly higher median, it suggests that they perform better on average.

**2. Assessing Variability**

* The **height of the box** represents the **interquartile range (IQR)**, which is the range of the middle 50% of the data.
* The **whiskers** extend to capture most of the data points within 1.5 times the IQR.
* **Relevance to Hypothesis**:
  + If the variability (spread) is large in one group, it could indicate inconsistent performance.
  + A smaller spread (tighter box) for handwritten notes might suggest more consistent performance.

**3. Identifying Outliers**

* Points outside the whiskers are **outliers** (e.g., extreme CGPA values like a 2.0 or a 4.0).
* **Relevance to Hypothesis**:
  + If the digital group has more outliers on the lower end, it might indicate more students struggling compared to the handwritten group.
  + Outliers in the handwritten group (especially high CGPA outliers) might support the claim of better performance.

**4. Distribution Comparison**

* The **relative position** of the boxes along the CGPA axis helps compare the overall performance of both groups.
* **Relevance to Hypothesis**:
  + If the entire box for handwritten notes is shifted higher on the CGPA scale compared to digital notes, it strengthens the hypothesis.
  + Conversely, overlapping boxes or a higher box for digital notes would weaken the hypothesis.

**5. Quickly Spotting Trends**

* A boxplot provides a **high-level view** of the data, allowing you to instantly:
  + See if one group generally performs better than the other.
  + Identify if the data distribution supports the hypothesis.

**How the Boxplot Relates to the T-Test**

* While the **t-test** provides a **statistical measure** (p-value) of whether the difference in means is significant, the boxplot offers a **visual confirmation** of this difference.
* For example:
  + If the medians and boxes are clearly separated, the t-test is likely to confirm significance.
  + If the boxes overlap significantly, it suggests the groups are similar, and the t-test is less likely to reject the null hypothesis.

**Why the Boxplot is Useful for This Hypothesis**

* **Easy Comparison**: Allows direct comparison of CGPA distributions between the two groups.
* **Highlights Key Metrics**: Shows the median, spread, and outliers in a concise way.
* **Supports Statistical Analysis**: Provides visual evidence that complements the t-test results.

**Explanation of t\_stat and p\_value**

**1. t\_stat (T-Statistic)**

* The **t-statistic** is the test statistic calculated by the **T-test**. It measures the size of the difference between the means of two groups relative to the variability in the data.
* **Formula**:

t=Xˉ1−Xˉ2s12n1+s22n2t = \frac{\bar{X}\_1 - \bar{X}\_2}{\sqrt{\frac{s\_1^2}{n\_1} + \frac{s\_2^2}{n\_2}}}t=n1​s12​​+n2​s22​​​Xˉ1​−Xˉ2​​

Where:

* + Xˉ1,Xˉ2\bar{X}\_1, \bar{X}\_2Xˉ1​,Xˉ2​: Sample means of the two groups (handwritten and digital note-takers).
  + s12,s22s\_1^2, s\_2^2s12​,s22​: Variances of the two groups.
  + n1,n2n\_1, n\_2n1​,n2​: Sample sizes of the two groups.
* **Interpretation**:
  + A **larger t-statistic** indicates a larger difference between group means relative to the variability.
  + If the t-statistic is sufficiently large (compared to a critical value), it suggests the difference in means is statistically significant.

**2. p\_value**

* The **p-value** is the probability of observing a t-statistic as extreme as the one calculated (or more extreme) if the null hypothesis (H0H\_0H0​) is true.
* **Interpretation**:
  + A **small p-value** (e.g., less than 0.05) indicates strong evidence against the null hypothesis, leading to its rejection.
  + A **large p-value** suggests insufficient evidence to reject the null hypothesis.

**Threshold in the Code**

The threshold value (α\alphaα) is implicitly set in the code here:

python

Copy code

"Conclusion": "Reject Null Hypothesis" if p\_value < 0.05 else "Fail to Reject Null Hypothesis"

* **Threshold Value:** α=0.05\alpha = 0.05α=0.05
  + If p≤0.05p \leq 0.05p≤0.05, the null hypothesis is rejected, concluding a significant difference between the two groups.
  + If p>0.05p > 0.05p>0.05, the null hypothesis is not rejected, indicating no significant difference.