# A Hypothesis Testing for Survey Responses

This section outlines the hypothesis testing performed on the survey responses, using statistical tests to evaluate the significance of the responses for each question.

**1. Wilcoxon Signed-Rank Test** (for Q1 to Q9)

- **Null Hypothesis ($H_0$):** The median response for each question (Q1 to Q9) is less then or equal to 3.
- **Alternative Hypothesis ($H_1$):** The median response for each question (Q1 to Q9) is greater than 3.

**2.Binomial Test** (for Q10)

- **Null Hypothesis ($H_0$):** The proportion of binary responses is less than or equal to 0.5.
- **Alternative Hypothesis ($H_1$):** The proportion of binary responses greater than 0.5.

The hypotheses are determined based on the following conditions:

- If the p-value is less than 0.05, the null hypothesis will be rejected, indicating a significant difference.
- If the p-value is greater than 0.05, the null hypothesis cannot be rejected, suggesting no significant difference.

The results of the hypothesis tests for questions Q1 to Q9 (Wilcoxon Signed-Rank Test) and Q10 (Binomial Test) are presented in Table 3.

**Table 3.** Wilcoxon Signed-Rank test result for Q1-Q9 and Binomial test result for Q10.

| **Question** | **Test Statistic** | **p-value** | **Sample Size** | **Interpretation ($\alpha$=0.05)** |
|---|---|---|---|---|
| Q1 | 1523 | $< 0.0001$ | 103 | Significant (greater than 3) |
| Q2 | 1888 | $< 0.0001$ | 103 | Significant (greater than 3) |
| Q3 | 1335 | $< 0.0001$ | 104 | Significant (greater than 3) |
| Q4 | 1000 | $< 0.0001$ | 104 | Significant (greater than 3) |
| Q5 | 1794 | $< 0.0001$ | 103 | Significant (greater than 3) |
| Q6 | 1827 | $< 0.0001$ | 104 | Significant (greater than 3) |
| Q7 | 1490 | $< 0.0001$ | 103 | Significant (greater than 3) |
| Q8 | 2131 | $< 0.0001$ | 103 | Significant (greater than 3) |
| Q9 | 1883 | $< 0.0001$ | 103 | Significant (greater than 3) |
| Q10 | 0.8942 | $< 0.0001$ | 104 | Significant (greater than 0.5) |

For all questions Q1 to Q9, the p-values are less than 0.05, indicating that the null hypothesis is rejected, and the median response for each question is significantly greater than 3. This suggests that, on average, the respondents rated

each of these questions more favorably than the neutral midpoint (3). For Q10, the p-value is also less than 0.05, leading to the rejection of the null hypothesis. This means that the proportion of positive binary responses is significantly greater than 0.5, indicating a preference for the positive response.

## B Hypothesis Testing for Between-Group Comparison

This section describes the use of the **Mann-Whitney U Test** (for Q1 to Q10) for comparing the survey responses between two groups of students: weak students (Group 1) and strong students (Group 2). For the comparison between weak and strong students, the hypotheses are as follows:

- **Null Hypothesis ($H_0$):** There is no significant difference between the weak students (Group 1) and the strong students (Group 2) with respect to the survey responses, i.e., the distribution of survey responses for weak students is less than or equal to that of strong students.
- **Alternative Hypothesis ($H_1$):** The distribution of survey responses for weak students is greater than that of strong students.

The hypotheses are determined based on the following conditions:

- If the p-value is less than 0.05, the null hypothesis will be rejected, indicating a significant difference.
- If the p-value is greater than 0.05, the null hypothesis cannot be rejected, suggesting no significant difference.

The results of the Mann-Whitney U test comparing weak and strong students (Groups 1 and 2) for each survey question are summarized in Table 4. The interpretation of each result follows the decision rule outlined above.

**Table 4.** Comparison between weak and strong students (group 1 and group 2) with interpretations. Where n denotes the sample size of the respective group.

| **Survey Question** | **Test Statistic** | **p-value** | $\mathbf{n}_{group1}$ | $\mathbf{n}_{group2}$ | **Interpretation ($\alpha$=0.05)** |
|---|---|---|---|---|---|
| Q1 | 1687 | 0.0053 | 53 | 50 | Significant |
| Q2 | 1537 | 0.0717 | 53 | 50 | Not Significant |
| Q3 | 1561 | 0.0711 | 54 | 50 | Not Significant |
| Q4 | 1603 | 0.0350 | 54 | 50 | Significant |
| Q5 | 1482 | 0.1355 | 53 | 50 | Not Significant |
| Q6 | 1594 | 0.0471 | 54 | 50 | Significant |
| Q7 | 1354 | 0.4212 | 53 | 50 | Not Significant |
| Q8 | 1655 | 0.0109 | 54 | 49 | Significant |
| Q9 | 1515 | 0.0905 | 53 | 50 | Not Significant |
| Q10 | 1335 | 0.5751 | 54 | 50 | Not Significant |

From the Mann-Whitney U test results, significant differences between weak and strong students were observed for Q1, Q4, Q6, and Q8, where the p-values

were less than 0.05. This indicates that the survey responses of weak students were significantly higher than those of strong students for these questions. For the remaining questions (Q2, Q3, Q5, Q7, Q9, and Q10), the p-values exceeded 0.05, suggesting no significant differences in the responses between the two groups.

## C Survey Questions

This section includes the set of survey questions designed to gather valuable insights into various aspects of the AAG system. The goal is to understand students' perspectives, experiences, and feedback, which will inform future development and improvements.

### Survey Items

1. **Identifying Mistakes:** To what extent did the AI feedback help you identify the specific mistakes in your answers, compared to just seeing the solution alone?
   - No additional help in identifying mistakes
   - Very little additional help
   - Some additional help
   - Noticeable improvement in identifying mistakes
   - Much clearer identification of mistakes
2. **Clarity of Explanation for Errors:** Did the AI comment clarify why your answers were incorrect better than the standard solution alone could?
   - No added clarity
   - Slightly clearer but limited
   - Moderately clearer
   - Significantly clearer explanation
   - Very clear and detailed explanation
3. **Level of Customization:** How well was the AI feedback tailored to your specific answer compared to a general solution?
   - Not customized at all
   - Minimally customized
   - Somewhat customized
   - Well-customized
   - Very highly customized to my answer
4. **Benefit of Customization:** How much did the customized AI comment, tailored to your specific answer, help you address your personal learning needs compared to a general solution?
   - No additional help for my learning needs
   - Very minimal additional help
   - Somewhat helpful for my needs
   - Noticeably helpful for my needs
   - Extremely helpful and targeted to my learning needs

5. **Actionable Feedback for Improvement:** How effective was the AI feedback in providing specific steps or areas to improve, compared to the solution alone?
   - Not effective in providing improvement steps
   - Very limited effectiveness
   - Somewhat effective
   - Effective in providing actionable steps
   - Extremely effective with clear, actionable feedback
6. **Future Problem-Solving:** How prepared do you feel to handle similar problems in the future with AI feedback, as opposed to only seeing the solution?
   - Not more prepared than with solution alone
   - Slightly more prepared
   - Moderately more prepared
   - Significantly more prepared
   - Very confident in handling similar problems
7. **Depth of Insight into Learning Gaps:** To what extent did the AI feedback help you identify specific knowledge gaps, in contrast to the general solution alone?
   - No additional insight into my learning gaps
   - Minimal additional insight
   - Some additional insight
   - Clearer insight into learning gaps
   - Provided significant, valuable insight
8. **Encouragement for Learning:** How much more motivated did you feel to address your mistakes after receiving the AI comment, as compared to just the solution?
   - Not more motivated than with solution alone
   - Slightly more motivated
   - Somewhat more motivated
   - Noticeably more motivated to improve
   - Very motivated to address and learn from mistakes
9. **Comparative Learning Benefit:** To what extent did the AI feedback contribute to an improved learning experience compared to receiving only the solution?
   - No additional benefit compared to solution alone
   - Slight additional benefit
   - Moderate additional benefit
   - Significant benefit for learning
   - Greatly enhanced learning experience
10. **Preference Between Solution Types:** Overall, which did you find more helpful for understanding: the solution alone or the solution with the AI feedback?
    - Solution alone
    - Solution + AI Comment
11. **Suggestions for AI Feedback Improvement (Open-Ended):** What additional information or aspects in the AI feedback would make it even more helpful compared to the solution alone?