BU354 Dashboard: Comparative Usability Report - Formulas

Derek Song

Note. For clarity and readability, the formulas in this document avoid using the general notation n (total responses) and k (number of categories). Instead, each formula is written directly in terms of the specific variables used in the BU354 UX analysis (e.g., observed counts for dashboard A vs dashboard E).

Paired sample t-tests

In the BU354 Comparative Usability Test, a paired samples t-test is performed to compare the rankings of two dashboards. The Python code the uses `scipy.stats.ttest_rel` package. Internally it computes the difference of the observed difference between the scores of the two dashboards.

$$d_i = \text{dashboard A}_i - \text{dashboard E}_i$$

The mean of these paired differences is then calculated as:

$$\bar{d} = \frac{1}{\text{number of respondents}} \sum_{i=1}^{\text{number of respondents}} d_i$$

Since the data represent a sample of some students as opposed to the population of students, the sample standard deviation of the differences is used:

$$s_d = \sqrt{\frac{\sum_{i=1}^{\text{number of respondents}} (d_i - \bar{d})^2}{\text{number of respondents} - 1}}$$

The paired t-statistic computed by `ttest_rel` is equal to:

$$t = \frac{\bar{d}}{s_d / \sqrt{\text{number of respondents}}}, \qquad df = \text{number of respondents} - 1$$

This difference compares the observed mean difference to the standard error of the mean difference (we check the standard error here to find out how "messy" our data is.). If we have a small standard error (usually from larger sample sizes or low variability) we get a bigger t-value.

In addition to statistical significance, we also find out the effect size using Cohen's d for paired samples, which uses the standard deviation of the difference scores:

$$d_{\text{Cohen}} = \frac{\bar{d}}{s_d}$$

The effect size indicates how meaningful the difference is. For example, an effect size of $d = 0.21$ is considered small, suggesting that students did not show a strong preference between Dashboard A and Dashboard D. The Python code mirrors this formula directly by dividing the mean difference by the standard deviation of the paired differences.

Chi-square tests

The chi-square goodness of fit test is used to determine whether the observed distribution of responses differs from an expected distribution. Here we always assume

H_0 : All response categories are equally preferred. (e.g. students do not prefer X over Y)

and the alternative hypothesis is:

H_1 : At least one category is preferred more or less than expected

Under the equal-preference assumption, the expected count for each category is n/k. As we will always do a dual comparsion in this report, this simplifies to

$$E = \frac{\text{number of respondents}}{2}.$$

for each category. To find out the chi-square statistic, we compute: (with O being observed, and E being expected)

$$\chi^2 = \frac{(O_{\text{category 1}} - E_{\text{category 1}})^2}{E_{\text{category 1}}} + \frac{(O_{\text{category 2}} - E_{\text{category 2}})^2}{E_{\text{category 2}}}.$$

with degrees of freedom:

$$df = \text{number of categories} - 1.$$

To quantify the magnitude of the deviation from equal preference, we compute Cramér's V:

$$V = \sqrt{\frac{\chi^2}{\text{number of respondents}(\text{number of categories} - 1)}}.$$