Submission deadline: Homework solutions must be submitted within seven days, i.e., not later than on the following Monday, at 23:59 hours.

Late submission policy:
• 50% of the total marks deducted for submission up to 24 hours late
• 100% of the total marks deducted for submission more than 24 hours late
• You must use the submit button on the course wiki page; make sure to submit

Maximum amount of points is ten (10).

You must use the submit button on the course wiki page; make sure to submit
Task 1 [2 points]

File HW1-PairProgramming.xlsx contains a subset of the results from an experiment about pair programming (as presented in Lecture 1). The Spreadsheet shows the data for “Quality”, “Duration”, and “Effort” for Pairs and Individuals (grouped by “Junior”, “Intermediate”, and “Senior”).

To do:

a) Calculate mean and variance for Pairs and Individuals (All, only Juniors, only Intermediate, only Seniors) for each attribute (Quality, Duration, Effort)
b) Conduct four t-Tests to check whether Pairs or Individuals perform better; do this for each attribute and for each grouping (All, only Junior, only Intermediate, only Senior)
c) Search the internet for a definition of the effect size measure Cohen’s d (provide the definition and the URL where you found it) and calculate the effect sizes between the performance of Pairs and Individuals (again for each attribute and each grouping)

Show all calculations; you may use a desk calculator, Excel, R scripts or Python Scripts; if you use Excel, R or Python, provide the Spreadsheet/Scripts

Solution:

a) [1 point]

<table>
<thead>
<tr>
<th>TrT</th>
<th>Block</th>
<th>Mean</th>
<th>Mean</th>
<th>Mean</th>
<th>Variance</th>
<th>Variance</th>
<th>Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Quality</td>
<td>Duration</td>
<td>Effort</td>
<td>Quality</td>
<td>Duration</td>
<td>Effort</td>
</tr>
<tr>
<td>Pair</td>
<td>Junior</td>
<td>3.20</td>
<td>157.96</td>
<td>315.92</td>
<td>0.33</td>
<td>1054.21</td>
<td>4216.83</td>
</tr>
<tr>
<td>Ind</td>
<td>Junior</td>
<td>2.71</td>
<td>206.65</td>
<td>206.65</td>
<td>0.95</td>
<td>5963.84</td>
<td>5963.84</td>
</tr>
<tr>
<td>Pair</td>
<td>Intermediate</td>
<td>3.06</td>
<td>149.97</td>
<td>299.94</td>
<td>1.11</td>
<td>1143.38</td>
<td>4573.53</td>
</tr>
<tr>
<td>Ind</td>
<td>Intermediate</td>
<td>3.00</td>
<td>207.78</td>
<td>207.78</td>
<td>1.23</td>
<td>5530.89</td>
<td>5530.89</td>
</tr>
<tr>
<td>Pair</td>
<td>Senior</td>
<td>3.45</td>
<td>132.05</td>
<td>264.11</td>
<td>0.63</td>
<td>786.32</td>
<td>3145.29</td>
</tr>
<tr>
<td>Ind</td>
<td>Senior</td>
<td>3.19</td>
<td>177.42</td>
<td>177.42</td>
<td>1.08</td>
<td>2956.94</td>
<td>2956.94</td>
</tr>
<tr>
<td>Pair</td>
<td>All</td>
<td>3.24</td>
<td>145.06</td>
<td>290.12</td>
<td>0.74</td>
<td>1079.42</td>
<td>4317.68</td>
</tr>
<tr>
<td>Ind</td>
<td>All</td>
<td>2.98</td>
<td>196.38</td>
<td>196.38</td>
<td>1.10</td>
<td>4839.16</td>
<td>4839.16</td>
</tr>
</tbody>
</table>

Note: The measurements of attribute ‘Quality’ look like ordinal data and thus calculating the mean might not make much sense.

b) [0.5 points]

Note: the unequal sample sizes and the unequal variances render Student’s t-test non-applicable; use Welch’s t-test needs to be used; however, even Welch’s t-test would require normally distributed data (and data on interval/ratio scale).

We assume alpha=0.05; then difference is significant if p<0.05
c) [0.5 points]

Motivation for use of effect size:

Source: [http://staff.bath.ac.uk/pssiw/stats2/page2/page14/page14.html](http://staff.bath.ac.uk/pssiw/stats2/page2/page14/page14.html)

Most of the statistics you have covered have been concerned with null hypothesis testing: assessing the likelihood that any effect you have seen in your data, such as a correlation or a difference in means between groups, may have occurred by chance. As we have seen, we do this by calculating a p value -- the probability of your null hypothesis being correct; that is, p gives the probability of seeing what you have seen in your data by chance alone. This probability goes down as the size of the effect goes up and as the size of the sample goes up.

However, there are problems with this process. As we have discussed, there is the problem that we spend all our time worrying about the completely arbitrary .05 alpha value, such that \( p = .04999 \) is a publishable finding but \( p = .05001 \) is not. But there is also another problem: even the most trivial effect (a tiny difference between two groups’ means, or a miniscule correlation) will become statistically significant if you test enough people. If a small difference between two groups’ means is not significant when I test 100 people, should I suddenly get excited about exactly the same difference if, after testing 1000 people, I find it is now significant? The answer is probably no -- if it was a trivial effect with 100 people it’s still trivial with 1000: we don’t really care if something makes just a 1% difference to performance, even if it is statistically significant. So what is needed is not just a system of null hypothesis testing but also a system for telling us precisely how large the effects we see in our data really are. This is where effect-size measures come in.

A common measure of effect size is \( d \), sometimes known as Cohen's \( d \). This can be used when comparing two means, as when you might do a t-test, and is simply the difference in the two groups’ means divided by the average of their standard deviations*. This means that if we see a \( d \) of 1, we know that the two groups’ means differ by one standard deviation; a \( d \) of .5 tells us that the two groups’ means differ by half a standard deviation; and so on. Cohen suggested that \( d = 0.2 \) be considered a 'small' effect size, 0.5 represents a 'medium' effect size and 0.8 a 'large' effect size. This means that if two groups’ means don't differ by 0.2 standard deviations or more, the difference is trivial, even if it is statistically significant.

<table>
<thead>
<tr>
<th>TrT</th>
<th>Block</th>
<th>Quality</th>
<th>T-Test p-value</th>
<th>T-Test p-value</th>
<th>T-Test p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pair-Ind</td>
<td>Junior</td>
<td>0.02</td>
<td>2.77E-03</td>
<td>4.29E-07</td>
<td></td>
</tr>
<tr>
<td>Pair-Ind</td>
<td>Intermediate</td>
<td>0.83</td>
<td>2.25E-04</td>
<td>1.65E-06</td>
<td></td>
</tr>
<tr>
<td>Pair-Ind</td>
<td>Senior</td>
<td>0.25</td>
<td>4.24E-05</td>
<td>3.18E-09</td>
<td></td>
</tr>
<tr>
<td>Pair-Ind</td>
<td>All</td>
<td>0.05</td>
<td>6.69E-10</td>
<td>1.83E-18</td>
<td></td>
</tr>
</tbody>
</table>

Formula:
\[ d = \frac{M_1 - M_2}{SD_{pooled}} \]

with

\[ SD_{pooled} = \sqrt{\frac{\sum (X_1 - \bar{X}_1)^2 + \sum (X_2 - \bar{X}_2)^2}{n_1 + n_2 - 2}} \]

However, I used below the approximation:

\[ SD_{pooled} = \sqrt{\frac{SD_1^2 + SD_2^2}{2}} \]

<table>
<thead>
<tr>
<th>TrT</th>
<th>Block</th>
<th>Cohen's d</th>
<th>Cohen's d</th>
<th>Cohen's d</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pair-Ind</td>
<td>Junior</td>
<td>0.61</td>
<td>-0.82</td>
<td>1.53</td>
</tr>
<tr>
<td>Pair-Ind</td>
<td>Intermediate</td>
<td>0.05</td>
<td>-1.00</td>
<td>1.30</td>
</tr>
<tr>
<td>Pair-Ind</td>
<td>Senior</td>
<td>0.27</td>
<td>-1.05</td>
<td>1.57</td>
</tr>
<tr>
<td>Pair-Ind</td>
<td>All</td>
<td>0.28</td>
<td>-0.94</td>
<td>1.39</td>
</tr>
</tbody>
</table>

**Interpretation:**

<table>
<thead>
<tr>
<th>Effect size</th>
<th>abs(d)</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Very small</td>
<td>0.01</td>
<td>Sawilowsky, 2009</td>
</tr>
<tr>
<td>Small</td>
<td>0.20</td>
<td>Cohen, 1988</td>
</tr>
<tr>
<td>Medium</td>
<td>0.50</td>
<td>Cohen, 1988</td>
</tr>
<tr>
<td>Large</td>
<td>0.80</td>
<td>Cohen, 1988</td>
</tr>
<tr>
<td>Very large</td>
<td>1.20</td>
<td>Sawilowsky, 2009</td>
</tr>
<tr>
<td>Huge</td>
<td>2.0</td>
<td>Sawilowsky, 2009</td>
</tr>
</tbody>
</table>

Here is a link to webpage that calculates various effect size measures (including Cohen's d): [http://www.socscistatistics.com/effectsize/Default3.aspx](http://www.socscistatistics.com/effectsize/Default3.aspx)
Task 2 [3 points]

Assume you are conducting a survey that investigates the strengths and weaknesses of requirements elicitation techniques;

This the Research Question (RQ):

“What elicitation techniques are most effective?”

Assume you have a quasi gold standard of 25 relevant papers as listed in file HW1-Task2-GoldStandard.pdf.

Apply the following search strings to at least two of the following digital databases: Scopus, SpringerLink, ACM DL, IEEE Explore.

Search string (logical):

(elicitation OR “requirements gathering” OR “requirements acquisition”) AND (capture OR empirical OR experiment OR study OR review OR evaluation)

To do:

a) Transform the given search string to the format required by the databases you are using. Then calculate for each database to which you apply the search string the quasi-sensitivity and the quasi-precision. [Show your calculation!]

b) Based on the results, discuss the appropriateness of the search string and/or quasi gold standard.

Solution:

a) [2 points]

Sensitivity = (Number of relevant studies retrieved/Total number of relevant studies) * 100%

Precision = (Number of relevant studies retrieved/Number of studies retrieved) * 100%

Scopus query:
TITLE-ABS-KEY ( ( elicitation OR "requirements gathering" OR "requirements acquisition" ) AND ( capture OR empirical OR experiment OR study OR review OR evaluation ) )

13560 results.

Sensitivity = (11/25)*100 = 44%
Precision = (11/13560)*100 = 0.081%
ACM query:

( elicitation OR "requirements gathering" OR "requirements acquisition" ) AND ( capture OR empirical OR experiment OR study OR review OR evaluation )

1520 results.

Sensitivity= (1/25)*100= 4%
Precision= (/1520)*100= 0.065%

b) [1 point]

Sensitivity and Precision of this search string for both databases are quite low.

Sensitivity below 80% is generally unacceptable and the search string needs to be modified.
Task 3 [4 points]

Assume you are conducting a survey that investigates the effectiveness of requirements elicitation techniques (same as in Task 2).

Among many other studies you have found two relevant primary studies (S1 and S2):

To do:
Download the PDFs of studies S1 and S2 and do the following for each study:
a) Extract the following data:
   • Study Type (experiment, case study, survey, etc.)
   • Who were the participants in the study?
   • What was the setting of the study? (industry, academia)
   • What task(s) had the participants to perform?
   • Elicitation Techniques investigated (use the codes from the related taxonomy in file HW1-Task3-DataExtractionTaxonomies.pdf)
   • Response variables analyzed (use the codes from the related taxonomy in file HW1-Task3-DataExtractionTaxonomies.pdf)

b) Compare the requirements elicitation techniques with regards to all response variables that were analyzed in S1 and S2; What can you say about the techniques, i.e., is there evidence that one technique is better than the other? If the two papers use comparable techniques, are the results comparable (and if so, are they similar)?

For all answers say where exactly in the studies you found the information.

Solution:

a) [3 points]

Study Type (experiment, case study, survey, etc.):

S1: Experiment (completely randomized design, 3 groups)
S2: Experiment (no control group)

Who were the participants in the study?

S1: “45 non-faculty employees from two universities.”
S2: “54 practicing systems analysts who were recruited from organizations in the Baltimore metropolitan area. Twelve different organizations participated,
representing a variety of industry segments including banking, finance, insurance, construction, manufacturing, aerospace, government, research, and education. Only analysts with at least two years of experience in system development projects were eligible to participate in the study.

What was the setting of the study? (industry, academia)

S1: laboratory setting (academia)
S2: laboratory setting (academia)

What task(s) had the participants to perform?

S1: “The experiment utilized a short case describing a grocery company interested in developing an Internet-based food shopping system.” (cf. Figure 2)
S2: “a case scenario for the development of an online grocery shopping system”

Elicitation Techniques investigated:

S1:
T08 – syntactic interview (unstructured interview)
T09 – semantic interview (unstructured interview)
T10 – task characteristics interview (domain independent, structured interview)
S2:
T01 – unstructured interview

Response variables analyzed:

S1:
V03 – Number of rules
V04 – Number of clauses in rules
V11 – Number of goals
V12 – Number of processes
V13 – Number of tasks
V27 – Diversity
S2:
V03 – Number of rules
V27 – Diversity (breadth and depth of requirements)

b) [1 point]

S1:
V03 – Number of rules: T10 > T09 > T08
V04 – Number of clauses in rules: T10 > T09 > T08
V11 – Number of goals: no difference between T08, T09, and T10
V12 – Number of processes: T10 > T09 > T08
V13 – Number of tasks: T10 > T08
V27 – Diversity: no difference between T08, T09, and T10
S2:

V03 – Number of rules: no control for T01
V27 – Diversity (breadth and depth of requirements): no control for T01

The two papers investigate different techniques (albeit all techniques are interview techniques). Based on S1 one could argue that structured interview techniques perform better than unstructured interview techniques.
Task 4 [1 point]
You have conducted a survey about the use of software development frameworks; the survey used 5-point Likert scales where ‘1’ = “fully disagree” and ‘5’ = “fully agree.” The results can be found in file HW1-ProjectFrameworks.xlsx

a) Calculate the internal consistency (using Cronbach’s alpha – look it up in the internet) of the survey instrument
b) Calculate the inter-rater agreement using Fleiss’ Kappa (search the formula in the internet and report the formula and the URL where you found it)

[show all calculations]

Solution:

a) [0.5 points]

Use formula from: https://www.personality-project.org/r/html/score.items.html

\[ k = \text{number of items} (=> Q1, Q2, and Q3 in our case) \]
\[ av.r = \text{average correlation between items} \]
\[ \alpha = k * \frac{av.r}{1 + (k-1)*av.r} \]

Note: before the correlations are calculated, the Likert scales must be made to point in the same direction => Q3 is inversed

<table>
<thead>
<tr>
<th>Q1</th>
<th>Q2</th>
<th>Q3</th>
<th>Q1</th>
<th>Q2</th>
<th>Q3</th>
</tr>
</thead>
<tbody>
<tr>
<td>W 5</td>
<td>5</td>
<td>1</td>
<td>W 5</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>W 4</td>
<td>5</td>
<td>2</td>
<td>W 4</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>W 5</td>
<td>3</td>
<td>1</td>
<td>W 5</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>W 4</td>
<td>2</td>
<td>3</td>
<td>W 4</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>W 3</td>
<td>4</td>
<td>2</td>
<td>W 3</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>W 4</td>
<td>4</td>
<td>3</td>
<td>W 4</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>W 5</td>
<td>5</td>
<td>2</td>
<td>W 5</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>W 4</td>
<td>4</td>
<td>1</td>
<td>W 5</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>W 4</td>
<td>4</td>
<td>1</td>
<td>W 4</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>W 5</td>
<td>5</td>
<td>1</td>
<td>W 5</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>W 4</td>
<td>5</td>
<td>2</td>
<td>W 4</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>S 1</td>
<td>1</td>
<td>5</td>
<td>S 1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>S 2</td>
<td>2</td>
<td>4</td>
<td>S 1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>S 3</td>
<td>3</td>
<td>3</td>
<td>S 1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>S 2</td>
<td>2</td>
<td>4</td>
<td>S 2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>S 3</td>
<td>3</td>
<td>3</td>
<td>S 2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>S 2</td>
<td>2</td>
<td>5</td>
<td>S 3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>S 1</td>
<td>2</td>
<td>5</td>
<td>S 2</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>S 1</td>
<td>1</td>
<td>4</td>
<td>S 1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>S 1</td>
<td>2</td>
<td>4</td>
<td>S 1</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>S 2</td>
<td>1</td>
<td>5</td>
<td>S 2</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>X 2</td>
<td>2</td>
<td>5</td>
<td>X 2</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>X 1</td>
<td>1</td>
<td>5</td>
<td>X 2</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>X 1</td>
<td>2</td>
<td>4</td>
<td>X 1</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>X 2</td>
<td>2</td>
<td>4</td>
<td>X 2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>X 2</td>
<td>3</td>
<td>5</td>
<td>X 2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>X 3</td>
<td>2</td>
<td>5</td>
<td>X 2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>X 1</td>
<td>1</td>
<td>5</td>
<td>X 3</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>X 2</td>
<td>2</td>
<td>4</td>
<td>X 1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>X 1</td>
<td>1</td>
<td>5</td>
<td>X 2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>X 1</td>
<td>2</td>
<td>4</td>
<td>X 1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>X 1</td>
<td>2</td>
<td>4</td>
<td>X 2</td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>

Corr(Q1, Q2) = 0.84655171
Corr(Q1, Q3) = 0.87338185
Corr(Q2, Q3) = 0.84167374

av.r = 0.8538691
\[ \alpha = 0.94603212 \]
Thus, we have excellent internal consistency.

Or using R:

\[ \text{cronbach.alpha(T4\_new, standardized = FALSE, na.rm = TRUE)} \]

Cronbach's alpha for the 'T4\_new' data-set

Items: 3
Sample units: 33
alpha: 0.946

b) [0.5 points]

Use approach described in:
https://en.wikipedia.org/wiki/Fleiss%27_kappa

11 raters
9 subjects (W-Q1, W-Q2, W-Q3, S-Q1, S-Q2, S-Q3, X-Q1, X-Q2, X-Q3)
5 categories (likert scale: 1-5)

Agreement can be thought of as follows:
If a fixed number of people assign numerical ratings to a number of items then the kappa will give a measure for how consistent the ratings are.

The kappa, \( \kappa \), can be defined as,

\[ \kappa = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e} \]

The factor \( 1 - \bar{P}_e \) gives the degree of agreement that is attainable above chance, and, \( P - P_e \) gives the degree of agreement actually achieved above chance. If the raters are in complete agreement then \( \kappa = 1 \). If there is no agreement among the raters (other than what would be expected by chance), then \( \kappa \leq 0 \).
Let $N$ be the total number of subjects, let $n$ be the number of ratings per subject, and let $k$ be the number of categories into which assignments are made. The subjects are indexed by $i = 1, \ldots, N$ and the categories are indexed by $j = 1, \ldots, k$. Let $n_{ij}$ represent the number of raters who assigned the $i$-th subject to the $j$-th category.

First calculate $p_j$, the proportion of all assignments which were to the $j$-th category:

$$ p_j = \frac{1}{Nn} \sum_{i=1}^{N} n_{ij}, \quad 1 = \sum_{j=1}^{k} p_j $$

Now calculate $P_i$, the extent to which raters agree for the $i$-th subject (i.e., compute how many rater--rater pairs are in agreement, relative to the number of all possible rater-rater pairs):

$$ P_i = \frac{1}{n(n-1)} \sum_{j=1}^{k} n_{ij}(n_j - 1) $$

$$ = \frac{1}{n(n-1)} \sum_{j=1}^{k} (n_{ij}^2 - n_j) $$

$$ = \frac{1}{n(n-1)} \left[ \left( \sum_{j=1}^{k} n_{ij}^2\right) - (n) \right] $$

Now compute $P_{\bar{}}$, the mean of the $P_i$'s, and $P_{\bar{}}e$, which go into the formula for $\kappa$:

$$ P_{\bar{}} = \frac{1}{N} \sum_{i=1}^{N} P_i $$

$$ = \frac{1}{Nn(n-1)} \left( \sum_{i=1}^{N} \sum_{j=1}^{k} n_{ij}^2 - Nn \right) $$

$$ P_{\bar{}} e = \sum_{j=1}^{k} p_j^2 $$

In our application, this looks as follows:
kappa = 0.19 => This means that there is a slight agreement.

Use this table for interpretation:

<table>
<thead>
<tr>
<th>\kappa</th>
<th>Interpretation</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt; 0</td>
<td>Poor agreement</td>
</tr>
<tr>
<td>0.01 - 0.20</td>
<td>Slight agreement</td>
</tr>
<tr>
<td>0.21 - 0.40</td>
<td>Fair agreement</td>
</tr>
<tr>
<td>0.41 - 0.60</td>
<td>Moderate agreement</td>
</tr>
<tr>
<td>0.61 - 0.80</td>
<td>Substantial agreement</td>
</tr>
<tr>
<td>0.81 - 1.00</td>
<td>Almost perfect agreement</td>
</tr>
</tbody>
</table>

Or using R:

> T4t<-t(T4)
> kappam.fleiss(T4t, exact = FALSE, detail = FALSE)
Fleiss' Kappa for m Raters

Subjects = 9
Raters = 11
Kappa = 0.19

z = 8.18
p-value = 2.22e-16