CS 598 ACK
Experimental HCI & Interactive Technologies

Text Chapter 5

Statistics (2 of 5)
5.2.2 More than two conditions (analysis of variance)

If there are more than two conditions, then the analysis has to be done in two parts:

**STEP 1:** We need to determine whether the values of the independent variable have had any effect on performance over *all* conditions.

**STEP 2:** Once we have established this, we can determine which conditions performed better or worse than other conditions by comparing pairs of conditions.
Table 5.7: Form of data for ANOVA test (repeated measures), with example data: all participants have used k conditions\textsuperscript{13}

<table>
<thead>
<tr>
<th>Participant</th>
<th>Condition 1</th>
<th>Condition 2</th>
<th>Condition 3</th>
<th>\ldots</th>
<th>Condition k</th>
<th>Mean (over all conditions)</th>
</tr>
</thead>
<tbody>
<tr>
<td>p1</td>
<td>0.34</td>
<td>0.39</td>
<td>0.45</td>
<td></td>
<td>0.32</td>
<td>0.38</td>
</tr>
<tr>
<td>p2</td>
<td>0.23</td>
<td>0.61</td>
<td>0.26</td>
<td></td>
<td>0.64</td>
<td>0.52</td>
</tr>
<tr>
<td>p3</td>
<td>0.64</td>
<td>0.50</td>
<td>0.48</td>
<td></td>
<td>0.71</td>
<td>0.55</td>
</tr>
<tr>
<td>p4</td>
<td>0.55</td>
<td>0.29</td>
<td>0.48</td>
<td></td>
<td>0.46</td>
<td>0.46</td>
</tr>
<tr>
<td>\ldots</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>pn</td>
<td>0.35</td>
<td>0.57</td>
<td>0.61</td>
<td></td>
<td>0.32</td>
<td>0.42</td>
</tr>
<tr>
<td>Mean (over all participants)</td>
<td>0.43</td>
<td>0.44</td>
<td>0.52</td>
<td></td>
<td>0.39</td>
<td>0.42</td>
</tr>
</tbody>
</table>
For Step 1, an analysis of variance (ANOVA) statistical test can be used to determine whether the different conditions have had any effect on performance.

As with t-tests, there are different ANOVAs for repeated measures and independent measures data.

Unlike the t-test, there is no distinction between a one-tailed or two-tailed test for ANOVA calculations because the test indicates overall effect of all conditions, rather than whether some conditions are better or worse than others.
For Step 2, a post-hoc pairwise comparison test will show where the differences between the conditions lie.

Post-hoc tests are done after an initial successful ANOVA test, as a means of finding out more about the data.

If the ANOVA does not indicate any significance in the data, then there is no point in performing a post-hoc test.

If there are overall differences in the data, then a post-hoc test will indicate where the differences lie.
The ANOVA repeated measures test calculates a variety of statistics from the data such as the sums of the square of all data points, the square of the sum of all data for each condition, and various types of degrees of freedom. It calculates a final F-value statistic, usually expressed to three decimal points.

A “critical” F-value can then be read off a statistical table showing the values of the F distribution.

Two degrees of freedom are needed:

1. dof (Between Conditions) = the number of conditions less one;
2. dof (Error) = the number of participants less one multiplied by the number of conditions less one. The p-value is usually .05.
We can look at the means of the performance associated with the conditions and speculate which conditions produced “better” or “worse” performance.

However, the ANOVA test does not tell us this pairwise comparison.

The degrees of freedom values are 2 and 42.

Figure 5.7: Bar chart showing the overall error means for each condition.
5.2.3 Post-Hoc Pairwise Comparisons (Tukey)

An ANOVA test may reveal that the conditions have significantly affected performance (Step 1), but it will not indicate where the differences lie; that is, it will not say which of the conditions produced better or worse performance. An ANOVA result is seldom sufficient information, so pairwise comparison analyses need to be made (Step 2).\(^{18}\)

**STEP 2:** In the case of there being more than two conditions, determine which conditions performed better or worse than other conditions by comparing pairs of conditions.
Post-hoc tests calculate the differences between the means of all pairs of conditions and compare each of them with a critical difference, which, like the critical t-value and the critical F-value, is based on statistical tables.

First, the absolute value of the difference between the means of all pairs of conditions needs to be calculated, producing a pairwise diagonal matrix.

Note that we are not interested in the direction of the difference (because that can easily be determined from looking at the data or the overview bar charts). Instead, it is the size of the difference that matters. We therefore ignore any minus signs and take the absolute value of the differences.
Table 5.9: Matrix of pairwise condition differences

<table>
<thead>
<tr>
<th></th>
<th>Condition 1</th>
<th>Condition 2</th>
<th>Condition 3</th>
<th>Condition 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Condition 1</td>
<td></td>
<td>5.23</td>
<td>2.45</td>
<td>2.21</td>
</tr>
<tr>
<td>Condition 2</td>
<td>3.34</td>
<td></td>
<td>5.11</td>
<td></td>
</tr>
<tr>
<td>Condition 3</td>
<td></td>
<td></td>
<td></td>
<td>4.89</td>
</tr>
<tr>
<td>Condition 4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Each unshaded cell represents the absolute difference in the performance means between the two conditions.\(^\text{17}\)
Several post-hoc statistics can be used to determine the critical difference and a variety of views on when they are most appropriate.

The formulae for the Tukey’s Honestly Significant Difference (HSD) statistic, appears in Appendix A2, Tables A2.5 (repeated measures) and A2.6 (independent measures).

The calculation of the critical difference is based on both the actual data (using some of the numbers calculated from the ANOVA test) and the $q$-value.

The $q$-value can be read directly from a “Studentized range statistic” table using the p-value, the number of conditions, and the dof( Error): The number of participants less one multiplied by the number of conditions less one.
The critical difference is then compared with all cells in the pairwise matrix. Any cell that has a value greater than the critical difference indicates a pairwise significant difference.

A significant difference between the two conditions represented by the cell can then be stated as a conclusion (with confidence less than the p-value).

There is an important caveat to the process of performing pairwise comparisons. The Tukey HSD test is a single-step pairwise comparison method for which the p-value used is usually the typical .05.

The alternative to doing a single-step comparison method is to perform a series of t-tests between each pair of conditions. Because this entails repeated analysis of the same data, the required p-value for significance needs to be reduced by the number of t-tests performed. QUESTION: Why?
Because statistical methods are based on probabilities, if we analyze the data twenty times, then we are likely to find significant results with a p value = .05, even if the conditions have had no effect on the data.

Simply by doing the analysis repeatedly, the .05 probability will eventually work in our favor and produce a statistically significant result. To mitigate against this, the “Bonferroni adjustment” states that for each subsequent pairwise analysis in post-hoc comparison tests, the required p-value should be reduced.
However, because single-step post-hoc tests like Tukey automatically take the Bonferroni adjustment into account, .05 can be used per usual.

It is thus preferable to use a single-step method because p- value adjustments do not need to be made. In contrast, if a statistical package does not perform a single-step test like Tukey HSD, then the experimenter must make sure that appropriate adjustment is made to the p- value in pairwise t -test comparisons.

Such adjustments are called “Bonferroni Confidence Interval Adjustments.”

So if your data are normally distributed and you have been able to use parametric methods, then you would have answered your research question – you would be able to say whether the independent variable has had an effect on performance, and where the significant pairwise differences lie.

In contrast, if the data are not normally distributed, then nonparametric methods should be used.