CS 598 ACK
Experimental HCI & Interactive Technologies

Text Chapter 5

Statistics (4 of 5)
Preference data is a common type of quantitative data, which are collected by asking participants to rank or rate conditions or stimuli according to their preference.

Rating means associating a value with each item, for example, indicating the relative difficulty of tasks by giving each a number within a given range (e.g., 1 for “very easy,” 10 for “very difficult”): more than one item may have the same rating.

Ranking means putting the items in order to indicate how they differ, for example, sorting a set of conditions in order of complexity, and thus allocating a rank number to each condition: ideally, no two items may have the same rank (although in practice, ranks may be shared).
There is an important distinction between the nature of performance and preference data. Performance data are measured on an interval (or ratio) scale; that is, the numbers used to represent the data points are equally distant from each other, with the distance between consecutive numbers always representing the same value difference.

Preference data are not measured on an interval scale, as it is, by its nature, relative – the preference rank or rating given to a condition depends on the participant’s opinion of the other conditions.

There is no guarantee that the “preference distance” between consecutive preference ratings is the same – the difference in a participant’s preference between a condition given a 7 rating and one given an 8 rating is not necessarily the same as the difference in preference between conditions given ratings of 2 and 3.
Similarly, in a ranking task, if 1 represents “worst,” 2 represents “middle,” and 3 represents “best,” then it cannot be said that the difference in preference between “worst” and “middle” is always the same as the difference between “middle” and “best,” even though the difference between 1 and 2 is always the same as the difference between 2 and 3. For such non-interval (or non-ratio) scale data, nonparametric methods should be used.

When applying the nonparametric methods described previously, the performance data needed to be ranked before the methods were applied.

For ranked preference data, this initial ranking step has already been done, and nonparametric methods can be directly applied to the raw data. For rated preference data, the ratings need to be ranked as a first step (as with performance data).
## Table 5.24: Ranked data for fifteen web sites according to “good aesthetics”

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Procricket</td>
<td>14</td>
<td>6</td>
<td>12</td>
<td>7</td>
<td>14</td>
<td>2</td>
<td>7</td>
<td>7</td>
<td>2</td>
<td>10</td>
<td>5</td>
<td>8</td>
<td>12</td>
<td>15</td>
<td>10</td>
<td>11</td>
<td>8</td>
<td>3</td>
<td>12</td>
<td>180</td>
<td></td>
</tr>
<tr>
<td>Borders</td>
<td>11</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>7</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>2</td>
<td>5</td>
<td>11</td>
<td>61</td>
<td></td>
</tr>
<tr>
<td>Our Players</td>
<td>3</td>
<td>5</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>9</td>
<td>6</td>
<td>8</td>
<td>3</td>
<td>9</td>
<td>4</td>
<td>3</td>
<td>7</td>
<td>5</td>
<td>2</td>
<td>4</td>
<td>10</td>
<td>7</td>
<td>4</td>
<td>13 112</td>
</tr>
<tr>
<td>Google</td>
<td>2</td>
<td>13</td>
<td>3</td>
<td>10</td>
<td>3</td>
<td>12</td>
<td>6</td>
<td>3</td>
<td>1</td>
<td>14</td>
<td>12</td>
<td>13</td>
<td>14</td>
<td>11</td>
<td>7</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>11</td>
<td>8</td>
<td>140</td>
</tr>
<tr>
<td>UniSat</td>
<td>7</td>
<td>10</td>
<td>8</td>
<td>5</td>
<td>1</td>
<td>9</td>
<td>5</td>
<td>4</td>
<td>11</td>
<td>3</td>
<td>6</td>
<td>6</td>
<td>8</td>
<td>9</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>9</td>
<td>14</td>
<td>6</td>
<td>9</td>
</tr>
<tr>
<td>UOA News</td>
<td>15</td>
<td>15</td>
<td>15</td>
<td>12</td>
<td>5</td>
<td>15</td>
<td>11</td>
<td>9</td>
<td>14</td>
<td>15</td>
<td>15</td>
<td>10</td>
<td>11</td>
<td>9</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>9</td>
<td>5</td>
<td>15</td>
<td>2</td>
</tr>
<tr>
<td>ASB</td>
<td>6</td>
<td>7</td>
<td>6</td>
<td>6</td>
<td>8</td>
<td>8</td>
<td>3</td>
<td>8</td>
<td>10</td>
<td>4</td>
<td>1</td>
<td>3</td>
<td>4</td>
<td>6</td>
<td>7</td>
<td>6</td>
<td>7</td>
<td>2</td>
<td>10</td>
<td>2</td>
<td>7</td>
</tr>
<tr>
<td>SE</td>
<td>13</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>12</td>
<td>5</td>
<td>10</td>
<td>13</td>
<td>15</td>
<td>11</td>
<td>7</td>
<td>2</td>
<td>15</td>
<td>6</td>
<td>11</td>
<td>15</td>
<td>15</td>
<td>3</td>
<td>12</td>
<td>1</td>
<td>185</td>
</tr>
<tr>
<td>Te Mata</td>
<td>1</td>
<td>3</td>
<td>7</td>
<td>8</td>
<td>10</td>
<td>4</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>8</td>
<td>5</td>
<td>7</td>
<td>1</td>
<td>4</td>
<td>14</td>
<td>5</td>
<td>5</td>
<td>13</td>
<td>1</td>
<td>14</td>
<td>131</td>
</tr>
<tr>
<td>Cheese</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Foodtown</td>
<td>9</td>
<td>2</td>
<td>14</td>
<td>2</td>
<td>13</td>
<td>6</td>
<td>4</td>
<td>11</td>
<td>12</td>
<td>10</td>
<td>3</td>
<td>6</td>
<td>5</td>
<td>13</td>
<td>15</td>
<td>4</td>
<td>14</td>
<td>5</td>
<td>9</td>
<td>13</td>
<td>15</td>
</tr>
<tr>
<td>Seek</td>
<td>8</td>
<td>12</td>
<td>5</td>
<td>11</td>
<td>6</td>
<td>11</td>
<td>8</td>
<td>10</td>
<td>6</td>
<td>9</td>
<td>11</td>
<td>9</td>
<td>13</td>
<td>3</td>
<td>11</td>
<td>10</td>
<td>13</td>
<td>8</td>
<td>14</td>
<td>8</td>
<td>5</td>
</tr>
<tr>
<td>ECE</td>
<td>10</td>
<td>8</td>
<td>11</td>
<td>13</td>
<td>11</td>
<td>7</td>
<td>12</td>
<td>14</td>
<td>13</td>
<td>7</td>
<td>8</td>
<td>12</td>
<td>6</td>
<td>14</td>
<td>9</td>
<td>9</td>
<td>11</td>
<td>6</td>
<td>4</td>
<td>11</td>
<td>3</td>
</tr>
<tr>
<td>Telecom</td>
<td>12</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>10</td>
<td>15</td>
<td>15</td>
<td>4</td>
<td>12</td>
<td>4</td>
<td>11</td>
<td>10</td>
<td>11</td>
<td>13</td>
<td>12</td>
<td>12</td>
<td>1</td>
<td>15</td>
<td>10</td>
<td>10</td>
<td>213</td>
</tr>
<tr>
<td>Windows</td>
<td>5</td>
<td>14</td>
<td>13</td>
<td>14</td>
<td>15</td>
<td>14</td>
<td>13</td>
<td>5</td>
<td>9</td>
<td>5</td>
<td>14</td>
<td>15</td>
<td>15</td>
<td>12</td>
<td>2</td>
<td>13</td>
<td>3</td>
<td>12</td>
<td>13</td>
<td>7</td>
<td>4</td>
</tr>
<tr>
<td>Live</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gmail</td>
<td>4</td>
<td>11</td>
<td>10</td>
<td>15</td>
<td>4</td>
<td>13</td>
<td>14</td>
<td>12</td>
<td>5</td>
<td>13</td>
<td>13</td>
<td>14</td>
<td>12</td>
<td>10</td>
<td>10</td>
<td>14</td>
<td>6</td>
<td>12</td>
<td>6</td>
<td>9</td>
<td>214</td>
</tr>
<tr>
<td>Rank totals</td>
<td>120</td>
<td>120</td>
<td>120</td>
<td>120</td>
<td>120</td>
<td>120</td>
<td>120</td>
<td>120</td>
<td>120</td>
<td>120</td>
<td>120</td>
<td>120</td>
<td>120</td>
<td>120</td>
<td>120</td>
<td>120</td>
<td>120</td>
<td>120</td>
<td>120</td>
<td>120</td>
<td>120</td>
</tr>
</tbody>
</table>

Each cell is the unique ranking given by the participant to the web page.
Analysis of Preference Data and Further Analyses

Note that since the rankings have already been created, a Friedman test can be directly applied to this data to determine whether the web pages significantly differ in their judged aesthetic value, and a Nemenyi pairwise test could be directly applied to see if there are pairwise significant differences between them.
Preference data can also be analyzed to investigate whether they relate to performance data, asking such questions as, for example, “Do people perform better on the stimuli that they prefer?”

Or,

“Do people perform worse on those tasks that they perceive to be more difficult?”

It is no always the case that those stimuli or conditions that help people perform better are the ones that they like, or that their subjective perceptions of difficulty match the actual difficulty of the tasks as objectively measured by the performance data.
We can calculate the correlation coefficient between two sets of paired data: the result will range between –1 and 1, with 1 representing a perfect positive relationship, –1 representing a perfect negative relationship, and 0 representing no relationship. The statistical significance of a correlation coefficient depends on the number of participants, and can be read off a correlation statistical table.

There are two types of correlation coefficient than can be calculated: the Pearson correlation indicates whether there is a linear relationship between the two sets of values, and the Spearman’s rank correlation indicates whether the two sets of values follow the same increasing/decreasing trend (and, as its name suggests, uses ranked data).
There are additional analyses that may be useful or interesting, which this section covers. In particular, we look at what you might do if different performance measures give conflicting results, or if you believe that different experimental objects or tasks (or other factors) might have affected the performance data.

The first optional step in the analysis sequence is as follows:

**STEP 3:** In the case of there being more than one dependent variable, investigate any relationship between them.
In most cases, multiple dependent measures can be analyzed independently. However, sometimes the significant results from one analysis may contradict those of the other.

For example, the response time data may show that the Green condition produced significantly better performance than the Blue condition, whereas the error data show that the Green condition produced significantly worse performance than the Blue condition.
There is a contradiction here: Hierarchical is shown to be “better” than Orthogonal in terms of errors, but “worse” in terms of response time.
Analysis of Preference Data and Further Analyses

“Speed-Accuracy tradeoffs could be investigated by looking at the relationship between time and accuracy for data from the individual trials, with no aggregation according to condition, participant, task, or experimental object.

However, because the error data for each trial is only a 0 or a 1, and it is preferable to use continuous rather than binary data in correlation calculations, some aggregation may be appropriate. It seems appropriate to aggregate according to participant because some may have deliberately worked quickly and erroneously, whereas others took their time to get correct answers.

Thus, for each participant, we have an aggregated (usually the mean) value for response time, as well as an aggregated error value, for each condition. These paired time/error values are the ones used in the calculation of the correlation coefficient. The correlation coefficient will indicate whether there has been a trade-off between response time and errors.
Statistics (4 of 5)

Analysis of Preference Data and Further Analyses

The