NST 1B Experimental Psychology
Statistics practical 3

Difference tests (2): nonparametric

Rudolf Cardinal & Mike Aitken
22 / 23 January 2004; Department of Experimental Psychology
University of Cambridge

Handouts:
- Answers to Examples 3 (from last time)
- Handout 4 (diff. tests 2)
- Examples 4 (diff. tests 2)

pobox.com/~rudolf/psychology
These slides are on the web. No need to scribble frantically.

pobox.com/~rudolf/psychology
Nonparametric tests

Last time, we looked at the $t$ test, a parametric test — it made assumptions about parameters of the underlying populations (such as the distribution — e.g. assuming that the data are normally distributed).

If these assumptions are violated:

(a) we could transform the data to fit the assumptions better (NOT covered at Part 1B level)

or (b) we could use a nonparametric (‘distribution-free’) test that doesn’t make the same assumptions.

In general, if the assumptions of parametric tests are met, they are the most powerful. If not, we may need to use nonparametric tests. They may, for example, answer questions about medians rather than means. We’ll look at some nonparametric tests now that assume only that the data are measured on at least an ordinal scale.
The median is the value at or below which 50% of the scores fall when the data are arranged in numerical order.

(Can also be referred to as the 50th centile.)

If $n$ is odd, it’s the middle value (here, 17):

If $n$ is even, it’s the mean of the two middle values (here, 17.5):

\[
\text{The median is } \frac{17+18}{2} = 17.5
\]
The median

The **median** is the value at or below which 50% of the scores fall when the data are arranged in numerical order.

(Can also be referred to as the 50\textsuperscript{th} centile.)
Medians are less affected by outliers than means
Ranking removes ‘distribution’ information.

Whatever the distribution (normal, flat, skewed, bimodal...), the ranks are the same: 1, 2, 3, 4, 5, 6, 7.
How to rank data

Suppose we have ten measurements (e.g. test scores) and want to rank them. First, place them in ascending numerical order:

<table>
<thead>
<tr>
<th>X:</th>
<th>5</th>
<th>8</th>
<th>9</th>
<th>12</th>
<th>12</th>
<th>15</th>
<th>16</th>
<th>16</th>
<th>16</th>
<th>16</th>
<th>17</th>
</tr>
</thead>
<tbody>
<tr>
<td>rank:</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4.5</td>
<td>4.5</td>
<td>6</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>10</td>
<td></td>
</tr>
</tbody>
</table>

Then start assigning them ranks. When you come to a tie, give each value the mean of the ranks they’re tied for — for example, the 12s are tied for ranks 4 and 5, so they get the rank 4.5; the 16s are tied for ranks 7, 8, and 9, so they get the rank 8:
Nonparametric correlation

We’ve already seen this.
We met this in the first statistics practical. It’s a nonparametric version of correlation. You can use it when you obtain ranked data, or when you want to do significance tests on \( r \) but your data are not normally distributed (violating an assumption of the parametric \( t \) test that’s based on Pearson’s \( r \)).

- Rank the \( X \) values.
- Rank the \( Y \) values.
- Correlate the \( X \) ranks with the \( Y \) ranks. (You do this in the normal way for calculating \( r \), but you call the result \( r_s \).)
- To ask whether the correlation is ‘significant’, use the table of critical values of Spearman’s \( r_s \) in the *Tables and Formulae* booklet.
Nonparametric difference tests
Two unrelated samples: the Mann–Whitney $U$ test

**Logic**

- Suppose we have two samples with $n_1$ and $n_2$ observations in each (for a total of $n_1 + n_2 = N$ observations).
- We can rank all observations together, from 1 to $N$.
- If the two samples come from identical populations, the **sum of the ranks of ‘sample 1’ scores** is likely to be about the same as the **sum of the ranks of ‘sample 2’ scores**.
- But if sample 1 comes from a population with much lower values than sample 2, the sum of the ranks of ‘sample 1’ scores will generally be lower than the sum of the ranks of ‘sample 2’ scores.

**Null hypothesis:** the two samples were drawn from identical populations. [Unlike the unpaired $t$ test, whose null hypothesis was that the two samples came from populations with the same *mean*.] If we assume the distributions are similar, a significant Mann–Whitney test suggests that the **medians** of the two populations are different.
Calculating the Mann–Whitney $U$ statistic

1. Call the smaller group ‘group 1’, and the larger group ‘group 2’, so $n_1 < n_2$. (If $n_1 = n_2$, ignore this step.)
2. Calculate the sum of the ranks of group 1 ($= R_1$) and group 2 ($= R_2$).
3. \[ U_1 = R_1 - \frac{n_1(n_1 + 1)}{2} \]
4. \[ U_2 = R_2 - \frac{n_2(n_2 + 1)}{2} \]
5. The Mann–Whitney statistic $U$ is the smaller of $U_1$ and $U_2$.

Check your sums: verify that $U_1 + U_2 = n_1 n_2$ and $R_1 + R_2 = \frac{(n_1 + n_2)(n_1 + n_2 + 1)}{2}$.
Mann–Whitney \( U \) test: EXAMPLE 1

Mothers either received care from **first trimester** onwards — birth weights (in kg):
{1.68, 3.83, 3.11, 2.76, 1.70, 2.79, 3.05, 2.66, 1.40, 2.775}
or from the **third trimester** onwards — babies’ birth weights:
{2.94, 3.38, 4.90, 2.81, 2.80, 3.21, 3.08, 2.95}.

Is there a significant difference between the birthweights of the two groups?

- Third trimester group smaller, so is **group 1** \( (n_1 = 8) \). Other is **group 2** \( (n_2 = 10) \).

- **Ranks for group 1:** \{10, 16, 18, 9, 8, 15, 13, 11\}; **rank sum** \( R_1 = 100 \).
- **Ranks for group 2:** \{2, 17, 14, 5, 3, 7, 12, 4, 1, 6\}; **rank sum** \( R_2 = 71 \).

\[
U_1 = R_1 - \frac{n_1(n_1 + 1)}{2} = 100 - \frac{8 \times 9}{2} = 64
\]

\[
U_2 = R_2 - \frac{n_2(n_2 + 1)}{2} = 71 - \frac{10 \times 11}{2} = 16
\]

\[
\begin{align*}
U &= \frac{U_1 + U_2}{2} = \frac{64 + 16}{2} = 80 \\
n_1n_2 &= 8 \times 10 = 80 \quad \text{... good, they match} \\
R_1 + R_2 &= 100 + 71 = 171
\end{align*}
\]

\[
\frac{(n_1 + n_2)(n_1 + n_2 + 1)}{2} = \frac{(8 + 10)(8 + 10 + 11)}{2} = 171 \quad \text{... good, they do too}
\]
Distribution of the Mann–Whitney $U$ statistic (if $H_0$ is true)

A discrete (stepwise) distribution, rather than the continuous distributions we’ve looked at before.
Determining a significance level ($p$ value) from $U$

If $n_2 \leq 20$, look up the **critical value** for $U$ in your tables. (The critical value depends on $n_1$ and $n_2$.)

If your $U$ is **smaller** than the critical value, it’s significant (you reject the null hypothesis).

If $n_2 > 20$, the tables don’t give you critical values, but by this point the $U$ statistic is approximately **normally distributed**, so we can calculate a **$Z$ score** from $U$ and test that in the usual way, using tables of $Z$.

The formula for calculating $Z$ from $U$ is on the Formula Sheet.

Our example: $U = 16$, $n_1 = 8$, $n_2 = 10$; critical value of $U$ is 18 (from tables). Our $U$ less than this, so **birthweight difference was significant** ($p < 0.05$ two-tailed).
Mann–Whitney $U$ test: EXAMPLE 2 (a)

Transfer along a continuum practical (last year’s data, I’m afraid). Different groups of subjects were trained with {1 or 3} blocks of {easy or hard} discriminations before being tested on similar discriminations. Here are the test scores for the 3-block groups (high = good). Is there an effect of training difficulty? You could use either an unpaired $t$ test or a Mann–Whitney $U$ test; try the latter.

<table>
<thead>
<tr>
<th></th>
<th>Easy</th>
<th>30</th>
<th>40</th>
<th>42.5</th>
<th>45</th>
<th>50</th>
<th>50</th>
<th>55</th>
<th>55</th>
<th>65</th>
<th>65</th>
<th>75</th>
<th>80</th>
<th>82.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>3 Blocks</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></td>
<td>Easy</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></td>
<td>Hard</td>
<td>-10</td>
<td>-5</td>
<td>7.5</td>
<td>10</td>
<td>15</td>
<td>25</td>
<td>27.5</td>
<td>30</td>
<td>35</td>
<td>35</td>
<td>45</td>
<td>50</td>
<td>50</td>
</tr>
</tbody>
</table>

1. Call the smaller group ‘group 1’, and the larger group ‘group 2’, so $n_1 < n_2$. (If $n_1 = n_2$, ignore this step.)
2. Calculate the sum of the ranks of group 1 ($-R_1$) and group 2 ($-R_2$).
3. $U_1 = R_1 - \frac{n_1(n_1 + 1)}{2}$
4. $U_2 = R_2 - \frac{n_2(n_2 + 1)}{2}$
5. The Mann–Whitney statistic $U$ is the smaller of $U_1$ and $U_2$.

Check your sums: verify that $U_1 + U_2 = n_1 n_2$ and $R_1 + R_2 = \frac{(n_1 + n_2)(n_1 + n_2 + 1)}{2}$. 
### Mann–Whitney $U$ test: EXAMPLE 2 (b)

<table>
<thead>
<tr>
<th>3 Blocks Easy</th>
<th>15</th>
<th>20</th>
<th>30</th>
<th>40</th>
<th>42.5</th>
<th>45</th>
<th>50</th>
<th>50</th>
<th>55</th>
<th>55</th>
<th>65</th>
<th>65</th>
<th>75</th>
<th>80</th>
<th>82.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ranks</td>
<td>5.5</td>
<td>7</td>
<td>10.5</td>
<td>14</td>
<td>15</td>
<td>16.5</td>
<td>19.5</td>
<td>19.5</td>
<td>22.5</td>
<td>22.5</td>
<td>26.5</td>
<td>26.5</td>
<td>28</td>
<td>29</td>
<td>30</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>3 Blocks Hard</th>
<th>-10</th>
<th>-5</th>
<th>7.5</th>
<th>10</th>
<th>15</th>
<th>25</th>
<th>27.5</th>
<th>30</th>
<th>35</th>
<th>35</th>
<th>45</th>
<th>50</th>
<th>50</th>
<th>57.5</th>
<th>60</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ranks</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5.5</td>
<td>8</td>
<td>9</td>
<td>10.5</td>
<td>12.5</td>
<td>12.5</td>
<td>16.5</td>
<td>19.5</td>
<td>19.5</td>
<td>24</td>
<td>25</td>
</tr>
</tbody>
</table>

- Both groups same size. Arbitrarily, call the Easy group ‘group 1’ ($n_1 = 15$). Hard group is ‘group 2’ ($n_2 = 15$).
- **Rank sum** $R_1 = 5.5 + 7 + \ldots + 30 = 292.5$
- **Rank sum** $R_2 = 1 + 2 + \ldots + 25 = 172.5$

\[
U_1 = R_1 - \frac{n_1(n_1 + 1)}{2} = 292.5 - \frac{15 \times 16}{2} = 172.5
\]
\[
U_2 = R_2 - \frac{n_2(n_2 + 1)}{2} = 172.5 - \frac{15 \times 16}{2} = 52.5
\]
\[
U = \min(U_1, U_2) = \min(172.5, 52.5) = 52.5
\]

Critical $U$ ($n_1 = n_2 = 15$) for $\alpha = 0.05$ two-tailed is 65. So **significant**.

Check sums:

\[
U_1 + U_2 = 172.5 + 52.5 = 225
\]
\[
n_1n_2 = 15 \times 15 = 225
\]
\[
\text{... good, they match}
\]
\[
R_1 + R_2 = 292.5 + 172.5 = 465
\]
\[
\frac{(n_1 + n_2)(n_1 + n_2 + 1)}{2} = \frac{(15 + 15)(15 + 15 + 1)}{2} = 465
\]
\[
\text{... good, they do too}
\]
Time-saving tip…

If the ranks do not overlap at all, \( U = 0 \).

Example:

<table>
<thead>
<tr>
<th></th>
<th>55</th>
<th>65</th>
<th>75</th>
<th>80</th>
<th>82.5</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Group A Ranks</strong></td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
</tr>
<tr>
<td><strong>Group B Ranks</strong></td>
<td>10</td>
<td>15</td>
<td>39</td>
<td>40</td>
<td>48</td>
</tr>
</tbody>
</table>

\[ U = 0 \]
If you find a significant difference...

If you conduct a Mann–Whitney test and find a significant difference, **which group had the larger median and which group had the smaller median?**

<table>
<thead>
<tr>
<th>group 1</th>
<th>40</th>
<th>41</th>
<th>43</th>
<th>median = 41</th>
</tr>
</thead>
<tbody>
<tr>
<td>ranks 1</td>
<td>9</td>
<td>10</td>
<td>12</td>
<td>rank sum $R_1 = 31$</td>
</tr>
<tr>
<td>group 2</td>
<td>5</td>
<td>7</td>
<td>9</td>
<td>10</td>
</tr>
<tr>
<td>ranks 1</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
</tbody>
</table>

Here, $U = 2$. Significant (critical $U = 3$, $\alpha = 0.05$ two-tailed).

Group 1 has a significantly larger median (even though the rank sums convey the opposite impression).

**You have to calculate the medians.** But this is quick.
Two related samples: Wilcoxon matched-pairs signed-rank test

Logic
• Suppose we have a set of $n$ paired scores — for each subject, say, we have one score from condition 1 and one score from condition 2.
• We can calculate the difference score $condition_1 – condition_2$ for each pair. Then we rank the non-zero differences.
• If, on average, there is no difference between performance in condition 1 and condition 2, then the sum of the ranks of the positive differences should be about the same as the sum of the ranks of the negative differences.
• But if there is a difference between condition 1 and condition 2, the + and – rank sums should differ.

Null hypothesis: the distribution of differences between the pairs of scores is symmetric about zero. Since the median and mean of a symmetric population are the same, this can be restated as ‘the differences between the pairs of scores are symmetric with a mean and median of zero’.
Calculating the Wilcoxon $T$ statistic

Easy. From the Formula Sheet:

**Calculating the Wilcoxon matched-pairs signed-rank statistic, $T$**

1. Calculate the difference scores.
2. Ignore any differences that are zero.
3. Rank the difference scores, *ignoring their sign* (+ or –).
4. Add up all the ranks for difference scores that were positive; call this $T^+$.  
5. Add up all the ranks for difference scores that were negative; call this $T^-$.  
6. The Wilcoxon matched-pairs statistic $T$ is the smaller of $T^+$ and $T^-$.  

Check your sums: verify that $T^+ + T^- = \frac{n(n+1)}{2}$.  

Wilcoxon matched-pairs signed-rank test: EXAMPLE 1

Measure blood pressure (BP₁). Make subjects run a lot. Measure blood pressure again (BP₂) in the same subjects. Has their blood pressure changed?

<table>
<thead>
<tr>
<th>Before (BP₁):</th>
<th>130</th>
<th>148</th>
<th>170</th>
<th>125</th>
<th>170</th>
<th>130</th>
<th>130</th>
<th>145</th>
<th>119</th>
<th>160</th>
</tr>
</thead>
<tbody>
<tr>
<td>After (BP₂):</td>
<td>120</td>
<td>148</td>
<td>163</td>
<td>120</td>
<td>135</td>
<td>143</td>
<td>136</td>
<td>144</td>
<td>119</td>
<td>120</td>
</tr>
</tbody>
</table>

Difference (BP₁ – BP₂): 10 0 7 5 35 –13 –6 1 0 40

Rank of difference (ignoring zero differences and sign):

<table>
<thead>
<tr>
<th>‘Signed rank’</th>
<th>5</th>
<th>4</th>
<th>2</th>
<th>7</th>
<th>6</th>
<th>3</th>
<th>1</th>
<th>8</th>
</tr>
</thead>
</table>

Ranks of positive differences:

| Ranks of negative differences: | 6 | 3 |

Sum of positive ranks \( T^+ = 5 + 4 + 2 + 7 + 1 + 8 = 27 \)
Sum of negative ranks \( T^- = 6 + 3 = 9 \)
Wilcoxon \( T = \) the smaller of \( T^+ \) and \( T^- = 9 \). And \( n = 8 \).

Check sums:
\[
T^+ + T^- = 27 + 9 = 36
\]
\[
\frac{n(n+1)}{2} = \frac{8\times9}{2} = 36
\]
... good, they match
Distribution of the Wilcoxon $T$ statistic (if $H_0$ is true)

$f_n(T)$, where $n =$ number of non-zero difference scores

shaded area $= \alpha = P(T < T_{critical})$
Determining a significance level (p value) from $T$

If $n \leq 25$, look up the **critical value** for $T$ in your tables. (The critical value depends on $n$.)

If your $T$ is **smaller** than the critical value, it’s significant (you reject the null hypothesis).

If $n > 25$, the tables don’t give you critical values, but by this point the $T$ statistic is approximately **normally distributed**, so we can calculate a **Z score** from $T$ and test that in the usual way, using tables of $Z$.

The formula for calculating $Z$ from $T$ is on the Formula Sheet.

In our example, $T = 9$ and $n = 8$. Critical value of $T$ is 6 (for $\alpha = 0.05$ two-tailed); since our $T$ is **not** smaller than this, the BP difference was **not** significant.
Wilcoxon matched-pairs signed-rank test: EXAMPLE 2 (a)

**Proactive interference practical** (subset of last year’s data, I’m afraid). Subjects hear and repeat trigram (e.g. CXJ), perform distractor task, recall trigram. Compare trials 9 & 10 (after many similar trigrams) with trials 11 & 12 (after shift to new type of trigram, e.g. 925). Is there ‘release’ from proactive interference? Note very non-normal difference scores; parametric test unsuitable.

<table>
<thead>
<tr>
<th>Subject</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
</tr>
</thead>
<tbody>
<tr>
<td>% correct trials 9 &amp; 10</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>50</td>
<td>50</td>
<td>0</td>
<td>0</td>
<td>50</td>
<td>0</td>
<td>50</td>
<td>50</td>
</tr>
<tr>
<td>% correct trials 11 &amp; 12</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>50</td>
<td>50</td>
<td>100</td>
<td>50</td>
<td>100</td>
<td>100</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Subject</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
<th>17</th>
<th>18</th>
<th>19</th>
<th>20</th>
<th>21</th>
<th>22</th>
<th>23</th>
<th>24</th>
</tr>
</thead>
<tbody>
<tr>
<td>% correct trials 9 &amp; 10</td>
<td>50</td>
<td>50</td>
<td>50</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>50</td>
<td>50</td>
<td>50</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>% correct trials 11 &amp; 12</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>50</td>
<td>50</td>
<td>0</td>
<td>50</td>
<td>50</td>
<td>50</td>
</tr>
</tbody>
</table>

The procedure is:
1. Calculate the difference scores.
2. Ignore any differences that are zero.
3. Rank the difference scores, *ignoring their sign* (+ or −).
4. Add up all the ranks for difference scores that were positive; call this \( T^+ \).
5. Add up all the ranks for difference scores that were negative; call this \( T^- \).
6. The Wilcoxon matched-pairs statistic \( T \) is the smaller of \( T^+ \) and \( T^- \).

Check your sums: verify that \( T^+ + T^- = \frac{n(n+1)}{2} \).
Wilcoxon matched-pairs signed-rank test: EXAMPLE 2 (b)

Sum of positive ranks $T^+ = (4 \times 8) = 32$
Sum of negative ranks $T^- = (4 \times 17.5) + (11 \times 8) = 158$
Wilcoxon $T =$ the smaller of $T^+$ and $T^- = 32$. And $n = 19$.

Check sums:
\[
T^+ + T^- = 32 + 158 = 190
\]
\[
\frac{n(n+1)}{2} = \frac{19 \times 20}{2} = 190 \quad \text{... good, they match}
\]

From our tables, for $n = 19$ and $\alpha = 0.05$ two-tailed, the critical value of $T$ is 47. Our $T$ is smaller, so the difference is significant. (In fact, it’s significant at $\alpha = 0.01$ two-tailed.) Subjects did better on trials 11 & 12 than on trials 9 & 10.

Last year’s (2003) data
One sample: Wilcoxon signed-rank test with only one sample

Very easy.

**Null hypothesis:** the median is equal to $M$.

For each score $x$, calculate a difference score ($x - M$). Then proceed as for the two-sample Wilcoxon test using these difference scores.

(Logic: if the median is $M$, then the sum of the ranks of the positive differences — from scores where $x > M$ — should be the same as the sum of the ranks of the negative differences — from scores where $x < M$. If the median isn’t $M$, then the two rank sums should differ.)

---

**Calculating the Wilcoxon matched-pairs signed-rank statistic, $T$**

The procedure is:
1. Calculate the difference scores.
2. Ignore any differences that are zero.
3. Rank the difference scores, *ignoring their sign* (+ or −).
4. Add up all the ranks for difference scores that were positive; call this $T^+$. 
5. Add up all the ranks for difference scores that were negative; call this $T^−$. 
6. The Wilcoxon matched-pairs statistic $T$ is the smaller of $T^+$ and $T^−$. 
Comparison of parametric and non-parametric tests

<table>
<thead>
<tr>
<th>Parametric test</th>
<th>Equivalent nonparametric test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Two-sample unpaired t test</td>
<td>Mann–Whitney $U$ test</td>
</tr>
<tr>
<td>Two-sample paired t test</td>
<td>Wilcoxon signed-rank test with matched pairs</td>
</tr>
<tr>
<td>One-sample t test</td>
<td>Wilcoxon signed-rank test, pairing data with a fixed value</td>
</tr>
</tbody>
</table>
Vote, please!

We’re on hand now to answer questions about any aspect of statistics, e.g.
- this practical
- previous practicals
- example sheets so far

There’s one more statistics teaching practical (4/5 March, $\chi^2$). Then there’s a revision practical (29/30 April). We’ll go over past exam paper questions then.

There may be a free practical slot (either 3/4 Feb or 24/25 Feb).

Would you like an extra session — no new material, but a chance to go over problems or example questions?