Lecturer: Geoff Nicholls

Lecture 2: The Wilcoxon rank test; Location estimate.

Notes and Problem sheets are available at

http://www.stats.ox.ac.uk/~nicholls/sm2-sb1b/
(L1-7)

http://www.stats.ox.ac.uk/~caron/teaching/sb1b/
(L8-13)

and via the MSc weblearn pages.
Rank-based methods

- These are generally permutation tests which make a particular choice of test statistic.
- The tests statistic is a function of the ranking of the values in the data.
- By throwing out the values and basing the test on rank alone we get a test that give the same result for any monotone transformation of the data.
- The test is also relatively insensitive to outliers. This is the same sort of robustness we see in the median.
The Wilcoxon rank sum test

- This is a two sample test for location.
- Given $X \sim F$ and $Y - \Delta \sim F$ with $F, \Delta$ unknown.
- $\Delta$ - location shift

Definition: Wilcoxon rank sum test

\[ H_0 : \Delta = 0 \]

$R_k, k = 1 \ldots, n+m$: ranks of the combined sample $Z = (X, Y)$

Wilcoxon test statistic is defined as

\[ W = \sum_{i=n+1}^{n+m} R_i \]

This choice of test statistic is sensitive to variation in the location parameter $\Delta$. As $\Delta$ gets bigger, $W$ gets bigger.
- $F_0$: distribution of $W$ under $H_0$
- $w^\alpha$: $\alpha$-quantile of $F_0$ (ie, of $W$)
- We reject the null
  - if $W > w^{1-\alpha}$ under alternative $H_A: \Delta > 0$
  - if $W < w^\alpha$ under alternative $H_A: \Delta < 0$
  - if $W < w^{\alpha/2}$ or $W > w^{1-\alpha/2}$ under $H_A: \Delta \neq 0$.

To get quantiles $w^\alpha$ we need the null distribution of $W$. Under the null all permutations of entries of $Z$ are equally likely.

$$Pr(W(R(Z)) > w^{1-\alpha}) = \alpha$$

so we choose $w^{1-\alpha}$ so that

$$N!^{-1} \sum_{r \in \mathcal{P}_N} 1\{W(r) > w^{1-\alpha}\} = \alpha.$$  

The distribution of $W$ has “parameters” $m$ and $n$. 
Alternatively we can compute a $p$-value using a Monte Carlo permutation test with test statistic $T(Z) = W(Z)$. Suppose we observe ranks $X = x$ and $Y = y$. Let $z = (x, y)$ and $W_{obs} = W(R(z))$. Under H0, $R \sim \text{Unif}(\mathcal{P}_{m+n})$.

- Under $H_A : \Delta > 0$, compute
  
  $$p^+ = \Pr(W(R) > W_{obs})$$

  and reject H0 if $p^+ < \alpha$.

- Under $H_A : \Delta < 0$, compute
  
  $$p^- = \Pr(W(R) < W_{obs})$$

  and reject H0 if $p^- < \alpha$.

- Under $H_A : \Delta \neq 0$, compute $p^+, p^-$ and reject H0 if
  
  $$2 \min(p^-, p^+) < \alpha$$

  as the distribution of $W$ is symmetric about $m(n+m+1)/2$. 
The Wilcoxon test statistic $W(R(X, Y))$ and the Mann-Witney test statistic

$$U(X, Y) = \sum_{i=1}^{n} \sum_{j=1}^{m} 1\{Y_j > X_i\}$$

are related,

$$U = W - m(m + 1)/2$$

(see PS1) so they give the same test for the same $H_0, H_A, \alpha$. Gibbons and Chakraborti convention is MW tests $H_0$: $F_X = F_Y$ and Wilcoxon tests $H_0$: $\Delta = 0$ given $F_X(a) = F_Y(a-\Delta), a \in \mathbb{R}$.

In R `pwilcox(q, m, n,...)` etc is the distribution of $U$ not $W$ (so it has mean $mn/2$ not $m(m + n + 1)/2$ etc).

The Wilcoxon test is a powerful test for location.
Example: gene expression data

<table>
<thead>
<tr>
<th>Patient</th>
<th>Subtype A</th>
<th>Subtype B</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.99</td>
<td>1.14</td>
</tr>
<tr>
<td>2</td>
<td>0.33</td>
<td>1.06</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Gene 3500</th>
<th>Rank</th>
<th>Subtype A</th>
<th>Rank</th>
<th>Subtype B</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.99</td>
<td>2</td>
<td>0.33</td>
<td>1</td>
<td>1.14</td>
</tr>
<tr>
<td>1.14</td>
<td>4</td>
<td>1.06</td>
<td>3</td>
<td>3</td>
</tr>
</tbody>
</table>

Test statistic

\[ W = \sum_{i=n+1}^{n+m} R_i = 4 + 3 = 7. \]

Is this ‘suspiciously’ low/high?
Exact distribution
What is the distribution of $W$ for $n = m = 2$ under the null hypothesis?
All following outcomes are equally likely under $H_0$:

<table>
<thead>
<tr>
<th>patient</th>
<th>Subtype A</th>
<th>Subtype B</th>
<th>W</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rank</td>
<td>1 2</td>
<td>3 4</td>
<td>7</td>
</tr>
<tr>
<td>Rank</td>
<td>1 3</td>
<td>2 4</td>
<td>6</td>
</tr>
<tr>
<td>Rank</td>
<td>1 4</td>
<td>2 3</td>
<td>5</td>
</tr>
<tr>
<td>Rank</td>
<td>2 3</td>
<td>1 4</td>
<td>5</td>
</tr>
<tr>
<td>Rank</td>
<td>2 4</td>
<td>1 3</td>
<td>4</td>
</tr>
<tr>
<td>Rank</td>
<td>3 4</td>
<td>1 2</td>
<td>3</td>
</tr>
</tbody>
</table>

So $P_{H_0}(W \geq 7) = 1/6$.

Remark - we should consider all permutations, but permutations within groups don't change $W$ so the table would be expanded by a factor of four without changing our $p$-value.
Exact distribution Adding two observations on each side:

<table>
<thead>
<tr>
<th>Patient</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene</td>
<td>3500</td>
<td>0.99</td>
<td>0.33</td>
<td>0.41</td>
<td>1.01</td>
<td>1.14</td>
<td>1.06</td>
<td>1.21</td>
</tr>
<tr>
<td>Rank</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>6</td>
<td>5</td>
<td>7</td>
<td>8</td>
</tr>
</tbody>
</table>

Test statistic is $W = 6 + 5 + 7 + 8 = 26$. Is this significant? What is the distribution of $W$ under the null hypothesis now?
Exact distribution \((n + m)! = 40320\) possible permutations

Alternatively

\[
P_{H_0}(W \geq 26) = 1/\binom{n + m}{m} \approx 0.0142.
\]

- p-value of \(H_0\) against \(H_A: \Delta > 0\) is 0.0142
- p-value of \(H_0\) against \(H_A: \Delta \neq 0\) is \(2 \cdot 0.0142 = 0.0284\)

Both are well below the usual level \(\alpha = 0.05\)

Reject \(H_0\)
The Monte Carlo Wilcoxon (permutation) test Going through all possible permutations infeasible for large $n, m \rightarrow$.

Can easily be written as a standalone R-function:

```r
get_pvalue <- function(w_observed, n, m, nsim = 100000){
  ranks <- 1:(n+m)
  w_rand <- numeric(nsim)
  for (sim in 1:nsim){
    w_rand[sim] <- sum(sample(ranks, m, replace=FALSE))
  }

  prob_larger <- mean(w_observed <= w_rand)
  prob_smaller <- mean(w_observed >= w_rand)
  p_value <- 2* min(prob_smaller, prob_larger)
  return(p_value)
}
```

Notice this only needs $n$ and $m$ (ie not $X$ and $Y$) to determine the distribution of $W$ under the null.
Example: gene expression data

- Take gene expression values for gene 1706
- Test statistic is $W = \sum_{i=n+1}^{n+m} R_i = 628$

- 181 out of 100,000 random permutations is $W^0$ smaller than the actual observed $W$
- Approximate p-value is thus 0.00181 (one-sided) or 0.00362 (two-sided test)
Example: gene expression data

```r
> n <- 42
> m <- 25
> w_observed <- 628
> get_pvalue(w_observed,n,m)
[1] 0.00362

sum(sim$w<=628)
# [1] 181
181/100000
# [1] 0.00181
```
Example: gene expression data
### Example: thyroxine dataset

<table>
<thead>
<tr>
<th>Thyroxine</th>
<th>Group</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>34</td>
<td>Slight or No Symptoms</td>
<td>5</td>
</tr>
<tr>
<td>45</td>
<td>Slight or No Symptoms</td>
<td>6</td>
</tr>
<tr>
<td>49</td>
<td>Slight or No Symptoms</td>
<td>7</td>
</tr>
<tr>
<td>55</td>
<td>Slight or No Symptoms</td>
<td>8</td>
</tr>
<tr>
<td>58</td>
<td>Slight or No Symptoms</td>
<td>9</td>
</tr>
<tr>
<td>59</td>
<td>Slight or No Symptoms</td>
<td>10</td>
</tr>
<tr>
<td>61</td>
<td>Slight or No Symptoms</td>
<td>12</td>
</tr>
<tr>
<td>62</td>
<td>Slight or No Symptoms</td>
<td>13</td>
</tr>
<tr>
<td>86</td>
<td>Slight or No Symptoms</td>
<td>15</td>
</tr>
<tr>
<td>5</td>
<td>Marked Symptoms</td>
<td>1</td>
</tr>
<tr>
<td>8</td>
<td>Marked Symptoms</td>
<td>2</td>
</tr>
<tr>
<td>18</td>
<td>Marked Symptoms</td>
<td>3</td>
</tr>
<tr>
<td>24</td>
<td>Marked Symptoms</td>
<td>4</td>
</tr>
<tr>
<td>60</td>
<td>Marked Symptoms</td>
<td>11</td>
</tr>
<tr>
<td>84</td>
<td>Marked Symptoms</td>
<td>14</td>
</tr>
<tr>
<td>96</td>
<td>Marked Symptoms</td>
<td>16</td>
</tr>
</tbody>
</table>
For the group with marked symptoms:

\[ W = 1 + 2 + 3 + 4 + 11 + 14 + 16 = 51 \]

We can then carry out analysis in R:

```r
> n <- 9
> m <- 7
> w_observed <- 51
> get_pvalue(w_observed,n,m)
[1] 0.40526
```

20263 out of 100,000 random permutations has \( W^0 \) smaller than the actual observed \( W \)

- One-sided p-value is 0.20263
- Two-sided p-value is 0.40526
Example: thyroxine dataset
A normal approximation to the distribution of $W$

- For $n$ samples of $X$ and $m$ samples of $Y$
- $X \sim F$ and $Y \sim F$

$$E_{H_0}(W) = \frac{m(n + m + 1)}{2}$$

$$Var_{H_0}(W) = \frac{nm(n + m + 1)}{12}$$

$$\frac{W - E_{H_0}(W)}{\sqrt{Var_{H_0}(W)}} \xrightarrow{d} \mathcal{N}(0, 1) \quad n, m \to \infty$$

See Gibbons and Chakraborti, "Nonparametric Statistical Inference", CRC, (2011) for proof of this CLT.
Example: thyroxine dataset

```r
> thyroxine$rank <- rank(thyroxine$thyr)
> w_obs<-sum(thyroxine$rank[thyroxine$group=='MarkedSymptoms'])
> n <- 9
> m <- 7
> E <- m*(m+n+1)/2  #59.5 here
> V <- n*m*(m+n+1)/12
> p <- 2*pnorm(-abs((w_obs+0.5-E)/sqrt(V)))
> p
[1] 0.3971011
```
Why the $+0.5$? This is a continuity correction. The distribution of $W$ is discrete and moves in steps of one. We want to estimate $p = \Pr(W \leq w_{obs})$ here (as we are in the lower tail).

To get the best approximation to this out of the CDF $\Phi(z)$ of a standard normal (ie, a continuous approximation to a discrete rv) we should evaluate the cdf at the value corresponding to $W + 1/2$.

Exercise: suppose we had $w_{new}=68$ (so, in the upper tail). Explain why we would now subtract the continuity correction,

$$> p \leftarrow 2*pnorm(-abs((w_{new}-0.5-E)/sqrt(V))).$$
Wilcoxon in R

> wilcox.test(thyr~group,data=thyroxine,conf.int=T,correct=F)

Wilcoxon rank sum test

data:  thyr by group
W = 23, p-value = 0.4079
alternative hypothesis: true location shift is not equal to 0
95 percent confidence interval:
 -50   26
sample estimates:
difference in location
       -26

How are the confidence interval for $\Delta$ and the location estimator $\hat{\Delta}$ computed?
Estimate of location

- Suppose $X \sim F$ and $Y - \Delta \sim F$ and we want to estimate $\Delta$ without saying anything about $F$.
- Location estimate is estimator $\hat{\Delta}$ for $\Delta$

- Lehman-Hodges estimator: value of $\tilde{\Delta}$ making the two-sided test $H_0: \Delta = \tilde{\Delta}$ least significant (ie choose $\hat{\Delta}$ so p-value closest to 1/2)
LH-Location estimate for a z-test

- Suppose $X \sim \mathcal{N}(0, \sigma^2)$ and $Y \sim \mathcal{N}(\Delta, \sigma^2)$
- z-test for $H_0 : \Delta = \hat{\Delta}$ is based on the statistics
  \[ Z(\hat{\Delta}) = \frac{(Y - \hat{\Delta} - X)}{\sigma \sqrt{1/n + 1/m}}, \]
- Under $H_0$: $Z \sim \mathcal{N}(0, 1)$
- Location estimator is value $\hat{\Delta}$ for which $Z(\hat{\Delta}) = 0$
- This is $\hat{\Delta} = \bar{Y} - \bar{X}$
LH-Location estimate for Wilcoxon test

\[ E_{H_0}(W) = \frac{m(n + m + 1)}{2}, \]

- \( W(X, Y) \) - Wilcoxon statistic for \( X_i, \ i = 1, \ldots, n \) and \( Y_j, \ j = 1, \ldots, m \)
- \( W(X, Y - \Delta) \) - Wilcoxon statistic for \( X_i, \ i = 1, \ldots, n \) and \( Y_j - \Delta, \ j = 1, \ldots, m \)
- Lehmann-Hodges estimator \( \hat{\Delta} \) is given by

\[
\hat{\Delta} = \arg\min_{\tilde{\Delta}} \left| W(X, Y - \tilde{\Delta}) - \frac{m(n + m + 1)}{2} \right|
\]

since \( W = m(n + m + 1)/2 \) would give a p-value equal one half (this is the mean and the median of the \( W \)-distribution).
It is easily computed (PS1 exercise) using the fact that

\[ \hat{\Delta} = \text{median}\{Y_j - X_i, 1 \leq i \leq n, 1 \leq j \leq m\}. \]

This is an unbiased estimator of \( \Delta \) [Gibbons and Chakraborti Sec 6.6.2].
Example: gene expression data Using the gene expression example again with $n = m = 3$ as an example

<table>
<thead>
<tr>
<th>Patient</th>
<th>Subtype A</th>
<th>Subtype B</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Gene 3500</td>
<td>0.99</td>
<td>0.33</td>
</tr>
<tr>
<td>Rank</td>
<td>3</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$X_i$</th>
<th>$Y_j$</th>
<th>$Y_j - X_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.99</td>
<td>1.14</td>
<td>0.15</td>
</tr>
<tr>
<td>0.99</td>
<td>1.06</td>
<td>0.07</td>
</tr>
<tr>
<td>0.99</td>
<td>1.21</td>
<td>0.22</td>
</tr>
<tr>
<td>0.33</td>
<td>1.14</td>
<td>0.81</td>
</tr>
<tr>
<td>0.33</td>
<td>1.06</td>
<td>0.73</td>
</tr>
<tr>
<td>0.33</td>
<td>1.21</td>
<td>0.88</td>
</tr>
<tr>
<td>0.41</td>
<td>1.14</td>
<td>0.73</td>
</tr>
<tr>
<td>0.41</td>
<td>1.06</td>
<td>0.65</td>
</tr>
<tr>
<td>0.41</td>
<td>1.21</td>
<td>0.8</td>
</tr>
</tbody>
</table>

- z-test - $\overline{Y} - \overline{X} = 0.56$
- Wilcoxon test - median of the differences: 0.73
Example: gene expression data Assume a recording error:

<table>
<thead>
<tr>
<th>Patient</th>
<th>Subtype A</th>
<th>Subtype B</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1 2 3</td>
<td>1 2 3</td>
</tr>
<tr>
<td>Gene 3500</td>
<td>0.99 0.33 0.41</td>
<td>1.14 1.06 121</td>
</tr>
</tbody>
</table>

- z-test - $\bar{Y} - \bar{X} = 40.34$
- Wilcoxon test - median of the differences: 0.73