Rank Tests

Patrick Breheny

October 7
Permutation testing allows great freedom to use a wide variety of test statistics, all of which lead to exact level-$\alpha$ tests regardless of the distribution of the data.

However, not all test statistics are equally good – we want test statistics with high power.

It is not possible to develop tests that are uniformly most powerful regardless of the distribution of the data.

Still, we would like our tests to be robust, meaning that they have good power for a wide variety of distributions.
Another attractive feature is *invariance*, meaning that the test results do not change when the data is transformed in some way.

For example:
- The results of a \( t \)-test do not change when \( x \) is replaced by \( ax + b \), for any constants \( a \) and \( b \).
- The \( t \)-test is said to be *location-scale invariant*.

A stronger type of invariance is invariance to any monotone transformation:
- The results of a \( t \)-test change if \( x \) is replaced by \( \log(x) \).
- The \( t \)-test is not invariant to monotone transformations.
Any test that is based on the ranks of the data, however, is clearly invariant to monotone transformations, as such transformations do not affect the relative ranking of observations.

Thus, rank-based tests do not depend on whether the outcome is measured on the original scale or the log scale – or any other scale, for that matter.

The is a powerful motivation for rank-based tests.

Another important motivation is that, as we will see, rank-based tests tend to be robust.
One way of constructing powerful tests based on ranks is to find the locally most powerful rank test.

We will see how this test is constructed for the most common application: testing for a difference in location between two groups.

A test is *locally most powerful* among a class of tests $\mathcal{T}$ for $H : \Delta = 0$ versus $K : \Delta \neq 0$ if it is uniformly most powerful at level $\alpha$ for $H$ versus $K_\epsilon$, where $K_\epsilon = \{ |\Delta| \in (0, \epsilon) \}$.

If the above class of tests is the set of rank-based tests, then the test is said to be a *locally most powerful rank* (LMPR) test.
**Theorem:** Let \( x_i \sim f(x - \Delta g_i) \), where \( g_i \) denotes group membership. Then

\[
T(r) = \sum_i g(i) \mathbb{E} \left\{ \frac{-\partial \log f(X(i))}{\partial X(i)} \right\}
\]

defines the locally most powerful rank test of \( H_0 : \Delta = 0 \)
Homework: Show that

$$P_0(r) + \Delta \frac{\partial}{\partial \Delta} P_{\Delta}(r)|_{\Delta=0} = \frac{1}{n!} \{1 + \Delta T(r)\},$$

where $T(r)$ is defined on the previous slide.

To accomplish this, you will need to interchange differentiation and integration. This cannot always be done – in general, certain regularity conditions regarding $f$ need to hold. Assume that these conditions hold and that interchanging the two is possible.

Hint: You may wish to consult Section 5.4 of Casella & Berger to refresh your memory concerning joint densities of order statistics.
This may seem like a step backwards – we’re trying to develop hypothesis tests that don’t assume anything about the distribution, but in order to calculate $T(r)$, we need to assume things about $f$

Keep in mind that all permutation tests are valid (i.e., have the correct size $\alpha$) regardless of the test statistic.

However, the true distribution $f$ will affect the power that arises from various test statistics.

Choosing $f$ poorly (i.e. you choose an $f$ that looks nothing like the actual $f$) will not affect the validity of your hypothesis test, only its power.
Linear rank statistics for $H_0$

- For testing $H_0$, a test statistic of the form

$$T(r) = \sum z_i a(r_i)$$

is called a *linear rank statistic.*

- An equivalent definition is

$$T(r) = \sum z_{ri} a(i)$$

- Here, $z_i$ is a covariate of some kind – e.g., an indicator of group membership.
- The function $a$ is called a *score.*
Connection with LMPR tests

- Note that the LMPR tests we just derived are based on linear rank statistics.
- Once again, all permutation tests based on linear rank statistics are valid level-$\alpha$ tests.
- However, different scores will lead to tests that are more powerful in some situations than others.
Central limit theorem approximation

- The null distribution of $T(r)$ can always be evaluated/approximated by numerical/Monte Carlo means, as we discussed in the previous lecture.

- A less computer-intensive approach is to use $E(T)$ and $V(T)$, and base the test on the central limit theorem.

- For example, under $H_0$,
  
  $E(T) = \bar{a} \sum_i z_i$
  
  $V(T) = \sigma_a^2 \sum_i (z_i - \bar{z})^2$, where $\sigma_a^2$ is the sample variance of $\{a_i\}$

- For linear statistics, then, we can easily obtain an estimate of $ASL$ without relying on Monte Carlo approximation (relying instead on a different approximation).
Logistic distribution

Suppose $x$ follows a logistic distribution:

$$f(x) = \frac{e^{-x}}{(1 + e^{-x})^2} \quad F(x) = \frac{1}{1 + e^{-x}}$$

This distribution is particularly easy to work with, because

$$f(x) = F(x)\{1 - F(x)\}$$

Thus,

$$a(i) = \frac{2i}{n + 1} - 1$$
This is a linear function of $i$ and therefore equivalent to the test statistic

$$T = \sum_{i} z_{r_i} i,$$

In other words, the sum of the ranks in one of the groups – i.e., the Wilcoxon Rank Sum Test

Thus, the Wilcoxon Rank Sum Test is the locally most powerful rank test when the true distribution of $x$ is logistic.
Other LMPR tests of $H_0$

This exercise can be carried out for a number of other distributions, although most of them do not have a closed form solution like the logistic distribution does:

<table>
<thead>
<tr>
<th>Distribution</th>
<th>$a(i)$</th>
<th>Name*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal (exact)</td>
<td>$E X(i)$</td>
<td>Fisher-Yates</td>
</tr>
<tr>
<td>Normal (approx.)</td>
<td>$\Phi^{-1}\left(\frac{i}{n+1}\right)$</td>
<td>van der Waerden</td>
</tr>
<tr>
<td>Double exponential</td>
<td>$\text{sign}(i - \frac{n+1}{2})$</td>
<td>Median test</td>
</tr>
</tbody>
</table>

*Some care should be used with test names, as different tests often go by different names in different settings. For example, the Fisher-Yates test is also called the “normal scores” test. Meanwhile, the Median test is usually associated with using the $\chi^2$ distribution on the scores rather than the exact null distribution.
Testing $H_1$

• Similar proofs and derivations can be constructed for testing $H_1$

• Here, linear rank tests are of the form:

\[
T(r) = \sum_i s_i a^+(r_i^+) = \sum_i s_{r_i} a^+(i)
\]

• LMPR tests are of the form

\[
a^+(i) = \mathbb{E} \left\{ -\frac{\partial}{\partial |X(i)|} \log f(|X(i)|) \right\}
\]
## LMPR tests of $H_1$

Locally most powerful rank tests of $H_1$ for various distributions:

<table>
<thead>
<tr>
<th>Distribution</th>
<th>$a^+(i)$</th>
<th>Name*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal (exact)</td>
<td>$\mathbb{E}</td>
<td>X(i)</td>
</tr>
<tr>
<td>Normal (approx.)</td>
<td>$\Phi^{-1}\left(\frac{1}{2} + \frac{1}{2} \frac{i}{n+1}\right)$</td>
<td>van der Waerden</td>
</tr>
<tr>
<td>Logistic</td>
<td>$i$</td>
<td>Wilcoxon signed-rank</td>
</tr>
<tr>
<td>Double exponential</td>
<td>1</td>
<td>Sign test</td>
</tr>
</tbody>
</table>

**Homework:** Show that the sign test is the locally most powerful rank test when $X$ follows a double exponential distribution.
Testing $H_2$

- For $H_2$, the test based on

$$T(r) = \sum_i a_f(r_i)a_g(q_i),$$

where the scores are the same as they were for $H_0$, is the LMPR test.

- In principle, one could assign different scores to the ranks of $X$ than you assign to the ranks of $Y$, to obtain tests that are, say, locally most powerful when $X$ follows a logistic distribution and $Y$ follows a normal distribution.

- However, this is rare; usually, we just assign the same scores to the ranks of $X$ and the ranks of $Y$. 
LMPR tests of $H_2$

Locally most powerful rank tests of $H_2$ for various distributions:

<table>
<thead>
<tr>
<th>Distribution</th>
<th>$a(i)$</th>
<th>Name*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal (exact)</td>
<td>$EX(i)$</td>
<td>Fisher-Yates</td>
</tr>
<tr>
<td>Normal (approx.)</td>
<td>$\Phi^{-1}\left(\frac{i}{n+1}\right)$</td>
<td>van der Waerden</td>
</tr>
<tr>
<td>Logistic</td>
<td>$i$</td>
<td>Spearman rank</td>
</tr>
<tr>
<td>Double exponential</td>
<td>$\text{sign}(i - \frac{n+1}{2})$</td>
<td>Quadrant test</td>
</tr>
</tbody>
</table>
Multivariate hypotheses

- Linear rank statistics can also be extended to test multivariate hypotheses.
- The most famous of these tests is the Kruskal-Wallis test.
- The basic idea is that

$$y = \alpha + \beta_1 z_{1i} + \cdots + \beta_p z_{pi} + \epsilon,$$

where $\epsilon \sim f$, and we are interested in testing $H_0: \beta_1 = \cdots = \beta_p = 0$.
- We won’t go into much detail regarding these multivariate tests, but I will make a few comments.
These tests are based on vectors of linear rank statistics:

\[ u(r) = \sum_i z_i a(r_i), \]

where \( z_i \) is now a vector of covariates (in the case of testing for equality of means across \( K \) samples, \( z_i \) would be a vector of indicator functions), and

\[ T(r) = u'Vu, \]

where \( V = V_0(u) \)

\( ASL \) can be calculated/approximated using either exact, Monte Carlo, or central limit theorem means:

\[ u'Vu \xrightarrow{d} \chi_p^2 \]
Multivariate scores and optimality

- One can use the same scores $a(i)$ that we derived earlier.
- However, these scores do not ensure that the resulting test is LMPR, like we had in the univariate case.
- Our LMPR proof does not extend to the multivariate case – indeed, LMPR tests do not necessarily exist for testing multivariate null hypotheses.
When it comes to numerically calculating a \( p \)-value, there are three approaches: exact calculation, asymptotic calculation based on the central limit theorem, and Monte Carlo approximation.

We have covered the Monte Carlo approach already.

The other approaches are available in R via the functions `wilcox.test` (and `kruskal.test`) for \( a(i) = i \), and via the package `exactRankTests` for any linear scores.
Asymptotic $p$-values

- Asymptotic Wilcoxon rank-sum tests and Wilcoxon signed-rank tests are both available via `wilcox.test`, which can be accessed in one of two ways:
  ```r
  wilcox.test(x1,x2)
  wilcox.test(x~g)
  ```
- For other scores, asymptotic evaluation of $p$-values is available via `perm.test` in the `exactRankTests` package:
  ```r
  perm.test(a1,a2)
  perm.test(a~g)
  ```
  where you can supply any scores $a[i]$
Exact $p$-values

- Exact $p$-values are available in both of these methods by specifying `exact=TRUE`.

- Both of these methods use a technique called the shift algorithm to obtain exact answers much, much faster than would be possible by evaluating all $n!$ permutations (although `perm.test` makes slight approximations to non-integer scores before applying the algorithm, so its ASL isn’t always “exact”).

- The default of both functions is to calculate exact scores if $n < 50$, and otherwise use a normal approximation.
For a homework assignment, we will continue to look at the driving/illegal drug use data from the previous lecture.

**Homework:** Test the null hypothesis that the distribution of following distance is the same in both groups using (a) the Wilcoxon rank-sum test, (b) the van der Waerden test, and (c) the Median test. For all three, report exact (or “exact”) $p$-values.