Rank Tests

Patrick Breheny

September 27

Power

- Permutation testing allows great freedom to use a wide variety of test statistics, all of which lead to exact level-α 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

Invariance

- 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

Rank-based tests

- 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
- This is a strong motivation for rank-based tests
- Another important motivation is that, as we will see, rank-based tests tend to be robust

Locally most powerful rank tests

- 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 α for H versus K_ϵ, 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

Locally most powerful tests for two-group comparison

Theorem: Let $X_i \sim f(x - \Delta g_i)$, where g_i denotes group membership. Then

$$T(\mathbf{r}) = \sum_i g_{(i)} \mathbb{E}\left\{ \frac{-\partial \log f(X_{(i)})}{\partial X_{(i)}} \right\}$$

defines the locally most powerful rank test of $H_0 : \Delta = 0$

Homework

Homework: Show that

$$\mathbb{P}_0(\mathbf{r}) + \Delta \frac{\partial}{\partial \Delta} \mathbb{P}_\Delta(\mathbf{r})|_{\Delta=0} = \frac{1}{n!}\{1 + \Delta T(\mathbf{r})\},$$

where $T(\mathbf{r})$ 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.

Comment

- 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 $T(\mathbf{r})$, we need to assume things about f
- Keep in mind that all permutation tests are valid (*i.e.*, have the correct size α) 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

$$T(\mathbf{r}) = \sum_i z_i a(r_i)$$

  is called a *linear rank statistic*
- An equivalent definition is

$$T(\mathbf{r}) = \sum_i z_{r_i} a(i)$$

- 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-α 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 $T(\mathbf{r})$ 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 $\mathbb{E}(T)$ and $\mathbb{V}(T)$, and base the test on the central limit theorem
- For example, under H_0,
    - $\mathbb{E}(T) = \bar{a} \sum_i z_i$
    - $\mathbb{V}(T) = \sigma_a^2 \sum_i (z_i - \bar{z})^2$, where σ_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:

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

- This distribution is particularly easy to work with, because

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

- Thus,

$$a(i) = \frac{2i}{n+1} - 1$$

Logistic distribution (cont'd)

- This is a linear function of i and therefore equivalent to the test statistic

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

- If z_i is an indicator of group membership, this is simply 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:

Distribution	$a(i)$	Name*
Normal (exact)	$\mathbb{E}X_{(i)}$	Fisher-Yates
Normal (approx.)	$\Phi^{-1}\left(\frac{i}{n+1}\right)$	van der Waerden
Double exponential	$\text{sign}(i - \frac{n+1}{2})$	Median test

*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 χ^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:

$$\begin{aligned} T(\mathbf{r}) &= \sum_i s_i a^+(r_i^+) \\ &= \sum_i s_{r_i} a^+(i) \end{aligned}$$

Testing H_1 (cont'd)

Theorem: Suppose $X_i \overset{\text{iid}}{\sim} f(x - \Delta)$, where x is symmetric about 0 (*i.e.*, X is symmetric about Δ). Then $T(\mathbf{r}) = \sum_i s_{r_i} a^+(i)$, where

$$a^+(i) = \mathbb{E}\left\{ -\frac{\partial}{\partial \left|X\right|_{(i)}} \log f(\left|X\right|_{(i)}) \right\}$$

defines the locally most powerful rank test of $H_1 : \Delta = 0$.

LMPR tests of H_1

Locally most powerful rank tests of H_1 for various distributions:

Distribution	$a^+(i)$	Name
Normal (exact)	$\mathbb{E}\lvert X\rvert_{(i)}$	Fraser
Normal (approx.)	$\Phi^{-1}\left(\frac{1}{2}+\frac{1}{2}\frac{i}{n+1}\right)$	van der Waerden
Logistic	i	Wilcoxon signed-rank
Double exponential	1	Sign test

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, linear rank tests are of the form

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

- **Theorem:** Suppose $X_i - \Delta Z_i \overset{\text{iid}}{\sim} f(x)$ and $Y_i - \Delta Z_i \overset{\text{iid}}{\sim} g(x)$, where Z_i is unobservable and may follow any arbitrary distribution, provided that $\mathbb{E}Z$ and $\mathbb{V}Z$ are finite. Then $T(\mathbf{r}) = \sum_i a_f(r_i) a_g(q_i)$, where

$$a_f(i) = \mathbb{E}\left\{-\frac{f'(X_{(i)})}{f(X_{(i)})}\right\} \qquad a_g(i) = \mathbb{E}\left\{-\frac{g'(Y_{(i)})}{g(Y_{(i)})}\right\}$$

defines the locally most powerful rank test of $H_2 : \Delta = 0$.

Testing H_2 (cont'd)

- 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:

Distribution	$a(i)$	Name
Normal (exact)	$\mathbb{E}X_{(i)}$	Fisher-Yates
Normal (approx.)	$\Phi^{-1}\left(\frac{i}{n+1}\right)$	van der Waerden
Logistic	i	Spearman rank
Double exponential	$\text{sign}(i - \frac{n+1}{2})$	Quadrant test

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

$$y_i = \alpha + \beta_1 z_{1i} + \cdots \beta_p z_{pi} + \epsilon_i,$$

where $\epsilon \overset{\text{iid}}{\sim} f$, and we are interested in testing $H_0 : \beta_1 = \cdots = \beta_p = 0$

Multivariate test statistics

- These tests are based on vectors of linear rank statistics:

$$\mathbf{u}(\mathbf{r}) = \sum_i \mathbf{z}_i a(r_i) \\ = \mathbf{Z}'\mathbf{a},$$

  where $\mathbf{z}_i$ is now a vector of covariates (in the case of testing for equality of means across K samples, $\mathbf{z}_i$ would be a vector of indicator functions)
- To proceed with hypothesis testing, we need to form a (scalar) test statistic from $\mathbf{u}$; for example, we could use $T(\mathbf{u}) = \max_j u_j$
- However, the more common (and typically more powerful) approach is to use quadratic forms

Quadratic test statistics

- Considering $\mathbf{r}$ as a random variable, we have, under the null, that

$$\mathbb{E}_0(\mathbf{u}) = \bar{a}\mathbf{Z}'\mathbf{1}$$
$$\mathbb{V}_0(\mathbf{u}) = \sigma_a^2 \tilde{\mathbf{Z}}'\tilde{\mathbf{Z}},$$

  where $\tilde{\mathbf{Z}}$ is a centered version of $\mathbf{Z}$

- Our quadratic test statistic is thus

$$T(\mathbf{r}) = (\mathbf{u} - \mathbf{u}_0)'\mathbf{V}^{-1}(\mathbf{u} - \mathbf{u}_0),$$

  where $\mathbf{u}_0 = \mathbb{E}_0(\mathbf{u})$ and $\mathbf{V} = \mathbb{V}_0(\mathbf{u})$

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

$$(\mathbf{u} - \mathbf{u}_0)'\mathbf{V}^{-1}(\mathbf{u} - \mathbf{u}_0) \xrightarrow{\text{d}} \chi_p^2$$

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

Testing H_1 with multiple groups

- It is possible to extend these notions to test H_1 in the case of multiple groups: forming a quadratic test statistic out of a multivariate linear rank statistic and using either exact, Monte Carlo, or asymptotic approached to calculating the ASL (the quadratic form again converges to a χ^2 distribution), although we will skip the details
- Still, it is worthwhile to be aware of the fact that there is a k-sample version of the Wilcoxon Signed Rank test, and it is called the *Friedman test*

Asymptotic vs. exact vs. Monte Carlo

- 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 `coin` for general linear scores and for Monte Carlo evaluation of the ASL

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:

```
wilcox.test(x1,x2)
wilcox.test(x~g)
```

- For other scores, evaluation of p-values is available via `independence_test` in the `coin` package:

```
independence_test(a~g)
```

where you can supply any scores `a[i]`

Exact p-values

- Exact p-values are available in both of these methods (in `wilcox.test` by specifying `exact=TRUE`, in `independence_test` by specifying `distribution='exact'`)
- 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 (this is only possible when $a(i)$ is an integer, so exact solutions take much longer for general scores than they do in the Wilcoxon case)
- The default of `wilcox.test` is to calculate exact scores if $n < 50$, and otherwise use a normal approximation; the default of `independence_test` is to use a normal approximation

Homework

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 both asymptotic and exact p-values.