STATS 200: Introduction to Statistical Inference **Autumn 2016**

Lecture 9 — Rank sum test and permutation tests

9.1 Rank sum test

The idea of converting observed data values to just their ranks, so as to deal with heavy-tailed data and deviations from normality, can be extended to the two-sample setting. Consider two independent samples $X_1, \ldots, X_n \overset{IID}{\sim} f$ and $Y_1, \ldots, Y_m \overset{IID}{\sim} g$, where f and g are two arbitrary PDFs, and the testing problem

$$\begin{aligned} &H_0 : f = g \\ &H_1 : f \text{ stochastically dominates } g \end{aligned}$$

(Recall from Lecture 7 that this alternative is one way of saying that values drawn from f "tend to be larger" than values drawn from g.)

The **rank-sum statistic** T_Y is defined as follows:

1. Consider the **pooled sample** of all observations $X_1, \ldots, X_n, Y_1, \ldots, Y_m$. Sort these $m + n$ values in increasing order. Assign the smallest a rank of 1, the next smallest a rank of 2, etc., and the largest a rank of $m + n$.

2. Define T_Y as the sum of the ranks corresponding to only the Y_i values, i.e. the values from only the second sample.[1]

We expect T_Y to be smaller under H_1 than under H_0, because under H_1 the values of Y_i tend to have smaller ranks. The test that rejects for small values of T_Y is called the **Wilcoxon rank-sum test**, known alternatively as the Mann-Whitney U-test or the Mann-Whitney-Wilcoxon test. (If we are testing a general two-sided alternative

$$H_1' : f \neq g$$

then we would reject for both large and small values of T_Y.)

The following theorem states that T_Y has the same distribution under every $P \in H_0$, and provides a method for determining the null distribution and rejection threshold when n and m are both large. (For small n and m, we can determine the exact null distribution of T_Y by computing T_Y for all $\binom{n+m}{m}$ possible sets of ranks for the Y_i's.)

Theorem 9.1. *The distribution of T_Y is the same under any PDF $f = g$. For large n and m, this distribution is approximately $\mathcal{N}(\frac{m(m+n+1)}{2}, \frac{mn(m+n+1)}{12})$.*

We won't prove this result; let's just make the following comments:

[1] One may consider equivalently T_X (the sum of ranks of the X_i's) as $T_X + T_Y$ is a fixed constant.

- If $f = g$, then each ordering of $X_1, \ldots, X_n, Y_1, \ldots, Y_m$ is equally likely. Since T_Y depends only on this ordering, its distribution must be the same under every PDF $f = g$.

- Let $I_k = 1$ if the kth largest value in $X_1, \ldots, X_n, Y_1, \ldots, Y_m$ belongs to the second sample, and $I_k = 0$ otherwise. Then

$$T_Y = \sum_{k=1}^{m+n} k I_k.$$

  Under H_0, I_k indicates whether the k^{th} "individual" is selected in a simple random sample of size m (without replacement) from a population of size $m+n$. Then the same computations as in Lecture 1 yield formulas for $\mathbb{E}[I_k]$, $\text{Var}[I_k]$, and $\text{Cov}[I_j, I_k]$. Applying linearity of expectation and bilinearity of covariance, we may obtain $\mathbb{E}[T_Y] = \frac{m(m+n+1)}{2}$ and $\text{Var}[T_Y] = \frac{mn(m+n+1)}{12}$ as in the above theorem. (Details are provided in Rice, Section 11.2.3 Theorem A and Section 7.3.1 Theorems A and B.)

9.2 Permutation and randomization tests

The main idea behind the (one-sample) signed-rank test and the (two-sample) rank-sum test is to exploit a symmetry under H_0. For the signed-rank test, the symmetry is that it is equally likely to observe $\pm X_1, \ldots, \pm X_n$ for each of the 2^n combinations of $+/-$ signs. For the rank-sum test, the symmetry is that it is equally likely to observe each of the $(m+n)!$ permutations of the pooled sample $X_1, \ldots, X_n, Y_1, \ldots, Y_m$.

In fact, this idea of exploiting symmetry provides an alternative (and useful) simulation-based method of obtaining a null distribution for *any* test statistic T for these problems:

Example 9.2. Consider two samples $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$, and any test statistic $T(X_1, \ldots, X_n, Y_1, \ldots, Y_m)$. (For concreteness, you can think about $T = \bar{X} - \bar{Y}$.) For a null hypothesis H_0 which specifies that all data from both samples are IID from a common distribution, for example

$$H_0 : X_1, \ldots, X_n, Y_1, \ldots, Y_m \overset{IID}{\sim} f$$

for an unknown PDF f, the **permutation null distribution** of T is the distribution of $T(X_1^*, \ldots, X_n^*, Y_1^*, \ldots, Y_m^*)$ when we fix the observed values $X_1, \ldots, X_n, Y_1, \ldots, Y_m$ and let $(X_1^*, \ldots, X_n^*, Y_1^*, \ldots, Y_m^*)$ be a permutation of $X_1, \ldots, X_n, Y_1, \ldots, Y_m$ chosen uniformly at random from the set of all $(m+n)!$ possible permutations. (For $T = \bar{X} - \bar{Y}$, what this effectively means is that we randomly choose n of the observations to be $X_1^*, \ldots, X_n^*$, set the remaining m observations to be $Y_1^*, \ldots, Y_m^*$, and compute $\bar{X}^* - \bar{Y}^*$.)

Under H_0, each of these $(m+n)!$ possible values of T is equally likely to be observed. To perform a test that rejects for large values of T, we may use the following procedure:

1. Randomly permute the pooled data B times (say $B = 10000$), and compute the value of T each time.

2. Compute an approximate p-value as the fraction of the B simulations where we obtained a value of T larger than t_{obs}, the value for the original (unpermuted) data. (Reject at level-α if this p-value is at most α.)

For a two-sided test that rejects for both large and small values of T, we can compute the p-value by taking the fraction of simulations where T is larger than t_{obs} or the fraction where T is smaller than t_{obs} (whichever is smaller) and multiply by 2.

This is called a **permutation test** based on T. It is an example of a **conditional test**, because we are looking at the conditional distribution of the data under H_0 given the set (but not the ordering) of their values.

The utility of this idea is that it may be applied to test statistics T where we do not understand its (unconditional) distribution under H_0, and where this distribution may vary for different PDFs $f = g$. Consider the following example:

Example 9.3. Let $X_1, \ldots, X_n \in \mathcal{X}$ and $Y_1, \ldots, Y_m \in \mathcal{X}$ be two random samples of "objects" (e.g. images, websites, documents) represented in some data space $\mathcal{X}$. Suppose we have a function $d(x, y)$ that measures a "distance" between any two objects $x, y \in \mathcal{X}$.

To test whether $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ appear to come from the same distribution, the following might be a reasonable test statistic:

$$T_1 = \frac{2}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} d(X_i, Y_j) - \frac{1}{\binom{n}{2}} \sum_{1 \leq i < i' \leq n} d(X_i, X_{i'}) - \frac{1}{\binom{m}{2}} \sum_{1 \leq j < j' \leq m} d(Y_j, Y_{j'}).$$

In words, T_1 is twice the average distance between an object in sample 1 and an object in sample 2, minus the average distance between two objects in sample 1 and minus the average distance between two objects in sample 2. So T_1 measures whether, on average, objects from the same sample are more similar to each other than objects from different samples.

Or we might consider a "nearest-neighbors" statistic: For each of the $m + n$ data values, look at the k other data values closest to it (as measured by the distance d) and count how many of these come from the same sample as itself. Let T_2 be the average of this count across all $m + n$ data points. So T_2 measures whether the k closest other objects tend to come from the same sample.

The distributions of T_1 and T_2 under H_0 may be difficult to understand theoretically and may depend on the unknown common distribution of $X_1, \ldots, X_n, Y_1, \ldots, Y_m$, but we can still carry out a permutation test based on T_1 or on T_2.

A similar idea may be applied in the one-sample setting for testing the null hypothesis

$$H_0 : X_1, \ldots, X_n \overset{IID}{\sim} f, \text{ for some PDF } f \text{ symmetric about } 0$$

based on the symmetry underlying the Wilcoxon signed-rank test. You will explore this in Homework 4.