Python `perm_stat` Library

Sean P. Sullivan
Department of Economics & School of Law
University of Virginia

Michael B. Sullivan
Department of Electrical and Computer Engineering
University of Texas at Austin

January 17, 2011

1 Introduction

This is a small library designed to run general permutation tests in Python. Identity, signed-rank, and rank-order transformations of the sample observations are currently supported, and other tests may be defined by adding appropriate functions (e.g. trimmed mean) to the `transformations.py` source file [3]. In every case, the null distribution is constructed by exhaustive enumeration under standard permutation test assumptions.

2 Syntax

Syntax for the `perm_stat` library is simple: one or two arguments are passed to named functions. Arguments may be either list or tuple objects of appropriate shape. Examples are included in `example_use.py` and in the following code snippet:

```
from perm_stat import *

x = [-2,1,5]
y = [-1,4,5,8]
print fisher_perm_test(x)
print wilcoxon_perm_test(x)
print pitman_perm_test(x,y)
print wmw_perm_test(x,y)
```

Functions return a tuple of the form (lower-tail p-value, upper-tail p-value, two-tail p-value). The meaning of each p-value depends on the test conducted, as explained below.

3 One-Sample and Paired-Sample Tests

Consider the random sample $X_1, \ldots, X_n \overset{\text{iid}}{\sim} F_X$. The sample may be generated directly in a one-sample context, or may be defined as the vector-difference of paired samples such that $X_i = A_i - B_i$ for all $i = 1, \ldots, n$ with $A_1, \ldots, A_n \overset{\text{iid}}{\sim} F_A$ and $B_1, \ldots, B_n \overset{\text{iid}}{\sim} F_B$. In either case, the null hypothesis, H_0, is that F_X is symmetric about $\mu_0 \equiv 0$. Note the following:

1. In a one-sample setting, the null hypothesis places a symmetry assumption on the shape of the underlying distribution.
2. Under the null, $\mu_0 \equiv 0$ is identically the mean and median of F_X.
3. The null of symmetry around any $\mu_0 \neq 0$ may be incorporated by appropriate transformation of the observed sample(s).

Alternative hypotheses concern the measure of centrality. Three alternative hypotheses are of interest:

1. $\mu < 0$
2. $\mu > 0$
3. $\mu \neq 0$

3.1 Fisher's Permutation Test

Fisher's permutation test applies the identity transformation to the sample data [1]. The test statistic is as follows:

$$T = \sum_{i=1}^{n} X_i \tag{1}$$

Under the null hypothesis, every observation in the sample is equally likely to have a positive or negative sign. The null distribution of T can thus be computed by calculating the set, $\mathcal{T}_0$, of test statistics under each of the 2^n ways to permute the signs of the observed sample.

Let T^* be the value of the test statistic for the observed sample. P-values corresponding to each of the alternative hypotheses are calculated as the probability of observing a value of T equal to or more-extreme than the observed value, T^*:

1. $\mu < 0$: p-value $= P(T \leq T^* | H_0) = \#\{T \in \mathcal{T}_0 : T \leq T^*\}/2^n$
2. $\mu > 0$: p-value $= P(T \geq T^* | H_0) = \#\{T \in \mathcal{T}_0 : T \geq T^*\}/2^n$
3. $\mu \neq 0$ p-value $= P(|T| \geq |T^*| | H_0) = \#\{T \in \mathcal{T}_0 : |T| \geq |T^*|\}/2^n$

3.1.1 Comments

Note that T is directly proportional to the arithmetic mean of the sample:

$$T = \sum_{i=1}^{n} X_i \propto \frac{1}{n} \sum_{i=1}^{n} X_i = \overline{X}. \tag{2}$$

This makes Fisher's permutation test the permutation version of Student's t test. Fisher's permutation test uses more information from the sample than does Wilcoxon's signed-rank test, and may afford greater power when the underlying distribution is not too heavy-tailed.

3.2 Wilcoxon's Signed-Rank Test

Wilcoxon's signed-rank test operates on the signed-ranks of the sample observations [5]. Let the rank function $r : \mathbb{R} \to \mathbb{N}$ assign natural numbers $1, \ldots, n$ to observations in a sample in order of increasing absolute value,

$$r(X_i) = \sum_{j=1}^{n} I\{|X_i| \geq |X_j|\}, \tag{3}$$

where I is the indicator function valued 1 if the logical is true and 0 otherwise. Assume there are no ties in the sample so that every observation is assigned a unique rank. The signed-rank function $z : \mathbb{R} \to \mathbb{Z}$ is then defined as follows:

$$z(X_i) = \text{sgn}(X_i) r(X_i). \tag{4}$$

The signed-rank test operates on the transformed sample $Z_1, \ldots, Z_n$ defined as $Z_i = z(X_i)$ for $i = 1, \ldots, n$.

The signed-rank test statistic is as follows:

$$T = \sum_{i=1}^{n} Z_i. \tag{5}$$

Under the null hypothesis, every observation in the sample is equally likely to have a positive or negative sign. The null distribution of T can thus be computed from calculating the set, $\mathcal{T}_0$, of test statistics under each of the 2^n ways to permute the signs of the observed sample.

Let T^* be the value of the test statistic for the observed sample. P-values corresponding to each of the alternative hypotheses are calculated as the probability of observing a value of T equal to or more-extreme than the observed value, T^*:

1. $\mu < 0$: p-value $= P(T \leq T^* | H_0) = \#\{T \in \mathcal{T}_0 : T \leq T^*\}/2^n$
2. $\mu > 0$: p-value $= P(T \geq T^* | H_0) = \#\{T \in \mathcal{T}_0 : T \geq T^*\}/2^n$
3. $\mu \neq 0$ p-value $= P(|T| \geq |T^*| | H_0) = \#\{T \in \mathcal{T}_0 : |T| \geq |T^*|\}/2^n$

3.2.1 Ties

The rank function is ill-defined when there are ties in the observed sample: this motivates the assumption that there are no ties in the provided data. In theory, a sufficient condition for eliminating ties is that F_X be a continuous function. (This makes the collection of two observations with equal values a measure-zero event.) In practice, however, design and measurement limitation may make tied values unavoidable even when F_X is continuous.

When a recursive or table definition of the null distribution is used to compute p-values under the Wilcoxon signed-rank test, ties can be problematic. Techniques for eliminating ties can make the test either overly conservative, or anti-conservative, in small samples [3].

As programmed in the `perm_stat` library, however, ties are of modest concern to the signed-rank test. The library adopts the midrank convention of assigning a rank to tied observations equal to the arithmetic average of the ranks that would otherwise be assigned if tied values were differentiated by an arbitrarily small amount. Because the library conducts a pure permutation test, constructing the null distribution from the provided sample, the presence of ties creates no special problem in comparing the observed test statistic to the null distribution.

3.2.2 Comments

An alternative definition of the test statistic is as the sum of positive ranks:

$$T' = \sum_{i=1}^{n} Z_i I\{Z_i > 0\}. \tag{6}$$

Asymptotic theory provides a normal approximation to the null distribution of T' [3].

4 Two-Sample Tests

Consider the random samples $X_1, \ldots, X_n \overset{\text{iid}}{\sim} F_X$ and $Y_1, \ldots, Y_m \overset{\text{iid}}{\sim} F_Y$. The null hypothesis, H_0, is that $F_X = F_Y$, such that observations in each sample may be considered draws from a common distribution. Note the following:

1. Unlike the one sample case, there is no distributional requirement of symmetry.

2. Motivation for the null hypothesis is an experiment wherein subjects in group X act as a control and a treatment is applied to subjects in group Y; if subjects from a common pool are randomly assigned to either group X or Y, then the null hypothesis corresponds to the hypothesis of no treatment effect.

3. The null of dis-equal central tendency may be incorporated by assuming, e.g., a shift model of the form $F_X(x) = F_Y(x - \theta)$ for all x and some $\theta \neq 0$. Without loss of generality, this model can be converted to one of equal central tendency by an appropriate transformation to the observed samples.

Alternative hypotheses concern the difference in centrality between distributions. Let μ_X and μ_Y be measures of centrality for X and Y respectively. The definition of μ_X and μ_Y will depend on the specifics of a given application, but interpretation in terms of stochastic dominance is quite general. Three alternative hypotheses are of interest:

1. $\mu_X < \mu_Y$
2. $\mu_X > \mu_Y$
3. $\mu_X \neq \mu_Y$

4.1 Pitman's Permutation Test

Pitman's permutation test applies the identity transformation to the sample data [4]. The minimum sufficient test statistic is the sum of X observations:

$$T = \sum_{i=1}^{n} X_i \tag{7}$$

Under the null hypothesis, every observation in the combined sample is equally likely to have appeared in either the X or Y sample. The null distribution of T can thus be computed by calculating the set, $\mathcal{T}_0$, of test statistics under each of the $\binom{m+n}{n}$ ways to form an X sample of length n with combinations of observations from the combined observed samples.

Let T^* be the value of the test statistic for the observed sample. P-values corresponding to each of the alternative hypotheses are calculated as the probability of observing a value of T equal to or more-extreme than the observed value, T^*:

1. $\mu_X < \mu_Y$: p-value $= P(T \leq T^*|H_0) = \#\{T \in \mathcal{T}_0 : T \leq T^*\}/\binom{m+n}{n}$
2. $\mu_X > \mu_Y$: p-value $= P(T \geq T^*|H_0) = \#\{T \in \mathcal{T}_0 : T \geq T^*\}/\binom{m+n}{n}$
3. $\mu_X \neq \mu_Y$: p-value $= P(|T| \geq |T^*||H_0) = \#\{T \in \mathcal{T}_0 : |T| \geq |T^*|\}/\binom{m+n}{n}$

4.1.1 Comments

An alternative definition of the test statistic is as a difference of arithmetic means:

$$T' = \frac{1}{n}\sum_{i=1}^{n} X_i - \frac{1}{m}\sum_{i=1}^{m} Y_i = \overline{X} - \overline{Y}. \tag{8}$$

This makes Pitman's permutation test the permutation version of the two-sample t test. Pitman's permutation test uses more information from the sample than does the Wilcoxon-Mann-Whitney rank-order test test, and may afford greater power when the underlying distribution is not too heavy-tailed.

4.2 Wilcoxon-Mann-Whitney Rank-Order Test

The Wilcoxon-Mann-Whitney rank-order test operates on the ordinal ranks of the sample observations [5, 2]. Let the rank function $r : \mathbb{R} \to \mathbb{N}$ assign natural numbers $1, \ldots, n+m$ to observations in the combined sample in order of increasing absolute value:

$$r(W_i) = \sum_{j=1}^{n+m} I\{|W_i| \geq |W_j|\} \quad \text{for} \quad W = [X_1, \ldots, X_n, Y_1, \ldots, Y_M]. \tag{9}$$

Assume there are no ties in the sample so that every observation is assigned a unique rank. The rank-order test operates on the transformed X sample $R_1, \ldots, R_n$ defined as $R_i = r(X_i)$ for $i = 1, \ldots, n$.

The rank-order test statistic is as follows:

$$T = \sum_{i=1}^{n} R_i. \tag{10}$$

Under the null hypothesis, every observation in the combined sample is equally likely to have appeared in either the X or Y sample. The null distribution of T can thus be computed by calculating the set, $\mathcal{T}_0$, of test statistics under each of the $\binom{m+n}{n}$ ways to form an X sample of length n with combinations of observations from the combined observed samples.

Let T^* be the value of the test statistic for the observed sample. P-values corresponding to each of the alternative hypotheses are calculated as the probability of observing a value of T equal to or more-extreme than the observed value, T^*:

1. $\mu_X < \mu_Y$: p-value $= P(T \leq T^* | H_0) = \#\{T \in \mathcal{T}_0 : T \leq T^*\} / \binom{m+n}{n}$
2. $\mu_X > \mu_Y$: p-value $= P(T \geq T^* | H_0) = \#\{T \in \mathcal{T}_0 : T \geq T^*\} / \binom{m+n}{n}$
3. $\mu_X \neq \mu_Y$: p-value $= P(|T| \geq |T^*| | H_0) = \#\{T \in \mathcal{T}_0 : |T| \geq |T^*|\} / \binom{m+n}{n}$

4.2.1 Comments

Handling of ties is analogous that of Wilcoxon's signed-rank test: the discussion and comments of section 3.2.1 apply. The Wilcoxon-Mann-Whitney test has high asymptotic efficiency, particularly for samples from heavy-tailed distributions [3].

References

[1] Ronald A. Fisher. *The Design of Experiments*. Oliver & Boyd, Edinburgh, 1935.

[2] H. B. Mann and D. R. Whitney. On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, 18(1):50–60, 1947.

[3] Rupert G. Miller. *Beyond Anova: Basics of Applied Statistics*. Chapman & Hall/CRC, Boca Raton, 1997.

[4] E. J. G. Pitman. Significance Tests Which May Be Applied To Samples From Any Populations. *Supplement to the Journal of the Royal Statistical Society*, 4(1):119–130, 1937.

[5] Frank Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83, 1945.