Asymptotically Valid and Exact Permutation Tests Based on Two-Sample U-statistics: Formulae

July 26, 2018

1 Exact and Asymptotically Robust Permutation Tests: the two-sample case

Assume $X_1, \ldots, X_m$ are i.i.d. according to a probability distribution P, and independently $Y_1, \ldots, Y_n$ are i.i.d. Q. Let $N = n + m$ and write

$$Z = (Z_1, \ldots, Z_N) = (X_1, \ldots, X_m, Y_1, \ldots, Y_n)$$

Assume that $\lambda_m = m/N$ is such that $\lambda_m \to \lambda \in (0, 1)$ with $\lambda_m - \lambda = \mathcal{O}(N^{-1/2})$. Sample analogues are denoted with either bars or circumflexes, depending on the context.

1.1 Parameter comparisons

In this section we consider the general problem of inference from the permutation distribution when comparing parameters from two populations. The test statistics will be based on *the difference of estimators* that are asymotitically linear. We will consider three cases: differences in mean, medians, and variances.

Difference of means. Here, the null hypothesis is of the form $H_0 : \mu(P) - \mu(Q) = 0$, and the corresponding test statistic is given by

$$T_{m,n} = \frac{N^{1/2}\left(\bar{X}_m - \bar{Y}_n\right)}{\sqrt{\frac{N}{m}\sigma_m^2(X_1, \ldots, X_m) + \frac{N}{n}\sigma_n^2(Y_1, \ldots, Y_n)}} \tag{1}$$

where $\bar{X}_m$ and $\bar{Y}_n$ are the sample means from population P and population Q, respectively, and $\sigma_m^2(X_1, \ldots, X_m)$ is a consistent estimator of $\sigma^2(P)$ when $X_1, \ldots, X_m$ are i.i.d. from P. Assume consitency also under Q.

Difference of medians. Let F and G be the CDFs corresponding to P and Q, and denote $\theta(F)$ the median of F i.e. $\theta(F) = \inf\{x : F(x) \geq 1/2\}$. Assume that F is continuously differentiable at $\theta(P)$ with derivative F' (and the same with F replaced by G). Here, the null hypothesis is of the form $H_0 : \theta(P) - \theta(Q) = 0$, and the corresponding test statistic is given by

$$T_{m,n} = \frac{N^{1/2}\left(\theta(\hat{P}_m) - \theta(\hat{Q})\right)}{\hat{v}_{m,n}} \tag{2}$$

where $\hat{\upsilon}_{m,n}$ is a consistent estimator of $\upsilon(P,Q)$:

$$\upsilon(P,Q) = \frac{1}{\lambda}\frac{1}{4(F'(\theta))^2} + \frac{1}{1-\lambda}\frac{1}{4(G'(\theta))^2}$$

Choices of $\hat{\upsilon}_{m,n}$ may include the kernel estimator of Devroye and Wagner (1980), the bootstrap estimator of Efron (1992), or the smoothed bootstrap Hall et al. (1989) to list a few. For further details, see Chung and Romano (2013).

Difference of variances. Here, the null hypothesis is of the form $H_0 : \sigma^2(P) - \sigma^2(Q) = 0$, and the corresponding test statistic is given by

$$T_{m,n} = \frac{N^{1/2}\left(\hat{\sigma}^2_m(X_1, \ldots, X_{,}) - \hat{\sigma}^2_n(Y_1, \ldots, Y_n)\right)}{\sqrt{\frac{N}{m}\left(\hat{\mu}_{4,x} - \frac{(m-3)}{(m-1)}(\hat{\sigma}^2_m)^2\right) + \frac{N}{n}\left(\hat{\mu}_{4,y} - \frac{(n-3)}{(n-1)}(\hat{\sigma}^2_y)^2\right)}} \tag{3}$$

where $\hat{\mu}_{4,m}$ the sample analog of $\mathbb{E}(X-\mu)^4$ based on an iid sample $X_1, \ldots, X_m$ from P. Similarly for $\hat{\mu}_{4,n}$.

1.2 The parameter as a function of the joint distribution

In this section, the parameter of interest is a function of the joint distribution i.e. $\theta(P,Q)$ and not just the difference $\theta(P) - \theta(Q)$. For a thorough dicussion, we refer the reader to Chung and Romano (2016). We will consider four cases:

Lehmann (1951) two-sample U statistics. Consider testing $H_0 : P = Q$, or the more general hypothesis that P and Q only differ in location[1] against the alternative that the Y's are more spread out than the X's. Then the null hypothesis is of the form $H_0 : \mathbb{P}(|Y - Y'| > |X - X'|) = 1/2$, and the corresponding test statistic is given by

$$T_{m,n} = \frac{\frac{1}{(mn)^2}\sum_{i=1}^m \sum_{j=1}^m \sum_{k=1}^n \sum_{l=1}^n \left(1_{\{|Y_l - Y_k| > |X_j - X_i|\}} - \frac{1}{2}\right)}{V_{m,n}} \tag{4}$$

where

$$V^2_{m,n} = 4\left[\frac{1}{m-1}\sum_{i=1}^{m-1}\left(\hat{\zeta}_x(X_i) - \frac{1}{m-1}\sum_{i=1}^{m-1}\hat{\zeta}_x(X_i)\right)^2 + \frac{m}{n}\frac{1}{n-1}\sum_{k=1}^{n-1}\left(\hat{\zeta}_y(Y_k) - \frac{1}{n-1}\sum_{k=1}^{n-1}\hat{\zeta}_y(Y_k)\right)^2\right]$$

and

$$\hat{\zeta}_x(X_i) = \frac{2}{(m-i)n(n-1)}\sum_{j=i+1}^{m}\sum_{k=1}^{n-1}\sum_{l=k+1}^{n} 1_{\{|Y_k - Y_l| > |X_i - X_j|\}}$$

$$\hat{\zeta}_y(Y_k) = \frac{2}{(n-k)m(m-1)}\sum_{i=1}^{m-1}\sum_{j=i+1}^{m}\sum_{l=k+1}^{n} 1_{\{|Y_k - Y_l| > |X_i - X_j|\}}$$

Two-sample Wilcoxon statistic. The null hypothesis is of the form $H_0 : \mathbb{P}(X \leq Y) = 1/2$, and the corresponding test statistic is given by

$$T_{m,n} = \frac{\frac{1}{mn}\sum_{i=1}^m \sum_{j=1}^n 1_{\{X_i \leq Y_j\}} - \frac{1}{2}}{\sqrt{\frac{1}{m}\hat{\xi}_x + \frac{1}{n}\hat{\xi}_y}} \tag{5}$$

[1] That is, $P(x) = Q(x+\tau)$ for some τ.

where

$$\hat{\xi}_x = \frac{1}{m-1}\sum_{i=1}^{m}\left(\frac{1}{n}\sum_{j=1}^{n}1_{\{Y_j \le X_i\}} - \frac{1}{m}\sum_{i=1}^{m}\left(\frac{1}{n}\sum_{j=1}^{n}1_{\{Y_j \le X_i\}}\right)\right)^2$$

and

$$\hat{\xi}_y = \frac{1}{n-1}\sum_{j=1}^{n}\left(\frac{1}{m}\sum_{i=1}^{m}1_{\{X_i \le Y_j\}} - \frac{1}{n}\sum_{j=1}^{n}\left(\frac{1}{m}\sum_{i=1}^{m}1_{\{X_i \le Y_j\}}\right)\right)^2$$

are themselves rank statistics.

Two-sample Wilcoxon statistic without continuity assumption. The null hypothesis is of the form $H_0 : \mathbb{P}(X \le Y) = \mathbb{P}(Y \le X)$, and the corresponding test statistic is given by

$$T_{m,n} = \frac{\frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}1_{\{X_i<Y_j\}} + \frac{1}{2}1_{\{X_i=Y_j\}} - \frac{1}{2}}{\sqrt{\frac{1}{m}\hat{\xi}_x + \frac{1}{n}\hat{\xi}_y}} \tag{6}$$

where

$$\hat{\xi}_x = \frac{1}{m-1}\sum_{i=1}^{m}\left(\hat{\zeta}_x(X_i) - \frac{1}{m}\sum_{i=1}^{m}\hat{\zeta}_x(X_i)\right)^2$$

and

$$\hat{\xi}_y = \frac{1}{n-1}\sum_{j=1}^{n}\left(\hat{\zeta}_y(Y_j) - \frac{1}{n}\sum_{j=1}^{n}\hat{\zeta}_y(Y_j)\right)^2$$

for

$$\begin{aligned}\hat{\zeta}_x(X_i) &\equiv \frac{1}{n}\sum_{j=1}^{n}1_{\{Y_j<X_i\}} + \frac{1}{2}1_{\{Y_j=X_i\}}\\ \hat{\zeta}_y(Y_j) &\equiv \frac{1}{m}\sum_{i=1}^{m}1_{\{X_i<Y_j\}} + \frac{1}{2}1_{\{X_i=Y_j\}}\end{aligned}$$

Hollander (1967) two-sample U statistics. The null hypothesis is of the form $H_0 : \mathbb{P}(X + X' < Y + Y') = 1/2$, and the corresponding test statistic is given by

$$T_{m,n} = \frac{\frac{1}{(mn)^2}\sum_{i=1}^{m}\sum_{j=1}^{m}\sum_{k=1}^{n}\sum_{l=1}^{n}\left(1_{\{X_i+X_j<Y_k+Y_l\}} - \frac{1}{2}\right)}{V_{m,n}} \tag{7}$$

where

$$V_{m,n}^2 = 4\left[\frac{1}{m-1}\sum_{i=1}^{m-1}\left(\hat{\zeta}_x(X_i) - \frac{1}{m-1}\sum_{i=1}^{m-1}\hat{\zeta}_x(X_i)\right)^2 + \frac{m}{n}\frac{1}{n-1}\sum_{k=1}^{n-1}\left(\hat{\zeta}_y(Y_k) - \frac{1}{n-1}\sum_{k=1}^{n-1}\hat{\zeta}_y(Y_k)\right)^2\right]$$

and

$$\begin{aligned}\hat{\zeta}_x(X_i) &= \frac{2}{(m-i)n(n-1)}\sum_{j=i+1}^{m}\sum_{k=1}^{n-1}\sum_{l=k+1}^{n}1_{\{Y_k+Y_l-X_j<X_i\}}\\ \hat{\zeta}_y(Y_k) &= \frac{2}{(n-k)m(m-1)}\sum_{i=1}^{m-1}\sum_{j=i+1}^{m}\sum_{l=k+1}^{n}1_{\{X_i+X_j-Y_l<Y_k\}}\end{aligned}$$

2 Exact and Asymptotically Robust Permutation Tests: the k-sample case

Assume we observe k independent samples, drawn from populations P_i, $i = 1, \dots, k$. For every i, we have a random sample of size n_i i.e. $X_{i,1}, \dots, X_{i,n_i} \sim P_i$. Denote $n = (n_1, \dots, n_k)$. Then our sample is given by

$$X = (X_{1,1}, \dots, X_{1,n_1}, X_{2,1}, \dots, X_{2,n_2}, \dots, X_{k,1}, \dots, X_{k,n_k})$$

The problem of interest is to test the null hypothesis

$$H_0 : \theta(P_1) = \cdots = \theta(P_k)$$

against the alternative

$$H_1 : \theta(P_i) \neq \theta(P_j) \quad \text{for some} \quad i, j$$

The test statistic is given by

$$T_n = \sum_{i=1}^{k} \frac{n_i}{\hat{\sigma}^2_{n,i}} \left[\hat{\theta}_{n,i} - \frac{\sum_{i=1}^{k} n_i \hat{\theta}_{n,i} / \hat{\sigma}^2_{n,i}}{\sum_{i=1}^{k} n_i / \hat{\sigma}^2_{n,i}} \right]^2 \tag{8}$$

where $\hat{\theta}_{n,i} = \hat{\theta}_{n,i}(X_{i,1}, \dots, X_{i,n_i})$ is an estimator of the real-valued parameter $\theta(P_i)$, and $\hat{\sigma}_{n,i} \equiv \hat{\sigma}_{n,i}(X_{i,1}, \dots, X_{i,n_i})$ is a consistent estimator of $\sigma(P_i)$. Again, we will consider three cases: equality of means, medians, and variances, respectively

Difference of means. Here, the null hypothesis is of the form $H_0 : \mu(P_1) = \cdots = \mu(P_k)$, and the corresponding test statistic is given by (8) with

$$\hat{\theta}_{n,i} = \bar{X}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} X_{i,j}$$

$$\hat{\sigma}_{n,i} = \frac{1}{n_i} \sum_{j=1}^{n_i} (X_{i,j} - \bar{X}_i)^2$$

Difference of medians. Let F_i be the CDF corresponding to P_i, and denote $\theta(P_i)$ the median of F_i i.e. $\theta(F_i) = \inf\{x : F_i(x) \geq 1/2\}$. Assume that F_i is continuously differentiable at $\theta(P_i)$ with derivative F_i'. Here, the null hypothesis is of the form $H_0 : \theta(P_1) = \cdots = \theta(P_k)$, and the corresponding test statistic is given by (8) with $\hat{\theta}_{n,i}$ the sample meadian and $\hat{\sigma}_{n,i}$ a consistent estimator of $\upsilon(P_i)$, the variance of the median based on the i-th sample. Once again, choices of $\hat{\sigma}_{n,i}$ may include the kernel estimator of Devroye and Wagner (1980), the bootstrap estimator of Efron (1992), or the smoothed bootstrap Hall et al. (1989) to list a few. For further details, see Chung and Romano (2013).

Difference of variances. Here, the null hypothesis is of the form $H_0 : \sigma^2(P_1) = \cdots = \sigma^2(P_k)$, and the corresponding test statistic is given by (8) with

$$\hat{\theta}_{n,i} = \frac{1}{n_i} \sum_{j=1}^{n_i} (X_{i,j} - \bar{X}_i)^2$$

$$\hat{\sigma}_{n,i} = \hat{\mu}_{4,i} - \frac{(n_i - 3)}{(n_i - 1)} (\hat{\theta}_{n,i})^2$$

where $\hat{\mu}_{4,i}$ the sample analog of $\mathbb{E}(X_{1,i} - \mu(P_i))^4$ based on an iid sample $X_{i,1}, \dots, X_{i,n_i}$ from P_i.

References

Chung, E. and Romano, J. P. (2013). Exact and asymptotically robust permutation tests. *The Annals of Statistics*, 41(2):484–507.

Chung, E. and Romano, J. P. (2016). Asymptotically valid and exact permutation tests based on two-sample u-statistics. *Journal of Statistical Planning and Inference*, 168:97–105.

Devroye, L. P. and Wagner, T. J. (1980). The strong uniform consistency of kernel density estimates. In *Multivariate Analysis V: Proceedings of the fifth International Symposium on Multivariate Analysis*, volume 5, pages 59–77.

Efron, B. (1992). Bootstrap methods: another look at the jackknife. In *Breakthroughs in statistics*, pages 569–593. Springer.

Hall, P., DiCiccio, T. J., and Romano, J. P. (1989). On smoothing and the bootstrap. *The Annals of Statistics*, pages 692–704.

Hollander, M. (1967). Asymptotic efficiency of two nonparametric competitors of wilcoxon's two sample test. *Journal of the American Statistical Association*, 62(319):939–949.

Lehmann, E. L. (1951). Consistency and unbiasedness of certain nonparametric tests. *The Annals of Mathematical Statistics*, pages 165–179.