Mathematics Involving Non-Parametric Statistics

Yuk Tung Liu

2016-05-04, updated on 2017-11-15

These short notes provide mathematical proofs of the following properties involving non-parametric statistics covered in Stat 200.

Wilcoxon Rank Sum: When there are no ties in both groups A and B, the expected value and variance of the rank sum for group A are

$$E(R_A) = \frac{n_A(N+1)}{2} \quad , \quad V(R_A) = \frac{n_A n_B(N+1)}{12}, \tag{0.1}$$

where n_A and n_B are the number of observations in group A and B, respectively. The total number of observations $N = n_A + n_B$.

Relationship Between Wilcoxon Rank Sum and U Statistic: The Wilcoxon Rank Sum for group A, R_A, is related to the U statistic for group A, U_A, by

$$R_A = U_A + \frac{n_A(n_A+1)}{2}. \tag{0.2}$$

Spearman's Rank-Order Correlation Coefficient: Suppose $x = (x_1, x_2, x_3, \cdots, x_n)$ and $y = (y_1, y_2, \cdots, y_n)$. If there are no ties in both x and y, and x and y are uncorrelated, the expected value and variance of Spearman's rank-order correlation coefficient r_s are

$$E(r_s) = 0 \quad , \quad V(r_s) = \frac{1}{\sqrt{n-1}}. \tag{0.3}$$

1 Wilcoxon Rank Sum

To calculate the mean and variance of the rank sum R_A, we need to calculate several quantities.

We first want to calculate the values of two series: $1 + 2 + 3 + \cdots + n$ and $1^2 + 2^2 + \cdots + n^2$. The first one is an arithmetic series, which can be computed as follows.

$$S_n = 1 + 2 + 3 + \cdots n$$

We can rewrite S_n as

$$S_n = n + (n-1) + (n-2) + \cdots + 1$$

Adding the two expressions gives

$$2S_n = \underbrace{(n+1) + (n+1) + \cdots + (n+1)}_{n \text{ times}} = n(n+1)$$

Hence,

$$1 + 2 + 3 + \cdots + n = \sum_{i=1}^{n} i = \frac{n(n+1)}{2}. \tag{1.1}$$

The second series is a discrete version of the integral

$$\int_1^n x^2 dx$$

To calculate the sum, we consider the integral

$$\int_{i-\frac{1}{2}}^{i+\frac{1}{2}} x^2 dx = \frac{1}{3}\left[\left(i+\frac{1}{2}\right)^3 - \left(i-\frac{1}{2}\right)^3\right]$$

Now,

$$\begin{aligned}
\frac{1}{3}\left[\left(i+\frac{1}{2}\right)^3 - \left(i-\frac{1}{2}\right)^3\right] &= \frac{1}{3}\left[\left(i^3+\frac{3}{2}i^2+\frac{3}{4}i+\frac{1}{8}\right) - \left(i^3-\frac{3}{2}i^2+\frac{3}{4}i-\frac{1}{8}\right)\right] \\
&= i^2 + \frac{1}{12}
\end{aligned}$$

$$\Rightarrow \quad i^2 = -\frac{1}{12} + \frac{1}{3}\left[\left(i+\frac{1}{2}\right)^3 - \left(i-\frac{1}{2}\right)^3\right]$$

$$\begin{aligned}
\Rightarrow \sum_{i=1}^{n} i^2 &= -\frac{n}{12} + \frac{1}{3}\sum_{i=1}^{n}\left[\left(i+\frac{1}{2}\right)^3 - \left(i-\frac{1}{2}\right)^3\right] \\
&= -\frac{n}{12} + \frac{1}{3}\sum_{i=1}^{n}\left(i+\frac{1}{2}\right)^3 - \frac{1}{3}\sum_{i=1}^{n}\left(i-\frac{1}{2}\right)^3 \\
&= -\frac{n}{12} + \frac{1}{3}\sum_{i=1}^{n}\left(i+\frac{1}{2}\right)^3 - \frac{1}{3}\left(\frac{1}{2}\right)^3 - \frac{1}{3}\sum_{i=2}^{n}\left(i-\frac{1}{2}\right)^3 \\
&= -\frac{n}{12} + \frac{1}{3}\sum_{i=1}^{n}\left(i+\frac{1}{2}\right)^3 - \frac{1}{24} - \frac{1}{3}\sum_{i=1}^{n-1}\left(i+\frac{1}{2}\right)^3 \\
&= -\frac{n}{12} + \frac{1}{3}\left(n+\frac{1}{2}\right)^3 - \frac{1}{24} \\
&= -\frac{n}{12} + \frac{1}{3}\left(n^3+\frac{3}{2}n^2+\frac{3}{4}n+\frac{1}{8}\right) - \frac{1}{24} \\
&= \frac{1}{3}n^3 + \frac{1}{2}n^2 + \frac{1}{6}n \\
&= \frac{n(2n^2+3n+1)}{6} \\
&= \frac{n(n+1)(2n+1)}{6} \qquad (1.2)
\end{aligned}$$

Let $x =$ permutation of $(1, 2, 3, \cdots, n)$. The expected value of x_i is

$$E(x_i) = \frac{1}{n}\sum_{i=1}^{n} i = \frac{1}{n}\frac{n(n+1)}{2} = \frac{n+1}{2} \qquad (1.3)$$

The variance of x_i is

$$V(x_i) = E(x_i^2) - E^2(x_i)$$

$$
\begin{aligned}
&= \frac{1}{n}\sum_{i=1}^{n} i^2 - \left(\frac{n+1}{2}\right)^2 \\
&= \frac{(n+1)(2n+1)}{6} - \frac{(n+1)^2}{4} \\
&= \frac{n^2-1}{12},
\end{aligned} \tag{1.4}
$$

where we have used (1.2) in the third line.

The covariance $cov(x_i, x_j)$ for $i \neq j$ can be computed as follows.

$$
\begin{aligned}
cov(x_i, x_j) &= E(x_i x_j) - E(x_i)E(x_j) \\
&= \frac{1}{n(n-1)}\sum_{i\neq j} ij - \left(\frac{n+1}{2}\right)^2 \\
&= \frac{1}{n(n-1)}\sum_{i,j} ij - \frac{1}{n(n-1)}\sum_{i=1}^{n} i^2 - \frac{(n+1)^2}{4} \\
&= \frac{1}{n(n-1)}\left(\sum_{i=1}^{n} i\right)^2 - \frac{1}{n(n-1)}\frac{n(n+1)(2n+1)}{6} - \frac{(n+1)^2}{4} \\
&= \frac{1}{n(n-1)}\frac{n^2(n+1)^2}{4} - \frac{(n+1)(2n+1)}{6(n-1)} - \frac{(n+1)^2}{4} \\
&= -\frac{n+1}{12}
\end{aligned} \tag{1.5}
$$

after straightforward algebra.

We are now ready to calculate the expected value and variance of Wilcoxon rank sum. Let $x_1, x_2, \cdots, x_{n_A}$ be the ranks of the elements of group A, and $x_{n_A+1}, x_{n_A+2}, \cdots, x_N$ be the ranks of the elements of group B. In the absence of ties, $x_1, x_2, \cdots, x_N$ is a permutation of $1, 2, 3, \cdots, N$. The rank sum for group A is

$$
R_A = \sum_{i=1}^{n_A} x_i
$$

$$
\Rightarrow \quad E(R_A) = \sum_{i=1}^{n_A} E(x_i) = \sum_{i=1}^{n_A} \frac{N+1}{2} = \frac{n_A(N+1)}{2} \quad \blacksquare \tag{1.6}
$$

The variance can be computed as follows.

$$
\begin{aligned}
V(R_A) &= V\left(\sum_{i=1}^{n_A} x_i\right) \\
&= \sum_{i=1}^{n_A} V(x_i) + \sum_{i\neq j} cov(x_i, x_j)
\end{aligned}
$$

Using (1.4) and (1.5), we have

$$
V(R_A) = \frac{n_A(N+1)}{12} - \frac{N+1}{12} \underbrace{\sum_{i\neq j} 1}_{=n_A(n_A-1)}
$$

$$= \frac{n_A(N+1)}{12} - \frac{n_A(n_A-1)(N+1)}{12}$$
$$= \frac{n_A(N+1)(N-n_A)}{12}$$
$$= \frac{n_A n_B(N+1)}{12} \quad \blacksquare \tag{1.7}$$

2 Wilcoxon Rank Sum and U Statistic

Let R_A be the Wilcoxon Rank Sum for group A and U_A be the U statistic for group A. Then

$$R_A = U_A + \frac{n_A(n_A+1)}{2}. \tag{2.1}$$

The proof is very easy if there are no ties. It requires more algebra when there are ties. In the following, we first consider the case when there are no ties (not in group A at least). Then we tackle the general case when there are ties.

Case 1: No ties.

Sort the numbers in group A in ascending order: $A_1, A_2, \cdots, A_{n_A}$. Here A_i $(i = 1, 2, \cdots, n_A)$ denote the ith number in group A when sorted in ascending order. Let $B_1, B_2, \cdots, B_{n_B}$ be the numbers in group B sorted in ascending order. By no ties we mean that $A_i \neq A_j$ for all $i \neq j$ $[i, j \in (1, n_A)]$ **and** $A_i \neq B_j$ for all $i \in (1, n_A)$ and $j \in (1, n_B)$. In other words, all numbers in group A $A_1, A_2, \cdots, A_{n_A}$ are unequal **and** none of the numbers in group B $B_1, B_2, \cdots, B_{n_B}$ is equal to any number in group A (However, there could be numbers in group B that are equal).

Let r_i $(i = 1, 2, \cdots, n_A)$ be the rank of A_i. Since there are no other numbers equal to A_i, r_i is equal to one plus the number of numbers smaller than A_i. By definition, there are $i-1$ numbers in group A smaller than A_i and u_i numbers in group B smaller than A_i, where u_i is the U count. Hence we have

$$r_i = i + u_i. \tag{2.2}$$

Summing i from 1 to n_A gives

$$R_A = \sum_{i=1}^{n_A} r_i = \sum_{i=1}^{n_A} i + \sum_{i=1}^{n_A} u_i = \frac{n_A(n_A+1)}{2} + U_A. \quad \blacksquare \tag{2.3}$$

Case 2: There are ties, meaning that at least two numbers in $A_1, A_2, \cdots, A_{n_A}, B_1, B_2, \cdots, B_{n_B}$ are equal.

There can also be multiple sets of ties. That is, at least two numbers are equal to a particular number, say t_1; at least two numbers are equal to another particular number, say t_2, ...etc. We can look at one particular set of ties, say t_q. Suppose there are p numbers of t_q occurring in groups A and B. Of these p ties, k of them are in group A and $p-k$ of them are in group B, where k can be any integer between 0 and p. If $k = 0$ (i.e. all the p ties are in group B), we can disregard it because it has no effect on the rank sum R_A. So we focus on the case when k is between 1 and p. Suppose the k ties in group A are $A_j, A_{j+1}, \cdots A_{j+k-1}$. These k ties have the same rank $r_j = r_{j+1} = \cdots = r_{j+k-1} \equiv r_{t_q}$ and the same U count $u_j = u_{j+1} = \cdots = u_{j+k-1} \equiv u_{t_q}$. From the definition of the rank in the case of ties, r_{t_q} is equal to one plus the number of numbers smaller than t_q plus $(p-1)/2$. Now there are $j-1$ numbers smaller than t_q in group A. Let u'_j be the number of numbers smaller than t_q in group B. Then

$$r_j = r_{j+1} = \cdots = r_{j+k-1} = r_{t_q} = j + u'_j + \frac{p-1}{2}. \tag{2.4}$$

Recall that when there are ties, 1/2 is contributed to the U count for the numbers in group B that are equal to t_q. Since there are $p-k$ numbers in group B equal to t_q, the U count is given by

$$u_j = u_{j+1} = \cdots = u_{j+k-1} = u_{t_q} = u'_j + \frac{p-k}{2}. \tag{2.5}$$

Combining equations (2.4) and (2.5) gives

$$r_j = r_{j+1} = \cdots = r_{j+k-1} = j + u_j - \frac{p-k}{2} + \frac{p-1}{2} = j + u_j + \frac{k-1}{2}. \tag{2.6}$$

It follows that

$$\sum_{i=j}^{j+k-1} r_i = r_j + r_{j+1} + \cdots + r_{j+k-1} = kj + \frac{k(k-1)}{2} + ku_j. \tag{2.7}$$

Since

$$\sum_{i=j}^{j+k-1} i = j + (j+1) + (j+2) + \cdots + (j+k-1) = kj + (1 + 2 + \cdots k - 1) = kj + \frac{k(k-1)}{2}, \tag{2.8}$$

we can write

$$\sum_{i=j}^{j+k-1} r_i = \sum_{i=j}^{j+k-1} i + \sum_{i=j}^{j+k-1} u_i = \sum_{i=j}^{j+k-1} (i + u_i). \tag{2.9}$$

This is the equation for a particular set of ties. The corresponding equations for the other sets of ties are equal to the equation above by changing j and k appropriate to the sets of ties. Summing over all ties appearing in group A, we have

$$\sum_{i \in \text{ties}} r_i = \sum_{i \in \text{ties}} (i + u_i). \tag{2.10}$$

When A_i does not belong to any set of ties, equation (2.2) holds. Summing over numbers that don't belong to any set of ties, we have

$$\sum_{i \in \text{no ties}} r_i = \sum_{i \in \text{no ties}} (i + u_i). \tag{2.11}$$

Combining these two equations yield

$$\sum_{i=1}^{n_A} r_i = \sum_{i=1}^{n_A} (i + u_i) = \sum_{i=1}^{n_A} i + \sum_{i=1}^{n_A} u_i = \frac{n_A(n_A+1)}{2} + U_A. \quad \blacksquare \tag{2.12}$$

3 Spearman's Rank-Order Correlation Coefficient

Let $x = (x_1, x_2, \cdots, x_n)$ and $y = (y_1, y_2, \cdots, y_n)$. If there are no ties in both x and y, we can replace x by a permutation of $(1, 2, 3, \cdots, n)$ and replace y by another permutation of $(1, 2, 3, \cdots, n)$. If x and y are uncorrelated, we have $E(f(x_i)g(y_j)) = E(f(x_i))E(g(y_j))$, where f and g are arbitrary functions.

The Spearman's rank-order correlation coefficient is

$$r_s = \frac{1}{n} \sum_{i=1}^{n} Z_{x_i} Z_{y_i}, \tag{3.1}$$

where

$$Z_{x_i} = \frac{x_i - \bar{x}}{SD_x} \quad , \quad Z_{y_i} = \frac{y_i - \bar{y}}{SD_y}$$

are the Z-scores associated with x and y. The mean and standard deviation of x are

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \quad , \quad SD_x = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (x - x_i)^2} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} x_i^2 - \bar{x}^2} \tag{3.2}$$

Since x is a permutation of $(1, 2, 3, \cdots, n)$,

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} i = \frac{n+1}{2} \quad , \quad SD_x = \sqrt{\frac{1}{n}\sum_{i=1}^{n} i^2 - \frac{(n+1)^2}{4}} = \sqrt{\frac{n^2-1}{12}} \tag{3.3}$$

using the results (1.1) and (1.2). Similarly,

$$\bar{y} = \bar{x} = \frac{n+1}{2} \quad , \quad SD_y = SD_x = \sqrt{\frac{n^2-1}{12}}. \tag{3.4}$$

By construction, the expected value and variance of the Z-score are $E(Z_{x_i}) = E(Z_{y_i}) = 0$ and $V(Z_{x_i}) = V(Z_{y_i}) = 1$.

The expected value of r_s is

$$E(r_s) = E\left(\frac{1}{n}\sum_{i=1}^{n} Z_{x_i} Z_{y_i}\right) = \frac{1}{n}\sum_{i=1}^{n} E(Z_{x_i} Z_{y_i}) = \frac{1}{n}\sum_{i=1}^{n} E(Z_{x_i})E(Z_{y_i}) = 0 \quad \blacksquare$$

The variance can be calculated as follows.

$$\begin{aligned} V(r_s) &= \frac{1}{n^2} V\left(\sum_{i=1}^{n} Z_{x_i} Z_{y_i}\right) \\ &= \frac{1}{n^2}\left[\sum_{i=1}^{n} V(Z_{x_i} Z_{y_i}) + \sum_{i \neq j} cov(Z_{x_i} Z_{y_i}, Z_{x_j} Z_{y_j})\right] \end{aligned} \tag{3.5}$$

$$\begin{aligned} V(Z_{x_i} Z_{y_i}) &= E(Z_{x_i}^2 Z_{y_i}^2) - E^2(Z_{x_i} Z_{y_i}) \\ &= E(Z_{x_i}^2)E(Z_{y_i}^2) - [E(Z_{x_i})E(Z_{y_i})]^2 \\ &= E(Z_{x_i}^2)E(Z_{y_i}^2) \end{aligned}$$

It follows from $V(Z_{x_i}) = V(Z_{y_i}) = 1$ and $E(Z_{x_i}) = E(Z_{y_i}) = 0$ that $V(Z_{x_i}) = E(Z_{x_i}^2) - E^2(Z_{x_i}) = E(Z_{x_i}^2)$. So $E(Z_{x_i}^2) = E(Z_{y_i}^2) = 1$ and $V(Z_{x_i} Z_{y_i}) = 1$. Thus,

$$V(r_s) = \frac{1}{n^2}\sum_{i=1}^{n} 1 + \frac{1}{n^2}\sum_{i \neq j} cov(Z_{x_i} Z_{y_i}, Z_{x_j} Z_{y_j}) = \frac{1}{n} + \frac{1}{n^2}\sum_{i \neq j} cov(Z_{x_i} Z_{y_i}, Z_{x_j} Z_{y_j}) \tag{3.6}$$

$$\begin{aligned} cov(Z_{x_i} Z_{y_i}, Z_{x_j} Z_{y_j}) &= E(Z_{x_i} Z_{x_j} Z_{y_i} Z_{y_j}) - E(Z_{x_i} Z_{y_i})E(Z_{x_j} Z_{y_j}) \\ &= E(Z_{x_i} Z_{x_j})E(Z_{y_i} Z_{y_j}) - E(Z_{x_i})E(Z_{y_i})E(Z_{x_j})E(Z_{y_j}) \\ &= E(Z_{x_i} Z_{x_j})E(Z_{y_i} Z_{y_j}) \\ &= cov(Z_{x_i}, Z_{x_j})cov(Z_{y_i}, Z_{y_j}) \\ &= [cov(Z_{x_i}, Z_{x_j})]^2 \end{aligned} \tag{3.7}$$

since $cov(Z_{x_i}, Z_{x_j}) = cov(Z_{y_i}, Z_{y_j})$.

$$cov(Z_{x_i}, Z_{x_j}) = cov\left(\frac{x_i - \bar{x}}{SD_x}, \frac{x_j - \bar{x}}{SD_x}\right)$$

$$
\begin{aligned}
&= \frac{1}{SD_x^2} cov(x_i, x_j) \\
&= -\frac{12}{n^2-1}\frac{n+1}{12} \\
&= -\frac{1}{n-1} \qquad (3.8)
\end{aligned}
$$

Hence,

$$
\begin{aligned}
V(r_s) &= \frac{1}{n} + \frac{1}{n^2}\frac{1}{(n-1)^2}\sum_{i \neq j} 1 \\
&= \frac{1}{n} + \frac{1}{n^2}\frac{1}{(n-1)^2} n(n-1) \\
&= \frac{1}{n} + \frac{1}{n(n-1)} \\
&= \frac{1}{n}\left(1 + \frac{1}{n-1}\right) \\
&= \frac{1}{n}\frac{n}{n-1} \\
&= \frac{1}{n-1} \quad \blacksquare \qquad (3.9)
\end{aligned}
$$