3. Two-sample Location Problem

Two-sample data

- Two-sample data consist of observations on two sets of random variables, denoted by $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$, drawn from two populations.
- The total number of observations is $N = m + n$.
- The data are not paired: $m \neq n$ in general; even if $m = n$, $Y_i - X_i$ does not represent the difference from the same subject, since X_i and Y_i come from two independent subjects, not the same one as in paired data.
- A typical example is to compare two treatments on different patients, such as a new medicine on m randomly selected patients (referred to as the *treatment group*) and placebo on other n patients (*control group*).
- The problem of interest is whether there is a significant difference between the distributions of $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$, and what is the difference.
- The basic assumptions on $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ are listed below.

Assumption 3.1 (basic assumptions)

(i) $X_1,\ldots,X_m$ are independent and identically distributed (i.i.d.) with common cdf F; $Y_1,\ldots,Y_n$ are i.i.d. with common cdf G.

(ii) $X_1,\ldots,X_m$ and $Y_1,\ldots,Y_n$ are mutually independent.

(iii) $X_1,\ldots,X_m$ and $Y_1,\ldots,Y_n$ are continuous random variables; or equivalently, F and G are continuous distributions.

Problem formulation

- Let X and Y denote the representative random variables of $X_1,\ldots,X_m$ and $Y_1,\ldots,Y_n$, respectively, with $X \sim F$ and $Y \sim G$.

- To assess the difference between F and G, the two-sample location problem considers the following *location-shift* model:

$$G(t) = F(t-\Delta) \quad \text{for all } t \in \mathbb{R}, \tag{3.1}$$

where Δ is a real value termed *location shift* or *treatment effect*.

- Equivalently, model (3.1) can be expressed as

$$Y \sim X + \Delta \quad (Y \text{ has the same distribution as } X + \Delta) \tag{3.2}$$

  This does not mean $Y = X + \Delta$, where X and Y are independent.

- It is obvious that (3.2) implies $\mathrm{Var}(Y) = \mathrm{Var}(X + \Delta) = \mathrm{Var}(X)$.

- Assume X and Y to have unique medians θ_X and θ_Y, respectively. Then

$$\Pr(Y \le \theta_X + \Delta) = \Pr(X + \Delta \le \theta_X + \Delta) = \Pr(X \le \theta_X) = 0.5 \text{ by (3.1) or (3.2).}$$

  This shows that $\theta_X + \Delta$ is the median of Y, i.e., $\theta_Y = \theta_X + \Delta$ or $\Delta = \theta_Y - \theta_X$. Thus $\Delta = 0 \iff \theta_Y = \theta_X$, $\Delta > 0 \iff \theta_Y > \theta_X$ and $\Delta < 0 \iff \theta_Y < \theta_X$.

- Consequently, if the median represents the treatment effect, we may interpret $\Delta = 0$ as no difference in treatment effects between X and Y; $\Delta > 0$ ($\Delta < 0$) as Y having a greater (smaller) treatment effect than X.

- A stronger interpretation for $\Delta > 0$ to represent a greater effect of Y than X is in the sense of the *stochastic order* introduced earlier.

- Under model (3.1) or (3.2), $\Delta > 0$ implies

$$\Pr(X \le t) = \Pr(X + \Delta \le t + \Delta) = \Pr(Y \le t + \Delta) \ge \Pr(Y \le t) \quad \text{for all } t \in \mathbb{R},$$

  and $\Pr(X \le t) > \Pr(Y \le t)$ for some $t \in \mathbb{R}$. This means $X <_{\text{st}} Y$ (X is less than Y in stochastic order).

- The above arguments lead to

$$\begin{cases} \Pr(X \le t) = \Pr(Y \le t) \text{ for all } t \in \mathbb{R} \ (X \sim Y) \ \Leftrightarrow \ \Delta = 0; \\ X <_{\text{st}} Y \ \Leftrightarrow \ \Delta > 0; \quad \text{and} \quad Y <_{\text{st}} X \ \Leftrightarrow \ \Delta < 0. \end{cases} \tag{3.3}$$

- Thus $\Delta = 0$ represents "no difference" between the distributions of X and Y; and $\Delta > 0$ ($\Delta < 0$) means "X is stochastically less (greater) than Y".

- Therefore, a test of $H_0 : \Delta = 0$ against $\Delta > 0$, $\Delta < 0$ or $\Delta \ne 0$ can determine the treatment effect and whether one treatment is better than the other (in the stochastic order of the samples involved).

- A nonparametric test for $H_0 : \Delta = 0$ is introduced next.

3.1 Wilcoxon rank sum test

Null hypothesis: $H_0 : \Delta = 0$

***Y*-Ranks:** Order $N = m + n$ observations $X_1, \ldots, X_m, Y_1, \ldots, Y_n$ in ascending order. Let S_j denote the rank of Y_j, $j = 1, \ldots, n$. $S_1, \ldots, S_n$ are referred to as the *Y-ranks*. Assume no ties and rearrange the *Y*-ranks such that $S_1 < \cdots < S_n$. Then under H_0, Assumption 3.1 implies $\Pr((S_1, \ldots, S_n) = (s_1, \ldots, s_n)) = 1\big/\binom{N}{n}$ for any $s_1 < \cdots < s_n$ drawn from $\{1, 2, \ldots, N\}$.

Test statistic: The test statistic W of the Wilcoxon rank sum test is defined by

$$W = \sum_{j=1}^{n} S_j = S_1 + S_2 + \cdots + S_n \quad \text{(the sum of } Y\text{-ranks)} \tag{3.4}$$

The range of W is $\{M_1, M_1 + 1, \ldots, M_2\}$, where

$$M_1 = 1 + 2 + \cdots + n = \frac{n(n+1)}{2} \quad \text{and} \quad M_2 = \sum_{j=1}^{n} (m + j) = mn + M_1 = mn + \frac{n(n+1)}{2}$$

Exact distribution of W: Under H_0, $(S_1,\ldots,S_n)$ has an equal probability $1\Big/\binom{N}{n}$ to be any $(s_1,\ldots,s_n)$ with $s_1<\cdots<s_n$ taken from $\{1,2,\ldots,N\}$. Therefore, the exact distribution of W under H_0 is given by

$$\Pr(W=w)=\frac{\text{No. of }(s_1,\ldots,s_n):s_1+\cdots+s_n=w}{\binom{N}{n}},\quad M_1\le w\le M_2.$$

Example 3.1 Let $m=2$ and $n=3$. Then $N=2+3=5$, $\binom{N}{n}=\binom{5}{3}=\binom{5}{2}=10$, $M_1=n(n+1)/2=3\times 4/2=6$ and $M_2=mn+M_1=3\times 2+6=12$.

Hence W has a range $\{6,7,8,9,10,11,12\}$ with probabilities as follows:

w	6	7	8	9	10	11	12
(s_1,s_2,s_3)	$(1,2,3)$	$(1,2,4)$	$(1,2,5)$ $(1,3,4)$	$(2,3,4)$ $(1,3,5)$	$(2,3,5)$ $(1,4,5)$	$(2,4,5)$	$(3,4,5)$
$\Pr(W=w)$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{2}{10}$	$\frac{2}{10}$	$\frac{2}{10}$	$\frac{1}{10}$	$\frac{1}{10}$

Mean and variance of W:

While the mean and variance of W can be calculated from its exact distribution, this is not a convenient way as it requires combinatorial enumerations for each case of (m,n), especially if the sample sizes are large.

The following theorem provides a more efficient way to derive the mean and variance of W and some other statistics based on ranks to be used later.

Theorem 3.1 Given N numbers $(a_1,\ldots,a_N)$ (not necessarily all distinct), let $B=(b_1,\ldots,b_n)$ be drawn randomly from $a_1,\ldots,a_N$ without replacement in the same order as $a_1,\ldots,a_N$, $n \le N$. Define random variables:

- $X = a_i$ with probability $1/N$, $a_i \in \{a_1,\ldots,a_N\}$, $i=1,2,\ldots,N$; and
- $S = S(B) = b_1 + b_2 + \cdots + b_n$ if $B=(b_1,b_2,\ldots,b_n)$.

Then

$$\mathrm{E}[S] = n\mathrm{E}[X] \quad \text{and} \quad \mathrm{Var}(S) = n\frac{N-n}{N-1}\mathrm{Var}(X)$$

Proof. Let $\mathcal{B} = \mathcal{B}(n) = \{b = (b_1, \ldots, b_n) : b_1, \ldots, b_n \in \{a_1, \ldots, a_N\}\}$ be the range of B, with each $b \in \mathcal{B}$ following the order of $a_1, \ldots, a_N$ in the sense that $i < j \Leftrightarrow k < l$ for $(b_i, b_j) = (a_k, a_l)$. Then $\Pr(B = b) = 1 \big/ \binom{N}{n}$, $b \in \mathcal{B}$.

Given that a_i is an element in $b = (b_1, b_2, \ldots, b_n)$, the other $n-1$ elements of b can be any $n-1$ of the $N-1$ numbers in $(a_1, \ldots, a_{i-1}, a_{i+1} \ldots, a_N)$. Thus the number of all possible $b \in \mathrm{B}$ that contain a_i is

$$\binom{N-1}{n-1} = \frac{(N-1)!}{(n-1)!(N-n)!} = \frac{N!}{n!(N-n)!} \cdot \frac{n}{N} = \binom{N}{n} \frac{n}{N}, \quad i \in \{1, \ldots, N\}$$

It follows that

$$\sum_{b \in \mathcal{B}} (b_1 + \cdots + b_n) = \sum_{b \in \mathcal{B}} \sum_{j=1}^{n} b_j = \binom{N}{n} \frac{n}{N} \sum_{i=1}^{N} a_i \tag{3.5}$$

Consequently,

$$\mathrm{E}[S] = \sum_{b \in \mathcal{B}} S(b) \Pr(B = b) = \binom{N}{n}^{-1} \sum_{b \in \mathcal{B}} (b_1 + \cdots + b_n) = n \frac{1}{N} \sum_{i=1}^{N} a_i = n \mathrm{E}[X]$$

Next, similar to the above arguments, the number of $b = (b_1, \ldots, b_n) \in \mathcal{B}$ to contain each pair (a_i, a_j) with $i < j$ is

$$\binom{N-2}{n-2} = \frac{(N-2)!}{(n-2)!(N-n)!} = \frac{N!}{n!(N-n)!} \cdot \frac{n(n-1)}{N(N-1)} = \binom{N}{n} \frac{n(n-1)}{N(N-1)}$$

This leads to

$$\sum_{b \in \mathcal{B}} \sum_{i<j} b_i b_j = \binom{N}{n} \frac{n(n-1)}{N(N-1)} \sum_{i<j} a_i a_j \tag{3.6}$$

An illustration of (3.5) and (3.6) is shown in Example 3.2 below.

By (3.5) (with b_i^2 in place of b_i) and (3.6),

$$\sum_{b \in \mathcal{B}} (b_1 + \cdots + b_n)^2 = \sum_{b \in \mathcal{B}} \left(\sum_{i=1}^{n} b_i^2 + \sum_{i \neq j} b_i b_j \right) = \binom{N}{n} \frac{n}{N} \left[\sum_{i=1}^{N} a_i^2 + \frac{n-1}{N-1} \sum_{i \neq j} a_i a_j \right]$$

$$= \binom{N}{n} \frac{n}{N} \left[\frac{N-n}{N-1} \sum_{i=1}^{N} a_i^2 + \frac{n-1}{N-1} \left(\sum_{i=1}^{N} a_i^2 + \sum_{i \neq j} a_i a_j \right) \right] \tag{3.7}$$

It follows from (3.7) that

$$\mathrm{E}\left[S^2\right]=\sum_{b\in\mathcal{B}}(b_1+\cdots+b_n)^2\,\Pr(B=b)=\binom{N}{n}^{-1}\sum_{b\in\mathcal{B}}(b_1+\cdots+b_n)^2$$

$$=\frac{N-n}{N-1}n\frac{1}{N}\sum_{i=1}^{N}a_i^2+\frac{n(n-1)N}{N-1}\cdot\frac{1}{N^2}(a_1+\cdots+a_N)^2$$

$$=\frac{N-n}{N-1}n\mathrm{E}\left[X^2\right]+\frac{n(n-1)N}{N-1}(\mathrm{E}[X])^2$$

Thus $\mathrm{E}[S]=n\mathrm{E}[X]\ \Rightarrow$

$$\mathrm{Var}(S)=\mathrm{E}\left[S^2\right]-(\mathrm{E}[S])^2=n\frac{N-n}{N-1}\mathrm{E}\left[X^2\right]+\left[\frac{n(n-1)N}{N-1}-n^2\right]\left(\mathrm{E}[X]\right)^2$$

$$=n\frac{N-n}{N-1}\mathrm{E}\left[X^2\right]+\frac{n}{N-1}\left[(n-1)N-n(N-1)\right]\left(\mathrm{E}[X]\right)^2$$

$$=n\frac{N-n}{N-1}\mathrm{E}\left[X^2\right]-n\frac{N-n}{N-1}\left(\mathrm{E}[X]\right)^2=n\frac{N-n}{N-1}\mathrm{Var}(X)$$

Example 3.2 To understand equations (3.5) and (3.6) by a simple example, let $N = 5$ and $n = 3$. Then $\mathcal{B} = \mathcal{B}(3)$ consists of

$$\binom{N}{n} = \binom{5}{3} = \binom{5}{2} = \frac{5\times 4}{2} = 10 \text{ elements } (b_1, b_2, b_3) = (b_1, \ldots, b_n)$$

with b_1, b_2, b_3 taken from $(a_1, a_2, a_3, a_4, a_5) = (a_1, \ldots, a_N)$ without replacement (in the same order as $a_1, \ldots, a_5$):

$$(a_1, a_2, a_3),\ (a_1, a_2, a_4),\ (a_1, a_2, a_5),\ (a_1, a_3, a_4),\ (a_1, a_3, a_5),$$

$$(a_1, a_4, a_5),\ (a_2, a_3, a_4),\ (a_2, a_3, a_5),\ (a_2, a_4, a_5),\ (a_3, a_4, a_5).$$

Hence

$$\begin{aligned}\sum_{b\in\mathcal{B}} (b_1 + \cdots + b_n) &= (a_1 + a_2 + a_3) + (a_1 + a_2 + a_4) + \cdots + (a_3 + a_4 + a_5) \\ &= 6(a_1 + a_2 + a_3 + a_4 + a_5) = 10 \times \frac{3}{5}(a_1 + \cdots + a_5) \\ &= \binom{5}{3}\frac{3}{5}(a_1 + \cdots + a_5) = \binom{N}{n}\frac{n}{N}(a_1 + \cdots + a_N)\end{aligned}$$

Similarly,

$$\sum_{b \in \mathcal{B}} \sum_{i<j} b_i b_j = (a_1a_2 + a_1a_3 + a_2a_3) + (a_1a_2 + a_1a_4 + a_2a_4)$$

$$+(a_1a_2 + a_1a_5 + a_2a_5) + (a_1a_3 + a_1a_4 + a_3a_4)$$

$$+(a_1a_3 + a_1a_5 + a_3a_5) + (a_1a_4 + a_1a_5 + a_4a_5)$$

$$+(a_2a_3 + a_2a_4 + a_3a_4) + (a_2a_3 + a_2a_5 + a_3a_5)$$

$$+(a_2a_4 + a_2a_5 + a_4a_5) + (a_3a_4 + a_3a_5 + a_4a_5)$$

$$= 3(a_1a_2 + a_1a_3 + a_1a_4 + a_1a_5 + a_2a_3 + a_2a_4 + a_2a_5)$$

$$+3(a_3a_4 + a_3a_5 + a_4a_5)$$

$$= 3\sum_{i<j} a_i a_j = 10 \times \frac{3 \times 2}{5 \times 4} \sum_{i<j} a_i a_j = \binom{5}{3} \frac{3(3-1)}{5(5-1)} \sum_{i<j} a_i a_j$$

$$= \binom{N}{n} \frac{n(n-1)}{N(N-1)} \sum_{i<j} a_i a_j$$

By Theorem 3.1, we can easily derive the mean and variance of W under H_0. Take $(a_1,\ldots,a_N)=(1,2,\ldots,N)$. Then $W=S$ in Theorem 3.1. Hence

$$\mathrm{E}[X]=\frac{1}{N}\sum_{i=1}^{N} i=\frac{1}{N}\cdot\frac{N(N+1)}{2}=\frac{N+1}{2}\quad\Rightarrow$$

$$\mathrm{E}_0[W]=n\mathrm{E}[X]=\frac{n(N+1)}{2}=\frac{n(m+n+1)}{2}\qquad(3.8)$$

and

$$\mathrm{Var}(X)=\mathrm{E}\left[X^2\right]-\left(\mathrm{E}[X]\right)^2=\frac{1}{N}\sum_{i=1}^{N} i^2-\left(\mathrm{E}[X]\right)^2=\frac{(N+1)(2N+1)}{6}-\left(\frac{N+1}{2}\right)^2$$

$$=\frac{(N+1)[2(2N+1)-3(N+1)]}{12}=\frac{(N+1)(N-1)}{12}\quad\Rightarrow$$

$$\mathrm{Var}_0(W)=n\frac{N-n}{N-1}\mathrm{Var}(X)=nm\frac{N+1}{12}=\frac{mn(m+n+1)}{12}\qquad(3.9)$$

Example 3.3 Let $m = 2$ and $n = 3$. By the distribution obtained in Example 3.1, W takes values $6, 7, 8, 9, 10, 11, 12$ with probabilities 1/10, 1/10, 2/10, 2/10, 2/10, 1/10, 1/10, respectively, under H_0. Therefore,

$$\mathrm{E}_0[W] = \frac{6+7+2(8+9+10)+11+12}{10} = \frac{90}{10} = 9$$

$$\mathrm{Var}_0(W) = \frac{6^2+7^2+2\left(8^2+9^2+10^2\right)+11^2+12^2}{10} - 9^2 = 84 - 81 = 3$$

Alternatively, by (3.8) and (3.9),

$$\mathrm{E}_0[W] = \frac{n(m+n+1)}{2} = \frac{3(2+3+1)}{2} = 9$$

$$\mathrm{Var}_0(W) = \frac{mn(m+n+1)}{12} = \frac{2\times 3(2+3+1)}{12} = \frac{36}{12} = 3$$

Obviously, calculations of $\mathrm{E}_0[W]$ and $\mathrm{Var}_0(W)$ using equations (3.8) and (3.9) are more convenient than via the exact distribution of W.

Symmetry of W

Given two samples $X_1,\ldots,X_m$ and $Y_1,\ldots,Y_n$, for each outcome $(s_1,\ldots,s_n)$ of the Y-ranks $(S_1,\ldots,S_n)$ drawn from $\{1,2,\ldots,m+n\}$ with $s_1 < \cdots < s_n$, take

$$\tilde{s}_j = m+n+1-s_{n+1-j},\quad j=1,\ldots,n,\ \text{ with } \tilde{s}_1 < \cdots < \tilde{s}_n.$$

Then there is a 1 – 1 correspondence between $(s_1,\ldots,s_n)$ and $(\tilde{s}_1,\ldots,\tilde{s}_n)$. It follows that for each outcome $(s_1,\ldots,s_n)$ with $s_1 < \cdots < s_n$ and $s_1 + \cdots + s_n = w$, there is one outcome $(\tilde{s}_1,\ldots,\tilde{s}_n)$ with $\tilde{s}_1 < \cdots < \tilde{s}_n$ such that

$$\tilde{s}_1 + \cdots + \tilde{s}_n = n(m+n+1) - s_n - \cdots - s_1 = n(m+n+1) - w = M_1 + M_2 - w$$

Thus under $H_0 : \Delta = 0$, $\Pr\left(\{(s_1,\ldots,s_n)\}\right) = \Pr\left(\{(\tilde{s}_1,\ldots,\tilde{s}_n)\}\right) = 1\Big/\binom{N}{n} \;\Rightarrow$

$$\Pr(W = w) = \Pr(W = n(m+n+1) - w) = \Pr(W = M_1 + M_2 - w)$$

for every value w of W. Consequently, W is symmetric about

$$\frac{M_1 + M_2}{2} = \frac{n(m+n+1)}{2} = \mathrm{E}_0[W] = \text{Median of } W \text{ under } H_0$$

Rejection rule: Let $\Pr(W \geq w_\alpha) = \alpha$ under H_0 with integer w_α. The Wilcoxon rank sum test rejects $H_0 : \Delta = 0$ at the α level if

- $W \geq w_\alpha$ against $H_1 : \Delta > 0$;
- $W \leq n(m+n+1) - w_\alpha$ against $H_1 : \Delta < 0$;
- either $W \geq w_{\alpha/2}$ or $W \leq n(m+n+1) - w_{\alpha/2}$ against $H_1 : \Delta \neq 0$.

Asymptotic distribution of W: Under H_0, if n is large, then approximately

$$W^* = \frac{W - \mathrm{E}_0[W]}{\sqrt{\mathrm{Var}_0(W)}} = \frac{W - n(m+n+1)/2}{\sqrt{mn(m+n+1)/12}} \sim N(0,1) \qquad (3.10)$$

Approximate rejection rule: Reject $H_0 : \Delta = 0$ at the α level if

- $W^* \geq z_\alpha$ against $H_1 : \Delta > 0$;
- $W^* \leq -z_\alpha$ against $H_1 : \Delta < 0$;
- $|W^*| \geq z_{\alpha/2}$ against $H_1 : \Delta \neq 0$, where W^* is defined in (3.10).

Example 3.4 In Example 4.1 of the textbook (from page 119), Table 4.1 shows a portion of the data on the *Pd* values from a study of water transfer in placental membrane. The interest is to test $H_0: \Delta = 0$ against $H_1: \Delta < 0$.

In this example, $m = 10$, $n = 5$, $M_1 = 5 \times 6/2 = 15$ and $M_2 = 5 \times 10 + 15 = 65$. Thus the range of W is $\{15, 16, \ldots, 65\}$. The combined data are ordered as follows.

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
0.73	0.74	0.80	0.83	0.88	0.90	1.04	1.15	1.21	1.38	1.45	1.46	1.64	1.89	1.91
X	Y	X	X	Y	Y	X	Y	Y	X	X	X	X	X	X

This shows that the observed *Y*-ranks are 2, 5, 6, 8, 9. Hence the observed value of the Wilcoxon rank sum statistic is $W = 2 + 5 + 6 + 8 + 9 = 30$. Since

$$\binom{N}{n} = \binom{15}{5} = \frac{15 \times 14 \times 13 \times 12 \times 11}{5 \times 4 \times 3 \times 2} = 3003$$

is large, it is tedious and time-consuming to obtain w_α or the *p*-value $\Pr(W \le 30)$ for $H_1: \Delta < 0$ by counting the numbers of ordered $(b_1, \ldots, b_5)$ from $(1, \ldots, 15)$ such that $W = 15, 16, \ldots, 30$ (as in Example 3.1).

We can use the computer software R to find the p-value $\Pr(W \le 30) = 0.1272$. Hence H_0 is accepted at the 10% level of significance.

On the other hand, with $m = 10$ and $n = 5$, the normal approximation is good enough to carry out the test. By (3.8) and (3.9),

$$\mathrm{E}_0[W] = \frac{5(10+5+1)}{2} = \frac{5\times 16}{2} = 40 \quad \text{and} \quad \mathrm{Var}_0(W) = \frac{5\times 10\times 16}{12} = \frac{200}{3}$$

Hence the observed value of the normalized test statistic is

$$W^* = \frac{30-40}{\sqrt{200/3}} = \frac{-10}{\sqrt{66.667}} = -1.225$$

and so the approximate p-value is $\Pr(W^* \le -1.225) \approx 0.110 \;\Rightarrow$ accept H_0 at the 10% level. This can also be concluded from $W^* = -1.225 > -z_{0.10} = -1.282$.

For the interpretation of the test results and more details of this example, refer to Example 4.1 of the textbook.

Ties: If there are ties among $X_1,\ldots,X_m,Y_1,\ldots,Y_n$, then similar to signed ranks, the average rank will be assigned to tied values.

In such a case, the mean $\mathrm{E}_0[W]$ in (3.8) is unchanged, and the same argument as that for (2.12) shows that the variance $\mathrm{Var}_0(W)$ in (3.9) reduces to

$$\mathrm{Var}_0(W) = \frac{mn(N+1)}{12} - \frac{mn}{12N(N-1)}\sum_{j=1}^{g} t_j(t_j-1)(t_j+1), \qquad (3.11)$$

where g is the number of groups with tied ranks, and t_j is the number of tied points in group j, $j=1,\ldots,g$. In (3.11), we can ignore groups with $t_j = 1$.

For example, if $Z_1 < Z_2 = Z_3 < Z_4 < Z_5 = Z_6 = Z_7$ are ordered values of combined $X_1,\ldots,X_4,Y_1,\ldots,Y_3$, then the ranks of $(Z_1,\ldots,Z_7)$ are $(1,2.5,2.5,4,6,6,6)$. So we can take $g = 2$, $t_1 = 2$ and $t_2 = 3$ in (3.11).

The conditional distribution of W on ties under H_0 can be worked out similarly to the case with no ties.

Example 3.5 Let $m = 2$, $n = 3$. Conditional on $Z_1 < Z_2 = Z_3 = Z_4 < Z_5$ for the ordered values of X_1, X_2, Y_1, Y_2, Y_3, $(Z_1, \ldots, Z_5)$ have ranks $(1,3,3,3,5)$, $g = 1$ and $t_1 = 3$. The conditional distribution of W under H_0 is given by

$$\Pr(W = 7) = \Pr((1,3,3) \times 3) = 3/10 = 0.3$$

$$\Pr(W = 9) = \Pr((1,3,5) \times 3, (3,3,3)) = 0.4$$

$$\Pr(W = 11) = \Pr((3,3,5) \times 3) = 0.3$$

Hence $\mathrm{E}_0[W] = 7(0.3) + 9(0.4) + 11(0.3) = 9$ (same as from (3.8)) and

$$\mathrm{Var}_0(W) = 7^2(0.3) + 9^2(0.4) + 11^2(0.3) - 9^2 = 83.4 - 81 = 2.4$$

$$< 3 \text{ from (3.9) with no ties}$$

On the other hand, by (3.11) with $g = 1$ and $t_1 = 3$,

$$\mathrm{Var}_0(W) = \frac{2 \times 3(5+1)}{12} - \frac{2 \times 3}{12 \times 5(5-1)} 3(3-1)(3+1) = 3 - \frac{3}{5} = 2.4$$

This matches the result from the direct calculation using the distribution of W.

The Mann-Whitney statistic

An alternative and equivalent test statistic to the Wilcoxon rank sum W in (3.4) for the two-sample location problem is the *Mann-Whitney statistic*:

$$U = \sum_{i=1}^{m}\sum_{j=1}^{n} I_{\{X_i<Y_j\}} = W - \frac{n(n+1)}{2} \quad \text{(assume no ties)} \tag{3.12}$$

The range of U is $\{M_1 - n(n+1)/2, \ldots, M_2 - n(n+1)/2\} = \{0,1,2,\ldots,mn\}$.

Let $a_1,\ldots,a_M$ be M distinct real numbers and $R(a_j)$ the rank of a_j in $a_1,\ldots,a_M$. Then $R(a_j) = k \iff a_i < a_j$ for $k-1$ integers $i \in \{1,\ldots,M\}$. Hence

$$\sum_{i=1}^{M} I_{\{a_i<a_j\}} = R(a_j) - 1 \quad \text{or} \quad R(a_j) = \sum_{i=1}^{M} I_{\{a_i<a_j\}} + 1 \tag{3.13}$$

Let R_j denote the rank of Y_j in $\{Y_1,\ldots,Y_n\}$. Then by (3.13),

$$\sum_{j=1}^{n}\sum_{i=1}^{n} I_{\{Y_i<Y_j\}} + n = \sum_{j=1}^{n}(R_j - 1) + n = \sum_{j=1}^{n} R_j - n + n = \sum_{j=1}^{n} j = \frac{n(n+1)}{2} \tag{3.14}$$

Similarly, (3.13) implies that the Y-ranks in W can be expressed as

$$S_j = \sum_{i=1}^{m} I_{\{X_i < Y_j\}} + \sum_{i=1}^{n} I_{\{Y_i < Y_j\}} + 1, \quad j = 1, \ldots, n. \tag{3.15}$$

It follows from (3.14) – (3.15) that

$$W = \sum_{j=1}^{n} S_j = \sum_{j=1}^{n} \sum_{i=1}^{m} I_{\{X_i < Y_j\}} + \sum_{j=1}^{n} \sum_{i=1}^{n} I_{\{Y_i < Y_j\}} + n = U + \frac{n(n+1)}{2}$$

This proves (3.12). Next, by (3.8) – (3.9) and (3.12), under H_0,

$$\mathrm{E}_0[U] = \mathrm{E}_0[W] - \frac{n(n+1)}{2} = \frac{n(m+n+1)}{2} - \frac{n(n+1)}{2} = \frac{mn}{2} \tag{3.16}$$

$$\mathrm{Var}_0(U) = \mathrm{Var}_0(W) = \frac{mn(m+n+1)}{12} \tag{3.17}$$

and U is symmetric about $mn/2$, so that

$$\Pr(U \le mn - u) = \Pr(U \ge u), \quad u = 0, 1, 2, \ldots, mn. \tag{3.18}$$

Remark 3.1

- The distribution of the Mann-Whitney statistic U depends on the sizes m and n of the two samples, but not on which size is for X or Y sample.
- More specifically, if we switch $X_1,\ldots,X_m;Y_1,\ldots,Y_n$ to $\tilde{X}_1,\ldots,\tilde{X}_n;\tilde{Y}_1,\ldots,\tilde{Y}_m$ with $\tilde{X}_j = Y_j$, $j=1,\ldots,n$, $\tilde{Y}_i = X_i$, $i=1,\ldots,m$, then $\tilde{U} = mn-U$, where U and $\tilde{U}$ are defined by (3.12) based on X_i, Y_j and $\tilde{X}_j, \tilde{Y}_i$ respectively. Since U is symmetric about $mn/2$, $\tilde{U} = mn - U$ has the same distribution as U.
- The R program for the Wilcoxon rank sum statistic produces the distribution of U, not of the Wilcoxon rank sum W itself. The order of m and n in the R commands for the distribution of U does not matter.
- To obtain the distribution of W using R, we can use the relation in (3.12):

$$\Pr(W \le w) = \Pr\left(U + \frac{n(n+1)}{2} \le w\right) = \Pr\left(U \le w - \frac{n(n+1)}{2}\right)$$

and $w_\alpha = u_\alpha + n(n+1)/2$, where $\Pr(U \ge u_\alpha) = \alpha$.

Ties: If there are ties among $X_1,\ldots,X_m,Y_1,\ldots,Y_n$, then the Mann-Whitney statistic is defined by

$$U = \sum_{i=1}^{m}\sum_{j=1}^{n}\left(I_{\{X_i<Y_j\}} + \frac{1}{2}I_{\{X_i=Y_j\}}\right) \tag{3.19}$$

The relationship $U = W - n(n+1)/2$ in (3.12) remains valid if average ranks are assigned to tied values in computing W.

Note that ties within $X_1,\ldots,X_m$ or $Y_1,\ldots,Y_n$ do not affect the value of U in (3.19); neither they affect the value of W in (3.4) (but they affect their variances).

For example, if $(X_1,X_2,X_3,X_4) = (1,4,4,10)$ and $(Y_1,Y_2,Y_3) = (4,8,8)$, then

$$(X_1,X_2,X_3,Y_1,Y_2,Y_3,X_4) = (1,4,4,4,8,8,10) \text{ with ranks } (1,3,3,3,\ 5.5,\ 5.5,\ 7)$$

By (3.4), $W = 3+5.5+5.5 = 14$ and by (3.19), $X_1 < Y_1,Y_2,Y_3$; $X_2,X_3 < Y_2,Y_3$; and $X_2,X_3 = Y_1 \Rightarrow$

$$U = 3+4+0.5+0.5 = 8 = 14-6 = 14 - \frac{3(4)}{2} = W - \frac{n(n+1)}{2}$$

3.2 Estimation of the location shift

A nonparametric estimator of the location shift Δ is given by

$$\hat{\Delta} = \text{median}\left\{Y_j - X_i,\ \begin{matrix} i = 1,\ldots,m \\ j = 1,\ldots,n \end{matrix}\right\} = \begin{cases} U_{((mn+1)/2)} & \text{if } mn \text{ is odd;} \\ \dfrac{U_{(mn/2)} + U_{(mn/2+1)}}{2} & \text{if } mn \text{ is even.} \end{cases}$$

where $U_{(1)} \le U_{(2)} \le \cdots \le U_{(mn)}$ are the ordered values of $(Y_j - X_i)$'s. Let

$$C_\alpha = mn + 1 + \frac{n(n+1)}{2} - w_{\alpha/2} = mn + 1 - u_{\alpha/2} \tag{3.20}$$

Then a $100(1-\alpha)\%$ confidence interval for Δ is given by

$$(\Delta_L, \Delta_U) = \left(U_{(C_\alpha)}, U_{(mn+1-C_\alpha)}\right) = \left(U_{(C_\alpha)}, U_{(u_{\alpha/2})}\right) \tag{3.21}$$

For large m and n, by (3.16) – (3.17), C_α in (3.20) can be approximately by

$$C_\alpha \approx \frac{mn}{2} - z_{\alpha/2}\sqrt{\frac{mn(m+n+1)}{12}} \tag{3.22}$$

Example 3.6 For the data in Example 3.4, $Y_j - X_i$ values are ordered below:

$U_{(1)} \le U_{(2)} \le \cdots \le U_{(50)}$									
−1.17	−1.15	−1.03	−1.01	−1.01	−0.99	−0.90	−0.76	−0.76	−0.74
−0.74	−0.72	−0.71	−0.70	−0.68	−0.64	−0.58	−0.57	−0.56	−0.55
−0.50	−0.49	−0.48	−0.43	−0.31	−0.30	−0.30	−0.25	−0.24	−0.23
−0.17	−0.16	−0.14	−0.09	−0.06	0.01	0.05	0.07	0.08	0.10
0.11	0.15	0.17	0.17	0.32	0.35	0.38	0.41	0.42	0.48

Thus $mn = 50 \Rightarrow \hat{\Delta} = (U_{(25)} + U_{(26)})/2 = (-0.31 - 0.30)/2 = -0.305$.

By R, $\Pr(U \le 9) = 0.028$ and $\Pr(U \le 8) = 0.020$. Hence $u_{0.02} = 50 - 8 = 42$ and $C_{0.04} = 50 + 1 - 42 = 9$. Then by (3.21), an exact 96% confidence interval of Δ is

$$(\Delta_L, \Delta_U) = (U_{(9)}, U_{(50+1-9)}) = (U_{(9)}, U_{(42)}) = (-0.76, 0.15)$$

If we use (3.22), then $C_{0.05} \approx 50/2 - 1.96\sqrt{50(16)/12} = 9.00$. Thus an approximate 95% confidence interval of Δ is also given by $(U_{(9)}, U_{(42)}) = (-0.76, 0.15)$.

Proof of the confidence interval of Δ

Let $U_{(1)} \le U_{(2)} \le \cdots \le U_{(mn)}$ be the ordered mn values of $(Y_j - X_i)$'s. Then

$$U_{(k)} < 0 < U_{(k+1)} \Leftrightarrow Y_j - X_i < 0 \text{ for } k \text{ pairs } (i, j) \Leftrightarrow$$

$$Y_j - X_i > 0 \text{ for } mn-k \text{ pairs } (i, j) \Leftrightarrow U = \sum_{i=1}^{m}\sum_{j=1}^{n} I_{\{X_i < Y_j\}} = mn - k \qquad (3.23)$$

Since $Y_j - (X_i + \Delta) \sim \{Y_j - X_i \text{ under } H_0 : \Delta = 0\}$, (3.23) implies

$$\Pr\left(U_{(k)} < \Delta < U_{(k+1)}\right) = \Pr_0\left(U_{(k)} < 0 < U_{(k+1)}\right) = \Pr_0\left(U = mn - k\right)$$

Consequently,

$$\Pr\left(\Delta < U_{(k)}\right) = \sum_{l=0}^{k-1} \Pr\left(U_{(l)} < \Delta < U_{(l+1)}\right) = \sum_{l=0}^{k-1} \Pr_0\left(U = mn - l\right)$$

$$= \Pr_0\left(U \ge mn - k + 1\right), \qquad (3.24)$$

where $U_{(0)} = -\infty$ and $\Pr_0$ denotes the probability under $H_0 : \Delta = 0$.

Thus by (3.20) and (3.24),

$$\Pr\left(\Delta < U_{(C_\alpha)}\right) = \Pr_0\left(U \geq mn - C_\alpha + 1\right) = \Pr_0\left(U \geq u_{\alpha/2}\right) = \frac{\alpha}{2} \tag{3.25}$$

On the other hand, by (3.24) together with the symmetry of U in (3.18),

$$\Pr\left(\Delta < U_{(u_{\alpha/2})}\right) = \Pr_0\left(U \geq mn + 1 - u_{\alpha/2}\right) = 1 - \Pr_0\left(U < nm + 1 - u_{\alpha/2}\right)$$

$$= 1 - \Pr_0\left(U \leq nm - u_{\alpha/2}\right) = 1 - \Pr_0\left(U \geq u_{\alpha/2}\right) = 1 - \frac{\alpha}{2} \tag{3.26}$$

It follows from (3.25) – (3.26) that

$$\Pr\left(U_{(C_\alpha)} < \Delta < U_{(u_{\alpha/2})}\right) = \Pr\left(\Delta < U_{(u_{\alpha/2})}\right) - \Pr\left(\Delta < U_{(C_\alpha)}\right) = 1 - \frac{\alpha}{2} - \frac{\alpha}{2} = 1 - \alpha$$

This proves the confidence interval of Δ in (3.21).