Wilcoxon の順位和検定の検定統計量 W の分布について

二つのグループを $X_1 = (x_{11}, \cdots, x_{1n_1}),\ X_2 = (x_{21}, \cdots, x_{2n_2})$ とし、それぞれの順位の和を $W_1,\ W_2$ とする。検定量 W に採用されるほうを X_1 としたとき (すなわち $W_1 < W_2$ のとき)

- 同順位がない場合

$$\text{平均 } E[W] = \frac{n_1(n_1+n_2+1)}{2}$$
$$\text{分散 } V[W] = \frac{n_1 n_2(n_1+n_2+1)}{12}$$

- 同順位がある場合 (平均順位をつける)

$$\text{平均 } E[W] = \frac{n_1(n_1+n_2+1)}{2}$$
$$\text{分散 } V[W] = \frac{n_1 n_2(n_1+n_2+1)}{12} - \frac{n_1 n_2}{12(n_1+n_2)(n_1+n_2-1)} \sum_{j=1}^{2} {t_j}^3 - t_j$$

ただし、t_jは第 j グループでの同順位となったものの個数を表す。

となる。

[証明]

一般に、m 個の $y_1, y_2, \cdots, y_m$ から任意に一つ選んだものの順位 R を考える。
R は 1 から m までを一つずつとり、それを等しく $1/m$ の確率で選ぶので、その平均値は

$$E[R] = \frac{1}{m}\sum_{i=1}^{m} i = \frac{1}{m}\cdot\frac{m(m+1)}{2} = \frac{m+1}{2}$$

となる。このことから (n_1+n_2) 個の $x_{11}, x_{12}, \cdots, x_{1n_1}, x_{21}, \cdots, x_{2n_2}$ から任意に n_1 個選んだものの順位の和を W のとりうる値と考えると、その平均値は

$$E[W] = E\left[\sum_{i=1}^{n_1} R_i\right] = \sum_{i=1}^{n_1} E[R_i] = \frac{n_1(n_1+n_2+1)}{2}$$

となる。$E[R]$ の $\sum_{i=1}^{m} i$ は同順位がない場合の式であるが、同順位があっても平均順位なのでその和は変わらない。なぜなら、第 k 番目から $(k+l-1)$ 番目までの l 個が同順位であったとすると、この l 個に前から順にそのまま順位をつけた時の l 個の順位和は

$$\sum_{j=1}^{l}(k+j-1) \ = \ l(k-1) + \sum_{j=1}^{l} j \ = \ l(k-1) + \frac{l(l+1)}{2} \ = \ l\left(k + \frac{l-1}{2}\right)$$

となる。最後の式は、k 番目に $(l-1)/2$ を加えてできるこの l 個の平均順位に個数 l を掛けたものであり、まさにこの l 個の平均順位の和を表している。よって、同順位の有無に関わらず統計量 W の平均値は上の式で与えられる。

同じように、一般に、m 個の $y_1, y_2, \cdots, y_m$ から任意に一つ選んだものの順位 R を考え、その分散を求める。ただし、ここでは同順位はない場合とする。

$$
\begin{aligned}
V[R] = E[R^2] - (E[R])^2 \ &= \ \frac{1}{m}\sum_{i=1}^{m} i^2 - \left\{\frac{1}{m}\sum_{i=1}^{m} i\right\}^2 \\
&= \frac{1}{m}\cdot\frac{m(m+1)(2m+1)}{6} - \left\{\frac{1}{m}\cdot\frac{m(m+1)}{2}\right\}^2 \ = \ \frac{(m+1)(2m+1)}{6} - \left(\frac{m+1}{2}\right)^2 \\
&= \frac{(m+1)(m-1)}{12}
\end{aligned}
$$

これより (n_1+n_2) 個の $x_{11}, x_{12}, \cdots, x_{1n_1}, x_{21}, \cdots, x_{2n_2}$ から任意に n_1 個選んだものの順位の分散は

$$
\sum_{i=1}^{n_1} V[R_i] = \frac{n_1(n_1+n_2+1)(n_1+n_2-1)}{12}
$$

さらに、$y_1, y_2, \cdots, y_m$ から任意に異なる 2 つを選んだ時のそれぞれの順位 $R_i,\ R_j (R_i \neq R_j)$ の共分散を求めたい。まずはすべての組み合わせの積 R_iR_j の和 $\sum_{i=1}^{m}\sum_{i<j}^{m} ij$ を考える。

$$
\left(\sum_{i=1}^{m} i\right)^2 = \sum_{i=1}^{m} i^2 + 2\sum_{i=1}^{m}\sum_{i<j}^{m} ij \quad \text{より} \quad \sum_{i=1}^{m}\sum_{i<j}^{m} ij = \frac{1}{2}\cdot\left\{\left(\sum_{i=1}^{m} i\right)^2 - \sum_{i=1}^{m} i^2\right\}
$$

よって

$$
\sum_{i=1}^{m}\sum_{i<j}^{m} ij = \frac{1}{2}\cdot\left\{\left(\frac{m(m+1)}{2}\right)^2 - \frac{m(m+1)(2m+1)}{6}\right\} = \frac{m(m+1)(m-1)(3m+2)}{24}
$$

$R_i,\ R_j$ をとる確率はすべて $\dfrac{1}{{}_mC_2} = \dfrac{2!(m-2)!}{m!} = \dfrac{2}{m(m-1)}$ だから

$$
E[R_iR_j] = \frac{2}{m(m-1)}\sum_{i=1}^{m}\sum_{i<j}^{m} ij = \frac{(m+1)(3m+2)}{12}
$$

したがって、共分散 $Cov[R_i, R_j] = E[R_iR_j] - E[R_i]E[R_j]$ より

$$
Cov[R_i, R_j] = \frac{(m+1)(3m+2)}{12} - \left(\frac{m(m+1)}{2}\right)^2 = -\frac{m+1}{12}
$$

以上を用いて、同順位がない場合の統計検定量 W の分散を求めると、$V[X+Y] = V[X] + V[Y] + 2Cov[X, Y]$ より

$$
\begin{aligned}
V[W] &= V\left[\sum_{i=1}^{n_1} R_i\right] = \sum_{i=1}^{n_1} V[R_i] + 2\sum_{i<j} Cov[R_i, R_j] \\
&= \frac{n_1(n_1+n_2-1)(n_1+n_2+1)}{12} + 2\cdot\frac{n_1(n_1-1)}{2}\cdot\left(-\frac{n_1+n_2+1}{12}\right) = \frac{n_1n_2(n_1+n_2+1)}{12}
\end{aligned}
$$

となる。

同順位がある場合も、一般に m 個の $y_1, y_2, \cdots, y_m$ から任意に一つ選んだものの順位 R を考え、その分散を求める。このとき t_j は第 j グループでの同順位となったものの個数とする。平均を考えた時と同様に第 k 番目から $(k+l-1)$ 番目までの l 個が同順位であったとして、同順位の部分の 2 乗和を考える。この l 個に前から順にそのまま順位をつけた時の順位は $(k+j-1)_{(j=1,2,\cdots,l)}$, 平均順位は $\left(k+\frac{l-1}{2}\right)$ であったから、両者の二乗和の差は

$$
\begin{aligned}
&\sum_{j=1}^{l}(k+j-1)^2 - l\left(k+\frac{l-1}{2}\right)^2 \\
&\qquad = \cancel{l}k^2 + 2k\sum_{j=1}^{l}(j-1) + \sum_{j=1}^{l}(j-1)^2 - \cancel{l}k^2 - lk(l-1) - \frac{l(l-1)^2}{4} \\
&\qquad = 2k\cdot\frac{(l-1)l}{2} + \frac{(l-1)l(2l-1)}{6} - lk(l-1) - \frac{l(l-1)^2}{4} \\
&\qquad = \frac{(l-1)l(l+1)}{12}
\end{aligned}
$$

となる。すなわち、

$$
E[R^2] = \frac{1}{m}\sum_{i=1}^{m}\left(i^2 - \frac{(t_j-1)t_j(t_j+1)}{12}\right) = \frac{1}{m}\cdot\frac{m(m+1)(2m+1)}{6} - \frac{1}{12m}\sum_{j=1}^{2}({t_j}^3 - t_j)
$$

となるので、これより m 個の $y_1, y_2, \cdots, y_m$ から任意に一つ選んだものの順位 R の分散は

$$
\begin{aligned}
V[R] &= E[R^2] - (E[R])^2 \\
&= \frac{(m+1)(2m+1)}{6} - \frac{1}{12m}\sum_{j=1}^{2}({t_j}^3 - t_j) - \left(\frac{m+1}{2}\right)^2 \\
&= \frac{(m+1)(m-1)}{12} - \frac{1}{12m}\sum_{j=1}^{2}({t_j}^3 - t_j)
\end{aligned}
$$

また、$y_1, y_2, \cdots, y_m$ から任意に異なる 2 つを選んだ時のそれぞれの順位 $R_i,\ R_j (R_i \neq R_j)$ の共分散は同順位がない場合と同じ要領で、

$$
Cov[R_i, R_j] = -\frac{n_1+n_2+1}{12} + \frac{1}{12(n_1+n_2)(n_1+n_2-1)}\sum_{j=1}^{2}({t_j}^3 - t_j)
$$

よって、同順位がある場合の統計検定量 W の分散は

$$
\begin{aligned}
V[W] &= V\left[\sum_{i=1}^{n_1} R_i\right] = \sum_{i=1}^{n_1} V[R_i] + 2\sum_{i<j} Cov[R_i, R_j] \\
&= \frac{n_1(n_1+n_2-1)(n_1+n_2+1)}{12} - \frac{n_1}{12m}\sum_{j=1}^{2}({t_j}^3 - t_j) \\
&\quad + 2\cdot\frac{n_1(n_1-1)}{2}\cdot\left(-\frac{n_1+n_2+1}{12} + \frac{1}{12(n_1+n_2)(n_1+n_2-1)}\sum_{j=1}^{2}({t_j}^3 - t_j)\right) \\
&= \frac{n_1 n_2(n_1+n_2+1)}{12} - \frac{n_1 n_2}{12(n_1+n_2)(n_1+n_2-1)}\sum_{j=1}^{2}({t_j}^3 - t_j)
\end{aligned}
$$

となる。

Wilcoxon の W と Mann-Whitniy の U の関係

　Wilcoxon 順位和検定が W を平均順位を用いて決めるのに対し、Mann-Whitney U 検定は平均順位和の代わりに、相手にどれだけ勝っているか（大きいとき 1、同点は 0.5 で計算）の和で検定値 U を決める。

　ここで、W と U の関係を考える。二つのグループを $X_1 = (x_{11}, \cdots, x_{1n_1}),\ X_2 = (x_{21}, \cdots, x_{2n_2})$ とし、データはそれぞれ大きい順にソートされているものとする。

　はじめに、次の関数を定義する。

$$
\begin{aligned}
K_{1l}(j):&\quad x_{1j} < x_{1k}\text{となる } x_{1k}\text{の個数} \\
K_{1e}(j):&\quad x_{1j} = x_{1k}\text{となる } x_{1k}\text{の個数（ただし、}k \neq j\text{）} \\
K_{2l}(j):&\quad x_{1j} < x_{2k}\text{となる } x_{2k}\text{の個数} \\
K_{2e}(j):&\quad x_{1j} = x_{2k}\text{となる } x_{2k}\text{の個数}
\end{aligned}
$$

この関数と平均順位の関係は

$$
hrank(x_{1j}) = K_{1l}(j) + K_{2l}(j) + 1 + \frac{K_{1e}(j) + K_{2e}(j)}{2}
$$

である。

さらにこの関数と U の関係も考える。X_1 が W で採用されているとすると、X_1 のほうが順位和が小さい、すなわち数値の大きいグループである。Mann-Whitney で比較して採用されるほうは、数値の小さいほうのグループなので、X_2 となる。よって

$$
U = \sum_{j=1}^{n_1} K_{2l}(j) + \frac{K_{2e}(j)}{2}
$$

となる。

> Wilcoxon 順位和の小さいほう W（データ数 n_1）と、Mann-Whitney U 検定の U との関係は、
> $$U = W - \frac{n_1(n_1+1)}{2}$$

（証明） $K(j) = K_{1l}(j) + 1 + K_{1e}(j)/2$ を求める。まず、$x_{1(j-1)} > x_{1j} > x_{1(j+1)}$ のとき、

$$
K_{1l}(j) = j - 1, \quad K_{1e}(j) = 0, \quad K(j) = j
$$

となる。次に

$$x_{1(m-1)} > x_{1m} = \cdots = x_{1j} = \cdots = x_{1M} > x_{1(M+1)}$$

とすると、$m \leqq j \leqq M$ において、

$$K_{1l}(j) = m-1, \quad K_{1e}(j) = M-m\,(\text{自分自身を除いているから}), \quad K(j) = \frac{M+m}{2}$$

となる。ここで、

$$\sum_{j=m}^{M} K(j) = (M-m+1)\frac{M+m}{2} = \frac{M(M+1)}{2} - \frac{m(m+1)}{2} = \sum_{j=m}^{M} j$$

となるから、トータルで $\sum_{j=1}^{n_1} K(j) = \sum_{j=1}^{n_1} j$ となる。したがって、

$$\begin{aligned}
W &= \sum_{j=1}^{n_1} hrank(x_{1j}) \\
&= \sum_{j=1}^{n_1} \left\{ K_{1l}(j) + K_{2l}(j) + 1 + \frac{K_{1e}(j) + K_{2e}(j)}{2} \right\} \\
&= \sum_{j=1}^{n_1} \left(K_{2l}(j) + \frac{K_{2e}(j)}{2} \right) + \sum_{j=1}^{n_1} \left(K_{1l}(j) + \frac{K_{1e}(j)}{2} + 1 \right) \\
&= U + \sum_{j=1}^{n_1} K(j) \ = \ U + \frac{n_1(n_1+1)}{2}
\end{aligned}$$