2群の平均の差：ノンパラメトリック検定

渡邊直樹

2021年10月13日

1　マン・ホイットニーのU検定

マン・ホイットニーのU検定（Mann-Whitney U-test）は，2群の平均に有意な差があるかを確かめるために，広く用いられてきた統計的手法である．歴史的には，Frank Wilcoxonが1945年に開発したデータ処理の手続きと同等のものとして，Henry MannとDonald Whitneyが1947年にUという統計量を用いて統計学的に定式化した．ウィルコクソンは米国にかつてあった化学・医薬品会社アメリカン・サイアナミッド（American Cyanamid）の社員であり，彼が開発した手法はウィルコクソンの順位和検定（Wilcoxon rank-sum test）と呼ばれている．

これらの検定では，2群の比較対象が順序であっても連続量であってもよく，標本数（サンプルサイズ）が2群で異なっていてもよいのだが，特に，確率変数の母集団に正規性を仮定できないときに用いる．たとえば，表1ではA群（グループA）とB群（グループB）のメンバーの体重をリストしており，体重そのものは連続量である．各グループのメンバーの体重が正規分布に従うかどうかはサンプルサイズが小さいこともあり，正確にはわかりにくい[1]．グループAとグループBのメンバー間で彼らの体重には平均的には差がありそうだが，その差は統計的に有意なものであるといえるだろうか？

表1の例をマン・ホイットニーのU検定で検討してみよう．まず，グループ$i = A, B$のメンバーの体重について，全体での順位を足し合わせたもの（順位和）は$R_A = 3+5+7+8+9+10+15+16+17+18 = 108$，$R_B = 1+2+4+6+11+12+13+14 = 63$である．グループAのサン

[1]一般的には，コルモゴロフ検定シャピロ・ウィルク検定などを適用して，確率変数が正規分布に従うかどうかを確認する．

表 1: A 群と B 群のメンバーの体重とその順位

順位	1	2	3	4	5	6
体重	56.6	57.3	58.7	59.0	62.2	62.9
群	B	B	A	B	A	B
順位	7	8	9	10	11	12
体重	64.3	64.5	65.1	65.3	65.5	66.2
群	A	A	A	A	B	B
順位	13	14	15	16	17	18
体重	69.0	70.2	71.5	72.7	75.0	75.8
群	B	B	A	A	A	A

プルサイズは 10，グループ B のサンプルサイズは 8 なので，グループ A のメンバーの平均順位は 10.8，グループ B のそれは 7.875 となり，グループ A のメンバーの体重の方が平均的に重いようには見える．

グループ i のサンプルサイズを n_i とすると，グループ i の U の値は

$$U_i = n_A n_B + \frac{n_i(n_i+1)}{2} - R_i$$

と定義される．グループ A のサンプルサイズは 10，グループ B のサンプルサイズは 8 なので，グループ A の U の値は $U_A = 80 + 55 - 108 = 27$，グループ B のそれは $U_B = 80 + 36 - 63 = 53$ と計算される．検定に用いられる U の値は $\min(U_A, U_B) = 27$ である．両グループのサンプルサイズ n_A と n_B が大きくなると，U が従う確率分布は平均 $n_A n_B/2$，分散 $n_A n_B(n_A + n_B + 1)/12$ の正規分布に漸近することが知られている．この性質を使って，グループ A とグループ B のメンバー間で彼らの体重に平均的には差がない確率を計算することができる．その確率が p 値であり，上記の例では 0.03435 となる．よって，有意水準を 0.05 に設定すると，確かに，グループ A のメンバーの体重の方が平均的には重いのだろう．

ここで,「重いのだろう」という曖昧な表現をした理由は，U 検定では連続量で表される体重を順位に直して検討しているためである．順位による評価は量に対して直接の言及はできない．アンケート調査などにおける「良くない、普通、良い」といった順序付けが最初からなされている場合でないと，明確な表現で平均値に有意な差があると表現することには慎重であるべきである．

上記の例では同順位（タイ）となる体重の値がなかったことに注意しよう．たとえば，11 番目の順位に同じ数値が 3 つあったとすると，11 番目から 13 番目の平均順位である 12.0 を R_i の算出における該当する体重の順位として加える．これにより，正規分布による近似の精度はやや落ちる．R または EZR (Easy R) では，「タイがあるため，正確な p 値を計算することができません」というアラートが結果に対して付される．同順位が 1 つもなければ，正規分布での近似ではなく，U から厳密な p 値を計算するための漸化式があることが知られている．

補論 1：EZR でウィルコクソンの順位和検定を行う

Excel アドインにはマン・ホイットニーの U 検定もウィルコクソンの順位和検定も実装されていないが，2 群の確率変数がそれぞれ正規分布に従う場合には，（確率分布の等分散性を仮定する）t 検定，（等分散性を仮定しない）ウェルチの t 検定を行うことはできる．以下では，EZR におけるウィルコクソンの順位和検定の手順を説明する．

- EZR はメニューで選ばれたデータ処理を R に変換して実行している．その R のコードは「R スクリプト」画面に表示される．もちろん，この画面にカーソルを持っていき，そこから R のコードを直接入力し，画面右下にある「実行」ボタンを押すことによっても，そのコマンドを実行するもできる．

データセットの作成には使い慣れた Excel のスプレッドシートが便利なので，EZR で直接読み込むか，csv ファイルとして保存して，それを読み込んだ方がよい．ただ，R スクリプトからのコード入力に慣れるために，ここでは表 1 のデータを EZR に直接書き込んでみよう．

```
A = c(58.7, 62.2, 64.3, 64.5, 65.1, 65.3, 71.5, 72.7, 75.0, 75.8)
B = c(56.6, 57.3, 59.0, 62.9, 65.5, 66.2, 69.0, 70.2)
wilcox.test(A,B)
```

- c( ) の中にカンマで区切って数値を並べると，ベクトルを生成することができる．ここで，=は数学でいう等号ではなく，それは A という変数（この場合はベクトル）に c( ) の中に書かれている数値を代入（格納）することを意味する．（数学でいう等号は==で表される．）

- 上記のコードを実行するには，1 行ごとにカーソルを行末に持っていき,「実行」ボタンを押す必要がある．出力結果は次の通りである．

```
Wilcoxon rank sum test

data:  A and B
W = 53, p-value = 0.2743
alternative hypothesis: true location shift is not equal to 0
```

なお，ウィルコクソンの順位和検定における W 統計量はサンプル数が少ない方の群における数値の順位和なので，マン・ホイットニーの U 検定の説明で用いた記法で書くと，入力したデータでは R_B に相当する．W が従う分布は正規分布などで近似されることが多いが，より正確な p 値を得るために並べ替え（permutation）による計算が行われることもあり，ソフトウェアによって出力される値が少しずつ異なる．一方，外れ値に対して検定結果が頑健であり，計算が比較的容易であることなどから，現在でもよく使われる．R とは異なる統計ソフトウェアである SPSS ではマン・ホイットニーの U 検定が実装されている．EZR にはマン・ホイットニーの U 検定は実装されていないので，それを実行するには「R による統計処理」（オーム社）の著者である青木繁伸氏のウェブサイトで公開されているコードを使用する．

http://aoki2.si.gunma-u.ac.jp/R/u-test.html

2　ブルンナー・ムンツェル検定

2 群の確率変数がそれぞれ正規分布に従うならば，マン・ホイットニーの U 検定よりも，t 検定を用いた方が一般に検定力が高い．検定力とは分析者が棄却したい仮説（帰無仮説）が偽であるときに正しくそれを棄却する確率のことである．連続量の平均の差については，順位に直すことで失われる情報もないため，t 検定を適用することで前章末で触れたような不明確な言及を回避できる．しかし，t 検定にせよマン・ホイットニーの U 検定にせよ，2 群の確率変数の分散が等しいことを前提とした検定方法である．この前提を保証することが困難であるとしても，等分

散を前提とせずに適用できる統計的手法がブルンナー・ムンツェル検定（Brunner-Munzel test）である[2].

ブルンナー・ムンツェル検定では，2 群の確率変数の分布に関する前提をおかず，各群から 1 つずつサンプルを取り出したとき，どちらかが他方よりも大きくはならない確率をそれぞれ求め，それが 0.5 となるという帰無仮説を検定することを考える．つまり，

$$p = \mathrm{Prob}(X_A < X_B) + \frac{1}{2}\mathrm{Prob}(X_A = X_B) \tag{1}$$

において，$p = 1/2$ となるという仮説を検定する．

表 1 の例でグループ A のメンバーの順位和を R_A，グループ B のそれを R_B とした．サンプルサイズで除した平均順位をそれぞれ $\bar{R}_A = R_A/n_A$，$\bar{R}_B = R_B/n_B$ とすると，

$$\hat{p} = \frac{\bar{R}_B - \bar{R}_A}{n_A + n_B} + \frac{1}{2} \tag{2}$$

が (1) の不偏推定量となる．つまり，計算を通じて，(2) の期待値 $E[\hat{p}]$ がちょうど (1) になると判る．ここで，各グループ $i = A, B$ 内での順位和の平均は，1 から n_i までの和をサンプルサイズ n_i で除したものなので，$(n_i + 1)/2$ である[3]．グループ i のメンバーの体重が全体での順位で k 番目であることを R_{ik}，グループ内での順位を $R_{ik}^{(i)}$ と表わそう．すると，

$$\begin{aligned} S_i^2 &= \frac{1}{n_i - 1}\sum_{k=1}^{n_i}(R_{ik} - R_{ik}^{(i)} - \bar{R}_i + \frac{n_i + 1}{2})^2 \\ \hat{\sigma}_i &= \frac{S_i^2}{((n_A + n_B) - n_i)^2} \end{aligned}$$

としたとき，n_A と n_B が大きくなると

$$W^{BM} = \frac{1}{(n_A + n_B)^{1/2}}\frac{\bar{R}_B - \bar{R}_A}{\hat{\sigma}_N}$$

は平均 0，分散 1 の標準正規分布に漸近する．ここで，

$$\hat{\sigma}_N^2 = (n_A + n_B)(\frac{\hat{\sigma}_A^2}{n_A} + \frac{\hat{\sigma}_B^2}{n_B})$$

である．n_A と n_B が小さいときには，W^{BM} はある公式によって決まる自由度の t 分布の方が近似の精度が高いとされている．

[2] 2 群の確率変数が正規分布に従うとき，Welch の t 検定（Welch's t-test）はそれらの等分散を前提とせずに 2 群の平均に有意な差があるかを検証するために用いられる．

[3] 各グループ $i = A, B$ 内での順位和は $n_i(n_i + 1)/2 = n_i(n_i + 1)/2$ なので，これを n_i で除したものは $(n_i + 1)/2$ である．

補論 2：EZR でブルンナー・ムンツェル検定を行う

以下では，EZR におけるブルンナー・ムンツェル検定の手順をノートする．補論 1 のウィルコクソンの順位和検定で用いたデータを使ってみよう．R では，lowstat と呼ばれるパッケージを library から呼び出して，ブルンナー・ムンツェル検定を行う．

```
>library(lawstat)
>brunner.munzel.test(A,B)
```

出力結果は次の通りである．

```
Brunner-Munzel Test

data:  A and B
Brunner-Munzel Test Statistic = -1.1673, df = 15.425, p-value = 0.2608
95 percent confidence interval:
 0.04148739 0.63351261
sample estimates:
P(X<Y)+.5*P(X=Y)
          0.3375
```