岡本安晴　2001.10

差の検定

独立なデータ：Mann-Whitney U Test

信号の光がついたらすぐにスィッチを押すということを何回か行った場合に、各回における光が点いたときからスゥッチが押されるまでの時間を、AとBの２人の人について測定した結果が、表１のようであったとする。

表 1　AさんとBさんの反応時間
（ミリ秒単位）

A	B
168	202
159	257
249	203
179	204
239	241
174	221
172	202
157	209
165	225
208	192
167	187
212	210
213	218
223	182
170	229
	245
	235
	259

Aの測定値は１５個（１５回分）、Bの測定値は１８個ある。Aの反応時間と、Bの反応時間に、差があるかどうかの検定を行うことを考える。反応時間の分布は正規分布に従っているとは考え難いので、ｔ検定ではなく、データ値の順序情報を用いるMann-WhitneyのUテストと呼ばれている方法で行うことにする。

いま、Uテストを一般的に説明するため、m個のデータ、X_1、･･･、X_m、と、n個のデータ、Y_1、･･･、Y_nがあるとする。Uテストでは、X_iの分布とY_jの分布の差をみるために、X_iとY_jの大小関係によって値が決まる、次のD_{ij}という変数を導入する

（Gibbons,1971）。

$$D_{ij} = \begin{cases} 1 & X_i > Y_j \\ 0 & X_i < Y_j \end{cases}$$

　このＤｉｊの和としてＵを定義する。

$$U = \sum_{i=1}^{m} \sum_{j=1}^{n} D_{ij} \qquad （１）$$

　すなわち、Ｕは、ＸｉとＹｊを比べたとき、Ｘｉの方がＹｊより大きい組み合わせの数である。従って、Ｘｉの方がＹｊより全体として値が小さいときＵの値は小さくなり、 Ｘｉの方がＹｊより全体として値が大きいときＵの値は大きくなる。Ｕ検定では、このＵの値にもとづいて検定が行われる。

　帰無仮説「X_iの分布とY_jの分布に差はない」のもとで、Ｕの値が小さ過ぎないか、あるいは大き過ぎないかを調べる。帰無仮説のもとでのＵの分布は、以下に説明するように漸化式で与えることができる（Gibbons,1971）。

　２つの独立なサンプル、Ｘ１、…、ＸｍとＹ１、…、Ｙｎ、に対して、（１）式によって与えられるＵの値がｕである確率を、$p_{m,n}(u)$で表わす。

　Ｕの値は、ｍ個の文字Ｘとｎ個の文字Ｙを混ぜて一直線に並べたときに、これらの mn 個の文字の並びから２つの文字（隣り合う文字である必要はない）をとったときに、左側の文字がＸで右側の文字がＹである組み合わせの対の総数とみなすことができる。つまり、大きいものから順番に、左から並べられているとみなす。ｍ個のＸとｎ個のＹの並べ方は$\binom{m+n}{m}$通りあるが、ＸｉとＹｊの分布が同じであるという仮説のもとでは、これらの並べ方の生起確率はいずれも等しいと考えられる。従って、ｍ個のＸとｎ個のＹを並べたときに、Ｕの値がｕになる並べ方の総数を$r_{m,n}(u)$とおけば、

$$p_{m,n}(u) = \frac{r_{m,n}(u)}{\binom{m+n}{m}}$$

となる。

ＸとＹの文字の並べ方において、左端がＹのものとＸのものとに分けて考えれば、次の関係式の成り立つことがわかる。

$$r_{m,n}(u) = r_{m,n-1}(u) + r_{m-1,n}(u-n) \qquad (２)$$

$r_{m,n-1}(u)$は、左端がＹであるときの並べ方の数を表わす。この場合は、左端のＹを無視して、（ｎ－１）個のＹがあると考えて数える。

$r_{m-1,n}(u-n)$は、左端がＸであるときの並べ方の数を表わす。（ｍ－１）個のＸと、ｎ個のＹを並べてから、左端にＸを付け加えると、ｕの値はｎ増えることに注意して、数える。

式（２1）より、$p_{m,n}(u)$についての次の関係式が得られる。

$$p_{m,n}(u) = \frac{r_{m,n}(u)}{\begin{pmatrix} m+n \\ m \end{pmatrix}}$$

$$= \frac{r_{m,n-1}(u)}{\begin{pmatrix} m+n \\ m \end{pmatrix}} + \frac{r_{m-1,n}(u-n)}{\begin{pmatrix} m+n \\ m \end{pmatrix}}$$

$$= \frac{n}{m+n} \cdot \frac{r_{m,n-1}(u)}{\begin{pmatrix} m+n-1 \\ \mathrm{m} \end{pmatrix}} + \frac{m}{m+n} \cdot \frac{r_{m-1,n}(u-n)}{\begin{pmatrix} m-1+n \\ m-1 \end{pmatrix}}$$

$$= \frac{np_{m,n-1}(u) + mp_{m-1,n}(u-n)}{m+n}$$

確率$p_{m,n}(u)$は、上式により再帰的に求めることができる。

分布関数は、次式により与えられる。

$$P(u \le U) = \sum_{u=0}^{U} p_{m,n}(u)$$

Ｕの値の計算においては、次の関係が利用される（Siegel,1956）。

X_1、･･･、X_mとY_1、･･･、Y_nを一緒にして、小さいものから順番に１、２、･･･、ｍ＋

ｎと順位付けを行ったときのＹ$_j$の順位を、R_j^{XY}で表わす。Ｙ$_j$を、Ｙ$_1$、･･･、Ｙ$_n$のなかで、１からｎまでの順位付けを行ったときの順位は、R_j^{Y}で表わす。Ｙ$_j$より小さいX_iの総数は、

$$m-\sum_{i=1}^{m} D_{ij}$$

で与えられるので、次式が成り立つ。

$$R_j^{XY} = m-\sum_{i=1}^{m} D_{ij} + R_j^{Y}$$

上式の両辺の、ｊについての和をとれば、次式となる。

$$\sum_{j=1}^{n} R_j^{XY} = mn-\sum_{j=1}^{n}\sum_{i=1}^{m} D_{ij} + \sum_{j=1}^{n} R_j^{Y}$$

すなわち

$$\sum_{j=1}^{n} R_j^{XY} = mn - U + \frac{n(n+1)}{2}$$

$$U = mn + \frac{n(n+1)}{2} - \sum_{j=1}^{n} R_j^{XY} \qquad (3)$$

となる。

いま、

$$D'_{ij} = \begin{cases} 1 & \text{if} \quad Y_j > X_i \\ 0 & \text{if} \quad Y_j < X_i \end{cases}$$

$$U' = \sum_{i=1}^{m}\sum_{j=1}^{n} D'_{ij}$$

とおき、Ｘ$_1$、･･･、Ｘ$_m$とＹ$_1$、･･･、Ｙ$_n$を一緒にして順位付けしたときのＸ$_i$の順位をR_i^{XY}とおいて、Ｘ$_i$とＹ$_j$を入れ替えて考えれば

$$U' = mn + \frac{m(m+1)}{2} - \sum_{i=1}^{m} R_i^{XY}$$

の成り立つことが分かる。

したがって、

$$U + U' = 2mn + \frac{n(n+1)}{2} + \frac{m(m+1)}{2} - \sum_{j=1}^{n} R_j^{XY} - \sum_{i=1}^{m} R_i^{XY}$$

$$= 2mn + \frac{n(n+1)}{2} + \frac{m(m+1)}{2} - \frac{(m+n)(m+n+1)}{2}$$

$$= mn \qquad (4)$$

となる。

プログラム PMWU.dpr では、式（３）および式（４）が用いられている。さらに、Uの値を変換して、正規分布による検定も行えるようになっている。

Uの平均μと標準偏差σは、X_iの分布とY_jの分布に差が無いという仮説のもとでは次式で与えられる。

$$m = \frac{mn}{2}$$

$$s = \sqrt{\frac{mn(m+n+1)}{12}}$$

サンプル数が多いとき、次式

$$z = \frac{U - m}{s}$$

で与えられるｚは、平均０、分散１の標準正規分布に従う、として検定が行わる。

X_iとY_jを一緒にしたときの順位付けにおいて、同じ順位のもの（タイ・スコア）があるときには、σの値は次のように修正される。すなわち、順位ｒのものがt_r個あったときのσの修正値s_cは、次式で与えられる（Siegel & Castellan,1988）。

$$s_c = \sqrt{\frac{mn}{N(N-1)}\left(\frac{N^3 - N}{12} - \sum_{r=1}^{N}\frac{t_r^3 - t_r}{12}\right)}$$

ここで、

$$N = m + n$$

である。

Uの値の計算のためのプログラム PMWU.dpr を実行すると､図１のフォームが表示される。

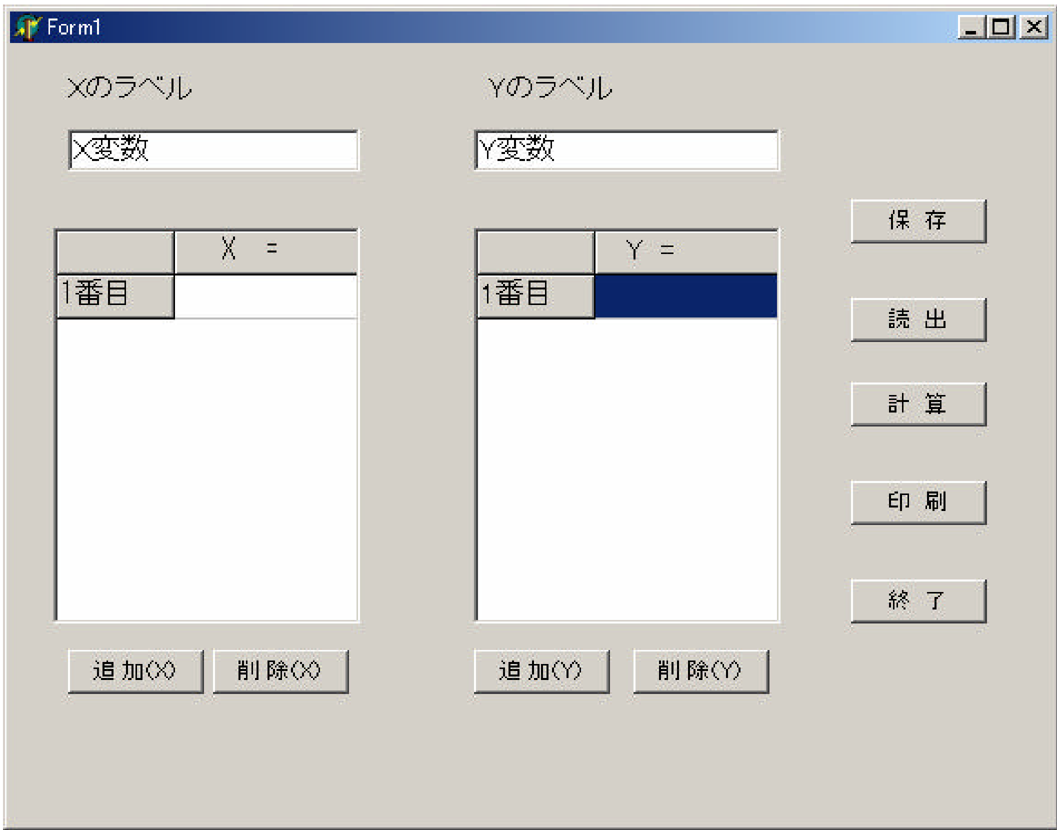

図１　起動時のフォーム

対応のないデータの分析なので、左側と右側の 2 つのデータ設定用 StringGrid、及びそれぞれ独立に行数を設定するための「追加」ボタンと「削除」ボタンが用意されている。「追加」ボタンをクリックすると、その上の StringGrid 内のアクティブなセルの下に行が新しく追加・挿入される。セルをアクティブにするには、そのセルをクリックする。「削除」ボタンをクリックすると、その上の StringGrid 内のアクティブなセルを含む行が削除される。

「印刷」ボタンをクリックすると、設定されているデータがプリンタに出力される。

図２　データの設定（表１のデータ）

図２のようにデータを設定後、「計算」ボタンをクリックするとＵの値の計算が始る。「計算」ボタンをクリックする前に、StringGrid には空白のセルがないようにし、余分なセルがあれば「削除」ボタンをクリックして削除しておく。

「計算」ボタンをクリックすると、まず図３のダイアログボックスが表示される。

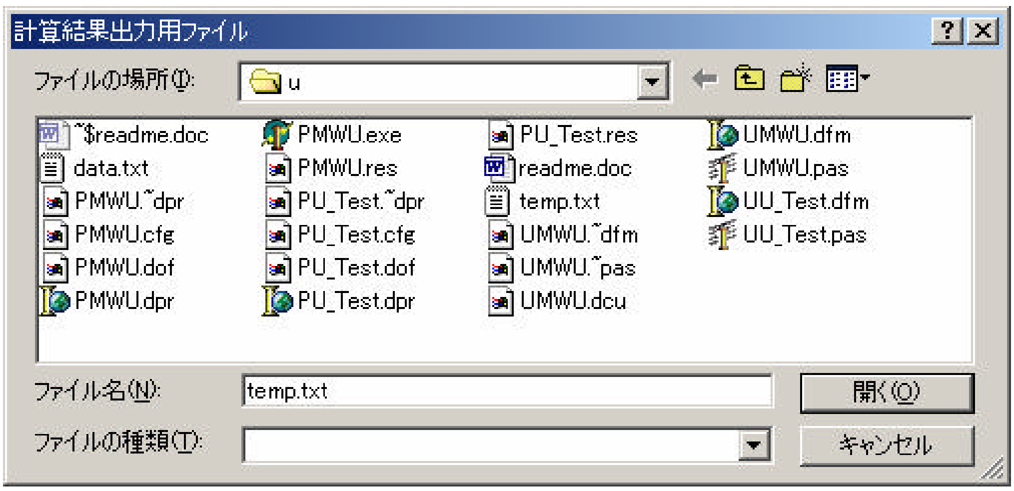

図３　計算結果出力用ファイル名の設定

ファイル名の設定後、「開く」ボタンをクリックすると計算が始る。計算結果は、図３で設定した名前のファイルにテキストファイルとして書き出されるので、プログラムの実行終了後、エディタなどで開いて見ることができる。

計算が終了すると、図４のように「計算を終了しました」のメッセージがフォームに表示される。

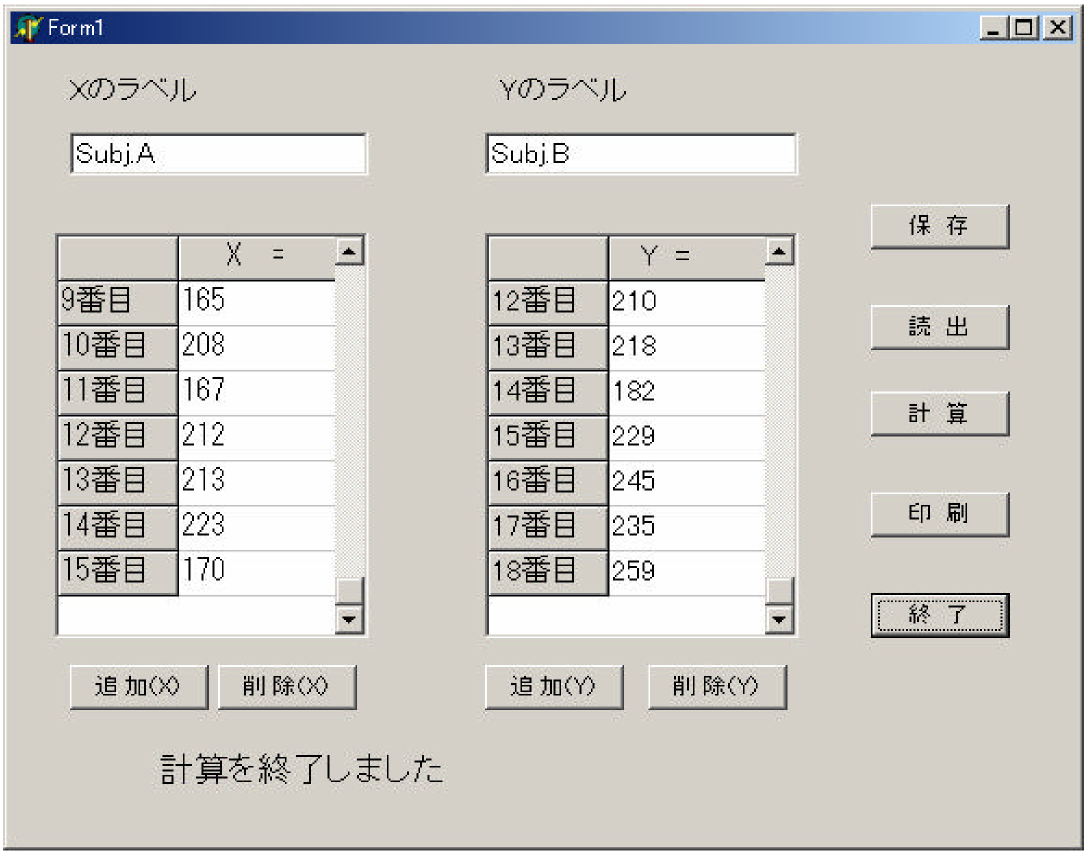

図４　計算終了時のフォーム

「終了」ボタンのクリックでプログラムの実行終了となる。

プログラムの実行終了後、図３で設定した名前の出力ファイルをエディタなどで開いて見ることができる。図２のデータの場合の計算結果の出力は、リスト１のようになっている。

リスト１　計算結果の出力内容

```
データ１ =
          168.00
          159.00
          249.00
          179.00
          239.00
          174.00
          172.00
          157.00
          165.00
          208.00
          167.00
          212.00
          213.00
          223.00
          170.00

中央値 =            174.00

データ２ =
          202.00
          257.00
          203.00
          204.00
          241.00
          221.00
          202.00
          209.00
          225.00
          192.00
          187.00
          210.00
          218.00
          182.00
          229.00
          245.00
          235.00
          259.00

中央値 =            214.00


U =  66.0      m = 15      n = 18

Z              =  -2.495
Z(corrected)   =  -2.495
```

リスト１におけるＵの値に対する累積確率は

$$P(u \le 66) \approx 0.0059 < 0.05 = 5\%$$

となる。Ｕの変換値ｚに対する累積確率は

$$P(z \le -2.495) \approx 0.0063 < 0.05 = 5\%$$

となっている。図２のデータの場合、有意水準５%で差が認められる。

図２のように設定されているデータは、「保存」ボタンのクリックでファイルに保存することができる。「保存」ボタンをクリックすると図５のダイアログボックスが表示される。

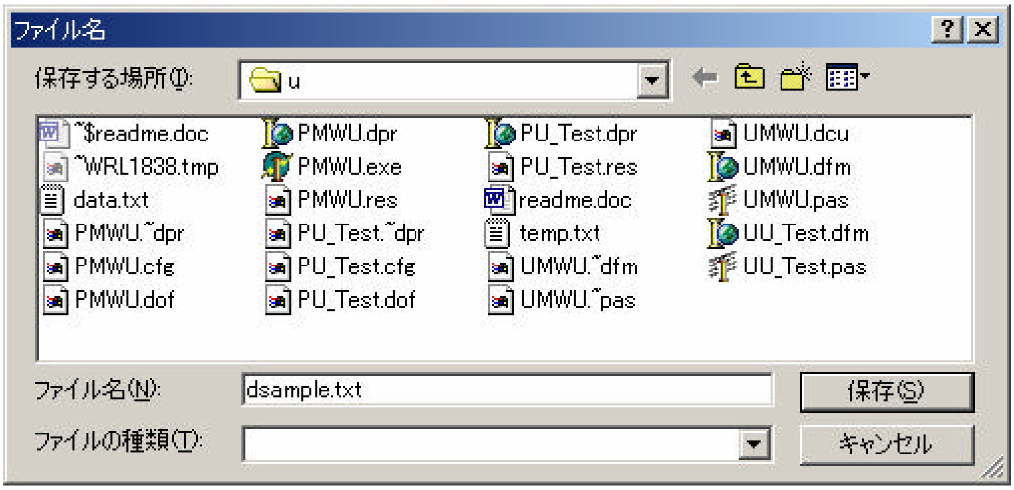

図５　データ保存用ファイル名の設定

データを保存するファイルの名前を設定して、図５のダイアログボックスの「保存」ボタンをクリックすると、設定した名前のファイルにデータが保存される。

保存したデータは「読出」ボタンのクリックで読み込むことができる。「読出」ボタンをクリックすると、図６のダイアログボックスが表示される。

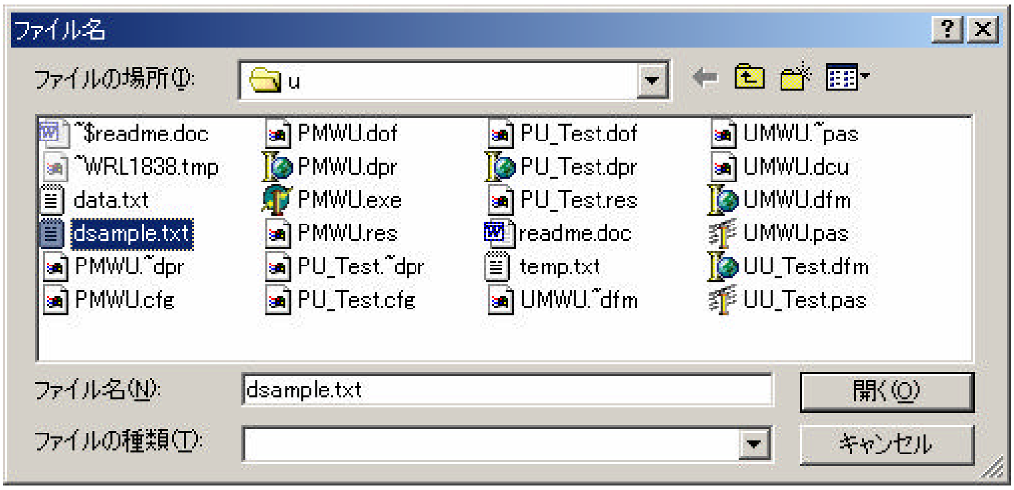

図６　データ読み出し用ファイル名の設定

データを読み出すファイルの名前を設定後、「開く」ボタンをクリックすると、設定した名前のファイルからデータが読み込まれる。

参考文献

Gibbons,J.D.(1971).　*Nonparametric statistical inference.* McGraw-Hill,Inc.

岡本安晴 (1998). Delphi **で学ぶデータ分析法.**　CQ **出版社.**

Siegel,S (1956). *Nonparametric Statistics: For the Behavioral Sciences.* McGraw-Hill Book Company,Inc.

Siegel,S. & Castellan,N.J.,Jr.(1988).　*Nonparametric statistics for the behavioral sciences. 2nd. Ed.* McGraw-Hill,Inc.