1.例子引入

问题：如何检验下面两组数据是否有显著性差异?

```
X = c(698, 688, 675, 656, 655, 648, 640, 639, 620)
Y = c(780, 754, 740, 712, 693, 680, 621)
```

单样本方法：使用Wilcoxon符号秩检验。为了能使数据配对，我们随机从X中剔除两个数据，使X和Y的样本量相同。假设我们剔除639和620。随后我们构造数据$Z = X - Y$, 数据就变成了:

```
> Z
[1] -82 -66 -65 -56 -38 -32  19
```

随后，我们对Z作出假设:
$H_0 : M_e = 0 \leftrightarrow H_1 : M_e \neq 0$
如果原假设成立，说明X和Y没有显著性差异。

然而，这种方法的弊端十分明显:

- 要求两样本的数据量相同
- 配对后样本量缩减到原来的$\frac{1}{2}$

双样本检验正是为了解决这一问题。它们的核心优势是不要求两个样本的样本量一样。Mann-Whitney就针对Wilcoxon符号秩检验进行了双样本的拓展，后经证明他们的统计量和Wilcoxon**符号秩**检验统计量有密切关系，因此也被称作Wilcoxon**秩和**检验。

所谓**符号秩**和**秩和**，本质上只是对数据方向的转换不同。符号秩检验用数据的**原始符号**作为其方向；而**秩和检验**是用它**从属的样本**作为方向。

2.基本思想

双样本检验的基本操作就是“先混合，后标注”。例如上述样本:

```
# 一、先混合
combine = c(x, y)

# 记录当前数据的标签
labels = c(rep("A", length(x)), rep("B", length(y)))

# 二、后排序
indexs = order(combine)
sort_combine = combine[indexs]
sort_labels = label[index]

data.frame(
    `样本` = sort_combine,
    `标签` = sort_labels,
    `秩` = rank(sort_combine)
  )
```

输出数据：

```
   样本 标签 秩
1   620    A  1
2   621    B  2
3   639    A  3
4   640    A  4
5   648    A  5
6   655    A  6
7   656    A  7
8   675    A  8
9   680    B  9
10  688    A 10
11  693    B 11
12  698    A 12
13  712    B 13
14  740    B 14
15  754    B 15
16  780    B 16
```

如果两组样本不存在显著差异，对混合样本进行排序后，X和Y应该能**充分地混合**在一起。

所谓充分地混合，就是认为X和Y均匀地分布在混合样本中，就好像它们本身就是从这个混合样本抽取的一样。

表现在数学上就是，X和Y的**秩和**应该和它们的**样本量**之间呈现合理关系。例如，X有10个样本，Y有20个样本。但如果X和Y的秩和相近，这就很不合理。X的样本量少，理应秩和比Y小得多。此时我们应该拒绝原假设，认为X中的数据显著地比Y中的数据要大。

3.统计量

由上述基本思想不难得到，检验统计量应该是X的秩和或Y的秩和。记作W_x或W_y。同时，我们将X和Y的样本量分别记作m和n，假设$m > n$。那么拒绝域应当形如$W_x < c$，并且c和n相关。

然而，Mann-Whitney构造的统计量是W_{xy}和W_{yx}。他们和W_x和W_y关系很密切。我们用一个例子说明其定义，核心点就是"组合计数"。

- 原始样本：$X = \{1, 2\}, Y = \{3, 4\}$。
- 组合：$Z = \{(1, 3), (1, 4), (2, 3), (2, 4)\}$。
- 计数(W_{xy}表示当$X < Y$时计数加1)：因为在Z中，所有(x, y)的组合都满足$x < y$所以$W_{xy} = 4$；同理，$W_{yx} = 0$。

4.拒绝域

通过W_{xy}的直观定义，我们知道，当W_{xy}很大，说明有很多X小于Y。反之亦然。记X和Y的中心位置分别为θ_x和θ_y，那么：

- 左侧检验($\theta_x < \theta_y$)：$W_{xy} > c_0$
- 右侧检验($\theta_x > \theta_y$): $W_{xy} < c_1$
- 双侧检验($\theta_x \neq \theta_y$)：$|W_{xy}| < c_2$

注意, c_0, c_1, c_2均和X和Y样本量相关，理由如基本思想所述。

5.实际计算(重要)

注意，对于统计量W_{xy}和W_{yx}，按照定义式去算，计算复杂度将很高。例如一个n和m的样本，组合后样本Z就高达nm个，复杂度是$O(n*m)$。相反，W_x和W_y就好算得多。我们只需要一次排序，复杂度为$O(N\log N)$，其中$N=n+m$。因此，如果我们能建立W_x和W_{xy}之间的关系，就有可能显著降低计算复杂度。

5.1.无结 + 小样本

事实是，无结情况下，我们有关系式：

$$W_{xy}=W_y-\frac{n(n+1)}{2}$$

W_{yx}同理。我们先算W_y和W_x，再通过关系式算出W_{xy}和W_{yx}。然后，通过查表，我们就能知道W_{xy}的合理区间在哪里了。

5.2.有结 or 大样本

有结情况下，W_{xy}和W_y之间的精确关系就显得格外复杂了。此时如果要计算精确的W_{xy}最好还是依照定义式。

但是，实际计算中，我们通常不计算精确的W_{xy}值。这是因为即便是小样本的有结情况下，W_y的精确分布依旧非常复杂(需要考虑所有打结情况，情况异常多)，制表不现实，因此"完美"的精确计算是不可能的了。

此时，我们的做法是, 继续套用无结情况下的关系式，计算得到W_{xy}。然后使用正态近似，同时利用结的信息进行正态修正。

5.3.一些结论

无结情况下：

$$E(W_{xy})=\frac{mn}{2}$$

$$D(W_{xy}) = \frac{mn(m+n+1)}{12}$$

有结情况下：
$E(W_{xy})$根据Z值的正负左右微调0.5
$D(W_{xy})$**减去**修正项k，τ为结长, g为打结数：

$$k = \frac{mn\left[\sum_{i=1}^{g}(\tau^3-\tau)\right]}{12(m+n)(m+n-1)}$$

5.4 思考

说实话，读到这里，不知道你有没有发现，直接用W_x和W_y做检验更简单直接，而且是确切可行的。那么我们为什么还要舍近求远地去计算W_{xy}和W_{yx}呢？这大概是为了纪念Mann-Whitney作出的贡献吧。

另外，标准教科书上只为W_{xy}和W_{yx}统计量无结的情况进行了精确分布制表，所以想用好已有资源就乖乖去算W_{xy}和W_{yx}吧。

ps: 前面公式都没写两条。下面我就将火力全开，推导这个W_{yx}和W_{xy}的一些性质。这是非应用层的内容，不感兴趣请略过。

6.关系式推导

6.1.定义阐述

要理解W_x和W_{xy}之间的关系，首先要看看W_{xy}是如何用数学定义的。记$I_m = \{1, 2, ..., m\}$, $I_n = \{1, 2, ..., n\}$

$$W_{xy} = \#(X_i < Y_j, i \in I_m, j \in I_n) = \sum_{i=1}^{m}\sum_{j=1}^{n} I(X_i < Y_j)$$

仿照此定义，同理可得W_{yx}, W_{xx}, W_{yy}. 同时为了方便，我们定义:

$$W_{x=y}=\#(X_i=Y_j,i\in I_m,j\in I_n)=\sum_{i=1}^{m}\sum_{j=1}^{n}I(X_i=Y_j)$$

同理我们可以定义$W_{x=x}$和$W_{y=y}$，我们马上可以发现：

$$\sum_{i=1}^{m}\sum_{j=1}^{n}(I(X_i<Y_j)+I(X_i>Y_j)+I(X_i=Y_j))=mn$$

$$\Longrightarrow\ W_{xy}+W_{yx}+W_{x=y}=mn \tag{1}$$

这是一个非常直观的结论，我们在定义时顺手就发现了。

6.2.无结情况(简单)

无结情况下，$W_{x=y}=0$，于是$W_{xy}+W_{yx}=nm$
让我们仿照W_{xy}的定义等价变换W_x。原始定义：

$$W_x=\sum_{i=1}^{m}R(X_i)$$

不难发现，此时，所谓的秩，就是所有小于自己的数。比如:

$$\{32,45,67,89\}$$

67的秩为3，是因为它比32和45大。同时，再进行一个简单的加一操作，就是3。所以，此时的秩可以等价为：

$$R(X_i)=\#(Z_j<X_i,j\in I_{m+n})+1$$

特别的，Z_j又可以拆分成X和Y两部分。于是:

$$R(X_i)=\#(X_j<X_i,j\in I_m)+\#(Y_j<X_i,j\in I_n)+1$$

所以：

$$
\begin{aligned}
W_x &= \sum_{i=1}^{m} R(X_i) \\
&= \sum_{i=1}^{m} [\#(X_j < X_i, j \in I_m) + \#(Y_j < X_i, j \in I_n) + 1] \\
&= \sum_{i=1}^{m} \#(X_j < X_i, j \in I_m) + W_{yx} + m \\
&= \binom{m}{2} + W_{yx} + m \\
&= \frac{(1+m)m}{2} + W_{yx}
\end{aligned}
$$

6.3.有结情况(稍复杂)

6.3.1.重新定义秩

在有结情况下，我们需要重新定义$R(X_i)$，并且将其与计数相关联。我们用一个例子观察：

秩	混合数据	所属样本
1	621	Y
2.5	640	X
2.5	640	X

$R(640) = 2.5$可以拆分两组计数。第一组是小于640的，计数1，记为A; 第二组是等于640的，计数2，记为B。但秩是如何由这两部分组成的?

可以看到，$R(640)$如果不取平均的话，最小为2，即$(A+1)$，最大为3，即$(A+B)$。平均秩正是两者的平均，即$\frac{2A+B+1}{2} = A + \frac{B+1}{2}$

因此，$R(X_i) = A_{x_i} + \frac{B_{x_i}+1}{2}$，其中：

$$
\begin{aligned}
A_{x_i} &= \#(Z_j < X_i, j \in I_{m+n}) \\
&= \#(X_a < X_i, a \in I_m) + \#(Y_b < X_i, b \in I_n) \\
&= \sum_{a=1}^{m} I(X_a < X_i) + \sum_{b=1}^{n} I(Y_b < X_i)
\end{aligned}
$$

同理可得：

$$
B_{x_i} = \sum_{a=1}^{m} I(X_a = X_i) + \sum_{b=1}^{n} I(Y_b = X_i)
$$

所以W_x和W_{xy}之间就联系起来了：

$$
\begin{aligned}
W_x &= \sum_{i=1}^{m} (A_{x_i} + \frac{B_{x_i} + 1}{2}) \\
&= \sum_{i=1}^{m} A_{x_i} + \frac{1}{2} \sum_{i=1}^{m} B_{x_i} + \frac{m}{2} \\
&= (W_{xx} + W_{yx}) + \frac{1}{2}(W_{x=x} + W_{x=y}) + \frac{m}{2} \quad (2)
\end{aligned}
$$

W_y的计算同理。

6.3.2.计算W_{xx}、$W_{x=x}$和$W_{x=y}$

观察关系式，我们发现W_{xx}、$W_{x=x}$和$W_{x=y}$这三个式子还是需要数据配对才能计算。如果仅通过排序后结的信息表示它们呢?
先看W_{xx}，假设$X_1 = \{1, 2, 3, 4, 5\}$。由于我们对序列进行了升序排列，所以：

$$
W_{xx} = len(\{(1, 2), (1, 3), (1, 4), ..., (4, 5)\}) = \binom{5}{2}
$$

然而，这是在X_1内部无结的情况。有结呢？不难发现，假设$X_2 = \{1, 2, 2, 2, 5\}$，原本X_1中的组合$\{(2, 3), (2, 4), (3, 4)\}$全部被$(2, 2)$所替代，不再计数。这一个结长为3的结减掉的组合有$\binom{3}{2}$个，推广一下就是结长为$\tau$的结减去$\binom{\tau}{2}$个计数。但$\tau = 1$时这个式子没有意义，而我们期望其值为0.我们不妨直接将这个式子先直接展开：

$$
\binom{\tau}{2} = \frac{\tau(\tau - 1)}{2}
$$

当$\tau = 1$时$\frac{\tau(\tau-1)}{2} = 0$，真是妙不可言！所以:

$$W_{xx} = \binom{m}{2} - \sum_{i=1}^{g_x} \frac{\tau_{x_i}(\tau_{x_i} - 1)}{2}$$

而对于$W_{x=x}$，我们很容易发现:

$$W_{x=x} = \sum_{i=1}^{g_x} \tau_{x_i}^2$$

最后剩一个$W_{x=y}$。由于它需要x和y之间作比较，因此必须依赖混合样本(Z)结的信息。现在，假设混合样本共有G个结，每个结长为τ，同理可得:

$$W_{z=z} = \sum_{i=1}^{G} \tau_i^2$$

另一方面，$W_{z=z}$中的有序数对可以分为三类:

- X 内部的相等对： $W_{x=x}$
- Y 内部的相等对： $W_{y=y}$
- X 和 Y 之间的相等对：包括有序对 (X_i, Y_j) 和 (Y_j, X_i)，但 $I(X_i = Y_j) = I(Y_j = X_i)$，所以 X 和 Y 之间的相等对总数为 $2W_{x=y}$.

因此，有:

$$\sum_{k=1}^{G} \tau_i^2 = W_{x=x} + W_{y=y} + 2W_{x=y}.$$

解出$W_{x=y}$:

$$W_{x=y} = \frac{\sum_{i=1}^{G} \tau_i^2 - W_{x=x} - W_{y=y}}{2}$$

6.3.3.结论

$$W_x = (W_{xx} + W_{yx}) + \frac{1}{2}(W_{x=x} + W_{x=y}) + \frac{m}{2}$$

其中:

- $W_{xx}=\binom{m}{2}-\sum_{i=1}^{g_x}\frac{\tau_{x_i}(\tau_{x_i}-1)}{2}$
- $W_{x=x}=\sum_{i=1}^{g_x}\tau_{x_i}^2$
- $W_{y=y}=\sum_{i=1}^{g_y}\tau_{y_i}^2$
- $W_{x=y}=\frac{\sum_{i=1}^{G}\tau_i^2-W_{x=x}-W_{y=y}}{2}$

6.3.4.验算

$X=\{2,2,5\}$和$Y=\{1,3,3\}$
混合样本：$Z=\{1,2,2,3,3,5\}$，秩为$\{1,2.5,2.5,4.5,4.5,6\}$
由定义式可以得到：$W_{yx}=5, W_x=2.5+2.5+6=11$
由关系式可得：

- $W_{xx}=3-(1+0)=2$
- $W_{x=x}=2^2+1^2=5$
- $W_{y=y}=1^2+2^2=5$
- $W_{x=y}=\frac{(1^2+2^2+2^2+1^2)-5-5}{2}=0$
- $W_x=2+W_{yx}+\frac{1}{2}(5+0)+\frac{3}{2}=6+W_{yx}$

所以，$W_{yx}=W_x-6=11-6=5$，符合定义式的计算结果。该结论完全是个人推导，更多情况的验证请自行尝试!

7.精确分布

有结的情况异常复杂，因为我们需要考虑所以可能的结数和结长，其计算量相当惊人。而对于无结的情况，很容易能想到：

$$P(W_x=k)=\frac{\#(分配方式使W_x=w)}{\binom{m+n}{m}}$$

简化计算思路有递推公式法，生成函数法等。当然最简单地还是暴力枚举，因为总数也就 $\binom{m+n}{m}$。现在有计算机的情况下，通过暴力枚举给小样本制表还是相当轻松的。

8.近似分布

设第一个样本有 m 个观测值，第二个样本有 n 个观测值，总观测数为 $N = m + n$。将所有观测值合并排序后分配秩（从 1 到 N），令 W_x 表示第一个样本的秩和。在零假设下（两个样本来自同一分布），W_x 的方差推导如下。

8.1.大体思路

- W_x 是第一个样本的秩和，由于在零假设下所有可能的秩分配是等概率的（即从 N 个秩中随机选择 m 个秩分配给第一个样本），因此 W_x 是一个随机变量。
- 方差推导基于指示随机变量和协方差计算，考虑抽样 without replacement 的影响。
- 最终得到方差公式为 $\mathrm{Var}(W_x) = \frac{mn(N+1)}{12}$，其中 $N = m + n$。

8.2.推导步骤

1. **定义指示随机变量**：令 I_i 表示秩 i 是否被分配到第一个样本的指示变量，即：

$$I_i = \begin{cases} 1 & \text{if rank } i \text{ is in sample 1} \\ 0 & \text{otherwise} \end{cases}$$

   则 $W_x = \sum_{i=1}^{N} i I_i$.
2. **计算期望 $E[W_x]$**：
   每个 I_i 的期望为 $E[I_i] = \frac{m}{N}$，因此：

$$E[W_x] = \sum_{i=1}^{N} i E[I_i] = \frac{m}{N} \sum_{i=1}^{N} i = \frac{m}{N} \cdot \frac{N(N+1)}{2} = \frac{m(N+1)}{2}.$$

3. **计算方差 $\mathrm{Var}(W_x)$**：
   方差公式为：

$$\mathrm{Var}(W_x) = \sum_{i=1}^{N} i^2 \mathrm{Var}(I_i) + \sum_{i \neq j} ij \mathrm{Cov}(I_i, I_j).$$

- **计算** $\mathrm{Var}(I_i)$:
  $\mathrm{Var}(I_i) = E[I_i^2] - (E[I_i])^2 = \frac{m}{N} - \left(\frac{m}{N}\right)^2 = \frac{m}{N}\left(1 - \frac{m}{N}\right) = \frac{mn}{N^2}$.
- **计算** $\mathrm{Cov}(I_i, I_j)$ **for** $i \neq j$:

$$\mathrm{Cov}(I_i, I_j) = E[I_i I_j] - E[I_i]E[I_j].$$

其中 $E[I_i I_j] = P(I_i = 1, I_j = 1) = \frac{\binom{N-2}{m-2}}{\binom{N}{m}} = \frac{m(m-1)}{N(N-1)}$, 且 $E[I_i]E[I_j] = \left(\frac{m}{N}\right)^2$。

因此:

$$\mathrm{Cov}(I_i, I_j) = \frac{m(m-1)}{N(N-1)} - \left(\frac{m}{N}\right)^2 = \frac{m(m-1)N - m^2(N-1)}{N^2(N-1)} = \frac{m(m-N)}{N^2(N-1)} = -\frac{}{N^2}$$

因为 $N = m + n$, 所以 $m - N = -n$.

4. **代入方差公式**:

$$\mathrm{Var}(W_x) = \sum_{i=1}^{N} i^2 \cdot \frac{mn}{N^2} + \sum_{i \neq j} ij \cdot \left(-\frac{mn}{N^2(N-1)}\right).$$

令 $S_1 = \sum_{i=1}^{N} i = \frac{N(N+1)}{2}$, $S_2 = \sum_{i=1}^{N} i^2 = \frac{N(N+1)(2N+1)}{6}$。
注意:

$$\sum_{i \neq j} ij = \sum_{i=1}^{N}\sum_{j=1}^{N} ij - \sum_{i=1}^{N} i^2 = S_1^2 - S_2.$$

因此:

$$\mathrm{Var}(W_x) = \frac{mn}{N^2} S_2 - \frac{mn}{N^2(N-1)}(S_1^2 - S_2) = \frac{mn}{N^2}\left[S_2 - \frac{1}{N-1}(S_1^2 - S_2)\right].$$

简化括号内表达式：

$$S_2 - \frac{1}{N-1}(S_1^2 - S_2) = \frac{(N-1)S_2 + S_2 - S_1^2}{N-1} = \frac{NS_2 - S_1^2}{N-1}.$$

所以：

$$\mathrm{Var}(W_x) = \frac{mn}{N^2(N-1)}(NS_2 - S_1^2).$$

计算 $NS_2 - S_1^2$：

$$NS_2 = N \cdot \frac{N(N+1)(2N+1)}{6} = \frac{N^2(N+1)(2N+1)}{6},$$

$$S_1^2 = \left(\frac{N(N+1)}{2}\right)^2 = \frac{N^2(N+1)^2}{4}.$$

因此：

$$NS_2 - S_1^2 = N^2(N+1)\left(\frac{2N+1}{6} - \frac{N+1}{4}\right) = N^2(N+1) \cdot \frac{4(2N+1) - 6(N+1)}{24} =$$

代入方差：

$$\mathrm{Var}(W_x) = \frac{mn}{N^2(N-1)} \cdot \frac{N^2(N+1)(N-1)}{12} = \frac{mn(N+1)}{12}.$$

由于 $N = m + n$，最终结果为：

$$\mathrm{Var}(W_x) = \frac{mn(m+n+1)}{12}.$$