第三章　两独立样本数据的位置和尺度推断

在单一样本的推断问题中, 引人关注的是总体位置的估计问题. 在实际应用中, 常常涉及两不同总体的位置参数或尺度参数的比较问题, 比如, 两支股票中哪一支股的红利更高, 两种汽油中哪一种对环境的污染更少, 两种市场营销策略哪种更有效等等.

假定两独立样本

$$X_1, X_2, \cdots, X_m \overset{\text{i.i.d.}}{\sim} F_1\left(\frac{x-\mu_1}{\sigma_1}\right), \quad Y_1, Y_2, \cdots, Y_n \overset{\text{i.i.d.}}{\sim} F_2\left(\frac{x-\mu_2}{\sigma_2}\right).$$

而且$X_1, X_2, \cdots, X_m, Y_1, Y_2, \cdots, Y_n$相互独立. 其中$\mu_1, \mu_2$是位置参数, σ_1, σ_2是尺度参数, 有关μ_1和μ_2 的估计和检验问题称为两样本的位置参数问题。有关σ_1和σ_2的估计和检验问题称为两样本的尺度参数问题。

对位置参数问题, 本章只考虑如下简单的情况：

$$X_1, X_2, \cdots, X_m \overset{\text{i.i.d.}}{\sim} F_1(x) = F(x), Y_1, Y_2, \cdots, Y_n \overset{\text{i.i.d.}}{\sim} F_2(x) = F(y-\mu).$$

两样本具有相似的分布。这时典型的假设检验问题表示如下：

$$H_0 : \mu = 0 \leftrightarrow H_1 : \mu > 0.$$

这时, 两样本的位置比较相当于中位数之间的比较, 即如果$\mu > 0$,则Y的取值平均来讲比X大。假设分布函数是连续的，在分布函数上的表现为:给定c，如果$F_1(c) \geqslant F_2(c)$, 那么$1 - F_1(c) \leqslant 1 - F_2(c)$, 有$P(X > c) \leqslant P(Y > c)$. 这也就是说：

$$\begin{aligned} P(Y < X) &= \int_{-\infty}^{+\infty} \int_{-\infty}^{x} \mathrm{d}[F(y-\mu)F(x)] \\ &= \int_{-\infty}^{+\infty} F(x-\mu)\ \mathrm{d}F(x) \\ &\leqslant \int_{-\infty}^{+\infty} F(x)\ \mathrm{d}F(x) = \frac{1}{2}. \end{aligned}$$

对于两样本中位数位置检验,本章将介绍两种常用的分析方法：Brown-Mood中位数检验和Mann-Whitney秩和检验.讨论ROC曲线和Mann-Whitney统计量的关系，引入置换检验的相关概念。

对尺度参数问题, 假设

$$X_1, X_2, \cdots, X_m \sim F\left(\frac{x-\mu_1}{\sigma_1}\right), Y_1, Y_2, \cdots, Y_n \sim F\left(\frac{x-\mu_2}{\sigma_2}\right).$$

F处处连续, 且$X_1, X_2, \cdots, X_m, Y_1, Y_2, \cdots, Y_n$相互独立.

假设检验问题为

$$H_0 : \sigma_1 = \sigma_2 \leftrightarrow H_1 : \sigma_1 \neq \sigma_2.$$

对于两样本尺度参数的检验, 本章将介绍两种方法：Mood方法和Moses方法.

§3.1 Brown-Mood中位数检验

1. 假设检验问题

Brown-Mood中位数检验是由布朗(Brown,1948∼1951)和沐德(Mood,1950)提出的，该方法用于检验两组数据的中位数是否相同，该检验有时也称为Westernberg-Mood 检验，也可以视作是Fisher 精确性检验的一种特殊形式。假设$X_1, X_2, \cdots, X_m$, $Y_1, Y_2, \cdots, Y_n$ 是两组相互独立的样本，来自两个分布$F(x)$ 和$F(y-\mu)$，有相应的中位数med_X和med_Y. 假设检验问题为

$$H_0 : \mathrm{med}_X = \mathrm{med}_Y \leftrightarrow H_1 : \mathrm{med}_X > \mathrm{med}_Y. \tag{3.1}$$

在零假设之下, 如果两组数据有相同的中位数, 则将两组数据混合后, 两组数据的混合中位数med_{XY}与med_X 和med_Y相等, 两组数据应该比较均匀地分布在med_{XY}两侧。因此, 与符号检验类似, 检验的第一步是找出混合数据的样本中位数M_{XY}, 将X和Y按照分布在M_{XY}的左右两侧分为四类, 对每一类计数, 形成2×2列联表, 如表3.1所示.

表3.1 X和Y按照分布在M_{XY}两侧计数表

	X	Y	总和
$> M_{XY}$	A	B	t
$< M_{XY}$	C	D	$(m+n)-(A+B)$
总和	m	n	$m+n \equiv A+B+C+D$

令A, B, C, D表示上述列联表中4个类别的样本点数, A表示左上角取值，即X样本中大于M_{XY}的点数. t表示混合样本中大于M_{XY}的样本点的个数, 它依赖于$m+n$的奇偶性。当m, n 和t 固定时, A 的分布在零假设下是超几何分布：

$$P(A=k) = \frac{\begin{pmatrix} m \\ k \end{pmatrix}\begin{pmatrix} n \\ t-k \end{pmatrix}}{\begin{pmatrix} m+n \\ t \end{pmatrix}}, k \leqslant \min\{m, t\}.$$

在给定m,n 和t 时, 若A的值太大, 可以考虑拒绝零假设, 接受单边检验($H_1: M_X > M_Y$). 同理, 可以得到另外一个单边检验($H_1: M_X < M_Y$)和双边检验的解决方案, 如表3.2所示.

表3.2 Brown-Mood中位数检验的基本内容

零假设：H_0	备择假设：H_1	检验统计量	p 值
$H_0: M_X = M_Y$	$H_1: M_X > M_Y$	A	$P_{\text{hyper}}(A \geqslant a)$
$H_0: M_X = M_Y$	$H_1: M_X < M_Y$	A	$P_{\text{hyper}}(A \leqslant a)$
$H_0: M_X = M_Y$	$H_1: M_X \neq M_Y$	A	$P_{\text{hyper}}(A \leqslant c) + P_{\text{hyper}}(A \geqslant c')$
对水平α, 如果p 值$< \alpha$, 拒绝H_0; 否则，不能拒绝			

例3.1 为研究两不同品牌同一规格显示器在某市不同商场的零售价格是否存在差异, 收集了出售A品牌的9家商场的零售价格数据(单位：元)和出售B品牌的7家商场的零售价格数据, 列表如下(见表3.3).

表3.3 两种同品牌显示器在不同商场的零售价格

A品牌:	698	688	675	656	655	648	640	639	620
B品牌:	780	754	740	712	693	680	621		

解 首先计算混合样本中位数： $M_{XY} = 676.5$, 得到如表3.4所示列联表：

表3.4 两种显示器价格按分布在零售价格中位数两侧的计数表

	X样本	Y 样本	总和
观测值大于M_{XY} 的数目	2	6	8
观测值小于M_{XY} 的数目	7	1	8
总和	9	7	16

在比较不同商场显示器零售价格的例3.1中, $A = 2$, 备择检验是$H_1: M_X < M_Y$. 作单边检验时, p值为$P(A \leqslant 2) = 0.0203$. 这个$p$值相当小, 因而拒绝零假设. 对于两个方差相等的正态总体, 该检验相对于t检验的ARE为$2/\pi = 0.637$,对比符号检验相对于t检验的ARE $= 2/\pi$,二者相等, 这表明它和单样本情况的符号检验效率相当.

这个检验统计量也常常用于构造$\theta = M_X - M_Y$的置信区间。如果假设X与$Y - \theta$独立同分布，这表示在位置漂移(location shifting)假设成立的条件下，θ的置信水平为$1-\alpha$的置信区间可以从下列区间产生：

$$Y_{(t-c'+1):n} - X_{c':m} \leqslant \theta \leqslant Y_{t-c:n} - X_{(c+1):m}.$$

这里的c和c'满足$\Pr[A \leqslant c] + \Pr[A \geqslant c'] = \alpha$.

2. 大样本检验

大样本的时候, 在零假设下, 可以利用超几何分布的正态近似进行检验：

$$Z = \frac{A - mt/(m+n)}{\sqrt{mnt(m+n-t)/(m+n)^3}} \xrightarrow{\mathcal{L}} N(0,1).$$

小样本时, 也可以使用连续性修正

$$Z = \frac{A \pm 0.5 - mt/(m+n)}{\sqrt{mnt(m+n-t)/(m+n)^3}} \xrightarrow{\mathcal{L}} N(0,1).$$

例3.2(例3.1续)　用R语言编写的程序计算p值为0.02, 结论与用精确分布检验一致.

在R中编写计算Brown-Mood中位数检验的程序：

```
BM.test<-function(x, y, alt)
{   xy    <- c(x, y)
    md.xy <- median(xy)
    t     <- sum(xy> md.xy)
    lx    <- length(x)
    ly    <- length(y)
    lxy   <- lx + ly
    A     <- sum(x > md.xy)
    if(alt == "greater")
        { w <- 1-phyper(A, lx, ly, t)}
    else if (alt == "less")
        { w <- phyper(A, lx, ly, t) }
    conting.table_matrix(c(A, lx-A, lx, t-A, ly-(t-A),
    ly, t, lxy-t, lxy), 3, 3)
    col.name<-c("X", "Y", "X+Y")
    row.name<-c(">MXY", "<MXY", "TOTAL")
    dimnames(conting.table)<-list(row.name,col.name)
    list(contingency.table=conting.table, p.value = w)
}
```

输出结果如下：

```
 > BM.test(X,Y,"less")
$contingency.table:
      X Y X+Y
```

```
 >MXY 2 6   8
 <MXY 7 1   8
TOTAL 9 7  16

$p.value:
[1] 0.02027972
```

值得注意的是, 我们这里虽然只给出了中位数的检验, 但是任意p分位数M_p的检验都是类似的, 只是大于M_p 的t不再是$\frac{m+n}{2}$,而是$(m+n)(1-p)$.其他结果都是类似的, 请读者试完成习题。

§3.2 Wilcoxon-Mann-Whitney秩和检验

1. 无结点Wilcoxon-Mann-Whitney秩和检验

前面的Brown-Mood检验与符号检验的思想类似, 仅仅比较了两组数据的符号, 与单样本的Wilcoxon符号秩检验类似, 也可利用更多的样本信息。这里假定两总体分布有类似形状, 不假定对称, 即样本$X_1, X_2, \cdots, X_m \sim F(x-\mu_1)$ 和$Y_1, Y_2, \cdots, Y_n \sim F(x-\mu_2)$,检验问题为

$$H_0 : \mu_1 = \mu_2 (\mu = \mu_1 - \mu_2 = 0) \leftrightarrow H_1 : \mu_1 \neq \mu_2 (\mu = \mu_1 - \mu_2 \neq 0). \tag{3.2}$$

把样本$X_1, X_2, \cdots, X_m$ 和$Y_1, Y_2, \cdots, Y_n$ 混合在一起, 将$m+n$ 个数按照从小到大的顺序排列起来. 每一个Y观测值在混合排列中都有自己的秩. 令R_i 为Y_i 在这N个数中的秩(即Y_i 是第R_i小的). 令I_m和I_n分别表示两样本的指标集, 则

$$R_i = \#\{X_j < Y_i, j \in I_m\} + \#\{Y_k \leqslant Y_i, k \in I_n\}.$$

显然如果这些秩的和 $W_Y = \sum_{i=1}^{n} R_i$过小, 则Y样本的值从平均的意义上来看偏小, 这时可以怀疑零假设。同样, 对于X样本也可以得到W_X. 称W_Y或W_X为 Wilcoxon**秩和统计量**(Wilcoxon rank-sum statistics)。

根据单样本的Wilcoxon符号秩检验可知

$$W_Y = \sum_{i=1}^{n} R_i = \#\{X_j < Y_i, j \in I_m, i \in I_n\} + \frac{n(n+1)}{2}.$$

记

$$W_{XY} = \#\{X_j < Y_i, j \in I_m, i \in I_n\},$$
$$W_{YX} = \#\{Y_i < X_j, j \in I_m, i \in I_n\}.$$

W_{XY} 表示混合样本中Y 观测值大于X观测值的个数. 它是对Y相对于X的秩求和.

$$W_Y = W_{XY} + \frac{n(n+1)}{2}, \tag{3.3}$$

$$W_X = W_{YX} + \frac{m(m+1)}{2}. \tag{3.4}$$

而$W_X + W_Y = \dfrac{(n+m)(n+m+1)}{2}$, 于是有

$$W_{XY} + W_{YX} = nm.$$

在零假设之下, W_{XY}与W_{YX}同分布, 它们称为Mann-Whitney **统计量**. 从式(3.3)和式(3.4)中我们发现，Wilcoxon秩和统计量与Mann-Whitney 统计量是等价的. 事实上, Wilcoxon秩和检验于1945年首先由威尔科克森(Wilcoxon)提出, 主要针对两样本量相同的情况. 1947年, 曼(Mann)和惠特尼(Whitney)又在考虑到不等样本的情况下补充了这一方法. 因此, 也称两样本的秩和检验为Wilcoxon-Mann-Whitney检验(简称W-M-W检验). 事实上, Mann-Whitney 检验还被称为Mann-Whitney U检验, 原因是W_{XY}可以化为U统计量. 为了解零假设下W_Y或W_X的分布性质, 给出有关R_i的以下定理.

定理3.1 在零假设下,

$$P(R_i = k) = \frac{1}{n+m},\ k = 1, 2, \cdots, n+m;$$

和

$$P(R_i = k, R_j = l) = \begin{cases} \dfrac{1}{(n+m)(n+m-1)}, & k \neq l, \\ 0, & k = l. \end{cases}$$

由此容易得到

$$E(R_i) = \frac{n+m+1}{2},$$

$$\mathrm{var}(R_i) = \frac{(n+m)^2 - 1}{12},$$

$$\mathrm{cov}(R_i, R_j) = -\frac{n+m+1}{12},\ i \neq j.$$

由于$W_Y = \sum_{i=1}^{n} R_i$以及$W_Y = W_{XY} + n(n+1)/2$, 有

$$E(W_Y) = \frac{n(n+m+1)}{2},\ \mathrm{var}(W_Y) = \frac{mn(n+m+1)}{12}.$$

及

$$E(W_{XY})=\frac{mn}{2},\ \mathrm{var}(W_{XY})=\frac{mn(n+m+1)}{12}.$$

这些公式是计算Mann-Whitney-Wilcoxon统计量的分布和p 值的基础.

定理3.2 在零假设下, 若$m,n\to+\infty$, 且$\dfrac{m}{m+n}\to\lambda,(0<\lambda<1)$, 有

$$Z=\frac{W_{XY}-\dfrac{mn}{2}}{\sqrt{\dfrac{mn(m+n+1)}{12}}}\xrightarrow{\mathcal{L}}N(0,1),\tag{3.5}$$

$$Z=\frac{W_X-\dfrac{m(m+n+1)}{2}}{\sqrt{\dfrac{mn(m+n+1)}{12}}}\xrightarrow{\mathcal{L}}N(0,1).\tag{3.6}$$

对于双边检验, 令$K=\min\{W_X,W_Y\}$, 此时, K可以通过正态分布$N(a,b)$求得任意点的分布函数, a,b由式(3.5)和式(3.6)确定. 在显著性水平为α下，检验的拒绝域为

$$2P_{\mathrm{norm}}(K<k|a,b)\leqslant\alpha.$$

式中， k是满足上式的最大的k。 也可以通过计算统计量K的p值做决策, 即p值$=2P_{\mathrm{norm}}(K<k|a,b)$.

例3.3 研究不同饲料对雌鼠体重增加是否有差异,数据表如表3.5所示.

表3.5 喂不同饲料的两组雌鼠在8周内增加的体重

饲料	鼠数	各鼠增加的体重/g											
高蛋白	12	134	146	104	119	124	161	107	83	113	129	97	123
低蛋白	7	70	118	101	85	112	132	94					

解 假设检验问题如下：

$$H_0:\mu_1=\mu_2\leftrightarrow H_1:\mu_1\neq\mu_2.\tag{3.7}$$

先将两组数据混合从小到大排列, 并注明组别与秩，如表3.6 所示.

表3.6 两样本W-M-W秩和检验表

体重/g	70	83	85	94	97	101	104	107	112	113
组别	低	高	低	低	高	低	高	高	低	高
秩	1	2	3	4	5	6	7	8	9	10
体重/g	118	119	123	124	129	132	134	146	161	
组别	低	高	高	高	高	低	高	高	高	
秩	11	12	13	14	15	16	17	18	19	

令Y为低蛋白组, $n=7$, X为高蛋白组, R_i是低蛋白组在混合样本中的秩：

$$W_Y=\sum_{i=1}^{m}=1+3+4+6+9+11+16=50.$$

根据式(3.3), 可计算出$W_{XY}=W_Y-\dfrac{n(n+1)}{2}=50-7\times 8/2=22$. 当$m=12$,$n=7$时正态分布的临界值$q_{0.05}$为46, 或直接计算$W_{XY}=22$的$p$值，R程序计算后可得：$p=0.1003>0.05$, 没有显著性差异. 下面是R程序和输出：

```
weight.low=c(134,146,104,119,124,161,107,83,113,129,97,123)
weight.high=c(70,118,101,85,112,132,94)
wilcox.test(weight.high, weight.low)
        Wilcoxon rank sum test
data:  weight.high and weight.low
W = 22, p-value = 0.1003
alternative hypothesis: true location shift is not equal to 0
```

例3.4 Richard 2005年给出一个例子, 是关于服用某类药物对被试者视觉刺激反应时间的影响研究. 研究者随机将8名被试者放在实验条件下, 7名放在控制条件下, 用毫秒记录被试者的视觉反应时间. 这里测量上的问题是, 反应时间受其他不可测且不能忽略的因素(比如个体潜在反应时间或个体解决问题的时间差异)影响, 于是我们的测量可能会是有偏的. 数据有偏的直接结果是反应时间虽然不可能小于0但可能会无穷大. 这是信息不充分的典型情况. 测量数据如表3.7所示.

表3.7 计算Mann-Whitney U统计量

实验组		控制组		
时间/ms	秩	时间/ms	秩	
140	4	130	1	
147	6	135	2	
153	8	138	3	$U=mn+\dfrac{m(m+1)}{2}-R_1$
160	10	144	5	$=8\times 7+\dfrac{8\times(8+1)}{2}-81$
165	11	148	7	$=56+\dfrac{72}{2}-81$
170	13	155	9	$=56+36-81$
171	14	168	12	$=11$
193	15			
$R_1=81$		$R_2=39$		
$m=8$		$n=7$		

零假设：两组秩之间的差异是偶然产生的.

备择假设：两组秩之间的差异不是偶然产生的.

检验统计量：Mann-Whitney U检验统计量.

显著性水平：$\alpha = 0.05$.

样本量：$m = 8, n = 7$.

拒绝零假设的临界值：$U \leqslant 11$ 或$U \geqslant 46$. 如果U在两个临界值以外, 就拒绝零假设. 因为本例中$U = 11$,所以拒绝零假设.

2. 带结点时的计算公式

当X和Y中有相同数值时, 也就是说数据有结, 如用$(\tau_1, \tau_2, \cdots, \tau_g)$表示混合样本的结, 则相同的数据采用平均秩(如果数字相同则取平均秩). 此时, 大样本近似的Z应修正为

$$Z = \frac{W_{XY} - mn/2}{\sqrt{\dfrac{mn(m+n+1)}{12} - \dfrac{mn\left(\sum\limits_{i=1}^{g}\tau_i^3 - \sum\limits_{i=1}^{g}\tau_i\right)}{12(m+n)(m+n-1)}}}.$$

式中，τ_i是第i个结的结长；而g是所有结的个数。

关于Wilcoxon秩和检验(Mann-Whitney 检验), 可总结如表3.8 所示.

表3.8 Wilcoxon秩和检验(Mann-Whitney 检验)表

零假设：H_0	备择假设：H_1	检验统计量(Z)	p 值
$H_0: M_X = M_Y$	$H_1: M_X > M_Y$	W_{XY}或W_Y	$P(Z \leqslant z)$
$H_0: M_X = M_Y$	$H_1: M_X < M_Y$	W_{YX}或W_X	$P(Z \leqslant z)$
$H_0: M_X = M_Y$	$H_1: M_X \neq M_Y$	$\min(W_{YX}, W_{XY})$或$\min(W_X, W_Y)$	$2P(Z \leqslant z)$
大样本时, 用上述近似正态统计量计算p值			

这里虽然从表面看上去是按照备择假设的方向选择W_X或W_Y 作为检验统计量, 但是, 实际上往往是按照实际观察的W_X和W_Y的大小来确定备择假设. 在选定备择假设之后, 比如$H_1: M_X > M_Y$,我们之所以选W_Y或W_{XY}作为检验统计量, 是因为它们的观测值比W_X或W_{YX}的小, 因而计算或查表(表只有一个方向)要方便些. 如果利用大样本正态近似, 则可以选择任意一个作为检验统计量.

3. $M_X - M_Y$ 的点估计和区间估计

$M_X - M_Y$的点估计很简单, 只要把X和Y的观测值成对相减(共有mn对), 然后求它们的中位数即可. 就例3.3来说, 差$M_X - M_Y$的点估计为18.5.

如果想求$\theta \equiv M_X - M_Y$ 的$100(1-\alpha)\%$置信区间, 有以下两种方法.

(1) 将$\theta = M_X - M_Y$作为待估计参数, 用Bootstrap方法分别估计M_X和M_Y, 取得二者的差, 得到Bootstrap $\hat{\theta}^*$, 求出$\hat{\theta}$的方差, 再用第2章的方法求解. 以下给出求$M_X - M_Y$的$100(1-\alpha)\%$置信区间的R参考程序:

```
x1=firstsample
x2=secondsample
n1=length(x1)
n2=length(x2)
th.hat=median(x2)-median(x1)
B=1000
Tboot=   #vector of length Bootstrap
for (i in 1:B)
{
xx1=  #sample of size n1 with replacement from x1
xx2=  #sample of size n2 with replacement from x2
Tboot[i]=median(xx2)-median(xx1)
}
se=sd(Tboot)
Normal.conf=c(qnorm(Tboot,0.025),qnorm(Tboot,0.025))
Percentile.conf=c(quantile(Tboot,0.025),quantile(Tboot,0.975))
Provotal.conf=(2*th.hat+quantile(Tboot,0.025),
+2*th.hat-quantile(Tboot,0.025))
```

(2) 计算X与Y的差, 求排序后的中位数, 具体步骤如下:

① 得到所有mn个差$X_i - Y_j$.

② 记按升幂次序排列的这些差为$D_1, D_2, \cdots, D_N,\ N = mn$.

③ 从表中查出$W_{\alpha/2}$, 它满足$P(W_{XY} \leqslant W_{\alpha/2}) = \alpha/2$, 则所要的置信区间为$(D_{W_{\alpha/2}},\ D_{mn+1-W_{\alpha/2}})$.

在例3.3中$(N = 12 \times 7 = 84)$, 如果要求$\Delta = M_X - M_Y$的95%置信区间, 有$\alpha/2 = 0.025$; 对于$m = 12, n = 7$,查置信区间表得$W_{0.025} = 10$. 再找出$D_{19} = -3$及$D_{84+1-19} = D_{66} = 42$. 因此, 区间$(-3, 42)$ 为所求的$\Delta = M_X - M_Y$的95%置信区间.

对于差异具有统计意义的两组呈正态分布的样本来说, W-M-W检验相对于两样本的t检验的渐近相对效率是0.955; 而对于总体为非正态分布(例如非对称分布)的样本来说, W-M-W检验比两样本t 检验的效率高得多, 事实上这时的渐近相对效率能高达无穷大, 所以W-M-W方法对于两样本的检验是十分适用的.

§3.3 Mann-Whitney U统计量与ROC曲线

在机器学习中常常要建立二分类学习器，常用测试数据检测学习器的性能。一

个学习器的学习性能用ROC曲线表示，ROC的全称为Receiver Operating Characteristic，也称做受试者操作特征。ROC曲线最早由彼得森(Peterson)和博兹奥(Birdsall)于1953年提出，并用于军事领域，后来逐步运用到医学等领域。它的基本原理是首先对学习器产生的得分从大到小排序，依次将每个测试数据点选为一个二分类阈值，得到一对(正确率(TPR), 假阳率(FPR))值，其中TPR 表示把正例(+) 预测为正例(+)的数据比例，而FPR 表示把负例(−)预测为正例(+) 的数据比例，将TPR 设为y轴，将FPR 设为x 轴。在坐标(0,0)处将所有的样例全部预测为负例，这时正确率(TPR)和假阳率均为0. 在由FPR 和TPR 构成的直角坐标系里，如果该曲线和横轴所夹的面积较大，那么表示该学习器的学习性能较好。这里有一个统计问题，就是如何计算该曲线与横轴所辖面积。假设测试数据的数据量为n, 其中正例为e笔，负例为e' 笔，$e + e' = n$. 只有当TPR 上升时，ROC曲线与横轴之间才会有新增面积。也就是说，新增面积只与正例作为阈值有关。如果将第i 个正例(+) 设置为当前阈值，将f_i 个负例(-)预测为正例(+)，此时新增面积如图3.1 所示(截图引自S.J.Mason,2002)：

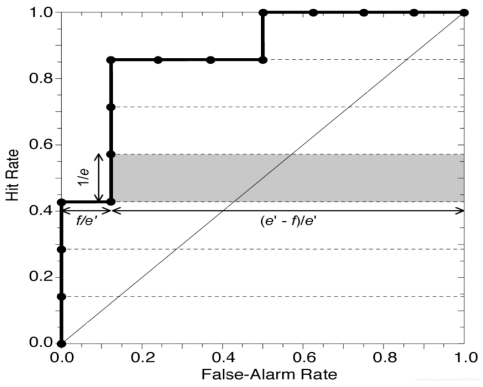

图 3.1 ROC面积计算图

由图3.1可知，新增ROC曲线下的面积为

$$\text{新增面积} = \frac{e' - f_i}{e'e};$$

整个ROC曲线下的面积AUC表示为

$$AUC = \frac{1}{e'e}\sum_{i=1}^{e}(e' - f_i) = 1 - \frac{1}{e'e}\sum_{i=1}^{e} f_i; \tag{3.8}$$

要想求出ROC下的曲线面积(公式3.8)，需要解出$F=\sum_{i=1}^{e}f_i$,，其计算公式如下：

$$F=\sum_{i=1}^{e}f_i=\sum_{i=1}^{e}(e'-r_i)=e'e-\sum_{i=1}^{e}r_i. \tag{3.9}$$

式(3.9)中r_i表示在该测试集中第i个正例得分高于负例得分的点数，也是每个正样本在得分混合序列中的相对秩，假设不存在得分相等的结，将这些信息代入到AUC的计算公式中得到：

$$AUC=1-\frac{1}{e'e}\sum_{i=1}^{e}f_i=1-\frac{1}{e'e}\sum_{i=1}^{e}r_i=\frac{U}{e'e}. \tag{3.10}$$

也就是说，AUC与Mann-Whitney U统计量的大小是等价的，AUC值越大就意味着Mann-Whitney U统计量的值越大，它们之间只差一个归一化参数$e'e$.

例3.5 ：假设已经得出一系列样本被划分为正类的概率（得分），按从大到小排序如表3.9所示，表中共有20个测试样例，“分类”一栏表示每个测试样例真实的类别标签（+表示正样例，– 表示负样例），“得分”表示每个测试样本属于正样的概率。

表3.9 测试样例的真实分类和得分数据表

ID	分类	得分	ID	分类	得分
1	+	0.9	11	+	0.4
2	+	0.8	12	−	0.39
3	−	0.7	13	+	0.38
4	+	0.6	14	−	0.37
5	+	0.55	15	−	0.36
6	+	0.54	16	−	0.35
7	−	0.53	17	+	0.34
8	−	0.52	18	−	0.33
9	+	0.51	19	+	0.30
10	−	0.505	20	−	0.10

接下来，从得分由高向低，依次将“得分”值所对应的样本作为阈值，当测试样本属于正样本的概率大于或等于这个阈值时，我们认为它是正样本，否则为负样本。举例来说，对于表3.9 的第4个样本，其得分值为0.6，那么样本1，2，3，4都被认为是正样本，因为它们的得分值都大于等于0.6，而其他样本则都认为是负样本。每次选取一个不同的样本点作为阈值，就可以得到一组FPR 和TPR，即ROC曲线上的一点。这样一来，共计可以产生20组FPR和TPR 的值，将它们画在ROC曲线上，结果如图3.2所示：

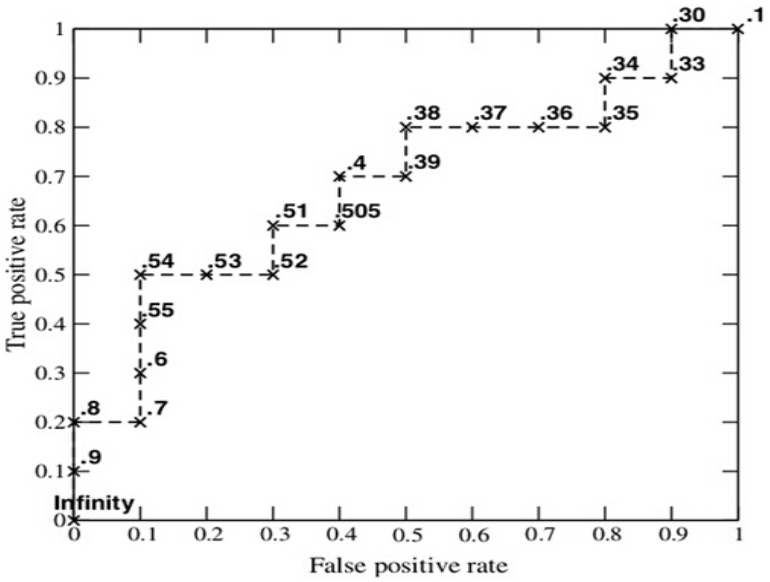

图3.2 ROC曲线示意图

由图3.2，首先由数据的得分秩根据两类的符号计算出正样例相对于负样例的秩和如下：

$$U = 1 + 2 + 5 + 6 + 7 + 9 + 9 + 9 + 10 + 10 = 68.$$

结合公式(3.10)，计算出AUC面积为0.68,表示此时该二分类器性能比较好。

§3.4 置换检验

置换检验(Permutation Test)是一种非参数检验，可以用来检验两个分布是否相同，它不基于大样本渐近理论，主要用于小样本。假设$X_1, X_2, \cdots, X_{n_1} \sim F_X$和$Y_1, Y_2, \cdots, Y_{n_2} \sim F_Y$是两个独立样本。零假设是两个样本来自同一个分布，比如，交通事故中驾驶员受伤程度与是否使用安全带的分布是否有不同。具体而言，这里的统计假设检验问题是：

$$H_0 : F_X = F_Y \Leftrightarrow H_1 : F_X \neq F_Y.$$

令$T(x_1, x_2, \cdots, x_{n_1}, y_1, y_2, \cdots, y_{n_2})$ 是一个检验统计量，常用两组数据的位置差来表示，

$$T(x_1, \cdots, x_{n_1}, y_1, \cdots, y_{n_2}) = |\bar{X}_{n_1} - \bar{Y}_{n_2}|.$$

考虑零假设，两组数据混合在一起就是一个分布，这样X和X之间，Y与Y 之间以及X 与Y 之间均可以互相置换，于是可以考虑由该数据形成的$(n_1 + n_2)!$ 种置

换，对每一种置换计算统计量T, 形成统计量的置换样本$T_1, T_2, \cdots, T_N$，其中$N = (n_1+n_2)!, T$ 取每种置换的可能性是$1/N$, 用$\mathbb{P}_p$表示T 的置换分布（permutation distribution）.令t 表示检验统计量的观测值，如果T很大，拒绝零假设，那么置换检验的p 值为：

$$p = \mathbb{P}_p(T > t_0) = \frac{1}{N!}\sum_{j=1}^{N!} I(T_j > t_0).$$

实际中，如果N比较大，把N!种不同的置换都试验一遍是不现实的，可以从置换集中随机抽取数据，计算近似的p 值，以下是置换检验p 值的Bootstrap 计算方法：

置换检验p值计算方法

(1).计算检验统计量的观测值$t_0 = T(X_1, X_2, \cdots, X_{n_1}, Y_1, Y_2, \cdots, Y_{n_2})$;
(2).随机置换数据，用置换数据再次计算检验统计量的值B 次，令$T_1, \cdots, T_B$ 表示置换样本后的T 的观察值;
(3).近似的p值为

$$\frac{1}{B}\sum_{j=1}^{B} I(T_j > t_0). \tag{3.11}$$

例3.6 西格尔(Siegel,1956)在家教育和学校教育是两种教育方式，两种教育方式对幼儿社会交往能力有怎样的差异并不明确。已有教育学理论支持将孩子送入幼儿园将有助于提升孩子的社会认知能力，设置如下假设：

零假设H_0：送入学校的孩子和在家教育的孩子的社会交往能力没有差异；
备择假设H_1：送入学校的孩子的社会交往认知能力高于在家教育的孩子的社会交往能力。

数据 ：实验对象是幼儿园适龄双胞胎，共计$n = 8$对，研究期初每一对随机选出一位送到幼儿园，另一位在家教育，研究期结束后，16 个孩子统一接受同一套认知能力测试，测试数据如表3.10 所示（分值高代表社会交往能力强）：

表3.10 两种不同教育方式认知能力测试得分表

幼儿园教育(x)	82	69	73	43	58	56	76	65
在家教育(y)	63	42	74	37	51	43	80	62
两者分值差异d	19	27	-1	6	7	13	-4	3

解 这是配对两样本位置检验问题，假设模型表达为：

$$d_i = me + \epsilon_i, i = 1, 2, \cdots, 8.$$

me是位置中心，用Wilcoxon检验p值=0.027.现在考虑计算置换检验的p 值, 程序如下：

```
> d=school-home
> dpm=c(d,-d)
> n=length(d)
> B=500
> dbs=matrix(sample(dpm,n*B,replace=TRUE),ncol=n)
> wilcox.teststat=function(x)wilcox.test(x)$statistic
> bs.teststat=apply(dbs,1,wilcox.teststat)
> mean(bs.teststat)>=wilcox.teststat(d))
[1] 0.0238
```

§3.5 Mood方差检验

对于尺度参数的检验, 它与两样本的位置参数有关, 如果不知道位置参数, 则一般很难通过秩检验判断两组数据的离散程度. 比如下面两组数据：

表3.11 两组独立样本实验数据

样本1	48	56	59	61	84	87	91	95
样本2	2	22	49	78	85	89	93	97

观察数据可以看出, 第二组数据比第一组数据分散, 但从秩的角度却很难区分. 所以Mood检验法假定两位置参数相等. 不失一般性, 假定为零. 于是有样本$X_1, X_2, \cdots, X_m \sim F\left(\dfrac{x}{\sigma_1}\right)$ 和$Y_1, Y_2, \cdots, Y_n \sim F\left(\dfrac{x}{\sigma_2}\right)$, 我们的检验问题为

$$H_0 : \sigma_1^2 = \sigma_2^2 \leftrightarrow \quad H_1 : \sigma_1^2 \neq \sigma_2^2.$$

F处处连续, 且$X_1, X_2, \cdots, X_m, Y_1, Y_2, \cdots, Y_n$相互独立,令$R_i$为$X_i$在混合样本中的秩, 当$H_0$成立时, $X_1, X_2, \cdots, X_m, Y_1, Y_2, \cdots, Y_n$独立同分布,

$$E(R_i) = \sum_{i=1}^{m+n} \frac{i}{m+n} = \frac{m+n+1}{2}.$$

当H_0成立时, 对样本X来说, 考虑秩统计量：

$$M = \sum_{i=1}^{m} \left(R_i - \frac{m+n+1}{2} \right)^2. \tag{3.12}$$

如果它的值偏大, 则X的方差也可能偏大. 可以对大的M 拒绝零假设. 这种方法由Mood于1954年提出, 称为 Mood**检验**.

在零假设H_0下, M的分布可以由秩的分布性质得出. 这里给出大样本近似. 在零假设下, 当$m,n\to\infty$ 并且$m/(m+n)$趋于常数时, 有

$$E(M)=m(m+n+1)(m+n-1)/12, \tag{3.13}$$

$$\mathrm{var}(M)=mn(m+n+1)(m+n+2)(m+n-2)/180, \tag{3.14}$$

$$Z=\frac{M-E(M)}{\sqrt{\mathrm{var}(M)}}\xrightarrow{\mathcal{L}}N(0,1). \tag{3.15}$$

当样本量比较小时, 比如$m+n<30$, 可以用连续性修正：

$$Z=\frac{M-E(M)\pm 0.5}{\sqrt{\mathrm{var}(M)}}\xrightarrow{\mathcal{L}}N(0,1), \tag{3.16}$$

也可以采用Laubscher等人于1968年的建议, 修正如下：

$$Z=\frac{M-E(M)}{\sqrt{\mathrm{var}(M)}}+\frac{1}{2\sqrt{\mathrm{var}(M)}}\xrightarrow{\mathcal{L}}N(0,1). \tag{3.17}$$

例3.7 假定有5位健康成年人的血液, 分别用手工(x)和仪器(y)两种方法测量血液中的尿酸浓度,测量结果如表3.12所示, 问：两种测量方法的精确度是否存在差异?

表3.12 两种不同血液测量方法的测量结果数据表

手工(x)	4.5	6.5	7	10	12
仪器(y)	6	7.2	8	9	9.8

解 假设检验：

H_0:两种尿酸浓度测量法的方差相同, 即 $\sigma_1^2=\sigma_2^2$;

H_1:两种尿酸浓度测量法的方差不同, 即 $\sigma_1^2\neq\sigma_2^2$.

统计分析：将两样本混合, 计算混合秩如表3.13所示。

表3.13 两样本混合之后的混合秩

尿酸浓度	4.5	6	6.5	7	7.2	8	9	9.8	10	12
秩	1	2	3	4	5	6	7	8	9	10
组别	x	y	x	x	y	y	y	y	x	x

设$m=n=5,(m+n+1)/2=(5+5+1)/2=5.5$。根据式(3.12), 有

$$\begin{aligned}M&=(1-5.5)^2+(3-5.5)^2+(4-5.5)^2+(9-5.5)^2+(10-5.5)^2\\&=61.25\end{aligned}$$

由附表9, $M_{0.025,5,5}=15.25$,$M_{0.975,5,5}=61.25$, $M=61.25$. 由于$15.25<M=61.25<65.25$,故不能拒绝H_0, 表示两种测量法的精度没有明显差异.

若用式(3.13)和式(3.14)分别计算, 则

$$\begin{aligned}E(M)&=m(m+n+1)(m+n-1)/12\\&=5(5+5+1)(5+5-1)/12\\&=41.25,\\\mathrm{var}(M)&=mn(m+n+1)(m+n+2)(m+n-2)/180\\&=5\times5(5+5+1)(5+5+2)(5+5-2)/180\\&=146.6667.\end{aligned}$$

代入式(3.16)得

$$\begin{aligned}Z&=\frac{1}{\sqrt{\mathrm{var}(M)}}\left[M-E(M)+\frac{1}{2}\right]\\&=\frac{1}{146.6667}[61.25-41.25+0.5]\\&=\frac{20.5}{12.1106}\\&=1.6927<Z_{0.05/2}=1.96\end{aligned}$$

所得结论与第一种方法相同.

§3.6 Moses方差检验

Moses于1963年提出了另一种检验两总体方差相等的方法, 该方法不需事先假设两分布平均值相等, 因此应用较广.

设$x_1,x_2,\cdots,x_m$为第一个分布的随机样本, 第一个总体的方差为σ_1^2. 设$y_1,y_2,\cdots,y_n$为第2个分布的随机样本, 第二个总体的方差为σ_2^2.

假设检验：

$$\begin{aligned}&H_0:\text{两分布方差相等, 即}\sigma_1^2=\sigma_2^2;\\&H_1:\text{两分布方差不等, 即}\sigma_1^2\neq\sigma_2^2.\end{aligned}$$

统计分析：

Moses检验法的统计值T求法如下.

(1) 将两样本各分成几组, 如第1组样本随机分成m_1组, 每组含k个观测值, 记为$A_1, A_2, \cdots, A_{m_1}$; 同理第2组样本随机分成$m_2$组, 每组含$k$个观测值,记为$B_1, B_2, \cdots, B_{m_2}$.

(2) 分别求各小组样本的离差平方和如下：

$$\mathrm{SSA}_r = \sum_{x_i \in A_r} (x_i - \bar{x})^2, \quad r = 1, 2, \cdots, m_1;$$
$$\mathrm{SSB}_s = \sum_{y_i \in B_s} (y_i - \bar{y})^2, \quad s = 1, 2, \cdots, m_2.$$

(3) 将两样本各小组的平方和SSA_r, $\mathrm{SSB}_s, r = 1, 2, \cdots, m_1, s = 1, 2, \cdots, m_2$混合, 排序按大小定秩.

(4) 计算第1组样本m_1组平方和的秩和, 用S表示, 则Moses的统计值T_{M}为

$$T_{\mathrm{M}} = \frac{S - m_1(m_1 + 1)}{2}.$$

如果两组数据的方差存在很大的差异, 从平均来看, 一组数据的平方和, 比另一组数据的平方和小, 因此查Mann-Whitney的W_α值表(见附表4), 若实际$T_{\mathrm{M}} < W_{0.025, m_1, m_2}$或$T_{\mathrm{M}} > W_{0.975, m_1, m_2} = m_1 m_2 - W_{0.025}$, 则不能拒绝$H_1$, 反之则接受$H_0$.

例3.8 设中风病人与健康成人血液中尿酸浓度如下：

表3.14 中风病人与健康成人血液中尿酸浓度数据表

病人(x)	8.2	10.7	7.5	14.6	6.3	9.2	11.9	5.6	12.8	5.2	4.9	13.5	$m = 12$
正常人(y)	4.7	6.3	5.2	6.8	5.6	4.2	6.0	7.4	8.1	6.5			$n = 10$

假设检验：

H_0 : 中风病人与健康成人血液的尿酸浓度的变异相同, 即$\sigma_1^2 = \sigma_2^2$;

H_1 : 中风病人与健康成人血液的尿酸浓度的变异不同, 即$\sigma_1^2 \neq \sigma_2^2$.

统计分析：现在将中风病人随机分成4组($m_1 = 4$), 每组3 人($K = 3$), 健康成人分成3组($m_2 = 3$), 每组3人($K = 3$), 多出1人去除. 各组尿酸浓度及其平方和如下.

表3.15 中风病人各组尿酸浓度及其平方和

中风病人(x)	观测值			平方和(SSA)	秩
1	8.2	14.6	11.9	20.65	5
2	10.7	6.3	5.2	16.94	4
3	7.5	5.6	12.8	27.85	6
4	9.2	4.9	13.5	36.98	7

表3.16 正常人各组尿酸浓度及其平方和

正常人(y)	观测值			平方和(SSB)	秩
1	4.7	6.8	6.0	2.25	2
2	6.3	5.6	7.4	1.65	1
3	5.2	4.2	8.1	8.21	3

如果取较小的$S = \min(\mathrm{SSA}, \mathrm{SSB}) = \min(22, 6) = 6$, 则

$$T_M = S - m_2(m_2+1) = 6 - 3(3+1)/2 = 0.$$

查附表4, $W_{0.025,4,3} = 0$, 统计量$T = 0 \leqslant W_{0.025} = 0$, 因此不能拒绝$H_1$. 由于$T = S_1 - m_1(m_1+1)/2 = 22 - 4(4+1)/2 = 12$,$W_{0.975} = m_1 m_2 - W_{0.025} = 4 \times 3 - 0 = 12$, $T = 12 > W_{0.975,4,3} = 12$, 所以接受$H_1$, 认为两组数据的方差不相等.

习题

3.1 在一项研究毒品对增强人体攻击性影响的实验中, 组A使用安慰剂, 组B使用毒品. 实验后进行攻击性测试, 测量得分(得分越高表示攻击性越强)显示如下表.

组A	组B
10	12
8	15
12	20
16	18
5	13
9	14
7	9
11	16
6	

(1) 给出这个实验的零假设.

(2) 画出表现这些数据特点的曲线图.

(3) 分析这些数据用哪种检验方法最合适.

(4) 用你选择的检验对数据进行分析.

(5) 是否有足够的证据拒绝零假设? 如何解释数据?

3.2 试针对例3.1进行如下操作：

(1) 给出0.25分位数的检验内容(包括假设、过程和决策);

(2) 应用(1)的结果分析比较两组数据的0.25分位数是否有差异, 对结果进行合理解释;

(3) 给出0.75分位数的检验内容(包括假设、过程和决策);

(4) 应用(3)的结果分析比较两组数据的0.75分位数是否有差异, 对结果进行合理解释.

3.3 一家大型保险公司的人事主管宣称在人际关系方面受过训练的保险代理人会给潜在顾客留下更好的印象. 为了检验这个假设, 从最近雇用的职员中随机选出22人, 一半人接受人际关系方面的课程训练, 剩下的11个人组成控制组. 在训练之后, 所有的22人都在一个与顾客的模拟会面中被观察, 观察者以20分制(1∼20) 对他们在建立与顾客关系方面的表现进行评级，得分越高, 评级越高. 数据在下表中列出.

受过人际关系的训练组	控制组
18	12
15	13
9	9
10	8
14	1
16	2
11	7
13	5
19	3
20	2
6	4

(1) 这项研究的零假设和备择假设各是什么?

(2) 画出表示这些数据特点的曲线图.

(3) 你认为分析这些数据用哪种检验方法最合适?

(4) 用你选择的检验方法对数据进行分析.

(5) 是否有足够的证据拒绝零假设? 如何解释数据?

3.4 两个不同学院教师一年的课时量分别为(单位：学时)：

A学院：221 166 156 186 130 129 103 134 199 121 265 150 158 242 243 198 138 117;

B学院：488 593 507 428 807 342 512 350 672 589 665 549 451 492 514 391 366 469.

根据这两个样本判断, 两个学院教师讲课的课时是否有不同? 估计其差别. 从两个学院教师讲课的课时来看, 教师完成讲课任务的情况是否类似? 给出检验和判断.

3.5 对A和B两块土壤有机质含量抽检结果如下, 试用Mood和Moses两种方法检验两组数据的方差是否存在差异.

A	8.8 8.2	5.6 4.9	8.9 4.2	3.6 7.1	5.5 8.6	6.3 3.9
B	13.0 14.5	16.5 22.8	20.7 19.6	18.4 21.3	24.2 19.6	11.7
	18.9 14.6	19.8 14.5				

3.6 根据第一章问题1的数据，请选择合适的方法进行中位数检验，比较两者的结果

案例与讨论：等候还是离开？

案例背景

一家中餐厅“师徒帮帮带”坐落在一所著名高等学府内，其提供的套餐虽然只有两类：套餐C 和套餐H，但因为营养全面、质量上乘、口感俱佳而远近闻名，每当午餐时间总是有很多学生来此用餐。该餐厅为保证质量目前只有两个柜台窗口$W1$ 和$W2$，每个窗口指定一位师傅每次接受一个订单，现场独立加工，一个订单加工完成后允许接受下一个订单。过去有商学院的研究生根据餐厅提供的汇总数据分析发现$W1$ 窗口师傅单位时间内能够完成的数量比$W2$窗口的多，其服务更快一些，但新闻学院学生的亲身体验发现$W2$ 窗口比$W1$ 更快一些，对于排$W1$ 窗扣还是排$W2$窗扣的两种不同的排队策略，怎么看？

数据说明与约定

该案例的数据存储在一个“waitingline.xls”文件中，共有三个工作簿，分别涉及到服务时间、数据说明和到达人数。工作簿3为arrivnumber 数据，以20 分钟为单位，收集了连续5 个工作日餐馆午饭高峰期到达人数的数据。

为简化分析，有如下约定：

1. 顾客到达餐厅之前对选择C餐还是H餐是确定的，进入排队后选餐类型不再改变；

2. 每位顾客只点一份餐。如果有柜台没有顾客等待，顾客将优先选择此柜台。如果两个柜台都至少有一个顾客在等待服务，他可以选择进入排队或离开餐馆；

3. 顾客决定选择排队还是离开只受到队伍长度、两个队伍的排队时间，以及自身对排队等待时间的忍耐极限值三个因素决定。

研讨问题

问题1.根据案例背景和数据约定，请思考以下三个问题：

(1). 在不考虑顾客排队时间忍耐值的情况下，$W1$窗口中没有排队的概率，$W2$窗口中没有排队的概率，以及新到的顾客必须排队的概率。

(2). 套餐C和套餐H各自的服务时间有什么统计规律？

问题2. 如果$W1$窗口前有$CHCHC$，$W2$窗口前有$CCCHH$同等数量的顾客在等候接受服务，仅仅考虑排队时间，建议哪只队伍比较合适？理由是什么。

问题3. 如果遇到$W1$窗口的排队序列是$CCCCH$，$W2$窗口前排队序列是$CCHCC$。正好有位刚到的学生只有30分钟可以等待，那么你建议的排队策略是什么（可靠性为75%)。