第4章 两样本位置和尺度检验

两样本位置和尺度检验

假设样本：（ $X_1,\cdots,X_m$)~i.i.d. F $\left(\frac{x-\mu_1}{\sigma_1}\right), X=\mu_1+\sigma_1\varepsilon$

$(Y_1, Y_2, \ldots, Y_n)$~i.i.d. F $\left(\frac{x-\mu_2}{\sigma_2}\right), \mathrm{Y}=\mu_2+\sigma_2\varepsilon, \varepsilon\sim F(x)$

样本之间相互独立，μ_1,μ_2 为位置参数,σ_1,σ_2 称为尺度参数。

$H_0:\mu_1=\mu_2 \leftrightarrow H_1:\mu_1\neq\mu_2$　$H_0:\sigma_1=\sigma_2 \leftrightarrow H_1:\sigma_1\neq\sigma_2$

Brown-Mood 中位数检验	Mann-Whitney 秩和检验。
Mood检验	其他秩方法

位置参数的检验

Brown-Mood中位数检验 $(\sigma_1 = \sigma_2)$

假设($X_1, \cdots, X_m$)~i.i.d.$F(x)$,

$(Y_1, Y_2, \ldots, Y_n)$~i.i.d.$F(x-\mu)$

$$H_0 : med_X = med_Y \leftrightarrow H_1 : med_X > med_Y$$

$$\Leftrightarrow H_0 : \mu = 0 \leftrightarrow H_1 : u < 0$$

原理：在零假设成立时，如果数据有相同中位数，那么混合样本的中位数应该和混合前的相等。

计算和例子

首先将两个样本混合，找出混合样本中位数M_{XY}，将X和Y按照在M_{XY}两侧分类计数，即：

	$>M_{XY}$	$<M_{XY}$	总和
X	A	B	m
Y	C	D	n
总和	t	$m+n-t$	$m+n=A+B+C+D$

在给定m，n和t的时候，在零假设成立时, A的分布服从超几何分布：

$$P(A=k)=\frac{\binom{m}{k}\binom{n}{t-k}}{\binom{m+n}{t}}, k \le m$$

当A值太大时，考虑拒绝零假设。

检验基本内容

H_0	H_1	检验统计量	P-值
$M_x = M_y$	$M_x > M_y$	A	$P_{H_0}(A \geq a)$
$M_x = M_y$	$M_x < M_y$	A	$P_{H_0}(A \leq a)$
$M_x = M_y$	$M_x \neq M_y$	A	$2\min(P_{H_0}(A \geq a), P_{H_0}(A \leq a))$

对于水平 α，如果 p- 值小于 α，那么拒绝零假设

例 4.1: 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：人民币￥），和出售 B 品牌的 7 家商场的零售价格数据，列表如下：：

表 4.3. 两不同品牌显示器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： M_{XY}=676.5

	$>M_{XY}$	$<M_{XY}$	总和
X	2	7	9
Y	6	1	7
总和	8	8	16

比较不同商场显示器零售价格的例 4.1 中， a =2，备择检验是 H_1 : $M_X > M_Y$ 作单边检验时， $p-$ 值为 $\mathrm{P}(A \leq 2) = 0.02027972$. 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，该检验相当于 $t-$ 检验的 ARE 为 $2/\pi = 0.637$ 。这表明它和单样本情况的符号检验同属一类。

```
phyper(2,9,7,8)
[1] 0.02027972
```

大样本

In probability theory and statistics, the **hypergeometric distribution** is a discrete probability distribution that describes the number of successes in a sequence of *n* draws from a finite population *without* replacement, just as the binomial distribution describes the number of successes for draws *with* replacement.

The notation is illustrated by this contingency table:

	drawn	not drawn	total
successes	k	$m - k$	m
failures	$n - k$	$N + k - n - m$	$N - m$
total	n	$N - n$	N

Perhaps the easiest way to understand this distribution is in terms of urn models. Suppose you are to draw "n" balls without replacement from an urn containing "N" balls in total, "m" of which are white. The hypergeometric distribution describes the distribution of the number of white balls drawn from the urn.

A random variable X follows the hypergeometric distribution with parameters N, m and n if the probability is given by

$$P(X = k) = \frac{\binom{m}{k}\binom{N-m}{n-k}}{\binom{N}{n}}.$$

where the binomial coefficient $\binom{a}{b}$ is defined to be the coefficient of x^b in the polynomial expansion of $(1 + x)^a$

The probability is positive when k is between $\max(0, n + m - N)$ and $\min(m, n)$.

The formula can be understood as follows: There are $\binom{N}{n}$ possible samples (without replacement). There are $\binom{m}{k}$ ways to obtain k

Hypergeometric

parameters:	$N \in 0, 1, 2, \ldots$ $m \in 0, 1, 2, \ldots, N$ $n \in 0, 1, 2, \ldots, N$
support:	$k \in \max(0, n+m-N), \ldots, \min(m, n)$
pmf:	$\frac{\binom{m}{k}\binom{N-m}{n-k}}{\binom{N}{n}}$
cdf:	
mean:	$\frac{nm}{N}$
median:	
mode:	$\left\lfloor \frac{(n+1)(m+1)}{N+2} \right\rfloor$
variance:	$\frac{nm(N-n)(N-m)}{N^2(N-1)}$
skewness:	$\frac{(N-2m)(N-1)^{\frac{1}{2}}(N-2n)}{[nm(N-m)(N-n)]^{\frac{1}{2}}(N-2)}$
kurtosis:	$\left[\frac{N^2(N-1)}{n(N-2)(N-3)(N-n)}\right] \cdot \left[\frac{N(N+1)-6N(N-n)}{m(N-m)} + \frac{3n(N-n)(N+6)}{N^2} - 6\right]$

大样本检验

对于大样本情况下，可以使用超几何分布的正态近似进行检验：

$$Z=\frac{A-mt/(m+n)}{\sqrt{mnt(m+n-t)/\left((m+n)^2(m+n-1)\right)}}\to N(0,1)$$

Mann-Whitney秩和检验

假设样本 $X_1, X_2, \cdots, X_m \sim F(x-\mu_1)$

$Y_1, Y_2, \cdots, Y_n \sim F(x-\mu_2)$　　X与Y独立。

假设检验问题：

$$H_0: \mu_1 = \mu_2 \longleftrightarrow H_1: \mu_1 \neq \mu_2$$

将两个样本混合，Y_i在混合样本中的秩：

$$R_i = \sum_{j=1}^{m} I(X_j < Y_i) + \sum_{j=1}^{n} I(Y_j < Y_i) + 1$$

定义 $W_Y = \sum_{i=1}^{n} R_i$，同样可定义W_X，称为Wilcoxon秩和统计量。

W-M-W统计量

$$W_{XY}=\sum_{i,j}(X_j<Y_i) \qquad W_{YX}=\sum_{i,j}(Y_i<X_j)$$

称为Man-Whitney统计量。

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

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

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

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

在零假设情况下，W_{YX} 和W_{XY} 同分布，并且和Wilcoxon秩和统计量W_X 等价。

对单边假设问题 $H_0 : \mu_1 = \mu_2 \leftrightarrow H_1 : \mu_1 > \mu_2$

当统计量W_X 偏大的时候，考虑拒绝零假设。

定理 4.1: 在零假设下,

$$\begin{aligned} \mathrm{P}(R_i = k) &= \frac{1}{n+m},\ k = 1, ..., n+m; \\ \text{和}\ \ \mathrm{P}(R_i = k, R_j = l) &= \begin{cases} \frac{1}{(n+m)(n+m-1)}, & k \neq l, \\ 0, & k = l. \end{cases} \end{aligned}$$

由此容易得到

$$\begin{aligned} \mathrm{E}(R_i) &= \frac{n+m+1}{2}, \\ \mathrm{Var}(R_i) &= \frac{(n+m)^2 - 1}{12}, \\ \mathrm{Cov}(R_i, R_j) &= -\frac{n+m+1}{12},\ (i \neq j). \end{aligned}$$

在零假设下

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

$$D(W_X) = mn(m+n+1)/12$$

W_X的分布关于$m(m+n+1)/2$对称。

在原假设 H_0 为真时，

$$-X_1,\cdots,-X_m,-Y_1,\cdots,-Y_n \text{ i.i.d.},$$

$$R'=(N+1-R_1,\cdots,N+1-R_N) \underline{\underline{d}} R$$

$$W_X=\sum_{i=1}^{m}R_i \quad \underline{\underline{d}} \quad \sum_{i=1}^{m}R_i'=m(N+1)-W_X ,$$

所以 W_X 的分布关于 $m(m+n+1)/2$ 对称。

大样本检验

定理　　在零假设下：若 $m,n\to\infty$ ，且 $\frac{m}{m+n}\to\lambda$，时：

$$Z=\frac{W_X-m(m+n+1)/2}{\sqrt{mn(m+n+1)/12}}\to N(0,1)$$

$$Z=\frac{W_{XY}-mn/2}{\sqrt{mn(m+n+1)/12}}\to N(0,1)$$

对于打结的情况需要使用修正的公式。

$$Z=\frac{W_X-m(m+n+1)/2}{\sqrt{mn(m+n+1)/12-nm\sum_{i=1}^{g}(\tau_i^3-\tau_i)/\left(12(n+m)(n+m+1)\right)}}\to N(0,1)$$

例题

高蛋白和低蛋白饲料对雌鼠体重增加是否有差异（画圈的为低蛋白）

解：假设检验问题如下：

$$H_0 : \mu_1 = \mu_2 \leftrightarrow H_1 : \mu_1 \neq \mu_2$$

先将两组数据混合从小到大排列，并注明组别与秩如下表：

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：

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

当 $m = 12, n = 7$，**检验水平0.05下，秩和检验的双侧** 临界值为 46，则 $p > 0.05$，没有显著性差异。

假设($X_1,\cdots,X_m$)~i.i.d.$F(x)$,

(Y_1, Y_2, … ,Y_n)~i.i.d.$F(x-\mu)$

构造 μ 的置信区间。

mn个X_i-Y_j的顺序统计量记为$Z_{(1)}\le\cdots\le Z_{(mn)}$,

$W_{XY}=\sum_{i,j}I(X_i<Y_j)$，当 $\mu=0$ 时，$W_{XY}=\sum_{i,j}I(X_i<Y_j)$ 分布已知。找 k_1,k_2 使得 $P(k_1\le\sum_{i,j}I(Y_j-X_i-\mu)=W_{XY}\le k_2)=1-\alpha$ ，则

$P(Z_{(mn-k_2)}\le\mu\le Z_{(mn-k_1+1)})\ge1-\alpha$ 。

备择假设的其他提法

$$H_0 : X \overset{d}{=} Y \leftrightarrow H_1 : X > Y$$

H_1的其他提法

(1) $P(X > Y) > 1/2$

(2) $\forall c, F(c) < G(c)$

(3) $X + a \overset{d}{=} Y, a < 0$

(4) $Me_X > Me_Y$

则有

$(3) \Rightarrow (2), (2) \Rightarrow (1),$

$(2) \Rightarrow (4), (1) \not\Rightarrow (4)$

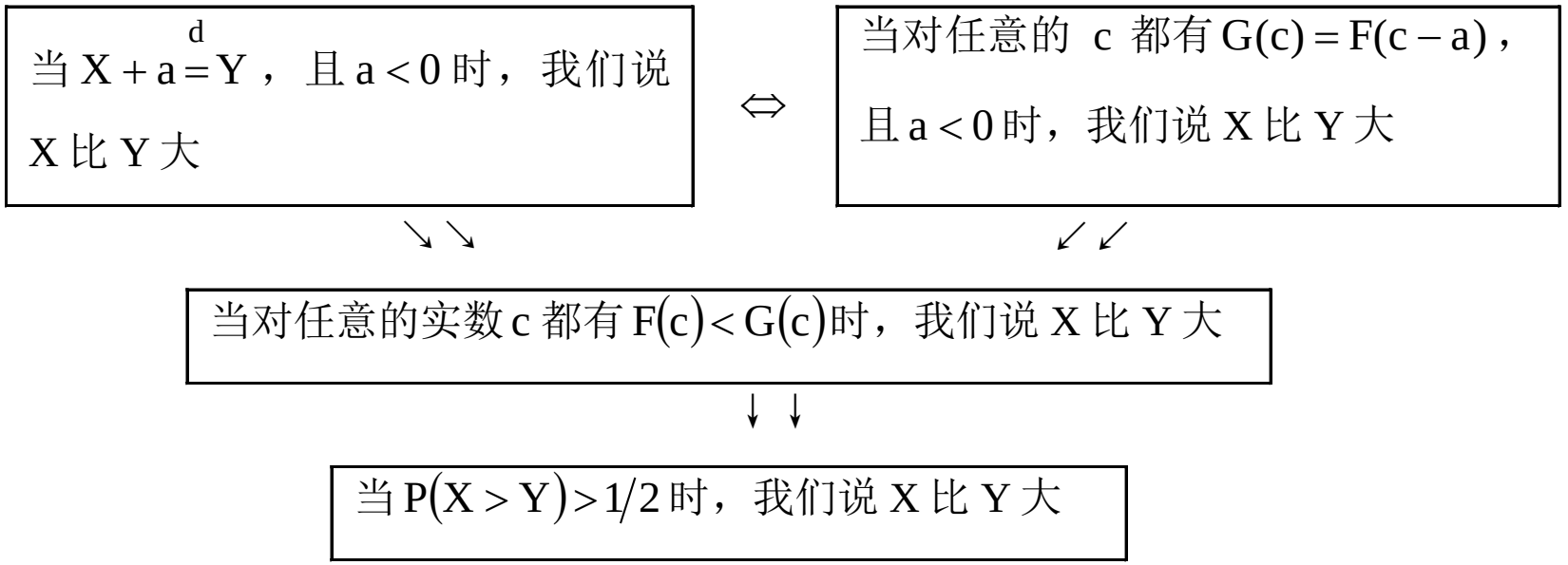

图 随机变量大小关系的定量描述方法之间的关系

尺度参数检验

检验问题以及原理

假定两分布位置参数相等，设 $X_1,\cdots,X_m \sim F(\frac{x}{\sigma_1})$ $Y_1,\cdots,Y_n \sim F(\frac{y}{\sigma_2})$，独立，检验问题：

$$H_0:\sigma_1^2=\sigma_2^2 \leftrightarrow H_1:\sigma_1^2>\sigma_2^2$$

令 R_i 表示 X_i 在混合样本之中的秩，在零假设成立的情况下，有：

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

秩方法的基本思想是，用 X_i 的秩 R_i 代替 X_i 作统计推断。R_i 可理解为 X_i 的得分。定义一个计分函数 $a(r)$，$r=1,2,\cdots,m+n=N$，在 X_i 的秩为 R_i 时，将 X_i 的得分定义为 $a(R_i)$。$\sigma_1^2>\sigma_2^2$ 时，X 样本倾向于排在两边，若计分函数 $a(r)$ 使得在 X_i 的秩 R_i 比较大和比较小的时候，X_i 的得分 $a(R_i)$ 都是比较大，$a(r)$ 是单谷函数；或者在 X_i 的秩 R_i 比较大和比较小的时候，X_i 的得分 $a(R_i)$ 都是比较小，计分函数 $a(r)$ 是单峰函数，统计量 $\sum_{i=1}^{m}a(R_i)$ 就可以作为尺度参数检验问题的检验统计量。

在 $a(r)$ 是单峰函数时，我们在 $\sum_{i=1}^{m} a(R_i)$ 比较小的时候拒绝原假设，认为 $\sigma_1^2 > \sigma_2^2$；而在 $a(r)$ 是单谷函数时，我们在 $\sum_{i=1}^{m} a(R_i)$ 比较大的时候拒绝原假设，认为 $\sigma_1^2 > \sigma_2^2$。类似地，若尺度参数 b 的检验问题的备择假设改为 $H_1 : \sigma_1^2 < \sigma_2^2$，那么在 $a(r)$ 是单峰函数时，我们在 $\sum_{i=1}^{m} a(R_i)$ 比较大的时候拒绝原假设，认为 $\sigma_1^2 < \sigma_2^2$；而在 $a(r)$ 是单谷函数时，我们在 $\sum_{i=1}^{m} a(R_i)$ 比较小的时候拒绝原假设，认为 $\sigma_1^2 < \sigma_2^2$。

Mood方差检验

令$R=(R_1,R_2,\cdots R_{m+n})$表示样本 $X_1,X_2,\cdots X_m,Y_1,Y_2,\cdots Y_n$的秩向量，

在零假设成立的情况下，有：$E(R_i)=\sum_{i=1}^{m+n}\frac{i}{m+n}=\frac{m+n+1}{2}$

考虑Mood秩统计量：$M_X=\sum_{i=1}^{m}(R_i-\frac{m+n+1}{2})^2$

如果X的方差偏大，那么

$\forall c>0, P(X>c)>P(Y>c),\qquad \forall c<0, P(X<c)>P(Y<c),$

$X_1,\cdots,X_m$在合样本中的秩居两端

那么 M_X 的值也应该偏大，对于大的M_X可以考虑拒绝零假设。

大样本近似

在$m,n\to\infty$,且 $\frac{m}{m+n}\to\lambda$, $0<\lambda<1$ 的时候，可以采用大样本近似：

$$Z=\frac{M-E(M)}{\sqrt{Var(M)}}\to N(0,1)$$

其中

$$E(M)=m(m+n+1)(m+n-1)/12$$

$$Var(M)=mn(m+n+1)(m+n+2)(m+n-2)/180$$

对于打结情况可以考虑用修正公式.

例：假定用手工和仪器分别各测量了5个健康成年人的尿酸浓度，测量结果如下，问这两种测量方法的精确度是否存在差异？

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 ： 两种尿酸浓度测定法的方差不同，即 $\sigma_1^2 \neq \sigma_2^2$

统计分析：将两样本混合，计算混合秩如下表。

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

$m = n = 5, m + n = 10$,

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

$$= (1-5.5)^2 + (3-5.5)^2 + (4-5.5)^2 + (9-5.5)^2 + (10-5.5)^2$$

$$= 61.25$$

$M_{0.025,5,5} = 15.25, M_{0.975,5,5} = 65.25, 15.25 < \mathrm{M} = 61.25 < 65.25,$

故不能拒绝 H_0,表示两种测量法的精度没有明显差异.

尺度参数检验问题的计分函数 $a(r)$ 的选取，通常有以下 4 种方法：

①Mood 检验：取 $a(r)$ 为单谷函数，$a(r)=\left(r-(N+1)/2\right)^2$，$r=1,2,\cdots,N$；

②Ansari－Bradley 检验；取 $a(r)$ 是单峰函数，$a(r)=(N+1)/2-\left|r-(N+1)/2\right|$，$r=1,2,\cdots,N$。

在 $N=2k$ 为偶数时，取 $a(r)=\begin{cases} r, & r=1,2,\cdots,k \\ N-r+1, & r=k+1,k+2,\cdots,N \end{cases}$

在 $N=2k+1$ 为奇数时，取 $a(r)=\begin{cases} r, & r=1,\cdots,k,k+1 \\ N-r+1, & r=k+2,k+3,\cdots,N \end{cases}$

例如，在 $N=8$ 时，

r	1	2	3	4	5	6	7	8
a(r)	1	2	3	4	4	3	2	1

在 $N=9$ 时，

r	1	2	3	4	5	6	7	8	9
a(r)	1	2	3	4	5	4	3	2	1

记 $A_X=\sum_{i=1}^{m}a(R_i)$；

③ Siegel – Turkey 检验：取 a(r) 为单谷函数，令 $a(1)=N$，$a(N)=N-1$，$a(N-1)=N-2$，$a(2)=N-3$，$a(3)=N-4$，$a(N-2)=N-5$，$a(N-3)=N-6$，$a(4)=N-7$，$a(5)=N-8$，…。例如，在 $N=9$ 时，

r	1	2	3	4	5	6	7	8	9
a(r)	9	6	5	2	1	3	4	7	8

记 $S_X=\sum_{i=1}^{m}a(R_i)$；

④Klotz 检验：记标准正态分布 $N(0,1)$ 的分布函数为 $\Phi(x)$，其反函数记为 $\Phi^{-1}(x)$。取 $a(r)=\left[\Phi^{-1}\left(r/(N+1)\right]\right]^2$，$r=1,2,\cdots,N$，$a(r)$ 为单谷函数。记 $K_X=\sum_{i=1}^{m}a(R_i)$。

我们将尺度参数检验问题的解总结成表 5.14。

表 尺度参数检验问题的解

原假设 H_0	备择假设 H_1	何种情况下拒绝原假设
X 和 Y 同分布	$b=\sigma_1^2/\sigma_2^2>1$	M_X 比较大；A_X 比较小；S_X 比较大；K_X 比较大
	$b<1$	M_X 比较小；A_X 比较大；S_X 比较小；K_X 比较小
	$b\neq 1$	M_X 比较大或比较小；A_X 比较大或比较小；S_X 比较大或比较小；K_X 比较大或比较小

练习：

1. 设样本 $X_1, X_2, \cdots, X_n \sim N(0,\sigma^2)$, $Y_1, Y_2, \cdots, Y_m \sim N(\mu,\sigma^2)$,

试用似然比检验法检验假设问题 $H_0 : \mu = 0 \leftrightarrow H_1 : u \neq 0$.

2. 设样本 $X_1, X_2, \cdots, X_n$ 来自连续型随机变量总体 X。记 $x_i\,(i=1,2,\cdots,N)$在样本中的秩为 R_i（$R_i=1,2,\cdots,N$）。称 $\sum_{i=1}^{N} c(i)a(R_i)$ 为线性秩统计量，其中 $a(r)\,(r=1,2,\cdots,N)$称为计分函数，$c(t)\,(t=1,2,\cdots,N)$称为回归系数。试证明线性秩统计量$\sum_{i=1}^{N} c(i)a(R_i)$的期望与方差分别为

$$E\left(\sum_{i=1}^{N} c(i)a(R_i)\right)=n\overline{ca}$$

$$Var\left(\sum_{i=1}^{N} c(i)a(R_i)\right)=\frac{1}{N-1}\sum_{i=1}^{N}\left(c(i)-\overline{c}\right)^2\sum_{i=1}^{N}\left(a(i)-\overline{a}\right)^2$$

其中 $\overline{a}=\sum_{i=1}^{N} a(i)\Big/N$，$\overline{c}=\sum_{i=1}^{N} c(i)\Big/N$。设有另一个线性秩统计量$\sum_{i=1}^{N} d(i)a(R_i)$，其计分函数仍为$a(r)$，而回归系数换为$d(t)$。试求$\sum_{i=1}^{N} c(i)a(R_i)$和$\sum_{i=1}^{N} d(i)a(R_i)$的协方差：$Cov\left(\sum_{i=1}^{N} c(i)a(R_i), \sum_{i=1}^{N} d(i)a(R_i)\right)$的值。

3.王静龙《非参数统计分析》上习题五之 1、3、4、7、8、10