第四章　多组数据位置推断

很多时候需要对多组数据的分布位置进行比较，传统的问题中需要通过试验组和对照组试验研究结构来采集数据,需要考虑到不同的组是否对响应变量的结果有影响。对于传统的假设检验，其方法是分析样本数据并通过抽样分布方法计算其$p-$ 值。如果该$p-$值小于某一显著性阈值，则可拒绝相应的原假设。然而，当许多假设一起检验时，总体错误率将随着原假设数量的增多而急剧上升。分析的主要工具是方差分析, 不同的试验设计选择不同的方差分析模型. 无论采用哪一种方差分析,在参数统计推断中,一般都需要组数据满足正态分布假定. 当先验信息或数据不足以支持正态假定, 就需要借助非参数方法解决. 另外，在高维问题上，多重检验与特征选择密切相关，特征选择问题可以描述为假设检验问题，即特定的数据特征是否为数据所描述问题相关的特征。书中还将介绍Bonferroni校正法、BH-FDR控制算法和稀疏弱信号HC 高阶鉴定检验算法。

本章中, 一般假定多个总体有相似的连续分布(除了位置不同外, 其他条件差异不大), 多组之间是独立样本. 形式上, 假定k个独立样本有连续分布函数$F_1, F_2, \cdots, F_k$, 假设检验问题可表示为

$$H_0 : F_1 = F_2 \cdots = F_k \leftrightarrow \quad H_1 : F_i(x) = F(x + \theta_i), i = 1, 2, \cdots, k.$$

这里F是某连续分布函数族, 各组之间位置的差异简化为位置参数θ_i可能不全相同. 本章主要介绍5种方法, 其中前两种主要基于完全随机设计之下的位置比较, 后三种针对完全区组和不完全均衡区组设计. 为此, 我们首先在4.1节简要回顾试验设计的基本概念.

4.1　试验设计和方差分析的基本概念回顾

在实际中, 经常需要比较多组独立数据均值之间的差异存在性问题. 例如, 材料研究中比较不同温度下试验结果的差异, 临床试验中比较不同药品的疗效, 产

品质量检测中比较采用不同工艺生产产品的强度, 市场营销中比较不同地区的产品销售量等. 如果差异存在, 还希望找出较好的. 在试验设计中, 称温度、药品、工艺和地区等影响元素为**因素**(factor), 因素不同的状态称为不同的**处理或水平**. 例如, 在200°C, 400°C, 160°C三个温度值下, 比较高度钢的抗拉强度, 1.0GPa, 1.2GPa, 1.5GPa就是三个处理或水平. 试验设计和方差分析的主要内容是研究不同的影响因素(也包括因子)如何影响试验的结果.

有时影响结果的因素不止一个, 比如还有催化剂, 考虑催化剂含量的0.5%和1.5%两个处理水平. 这样, 就要进行各种因素不同水平(level)的组合试验和重复抽样. 由于各种处理的影响, 因此抽样结果不尽相同, 总会存在偏差(bias), 这些偏差就是所谓试验误差. 试验误差若太大, 则不利于比较差异. 于是, 一种组合里不能允许有太多的样本. 另外, 还需要考虑一个组里的数据应该满足同质性, 在抽取数据时, 需要根据数据来源的随机性考虑如何更好地设计试验, 需要根据试验材料(如人、动物、土地)的性质、试验时间、试验空间(环境)及法律规章的可行性制定合理的试验方案, 用尽量少的样本和合适的方法分析试验观察值, 达到试验目的. 这都是试验设计中要考虑的基本问题.

在进行试验时, 一般试验者应遵循三个基本原则.

(1) 重复性原则：重复次数越多, 抽样误差越小, 但非抽样误差越大.

(2) 随机性原则：随机安排各处理, 消除人为偏见和主观臆断.

(3) 适宜性原则：采用合适的试验设计, 剔除外界环境因素的干扰.

多样本均值比较, 一般不能简单地用两样本t均值比较解决. 比如要比较三种处理之间的位置差异, 三种处理的两两比较共有$\binom{3}{2}=3$种, 假设两两处理比较的显著性水平为$\alpha=0.05$, 三次比较的显著性水平只有$1-(1-\alpha)^3=0.1426$. 也就是说, 只要拒绝一个检验, 就可能犯Ⅰ类错误, 第Ⅰ类错误的发生概率是14.26%, 而不是当初设定的0.05. 如果要比较的是8组, 犯Ⅰ类错误的发生概率是76.22%. 因此多总体均值的比较都采用方差分析法.

方差分析的基本原理是将不同因素之下的试验结果分解为两方面的因素作用, 即因素之间的差异和不明因素的随机误差两项. 先以单因素方差分析为例回顾参数方差分析的基本原理. 单因素方差分析模型由于没有区组影响, 因而有较简单的表达式：

$$x_{ij}=\mu+\mu_i+\varepsilon_{ij}, i=1,2,\cdots,k, j=1,2,\cdots,n_i. \tag{4.1}$$

其中x_{ij}表示第i个处理的第j个重复观测, n_i表示第i个处理的观测样本量. 假设

有k个总体$F(x-\mu_i), i=1,2,\cdots,k$,即$k$个处理, 在各总体为等方差正态分布以及观测值独立的假定下, 假设问题为

$$H_0:\mu_1=\cdots=\mu_k=\mu \leftrightarrow\ H_1:\exists i,j,\mu_i\neq\mu_j. \tag{4.2}$$

将观测值重新整理表达如下：

$$x_{ij}-\bar{x}..=(\bar{x}_{i\cdot}-\bar{x}..)+(x_{ij}-\bar{x}_{i\cdot}), i=1,2,\cdots,k; j=1,2,\cdots,n_i.$$

令x_{ij}表示第i个处理的第j个样本, 两边平方后为

$$\underbrace{\sum(x_{ij}-\bar{x}..)^2}=\underbrace{\sum n_i(\bar{x}_{i\cdot}-\bar{x}..)^2}+\underbrace{\sum(x_{ij}-\bar{x}_{i\cdot})^2}. \tag{4.3}$$

$$\text{SST(总平方和)}=\text{SSt(处理平方和)}+\text{SSE(误差平方和)} \tag{4.4}$$

在正态假定之下, 可以将平方和以及各自的平方和与自由度综合成方差分析表如表4.1所示.

表4.1　方差分析表

变异来源	自由度	平方和	均方	实际观测F值
处理	$k-1$	SSt	MSt	MSt/MSE
误差	$n-k$	SSE	MSE	
合计	$n-1$	SST		

对假设检验问题(4.2), 令检验统计量为

$$F=\frac{\text{MSt}}{\text{MSE}}=\frac{\sum_{i=1}^{k}n_i(\overline{x}_{i\cdot}-\bar{x})^2/(k-1)}{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij}-\overline{x}_{i\cdot})^2/(n-k)}.$$

这里$\overline{x}_{i\cdot}=\sum_{j=1}^{n_i}x_{ij}/n_i,\ \bar{x}=\sum_{i=1}^{k}\sum_{j=1}^{n_i}x_{ij}/n$. 若各处理数据假定为正态分布且等方差, 则$F$ 在H_0下的分布为自由度$(k-1,n-k)$的F分布.若$F=\text{MSt/MSE}>F_{(\alpha)}(k-1,n-k)$, 则考虑拒绝零假设：

$$H_0:\mu_1=\cdots=\mu_k\leftrightarrow H_1:\text{并非所有 }\mu_i\text{ 都相等.} \tag{4.5}$$

不同的试验设计有不同的方差分析方法, 下面分别说明.

1. 完全随机设计

先看一个例子.

例4.1 假设有A,B,C三种饲料配方用于北京鸭饲养, 比较采用不同饲料喂养对北京鸭体重增加的影响. 每种饲料设计重复观测4次, 需要12只鸭参与试验, 采用完全随机设计, 挑选12只体质相当的北京鸭. 比如采用体形相近且健康的北京鸭. 随机将三种饲料分配给不同的北京鸭进行试验, 2个月后北京鸭增加的体重(kg)如表4.2所示.

表4.2 北京鸭体重增加饲料比较数据表

B 2.0	C 2.8	B 1.8	A 1.5
A 1.4	B 2.4	C 2.5	C 2.1
C 2.0	A 1.9	A 2.0	B 2.2

这是一个典型的完全随机设计(completely randomized design, CRD)的例子, 是最简单的一种试验设计. 在这个例子中影响因素只有饲料一个, 因此分析这样的数据方法称为单因素方差分析.

为保证样本无偏性, 应用完全随机设计须具备以下两个条件:

(1) 试验材料(动物、植物、土地)为同质;

(2) 各处理(比如饲料配方)要随机安排试验材料.

假设检验问题为$H_0:\mu_1=\mu_2=\mu_3 \leftrightarrow H_1:\exists i,j,i\neq j,i,j=1,2,3,\mu_i\neq\mu_j$(至少有一对处理均值不等).

在进行方差分析之前通常需要将表4.3整理成如表4.3所示, 便于计算各项均值和方差.

表4.3 北京鸭体重增加饲料比较数据表

	处理				
重复	1	2	3	4	和
A	1.4	1.9	2.0	1.5	6.8
B	2.0	2.4	1.8	2.2	8.4
C	2.6	2.8	2.5	2.1	10.0
	6	7.1	6.3	5.8	25.2

各项平方和计算如下:

$$\text{总平方和}\quad \mathrm{SST}=1.4^2+\cdots+2.1^2-25.2^2/12=2.00,$$
$$\text{处理平方和}\quad \mathrm{SSt}=\frac{1}{4}(6.8^2+\cdots+10^2)-25.2^2/12=1.28,$$
$$\text{误差平方和}\quad \mathrm{SSE}=\mathrm{SST}-\mathrm{SSt}=2-1.28=0.72.$$

得出方差分析表如表4.4所示.

表4.4 方差分析表

因子	自由度	平方和	均方	F值	F_α 0.05	F_α 0.01
饲料(t)	2	1.28	0.64	8*	4.26	8.02
误差(E)	9	0.72	0.08			
总计(T)	11	2.00				

** 表示0.01显著性水平下显著, *表示0.05显著性水平下显著. 以下同.

结论：设$\alpha=0.05$, 如表4.4所示, $F=8>F_{0.05}(2,9)=4.26$, 接受H_1, 表示三种饲料在增加北京鸭体重方面存在差异.

以下是R软件中单因素方差分析的函数和结果输出：

```
*** Analysis of Variance Model ***
  >   aov(formula = y ~ x, data = xy, na.action = na.exclude)
  >                  x Residuals
 Sum of Squares    1.28    0.72
Deg. of Freedom     2         9
Residual standard error: 0.043
Estimated effects are balanced
          Df Sum of Sq Mean Sq   F Value     Pr(F)
        x  2     1.28  0.64           8      0.001
Residuals  9     0.72  0.08
```

2. 完全随机区组设计

在实践中, 除了处理之外,往往还有别的因素起作用. 假设需要对A, B, C, D四种处理血液凝固时间设计比较试验, 每种处理方法重复观测5次. 换句话说, 应该随机将20位正常人分为5组, 每组4人, 分别接受4种不同的处理, 共生成$4\times5=20$份血液, 供四种处理方法进行凝血试验比较. 由经验可知, 由于每个人体质不同, 血液自然凝固时间的差异可能比较大. 如果恰好自然凝血时间较短的人的血液都分配给较差的处理方法, 而凝血时间较长的血液分给较好的处理方法, 最后可能测不出哪一种处理方法更有效. 这是因为在血液凝固试验中, 不同

条件的人构成了另一个因素, 称为**区组**(block).如果只取5位正常人的血液, 每人分成4份随机分配4种处理方法, 这就是完全随机区组设计, 其中人为区组.

血液凝固时间见表4.5, 从表中可以看出, 影响结果的因素有各处理效应和区组(人)两个.

表4.5 四种血液凝血时间测量结果表

处理 \ 区组	x_{ij}					
	Ⅰ	Ⅱ	Ⅲ	Ⅳ	Ⅴ	处理和$x_{i\cdot}$
A	8.4	10.8	8.6	8.8	8.4	45.0
B	9.4	15.2	9.8	9.8	9.2	53.4
C	9.8	9.8	10.2	8.9	8.5	47.3
D	12.2	14.4	9.8	12.0	9.5	57.9
区组和$x_{\cdot j}$	39.8	50.3	38.4	39.5	35.6	203.6=$x_{\cdot\cdot}$

如果影响的因素有区组的影响, 则需要用两因素方差分析模型表示. 为简单起见, 这里只给出主效应的表示模型, 这表示处理因素与区组之间不考虑交互作用, 模型如下所示:

$$\begin{aligned} x_{ij} &= \mu + \tau_i + \beta_j + \varepsilon_{ij}, \\ & i = 1, 2, \cdots, k(\text{处理数}), \\ & j = 1, 2, \cdots, b(\text{区组数}). \end{aligned}$$

其中, x_{ij}表示第i个因子的第j个区组的观测, 每个因子的观测量为b, 每个区组的观测量为k, τ_i是第i个处理的效应, β_j是第j个区组的效应.

假设检验问题为$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 \leftrightarrow H_1 : \mu_i \neq \mu_j, \exists i, j$.

如果随机地把所有处理分配到所有的区组中, 使得总的变异可以分解为:

(1) 处理造成的不同;

(2) 区组内的变异;

(3) 区组之间的变异.

对于完全区组试验, 正态总体条件下的检验统计量为

$$F = \frac{\mathrm{MSt}}{\mathrm{MSE}} = \frac{\sum_{i=1}^{k} b(\bar{x}_{i\cdot} - \bar{x})^2/(k-1)}{\sum_{i=1}^{k}\sum_{j=1}^{b}(x_{ij} - \bar{x}_{i\cdot} - \bar{x}_{\cdot j} + \bar{x})^2/[(k-1)(b-1)]}.$$

式中, $\bar{x}_{i\cdot}=\sum_{j=1}^{b}x_{ij}/b,\ \bar{x}_{\cdot j}=\sum_{i=1}^{k}x_{ij}/k,\ \bar{x}=\sum_{i=1}^{k}\sum_{j=1}^{b}x_{ij}/n, n=kb$. 统计量$F$在零假设下为自由度为$(k-1,n-k)$的$F$分布. 如果要检验区组之间是否有区别, 只要把上面公式中的i和j交换、k和b 交换并考虑对称的问题即可.

各效应平方和计算如下：

总平方和

$$\begin{aligned}\mathrm{SST}&=\sum\sum(x_{ij}-\bar{x}_{\cdot\cdot})^2\\&=\sum\sum x_{ij}^2-x_{\cdot\cdot}^2/kb\\&=8.4^2+\cdots+9.5^2-203.6^2/20=68.672.\end{aligned}$$

区组平方和

$$\begin{aligned}\mathrm{SSB}&=k\sum_{}^{b}(x_{\cdot j}-\bar{x}_{\cdot\cdot})^2\\&=\sum x_{\cdot j}^2/k-x_{\cdot\cdot}^2/kb\\&=\frac{1}{4}(39.8^2+\cdots+35.6^2)-203.6^2/20\\&=31.427.\end{aligned}$$

处理平方和

$$\begin{aligned}\mathrm{SSt}&=b\sum_{}^{k}(\bar{x}_{i\cdot}-\bar{x}_{\cdot\cdot})^2\\&=\sum x_{i\cdot}^2/n-x_{\cdot\cdot}^2/kb\\&=\frac{1}{5}(45.0^2+\cdots+57.9^2)-203.6^2/20\\&=20.604.\end{aligned}$$

误差平方和

$$\mathrm{SSE}=\mathrm{SST}-\mathrm{SSB}-\mathrm{SSt}=16.641.$$

四种血液凝固处理结果如表4.6所示, 实际区组$F_b=5.6654>F_{0.01,4,12}=5.41$, 这表示区组(人)对血液凝固有显著差异; 处理$F_t=4.9524>F_{0.05,3,12}=3.49$, 表示不同的处理对凝血效果有差别. 图4.1给出四种凝血时间观测值分处理箱线图, 图中也显示了凝血效果处理间的差异. 其处理均值间存在差异, 到底是哪些处理之间存在差异还需要进一步的检验, 这里省略.

表4.6 双因素方差分析表

因素	自由度	平方和	均方	F值	F_α 0.05	0.01
区组(B)	$5-1=4$	31.427	7.8568	5.6654**	3.26	5.41
处理(t)	$4-1=3$	20.604	6.8680	4.9524*	3.49	5.95
误差(E)	$19-7=12$	16.641	1.3868			
总计(T)	$20-1=19$	68.672				

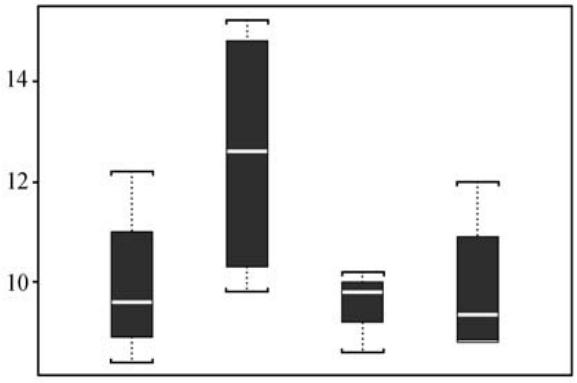

图 4.1: 四种凝血时间测量分处理箱线图

完全区组的试验设计的基本使用条件如下：

(1) 试验材料为异质, 试验者根据需要将其分为几组, 几个性质相近的试验单位成一区组(如一个人的血液分成四份, 此人即同一区组, 不同人为不同区组), 使区组内试验个体之间的差异相对较小, 而区组间的差异相对较大;

(2) 每一个区组内的试验个体按照随机安排全部参加试验的各种处理;

(3) 每个区组内的试验数等于处理数.

3. 均衡的不完全区组设计

以上介绍的完全随机区组设计要求每一个处理都出现在每一个区组中, 但在实际问题中, 不一定能够保证每一个区组都能有对应的样本出现. 此时就有了不完全区组设计. 当处理组非常大, 而同一区组的所有样本数又不允许太大时, 在一个区组中可能不能包含所有的处理, 此时只能在同一区组内安排部分处理, 即不是所有的处理都被用于各区组的试验中, 这种区组设计称为不完全区组设计(incomplete block). 在不完全区组设计中, 最常用的就是均衡不完全区组设计(balanced incomplete block design), 简称BIB随机区组设计. 具体而言, 每个区组安排相等处理数的不完全区组设计. 假定有k个处理和b个区组, 区组

样本量为$b \times n$(它表示区组中最多可以安排的处理个数), 均衡的不完全区组设计BIB(k, b, r, t, λ)满足以下条件:

(1) 每个处理在同一区组中最多出现一次;

(2) 区组样本量为t, t为每个区组设计的样本量, t小于处理个数k;

(3) 每个处理出现在相同多的r个区组中;

(4) 每两个处理在一个区组中相遇的次数一样(λ 次).

用数学的语言来说, 这些参数满足：

(1) $kr = bt$;

(2) $\lambda(k-1) = r(t-1)$;

(3) $b \geqslant r$ 或$k > t$.

如果$t = k, r = b$, 则为完全随机区组设计.

例4.2 比较4家保险公司A,B,C,D在Ⅰ,Ⅱ,Ⅲ,Ⅳ四个不同城市的保险经营业绩, 假设以当年签订保险协议的份数作为衡量业绩的标志. 由于4家保险公司未必在四所城市都有经营网点, 或即便有经营网点, 但分支机构的经营年限各有不同, 导致某些数据不可直接参与比较, 因此采取BIB设计, 得到如下数据(单位：万).

表4.7 不同城市保险公司绩效的BIB设计

	城市(区组)			
保险公司(处理)	Ⅰ	Ⅱ	Ⅲ	Ⅳ
A	34	28		59
B		30	36	45
C	36	44	48	
D	40		54	60

很容易看出BIB设计的均衡性质. 这里$(k, b, r, t, \lambda) = (4, 4, 3, 3, 2)$.

4.2 多重检验问题

1. FDR控制基本原理

考虑m个假设检验

$$H_{0j} : \mu_j = 0 \leftrightarrow H_{1j} : \mu_j \neq 0, \quad j = 1, 2, \cdots, m$$

令$p_1, p_2, \cdots, p_m$是这m个检验的p值，如果$p_j < \alpha/m$,则拒绝原假设，这就是Bonferroni校正法则.

定理4.1 Bonferroni法则的错误拒绝原假设的概率小于或等于α.

证明 令R表示至少有一个原假设被错误拒绝的事件，令R_j表示第j 个原假设被错误拒绝的事件，由式$\mathbb{P}(\bigcup_{j=1}^{m} R_j) \leqslant \sum_{j=1}^{m} \mathbb{P}(R_j)$,于是有

$$\mathbb{P}(R) = \mathbb{P}(\bigcup_{j=1}^{m} R_j) \leqslant \sum_{j=1}^{m} \mathbb{P}(R_j) = \sum_{j=1}^{m} \frac{\alpha}{m} = \alpha.$$

例4.3 在基因例子中，用$\alpha = 0.05$,根据Bonferroni法则有对应的检验水准为$0.05/12533 = 3.99\mathrm{E}-6$, 对任何一个$p-$ 值小于$3.99\mathrm{E}-6$的基因，就可以说两种病之间存在显著差异。

Bonferroni法则是比较保守的，因为它的出发点是力求不犯一个错拒原假设的错误。然而在实际中，比如在基因表达分析中，研究者需要能尽可能多地识别出表达了差异的少数基因。研究者能够容忍和允许在R次拒绝中发生少量的错误识别，只要相对于所有拒绝H_0的次数而言错误识别数足够少，这样的技术就值得被关注，于是产生了错误发现(false discovery) 的概念。到底什么才是错误率足够小呢？这就需要在错误发现V 和总拒绝次数R 之间寻找一种平衡，即在检验出尽可能多的候选基因的同时将错误发现控制在一个可以接受的范围内，Benjamini 和Hochberg(1995)的错误发现率为上述平衡提供了一种可能。

如表4.8给出了各种可能出现的检验类型：m_0和m_1分别表示在m次多重检验中真实H_0和非真实H_0 的个数，V表示在所有R次拒绝H_0的决定中拒绝了H_0的次数。

表4.8 m次多重检验中结果的类型

	不拒绝H_0	拒绝H_0	合计
H_0为真	U	V	m_0
H_0为假	T	S	m_1
合计	W	R	m

定义错误发现比率(FDP)如下：

$$\mathrm{FDP} = \begin{cases} V/R, & R \geqslant 0 \\ 0 & R = 0 \end{cases}$$

FDP是错误拒绝原假设的比例，注意到在表4.8中，除m、R和W外，其它量均是不能直接观察到的随机变量，于是需要估计所有R次拒绝中错误发现的期望比例$\mathrm{FDR} = \mathrm{E}(\mathrm{FDP})$。$Y$·本加米尼(Benjamini Y.)和Y· 哈克博格(Hochberg.Y)(1995)给

出了一个基于p 值逐步向下的FDR控制程序，称为BH-FDR检验，其控制程序如下：

Benjamini-Hochberg(BH-FDR)控制法

(1)令$p_{(1)} \leqslant p_{(2)} \leqslant \cdots \leqslant p_{(m)}$表示排序后的$p-$值;
(2)$H_{(i)}$是对应于$p_{(i)}$的原假设，定义Bonferroni-型多重检验过程：
(3) 令$k = \max\{i : p_{(i)} \leqslant \frac{i}{m}\alpha\}$.
(4) 拒绝所有的$H_{(i)}, i = 1, 2, \cdots, k$.

定理4.2 (Benjamini和Hochberg(1995))如果应用了上述控制过程，那么无论有多少原假设是正确的，也无论原假设不真时的$p-$值的分布是什么，都有

$$\mathrm{FDR} = \mathrm{E}(\mathrm{FDP}) \leqslant \frac{m_0}{m}\alpha \leqslant \alpha.$$

例4.4 假设有15个独立的假设检验得到如下由小到大排序的$p-$值如表4.9：

表4.9 15个独立检验由小到大排序的$p-$值

0.0024	0.0056	0.0096	0.0121	0.0201	0.0278	0.0298	0.0344
0.0349	0.3240	0.4262	0.5719	0.6528	0.7590	1.0000	

在$\alpha = 0.05$的显著性水平下，Bonferroni检验拒绝所有$p-$值小于$\alpha/15 = 0.0033$ 的假设，因此有1个假设被拒绝了。对于BH-FDR检验，发现使得$p_{(i)} < i\alpha/m$的最大的$i = 4$, 也就是说：

$$P(4) = 0.0121 < 4 \times 0.05/15 = 0.013.$$

因此拒绝前4个$p-$值最小的假设。

2.FDR的相关讨论

多重检验的目标是对整体检验错误率进行控制，Bonferroni和BH-FDR控制都是通过决定一个显著性水平的阈值，从而使检验结果犯第I类错误的概率整体被限制在某一固定水平α。除这两者之外，常用的还有族错误率测度(FWER). FWER定义为$P(V > 1)$,即错误拒绝原假设的概率.Bonferroni方法直接控制FWER$\leqslant \alpha$. 两者相较，可以证明控制FWER相当于控制FDR。

定理4.3 FWER$\geqslant$ FDR ,控制FDR相当于FWER的弱控制.

证明： 具体而言，考虑第一种情况下，当所有的原假设都为真时，$m_0 = m, S = 0, V = R$, 如果$V = 0$ 则$Q = 0$,如果$V > 0$, 则$Q = 1$. 于是，$\mathrm{E}(Q) = PV \geqslant 1)$, 这表明FWER 与FDR 等效。考虑第二种情况下，当原假设不都为真时，$m_0 < m$ 时，可以证明FDR<FWER, 此时如果$V > 0$, 则$V/R \leqslant 1$, 这样，$P(V \geqslant 1) \geqslant Q$,两边同时取期望，得到$P(V \geqslant 1) \geqslant E(Q)$. 于是，FDR$\leqslant$ FWER,因此，控制FWER 也一定控制FDR.

4.3 HC高阶鉴定法

对于多重检验，Y·本加米尼（Benjamini Y.)和Y· 哈克博格(Hochberg.Y) (1995) 提出的BH-FDR 方法是通过决定一个显著性水平的阈值，控制检验结果犯第I类错误的概率整体限制在某一固定水平。这个方法有一个隐含的假设是：数据中拒绝零假设的信号是很强的或者说大部分是很强的，由此可以直接用$p-$ 值恢复信号。BH-FDR 的关注点在于控制错误信号发现率，实际上还有一个关注点就是对错误信号发现率大与小的估计。如果有的信号误发现率小，有的信号误发现率大，仅仅是将假信号的比例控制在一定水平下，却对其误发现率不做出有效的估计，那么很有可能因为信号太弱而导致BH-FDR无法将信号检测出来。为理解这一点，可以假想有这么个例子：待检测的检验数量是100，其中有90 个检验的p- 值大于0.2,10 个检验的$p-$ 值在[0.001,0.01]之间，假设最小的三个检验$p-$值是(0.003,0.007,0.009), 选取$\alpha = 0.05$,BH-FDR 最小的阈值是$0.05/100 = 0.0005$. 用每个检验的观察$p-$ 值，没有一个$p-$值小于阈值($0.003 > 0.0005$)，BH-FDR未能成功地检出信号。但是如果不是90个检验的p-值大于0.2，而是与原假设数量相等的10 个检验的p- 值大于0.2，那么此时对应到3 个最小的$p-$ 值阈的值分别为$(0.0025, 0.005, 0.0075)$, BH-FDR 至少可以检测出3个信号。从这个例子中，我们发现BH-FDR 有两个明显的缺陷：一是阈值强烈地依赖于一个主观的值α，二是阈值与信噪比有关，如果噪声检验比较多，就会妨碍信号的有效检出，也就是说出现了“抑真效应”，有的时候检会测不出信号，有时会测出错误的信号。心理学上有个“破窗效应”与“抑真效应”在道理上有相近之处，说的是如果一栋大楼有扇窗户破了未受到重视得到及时修补，不久整栋楼所有窗户都会被人莫名其妙地打破，这表明噪声杂多的地方鉴别弱小的技术十分必要。

那么怎么才能在信号数量不多的情况下仍然可以将其鉴别出来，大卫·多诺霍(David Donoho)和金加顺(Jiashun Jin)(2004,2016) 提出了HC 高阶鉴定法理

论。这个理论建立了稀疏和弱信号的分析框架，其核心就是”HC“高阶鉴定法的概念和分析逻辑。

这个理论首先分析了多重检验中奈曼-皮尔逊(Neyman-Pearson)似然比检验中检验数量和信号强弱之间的关系，指出信号的强弱与稀疏度是决定数据分析进程进而决定方法预测性能的根本。假设有两个检验如下：

$$\begin{aligned} H_0 : X_1 &\overset{iid}{\sim} N(0,1); \\ H_1^{(i)} : X_i &\overset{iid}{\sim} (1-\epsilon_p)N(0,1) + \epsilon_p N(\tau_p, 1), \quad 1 \leqslant i \leqslant p. \end{aligned}$$

当$p \to \infty$, 用参数(ϵ_p, τ_p)表示信号的稀疏性和强弱性，他们和检验的数量p和信号的强度r 之间的关系表达如下：

$$\epsilon_p = p^{-\beta}, \qquad \tau_p = \sqrt{2r \log p}, \qquad 0 < \beta, r < 1$$

当$\epsilon_p << 1/\sqrt{p}$很小的时候，表明只有极少的非零均值，β 越大τ_p 越小信号越稀疏。当τ_p比较小，信号相对比较弱，此时r比较小。根据信号的稀疏性和信号的强弱性有β 和r 的如下相图(参见David Donoho&Jiashun Jin(2004) 文章)

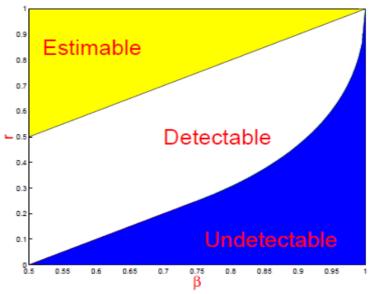

图 4.2: 信号强度r和β所刻画的参数估计区域

大卫·多诺霍和金加顺(2004)的论文中将参数的估计区域划分成了三个部分：可估计区域（Estimable）,信号可检测区域（Detectable）以及信号检测不出的区域（Undetectable）。这里使用的是概率强度的不同，横坐标β越大表示信号稀疏,纵坐标r 越大表示信号越强.β 较小信号强度r 较大的稠密区域是可估计区域，强度r 较大中等β 较小的稀疏区域是可检测区域，而β 比较大的信号稀疏而信号强度r 较小的是不可检测区域. 在可估区域，用现在常见的惩罚方法基本可以做到较好的恢复，能够实现分信号与噪音的离开；对于可检测的区域，虽然知道里

面有信号，但是几乎不可能将它们与噪音区分开，这也是BH-FDR失效的区域，如果是做信号检测、分类、聚类等工作，进行有效的推断还是仍然有可能的。此时进行推断的框架不是FDR，而是需要一个对稀疏和弱信号更敏感的框架，这个名字的来于约翰· 图基1976 年(John Tukey)stat411课程讲义的笔记——高阶鉴定法(Higher Criticism，简称HC鉴定法)。

这里给出HC检测法的经典算法如下：

高阶鉴定法HC的经典算法

(1)对每个检验计算一个统计量得分，根据统计量得分计算$p-$值;

(2)对p值进行排序$\pi_{(1)} < \pi_{(2)} < \cdots \pi_{(p)}$;

(3)计算第k个HC值，相当于算了一个二阶z-得分:

$$HC_{p,k} = \sqrt{p}\left[\frac{k/p - \pi_{(k)}}{\sqrt{\pi_{(k)}(1-\pi_{(k)})}}\right]$$

(4)取最大值，计算相应的$HC_p* = \max_{1\leqslant k\leqslant p\alpha_0}\{HC_{p,k}\}$, 找到对应的$\hat{k}$，前$k$可以认为是真显著的,拒绝所有的$H_{(i)}, i = 1, 2, \cdots, k$.

例4.5 (例4.4续)假设每组检验的样本量$n = 30$,编写程序。

（1）根据计算例4.4例题数据运用高阶鉴定法HC经典算法计算阈值和k值,对比BH-FDR 和HC 之间阈值的差别。

（2）调用chap4\HC数据，重新运行程序，比较BH-FDR和HC之间阈值的差别。

解 (1)程序略，可以计算出HC得分如表4.10所示

表4.10 15个独立检验按经典高阶鉴定法$p-$值依次排序的HC得分值

序号	1	2	3	4	5	6	7
	5.0869	6.6294	7.5626	9.0177	8.6442	8.7684	9.9507
8	9	10	11	12	13	14	15
10.6025	11.9253	2.8358	2.4054	1.7854	1.7398	1.5787	0.0000

从表4.10中可以看出HC得分先增后降，在第9个检验上达到最大，如图4.3(左)所示。HC 选择拒绝的检验数量是9，而BH检验拒绝的检验数量是4，观察$p-$值的分布，发现HC 的阈值正好选在了两组$p-$ 值间隔最大的位置，而BH的结果则比较保守而且随意。可以看出在信号比较强$p-$值相对稠密的情况下，HC鉴定法的效果比较理想。

(2)HC数据有80个检验的$p-$值，其中前15个检验与4.4相同，多出来的65个检验都是$p-$值较大的检验，运用BH-FDR在$\alpha = 0.05$ 水平下只能检出4个检验，功效有所下降，而HC则依然保持了较高的鉴别能力，最大值在第9个检验上取得。如图4.3(右)所示,k表示第k大的$p-$值。

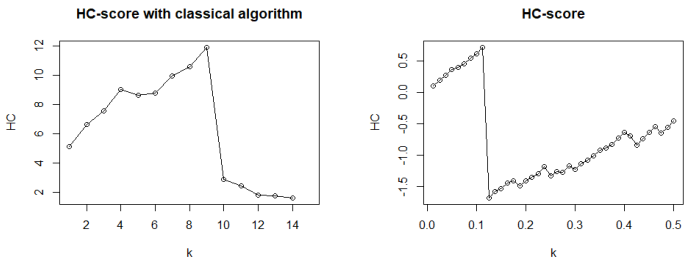

图 4.3: 例4.5高阶鉴定经典算法HC得分(左：15个检验；右：80个检验)

例4.6 该例子来自大卫·多诺霍和金加顺(2016) 文章。其中使用了美国哈佛医学院高登(G.J.Gordon)(2002) 提供的肺癌微阵列数据，共有181个组织样本，其中31个恶性胸膜间皮瘤样本（MPM,Malignant Pleual Mesothelioma）和150个肺癌样本（ADCA,Adenocarcinoma）每个样本包括12,533条基因，这个数据的主要目标是从12,533条基因中找到对识别两类疾病最有效的特征，文中使用KS检验输出$p-$值，对$p-$值排序，计算第k 个HC值，产生一个高阶鉴定法HC的经典算法的改进算法计算二阶z-得分如下:

$$HC_{p,k} = \frac{\sqrt{p}(k/p - \pi_{(k)})}{\sqrt{k/p + \max\{\sqrt{n}(k/p - \pi_{(k)}), 0\}}} \tag{4.6}$$

取最大值，计算相应的$HC_p* = \max_{1\leqslant k\leqslant p/2, \pi_{(k)} > (\log p)/p}\{HC_{p,k}\}$,找到对应的$\hat{k}$，前$k$个检验可以认为是真显著的,而且HC阈值$t_p^{HC}$是第$\hat{k}$大的KS- 统计量得分。图4.3绘制了KS 统计量得分（备注：D统计量与样本量$\sqrt{n}$的乘积）、$p-$值以及HC随实际检验的比例k/p 变化的曲线，从曲线上看，红线是阈值所在位置，$p-$值在0.01左右以下的检验在高阶鉴定法HC中得到拒绝，这个阈值是由第三张图HC 最大值所确定的，HC=5.0476，选择出来的特征数量为261, 错误检测率只有5例。

从图4.4和预测性能来看,高阶鉴定法HC不仅通过推断实施了有效的特征选择，而且在强弱信号的识别任务中展现了良好的区分能力。

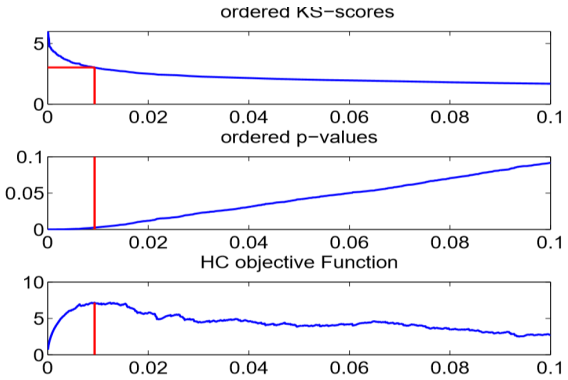

图 4.4: 多纳赫和金加顺在高登2002年的基因数据上运用HC方法的分析结果

4.4 Kruskal-Wallis单因素方差分析

1. Kruskal-Wallis检验的基本原理

Kruskal-Wallis检验是1952年由Kruskal和Wallis二人提出的. 它是一个将两样本W-M-W检验推广到3个或更多组检验的方法. 回想两样本中心位置检验的W-M-W检验：首先混合两个样本, 找出各个观测值在混合样本中的秩, 按各自样本组求和, 如果差异过大, 则可以认为两组数据的中心位置存在差异. 这里的想法是类似的, 如果数据取自完全随机设计, 先把多个样本混合起来求秩, 再按样本组求秩和. 考虑到各个处理的观测数可能不同, 可以比较各个处理之间的平均秩差异, 从而达到比较的目的. 在计算所有数据混合样本秩时,如果遇到有相同的观测值, 则像从前一样用秩平均法定秩. Kruskal-Wallis方法也称为H检验. H检验方法的基本前提是数据的分布是连续的, 除位置参数不同以外, 分布是相似的.

对检验问题(4.1),完全随机设计的数据如表4.11所示.

表4.11 完全随机设计数据形态

	总体1	总体2	$\cdots$	总体k
	x_{11}	x_{12}	$\cdots$	x_{1k}
重	x_{21}	x_{22}	$\cdots$	x_{2k}
复	$\vdots$	$\vdots$	$\vdots$	$\vdots$
测量	$x_{n_1 1}$	$x_{n_2 2}$	$\cdots$	$x_{n_k k}$

记x_{ij}代表第j总体的第i个观测值, n_j为第j个总体中样本的重复次数(replication).

现在将上表所有数据从大到小给秩, 最小值给秩1, 次小值给秩2, 依次类推, 最大值的秩为$n = n_1 + n_2 + \cdots + n_k$. 如果有相同秩, 则采取平均秩. 令$R_{ij}$为观测值$x_{ij}$的秩, 每个观察值的秩如表4.12所示.

表4.12 完全随机设计数据的秩

	总体1	总体2	$\cdots$	总体k
	R_{11}	R_{12}	$\cdots$	R_{1k}
重	R_{21}	R_{22}	$\cdots$	R_{2k}
复	$\vdots$	$\vdots$	$\vdots$	$\vdots$
测量	$R_{n_1 1}$	$R_{n_2 2}$	$\cdots$	$R_{n_k k}$
秩和	$R_{\cdot 1}$	$R_{\cdot 2}$	$\cdots$	$R_{\cdot k}$

假设检验问题为

$$H_0: k\text{个总体位置相同}\ (\text{即：}\mu_1 = \mu_2 = \cdots = \mu_k = \mu),$$
$$H_1: k\text{个总体位置不同}\ (\text{即：}\mu_i \neq \mu_j\ \ \text{for}\ \ i \neq j).$$

对每一个样本观察值的秩求和得到$R_{\cdot j} = \sum_{i}^{n_j} R_{ij},\ j = 1, 2, \cdots, k$. 第$j$组样本的秩平均为

$$\bar{R}_{\cdot j} = R_{\cdot j}/n_j.$$

观测值的秩从小到大依次为$1, 2, \cdots, n$, 则所有数据混合后的秩和为

$$R_{\cdot\cdot} = 1 + 2 + \cdots + n = n(n+1)/2.$$

下面分析$R_{\cdot j}$的分布. 假定有n个研究对象和k种处理方法, 把n个研究对象分配给第j种处理, 分配后的秩为$R_{1j}, R_{2j}, \cdots, R_{n_j j}$. 给定$n_j$后, 所有可能的分法为$\binom{n}{n_1, \cdots, n_k}$个, 这是多项分布的系数, 在零假设下, 所有可能的分法都是等可能的, 有

$$P_{H_0}(R_{ij} = r_{ij}, j = 1, 2, \cdots, k, i = 1, 2, \cdots, n_j) = \frac{1}{\binom{n}{n_1, n_2, \cdots, n_k}}.$$

定理4.4 在零假设下,

$$\begin{aligned} E(\bar{R}_{\cdot j}) &= \frac{n+1}{2}, \\ \operatorname{var}(\bar{R}_{\cdot j}) &= \frac{(n-n_j)(n+1)}{12n_j}, \\ \operatorname{cov}(\bar{R}_{\cdot i}, \bar{R}_{\cdot j}) &= -\frac{n+1}{12}. \end{aligned}$$

因而, 在H_0下, $\bar{R}_{\cdot j}$应该与$\dfrac{n+1}{2}$非常接近, 如果某些$\bar{R}_{\cdot j}$ 与$\dfrac{n+1}{2}$相差很远, 则可以考虑零假设不成立.

混合数据各秩的平方和为

$$\sum\sum R_{ij}^2 = 1^2 + 2^2 + \cdots + n^2 = n(n+1)(2n+1)/6.$$

因此混合数据各秩的总平方和为

$$\begin{aligned} \mathrm{SST} &= \sum_{j=1}^{k}\sum_{i=1}^{n_j}(R_{ij} - \bar{R}_{\cdot\cdot})^2 \\ &= \sum\sum R_{ij}^2 - R_{\cdot\cdot}^2/n \\ &= n(n+1)(2n+1)/6 - [n(n+1)/2]^2/n \\ &= \frac{1}{6}n(n+1)(2n+1) - \frac{1}{4}n(n+1)^2 \\ &= n(n+1)(n-1)/12. \end{aligned}$$

其总方差估值(总均方)为

$$\operatorname{var}(R_{ij}) = \mathrm{MST} = \mathrm{SST}/(n-1) = n(n+1)/12.$$

各样本处理间平方和为

$$\begin{aligned} \mathrm{SSt} &= \sum_{j}^{k} n_j(\bar{R}_{\cdot j} - \bar{R}_{\cdot\cdot})^2 \\ &= \sum_{j}^{k} R_{\cdot j}^2/n_j - R_{\cdot\cdot}^2/n \\ &= \sum R_{\cdot j}^2/n_j - n(n+1)^2/4. \end{aligned}$$

用处理间平方和除以总均方就得到Kruskal-Wallis的H值为

$$\begin{aligned} H &= \mathrm{SSt}/\mathrm{MST} \\ &= \frac{\sum R_{\cdot j}^2/n_j - n(n+1)^2/4}{n(n+1)/12} \\ &= \frac{12}{n(n+1)}\sum R_{\cdot j}^2/n_j - 3(n+1). \end{aligned} \tag{4.7}$$

在零假设下, H近似服从自由度$k-1$的$\chi^2(k-1)$ 分布.

结论：当统计量H的值$> \chi_\alpha^2(k-1)$, 拒绝零假设, 接受H_1假设, 表示处理间有差异.

当零假设被拒绝时应进一步比较哪两组样本之间有差异. Dunn于1964年提议可以用下列检验公式继续检验两两样本之间的差异：

$$d_{ij} = |\bar{R}_{\cdot i} - \bar{R}_{\cdot j}|/\mathrm{SE} \tag{4.8}$$

式中, $\bar{R}_{\cdot i}$与$\bar{R}_{\cdot j}$为第i和j处理平均秩, SE为两平均秩差的标准误差, 它的计算公式如下：

$$\begin{aligned} \mathrm{SE} &= \sqrt{\mathrm{MST}\left(\frac{1}{n_i}+\frac{1}{n_j}\right)} \\ &= \sqrt{\frac{n(n+1)}{12}\left(\frac{1}{n_i}+\frac{1}{n_j}\right)}, \forall i,j = 1,2,\cdots,k, i \neq j. \end{aligned} \tag{4.9}$$

当$n_i = n_j$时, 简化为

$$\mathrm{SE} = \sqrt{k(n+1)/6}. \tag{4.10}$$

若$|d_{ij}| \geqslant Z_{1-\alpha^*}$, 则表示第$i$与第$j$处理间有显著差异; 反之则表示差异不显著. 式中$\alpha^* = \alpha/k(k-1)$, α为显著水平, Z为标准正态分布的分位数值.

例4.7 为研究4种不同的药物对儿童咳嗽的治疗效果,将25个体质相似的病人随机分为4 组, 各组人数分别为8 人、4 人、7人和6 人, 各自采用A, B, C, D 4 种药进行治疗. 假定其他条件均保持相同, 5天后测量每个病人每天的咳嗽次数如表4.13 所示(单位：次数),试比较这4 种药物的治疗效果是否相同.

表4.13 4种药物治疗效果比较表

	A	秩	B	秩	C	秩	D	秩
	80	1	133	3	156	4	194	7
	203	8	180	6	295	15	214	9
重	236	10	100	2	320	16	272	12
	252	11	160	5	448	21	330	17
	284	14			465	23	386	19
复	368	18			481	25	475	24
	457	22			279	13		
	393	20						
处理内秩和$R_{\cdot j}$		104		16		117		88
处理内平均秩$\bar{R}_{\cdot j}$		13		4		16.7		14.7

解 假设检验问题为

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_4 = \mu,$$
$$H_1 : \text{至少有两个}\mu_i \neq \mu_j.$$

统计分析：由式(4.6), 有

$$\begin{aligned} H &= \frac{12}{25 \times (25+1)} \left[\frac{104^2}{8} + \frac{16^2}{4} + \frac{117^2}{7} + \frac{88^2}{6} \right] - 3 \times (25+1) \\ &= 8.072088 \end{aligned}$$

结论：$H = 8.072088 > \chi^2_{0.05,3} = 7.814728$, 故接受$H_1$, 显示4种药物疗效不等. 在R中可以调用Kruskal-Wallis检验程序如下：

```
> drug
  [1]  80 203 236 252 284 368 457 393 133 180 100 160 156
 [14] 295 320 448 465 481 279 194 214 272 330 386 475
 > gr.drug
  [1] 1 1 1 1 1 1 1 1 2 2 2 2 3 3 3 3 3 3 3 4 4 4 4 4 4
 > kruskal.test(drug, gr.drug)
     Kruskal-Wallis rank sum test
 data:  drug and gr.drug
 Kruskal-Wallis chi-square = 8.0721, df = 3, p-value = 0.0445
 alternative hypothesis: two.sided
```

既然得到4种药物疗效不同, 那么就可以利用Dunn方法进行两两之间的比较. 成对样本共有$k(k-1)/2 = 4(4-1)/2 = 6$组, 4种药物疗效的平均秩分别为

$$\bar{R}_{\cdot 1} = 13, \quad \bar{R}_{\cdot 2} = 4, \quad \bar{R}_{\cdot 3} = 16.7, \quad \bar{R}_{\cdot 4} = 14.7.$$

$$n_1 = 8, \quad n_2 = 4, \quad n_3 = 7, \quad n_4 = 6;$$

$$\alpha = 0.05, \quad \alpha^* = 0.05/4(4-1) = 0.0042;$$

$$Z_{1-0.0042} = Z_{0.9958} = 2.638.$$

由Dunn给出的SE计算公式(4.8)和式(4.9)得如下比较表4.14：

表4.14 Dunn两两比较表

比较式	$\lvert\bar{R}_{.i} - \bar{R}_{.j}\rvert$	SE	d_{ij}	$Z_{0.9958}$
A VS B	13−4=9	4.506939	1.9969207	2.638
A VS C	$\lvert 13 - 16.7\rvert = 3.7$	3.809059	0.9713686	2.638
A VS D	$\lvert 13 - 14.7\rvert = 1.7$	3.974747	0.4277002	2.638
B VS C	$\lvert 4 - 16.7\rvert = 12.7$	4.612999	2.7530896*	2.638
B VS D	$\lvert 4 - 14.7\rvert = 10.7$	4.750731	2.2522850	2.638
C VS D	$\lvert 14.7 - 16.7\rvert$=2	4.094615	0.4884464	2.638

由上表四种疗效比较结果可知, 仅B与C有显著性差别, 其他疗效之间都不存在显著性差异. 这也说明主要的差异在B 与C, 这与直观比较吻合.

2.有结点的检验

若各处理观测值有结点时, 则H校正如下式：

$$H_{\mathrm{c}} = \frac{H}{1 - \dfrac{\sum^{g}(\tau_j^3 - \tau_j)}{n^3 - n}} \tag{4.11}$$

式中, τ_j为第j个结的长度, g为结的个数.

当统计量H_{c}的值$> \chi^2_{\alpha,k-1}$, 则接受H_1假设, 表示处理间有差异, 这时Dunn用于检验任意两组样本之间的差异公式应调整为

$$\mathrm{SE} = \sqrt{\left(\frac{n(n+1)}{12} - \frac{\sum^{g}(\tau_i^3 - \tau_i)}{12(n-1)}\right)\left(\frac{1}{n_i} + \frac{1}{n_j}\right)} \tag{4.12}$$

若$|d_{ij}| \geqslant Z_{1-\alpha*}$, 则表示第$i$与第$j$处理间有显著差异; 反之则表示差异不显著. 式中$\alpha* = \alpha/[k(k-1)]$, α为显著性水平.

例4.8 表4.15所示为3个生产番茄的土地产量(kg), 试比较3种番茄品种的产量是否相同.

表4.15 番茄品种产量比较表

	A	B	C
	2.6(9)	3.1(14)	2.5(7.5)
	2.4(5.5)	2.9(11.5)	2.2(4)
	2.9(11.5)	3.2(16)	1.5(3)
	3.1(14)	2.5(7.5)	1.2(1)
	2.4(5.5)	2.8(10)	1.4(2)
		3.1(14)	
秩和$R_{\cdot j}$	45.5	73	17.5
重复	5	6	
秩平均$\bar{R}_{\cdot j}$	9.10	12.17	3.50

注：括号内为数据在混合样本中的秩.

解 假设检验问题为

$$H_0：三种番茄产量相同,$$

$$H_1：三种番茄产量不同.$$

统计分析：由式(4.6), 有

$$\begin{aligned} H &= \frac{12}{16\times(16+1)}\left[\frac{45.5^2}{5}+\frac{73^2}{6}+\frac{17.5^2}{5}\right]-3(16+1)\\ &= 9.1529. \end{aligned}$$

由式(4.10)得

$$H_{\mathrm{c}} = \frac{9.1529}{1-\dfrac{42}{16^3-16}} = 9.2482.$$

结论：由表4.15所示, $H_{\mathrm{c}} = 9.2482 > \chi^2_{0.05,2} = 5.991$, 因而接受$H_1$, 表示3种番茄产量不相等. 有关任意两种产量之间的差异比较留做作业.

表4.16 结点校正值计算表

同秩	5.5	7.5	11.5	14	和
τ_i	2	2	2	3	
τ_i^3	6	6	6	24	$\sum(\tau_i^3-\tau_i)=42$

通常传统处理这一类问题的参数方法是在正态假设下的F检验. 如果总体分布有密度f, 可以得到H对F检验的渐近相对效率为

$$\mathrm{ARE}(H,F) = 12\sigma^2\left(\int_{-\infty}^{\infty} f^2(x)\mathrm{d}x\right)^2.$$

它和前面提到的Wilcoxon检验对t检验的ARE相等, 这是合理的. 因为无论是单样本的Wilcoxon检验、两样本的Mann-Whitney 检验还是多样本的Kruskal-Wallis检验, 与之相关的估计量都是来源于混合样本秩和的比较方法, 而单样本和两样本的t检验、多样本的F检验都基于正态假设的同样考虑, 因而它们之间的渐近相对效率自然与样本组数无关.

4.5 Jonckheere-Terpstra检验

1. 无结点Jonckheere-Terpstra检验

正如一般的假设检验问题有双边检验和单边检验问题一样, 多总体问题的备择假设也可能是有方向性的, 比如：样本的位置显现出上升和下降的趋势, 这种趋势从统计上来看是否显著？

也就是说：假设k个独立样本$X_{11},\cdots,X_{1n_1};\cdots;X_{k1},\cdots,X_{kn_k}$ 分别来自有同样形状的连续分布函数$F(x-\theta_1);\cdots;F(x-\theta_k)$, 我们感兴趣的是有关这些位置参数某一方向的假设检验问题：

$$H_0:\ \theta_1=\cdots=\theta_k \leftrightarrow H_1:\ \theta_1\leqslant\cdots\leqslant\theta_k,$$

H_1中至少有一个不等式是严格的. 如果样本呈下降趋势, 则H_1的不等式反号.

与Mann-Whitney检验类似, 如果一个样本中观测值小于另一个样本的观测值的个数较多或较少, 则可以考虑两总体的位置之间有大小关系. 这里的思路也是类似的.

第一步, 计算：

$$\begin{aligned}W_{ij}&=\text{样本}i\text{中观测值小于样本}j\text{中观测值的个数}\\&=\#(X_{iu}<X_{jv}\ \ u=1,2,\cdots,n_i,v=1,2,\cdots,n_j).\end{aligned}$$

第二步, 对所有的W_{ij}在$i<j$范围求和, 这样就产生了Jonckheere-Terpstra统计量：

$$J=\sum_{i<j}W_{ij}.$$

它从0到$\sum\limits_{i<j}n_in_j$变化, 利用Mann-Whitney统计量的性质容易得到如下定理.

定理4.5 在H_0成立的条件下,

$$E_{H_0}(J)=\frac{1}{4}\left(N^2-\sum_{i=1}^{k}n_i^2\right),$$
$$\mathrm{var}_{H_0}(J)=\frac{1}{72}\left[N^2(2N+3)-\sum_{i=1}^{k}n_i^2(2n_i+3)\right].$$

其中, $N=\sum_{i}^{k}n_i$. 类似于Wilcoxon-Mann-Whitney统计量, 当J大时, 应拒绝零假设. 从(n_1,n_2,n_3) 及检验水平α得到在零假设下的临界值c,它满足$P(J\geqslant c)=\alpha$.

当样本量大, 超过表的范围时, 可以用正态近似, 有下面定理.

定 理4.6 在H_0成 立 的 条 件 下, 当$\min\{n_1,n_2,\cdots,n_k\}\ \to\ \infty$时,而 且 $\lim\limits_{n_i\to+\infty}\dfrac{n_i}{\sum\limits_{i=1}^{k}n_i}=\lambda_i\in(0,1)$, 则

$$Z=\frac{J-\left(N^2-\sum_{i=1}^{k}n_i^2\right)/4}{\sqrt{\left[N^2(2N+3)-\sum_{i=1}^{k}n_i^2(2n_i+3)\right]/72}}\xrightarrow{\mathcal{L}}N(0,1).$$

这样, 在给定水平α, 如果$J\geqslant E_{H_0}(J)+Z_\alpha\sqrt{\mathrm{var}_{H_0}(J)}$,拒绝零假设.

例4.9 为测试不同的医务防护服的功能, 让三组体质相似的受试者分别着不同的防护服装, 记录受试者每分钟心脏跳动的次数, 每人试验5次, 得到5次平均数列于表4.17. 医学理论判断, 这三组受试的心跳次数可能存在如下关系：第一组⩽第二组⩽第三组. 下面用这些数据验证这一论断是否可靠.

表4.17 三组受试心跳次数测试数据

第一组	125	136	116	101	105	109		
第二组	122	114	131	120	119	127		
第三组	128	142	128	134	135	131	140	129

解 设$\theta_i,i=1,2,3$表示第i组的位置参数, 则假设检验问题为

$$H_0:\ \theta_1=\theta_2=\theta_3\leftrightarrow H_1:\ \theta_1\leqslant\theta_2\leqslant\theta_3.$$

因此采用Jonckheere-Terpstra检验,计算W_{ij}如下：

$$W_{12}=25,\quad W_{13}=42,\quad W_{23}=44.5.$$

因此, $J = W_{12} + W_{13} + W_{23} = 111.5$. 经图4.5计算得$P(J \geqslant 111.5) = 0.02/2 = 0.01$, 因此有理由拒绝零假设$H_0$, 认为医学临床经验在显著性水平$\alpha > 0.02$下是可靠的.

在大样本情况下, 因为$n_1 = n_2 = 6, n_3 = 8$, 则有$E(J) = 66, \sqrt{\mathrm{var}(J)} = 14.38$. 因此, $z = \dfrac{112 - 66}{14.38} = 3.198$, $P(z \geqslant 3.198) = 0.008$. 因此, 可以在水平$\alpha \geqslant 0.01$ 时拒绝零假设, 也就是说, 这三个总体的位置的确有上升趋势.

在R中, 需要加载软件包clinfun, 用其中的函数jonckheere.test求解JT统计量和计算p 值, jonckheere.test在R中的主要作用是判断组的位置参数是否有显著的大小顺序, 其中的p值就是用JT统计量的正态分布近似计算出来的.

```
{
>     G1=c(125,136,116,101,105,109)
>     G2=c(122,114,131,120,119,127)
>     G3=c(128,142,128,134,135,131,140,129)
>     G123 <- list(G1,G2,G3)
>     n <- c(length(G1),length(G2),length(G3))
>     group_label <- as.ordered(factor(rep(1:length(n),n)))
>     jonckheere.test(unlist(G123), group_label, alternative="increasing")
>     Jonckheere-Terpstra test
      data:
      JT = 111.5, p-value = 0.0007822
      alternative hypothesis: increasing
      Warning message:
      In jonckheere.test(unlist(G123), group_label, alternative = "increasing") :
      Sample size > 100 or data with ties
      p-value based on normal approximation. Specify nperm for permutation p-value
 }
```

另外, 作为比较, 也给出SPSS的结果如表所示.

Jonckheere-Terpstra Test [a]

	VAR00001
Number of Levels in VAR00002	3
N	20
Observed J-T Statistic	111.500
Mean J-T Statistic	66.000
Std. Deviation of J-T Statistic	14.375
Std. J-T Statistic	3.165
Asymp. Sig. (2-tailed)	0.002

a. Grouping Variable: VAR00002

各组数据的箱线图如图4.5所示.

2. 带结点的Jonkheere-Terpstra检验

如果有结出现,则W_{ij} 可稍微变形为

$$W_{ij}^{*}=\#(X_{ik}<X_{jl},\ \ k=1,2,\cdots,n_i,l=1,2,\cdots,n_j)+\frac{1}{2}\#(X_{ik}=X_{jl},\ \ k=1,2,\cdots,n_i,l=1,2,\cdots,n_j). \tag{4.13}$$

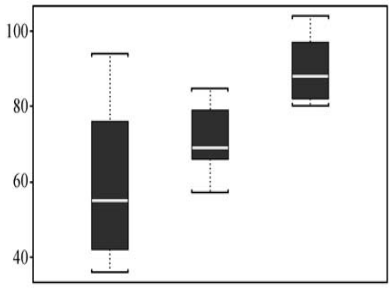

图 4.5: 医学防护服的效果比较箱线图

J也相应地变为

$$J^{*}=\sum_{i<j}W_{ij}^{*}. \tag{4.14}$$

类似于Wilcoxon-Mann-Whitney 统计量, 当J^{*}大时, 应拒绝零假设. 对于有结时Jonkheere-Terpstra统计量J^{*}的零分布, 由于它与结统计量有关, 因此造表比较困难. 但是当样本容量较大时, 可用如下的正态近似：当$\min(n_1,n_2,\cdots,n_k)\to$

∞时,

$$\frac{J^*-E_{H_0}(J^*)}{\sqrt{\mathrm{var}_{H_0}(J^*)}} \xrightarrow{\mathcal{L}} N(0,1).$$

其中

$$E_{H_0}(J^*)=\frac{N^2-\sum_{i=1}^{k}n_i^2}{4},$$

$$\begin{aligned}&\mathrm{var}_{H_0}(J^*)\\ =&\frac{1}{72}\left[N(N-1)(2N+5)-\sum_{i=1}^{k}n_i(n_i-1)(2n_i+5)-\sum_{i=1}^{k}\tau_i(\tau_i-1)(2\tau_i+5)\right]\\ &+\frac{1}{36N(N-1)(N-2)}\left[\sum_{i=1}^{k}n_i(n_i-1)(n_i-2)\right]\cdot\left[\sum_{i=1}^{k}\tau_i(\tau_i-1)(\tau_i-2)\right]\\ &+\frac{1}{8N(N-1)}\left[\sum_{i=1}^{k}n_i(n_i-1)\right]\cdot\left[\sum_{i=1}^{k}\tau_i(\tau_i-1)\right].\end{aligned}$$

其中, $\tau_1,\tau_2,\cdots,\tau_k$为混合样本的结统计量. 由大样本近似, 就可以对有结的情况进行检验.

例4.10 为研究三组教学法对儿童记忆英文单词能力的影响, 将18名英文水平、智力、年龄等各方面条件相当的儿童随机分成三组, 每组分别采用不同的教学法施教. 在学习一段时间后对三组学生记忆英文单词的能力进行测试, 测试成绩如下. 教学法的研究者经验认为三组成绩应该按A, B, C次序增加排列(两个不等号中至少有一个是严格的). 表4.18列出他们的测试成绩, 判断研究者的经验是否可靠.

表4.18 三组教学法的测验结果

A	40	35	38	43	44	41
B	38	40	47	44	40	42
C	48	40	45	43	46	44

解 本例的假设检验问题为:

$$H_0:\text{三组成绩相等}\leftrightarrow H_1:\theta_1\leqslant\theta_2\leqslant\theta_3.$$

易得$W_{12}^*=22$, $W_{13}^*=30.5$, $W_{23}^*=26.5$, 因此由式(4.13)得$J^*=79$. 查表得p值等于0.02306, 对水平$\alpha\geqslant 0.02306$能拒绝零假设. 如果用正态近似, 有p值等于0.0217, 结果和精确的比较一致.

附注：Jonckheere-Terpstra检验是由Terpstra(1952)和Jonckheere(1954)独立提出的, 它比Kruskal-Wallis检验有更强的势. Daniel(1978)和Leach(1979)对该检验进行过详细的说明.

4.6 Friedman秩方差分析法

前面的Kruskal-Wallis检验和Jonckheere-Terpstra检验都是针对完全随机试验数据的分析方法. 当各处理的样本重复数据存在区组之间的差异时, 必须考虑区组对结果的影响. 对于随机区组的数据, 传统的方差分析要求试验误差是正态分布的, 当数据不符合方差分析的正态前提时, Friedman(1937)建议采用秩方差分析法. Friedman检验对试验误差没有正态分布的要求, 仅仅依赖于每个区组内所观测的秩次.

1. Friedman检验的基本原理

假设有k个处理和b个区组, 数据观测值如表4.19所示.

表4.19 完全随机区组数据分析结构表(x_{ij})

		样本1	样本2	$\cdots$	样本k
区组	区组1	x_{11}	x_{12}	$\cdots$	x_{1k}
	区组2	x_{21}	x_{22}	$\cdots$	x_{2k}
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	区组b	x_{b1}	x_{b2}	$\cdots$	x_{bk}

与大部分方差分析的检验问题一样, 这里关于位置参数的假设检验问题为

$$H_0:\ \theta_1=\cdots=\theta_k \leftrightarrow H_1:\ \exists i,j\in 1,2,\cdots,k,\ \ \theta_i\neq\theta_j. \tag{4.15}$$

由于区组的影响, 不同区组中的秩没有可比性,比如要对比不同化肥的增产效果, 优质土地即便不施肥, 其产量也可能比施了优等肥的劣质土地的产量高. 但是, 如果按照不同的区组收集数据,那么同一区组中的不同处理之间的比较是有意义的, 也就是说,假设其他影响因素相同的情况下,在劣质土地上比较不同的肥料增产效果是有意义的. 因此,首先应在每一个区组内分配各处理的秩, 从而得到秩数据表4.20.

表4.20 完全随机区组秩数据表(R_{ij})

	样本1	样本2	$\cdots$	样本k	和$R_{i\cdot}$
区组1	R_{11}	R_{12}	$\cdots$	R_{1k}	$\frac{k(k+1)}{2}$
区组2	R_{21}	R_{22}	$\cdots$	R_{2k}	$\frac{k(k+1)}{2}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
区组b	R_{b1}	R_{b2}	$\cdots$	R_{bk}	$\frac{k(k+1)}{2}$
秩和$R_{\cdot j}$	$R_{\cdot 1}$	$R_{\cdot 2}$	$\cdots$	$R_{\cdot k}$	$\frac{k(k+1)}{2}$

如果R_{ij}表示第i个区组中第j处理在第i区组中的秩, 则秩按照处理求和为$R_{\cdot j}=\sum_{i=1}^{b}R_{ij},\ j=1,2,\cdots,k,\ \bar{R}_{\cdot j}=R_{\cdot j}/b.$

在零假设成立的情况下, 各处理的平均秩$\bar{R}_{\cdot j}$有下面的性质.

定理4.7 在零假设H_0下,

$$\begin{aligned}E(\bar{R}_{\cdot j})&=\frac{k+1}{2},\\ \operatorname{var}(\bar{R}_{\cdot j})&=\frac{k^2-1}{12b},\\ \operatorname{cov}(\bar{R}_{\cdot i},\bar{R}_{\cdot j})&=-\frac{k+1}{12}.\end{aligned}$$

证明 易知

$$\begin{aligned}R_{\cdot\cdot}&=b(1+2+\cdots+k)=bk(k+1)/2,\\ \hat{R}_{\cdot\cdot}&=R_{\cdot\cdot}/bk=(k+1)/2.\end{aligned}$$

$$\operatorname{var}(\bar{R}_{ij})=\sum^{b}\sum^{k}(R_{ij}-\bar{R}_{\cdot\cdot})^2/bk$$

$$\begin{aligned}&=\frac{1}{bk}\left[\sum\sum R_{ij}^2-R_{\cdot\cdot}^2/bk\right]\\ &=\frac{1}{bk}\left[\frac{bk(k+1)(2k+1)}{6}-\frac{bk(k+1)^2}{4}\right]\\ &=\frac{(k+1)(k-1)}{12}.\end{aligned}$$

各处理间平方和为

$$\begin{aligned}\mathrm{SSt} &= n\sum(\bar{R}_{\cdot j}-\bar{R}_{\cdot\cdot})^2\\ &=\sum^{k} R_{\cdot j}^2/b - R_{\cdot\cdot}^2/bk\\ &=\sum R_{\cdot j}^2/b - bk(k+1)^2/4.\end{aligned}$$

Friedman的Q'公式为

$$Q'=\frac{\mathrm{SSt}}{\mathrm{var}(R_{ij})}=\frac{12}{(k+1)(k-1)}\left[\sum R_{\cdot j}^2/b - bk(k+1)^2/4\right].$$

Friedman建议用$(k-1)/k$乘Q'得校正式

$$\begin{aligned}Q&=\frac{12}{bk(k+1)}\sum R_{\cdot j}^2-\frac{12bk(k+1)^2(k-1)}{4(k+1)(k-1)k}\\ &=\frac{12}{bk(k+1)}\sum R_{\cdot j}^2-3b(k+1).\end{aligned}\tag{4.16}$$

Q值近似自由度$\nu=k-1$的χ^2分布.

当数据有相同秩时, Q值校正如下式：

$$Q_{\mathrm{c}}=\frac{Q}{1-\dfrac{\sum^{g}(\tau_i^3-\tau_i)}{bk(k^2-1)}}.\tag{4.17}$$

式中, τ_i为第i个结的长度, g为结的个数. 结论：若实测$Q<\chi^2_{0.05,k-1}$, 则不拒绝H_0, 反之则接受H_1.

例4.11 设有来自A, B, C, D 4个地区的四名厨师制作名菜京城水煮鱼, 想比较它们的品质是否相同. 经四位美食评委评分结果如表4.21所示, 试测验4个地区制作的水煮鱼这道菜品质有无区别.

解 由于不同评委在口味和美学欣赏上存在差异, 因此适合用Freidman检验方法比较.

表4.21 评委对四名厨师的评分数据表

美食	地 区				
评委	A	B	C	D	
1	85(4)	82(2)	83(3)	79(1)	
2	87(4)	75(1)	86(3)	82(2)	
3	90(4)	81(3)	80(2)	76(1)	
4	80(3)	75(1.5)	81(4)	75(1.5)	
秩和$R_{\cdot j}$	15	7.5	12	5.5	$R_{\cdot\cdot}=40$

注：表中括号内数据为每位评委品尝四种菜后所给评分的秩.

假设检验问题为

$$H_0：4个地区的京城水煮鱼品质相同,$$
$$H_1：4个地区的京城水煮鱼品质不同.$$

统计分析：$b=4$(区组数), $k=4$(处理数).

结点校正如表4.22所示.

表4.22 结点校正计算表

相同的秩	1.5	
τ_i	2	
$\tau_i^3-\tau_i$	6	$(\tau_i^3-\tau_i)=6$

由式(4.15), 有

$$\begin{aligned}Q=&\frac{12}{4\times4\times(4+1)}[15^2+7.5^2+12^2+5.5^2]-3\times4\times(4+1)\\=&8.325.\end{aligned}$$

由式(4.16), 结合表4.22有

$$\begin{aligned}Q_{\mathrm{c}}=&\frac{8.325}{1-\dfrac{6}{4\times4(4^2-1)}}\\=&8.5385.\end{aligned}$$

结论：实际测量$Q_{\mathrm{c}}=8.5385>\chi^2_{0.05,3}=7.814$, 接受$H_1$, 认为4个地区的菜品质上存在显著差异. 在R中进行Friedman检验的函数语法如下：

```
friedman.test(y, groups, blocks)
```

例4.11的运算程序如下：

```
> BeijingFish
  [1] 85 82 83 79 87 75 86 82 90 81 80 76 80 75 81 75
> treat.BF
  [1] 1 2 3 4 1 2 3 4 1 2 3 4 1 2 3 4
> block.BF
  [1] 1 1 1 1 2 2 2 2 3 3 3 3 4 4 4 4
> friedman.test(BeijingFish, treat.BF, block.BF)
```

```
Friedman rank sum test
data:  BeijingFish and treat.BF and block.BF
Friedman chi-square = 8.5385, df = 3, p-value = 0.0361
alternative hypothesis: two.sided
```

2. Hollander-Wolfe两处理间比较

当秩方差分析结果样本之间有差异时, Hollander-Wolfe(1973)提出两样本(处理)间的比较公式:

$$D_{ij} = |R_{.i} - R_{.j}|/\mathrm{SE} = \frac{|R_{.i} - R_{.j}|}{\sqrt{b^2 S^2_{R_{ij}}(\frac{1}{b} + \frac{1}{b})}}. \tag{4.18}$$

式中$R_{.i}$与$R_{.j}$为第i与第j样本(处理)秩和, $S^2_{R_{ij}}$是R_{ij}的方差的无偏估计，由于

$$S^2_{R_{ij}} = \tfrac{k}{k-1}\mathrm{var}R_{ij} == \frac{(k+1)(k-1)}{12} \times \frac{k}{k-1} = \frac{k(k+1)}{12},$$

$$\mathrm{SE} \qquad = \qquad \sqrt{\frac{b^2 k(k+1)}{12}\left(\frac{2}{b}\right)} = \sqrt{bk(k+1)/6},$$

若有相同秩, 则

$$\mathrm{SE} = \sqrt{\frac{bk(k+1)}{6} - \frac{b\sum_{i=1}^{g}(\tau_i^3 - \tau_i)}{6(k-1)}} \tag{4.19}$$

式中, τ_i为同秩观测值个数, g为同秩组数. 当实测$|D_{ij}| \geqslant Z_{1-\alpha^*}$时, 表示两样本间有差异, 反之则无差异. $\alpha^* = \alpha/k(k-1)$, α为显著水平, $Z_{1-\alpha^*}$为标准正态分布分位数.

例4.12 由例4.7知, 4个地区所做的水煮鱼品质上有显著差异, 成对样本比较有$k(k-1)/2 = 4(4-1)/2 = 6$种, 四种水煮鱼的秩和分别为

$$R_{.1} = 15, \quad R_{.2} = 7.5, \quad R_{.3} = 12, \quad R_{.4} = 5.5.$$

设

$$\alpha = 0.05, \quad \alpha^* = 0.05/4(4-1) = 0.0083,$$

$$Z_{1-0.0083} = Z_{0.9917} = 2.395.$$

由式(4.18)得

$$\mathrm{SE} = \sqrt{\frac{4 \times 4(4+1)}{6} - \frac{4 \times 6}{6(4-1)}} = 3.464.$$

再利用式(4.17)得比较表4.23.

表4.23 两两处理的Hollander-Wolfe计算表

比较式	$\lvert R_{.i} - R_{.j}\rvert$	SE	D_{ij}	$Z_{0.9917}$
A VS B	15−7.5=7.5	3.464	2.165	2.395
A VS C	15−12=3	3.464	0.866	2.395
A VS D	15−5.5=9.5	3.464	2.742*	2.395
B VS C	—7.5−12—=4.5	3.464	1.299	2.395
B VS D	7.5−5.5=2	3.464	0.577	2.395
C VS D	12−5.5=6.5	3.464	1.876	2.395

由表4.23四种水煮鱼品质比较结果可知, 仅A与D有差别,其他水煮鱼品质间差异不显著.

4.7 随机区组数据的调整秩和检验

当随机区组设计数据的区组数较大或处理组数较小时,Friedman检验的效果就不是很好了. 因为Friedman检验的编秩是在每一个区组内进行的, 这种编秩的方法仅限于区组内的效应(response), 不同区组间效应的直接比较是无意义的. 为了去除区组效应, 可以用区组的平均值或中位数作为区组效应的估计值, 然后用每个观测值与估计值相减来反映处理之间的差异, 这样做就可能消除区组之间的差异.

于是Hodges和Lehmmann于1962年提出了调整秩和检验(aligned ranks test), 也称为Hodges-Lehmmann检验, 简记为HL检验. 对于假设检验问题:

$$H_0:\ \theta_1 = \cdots = \theta_k \leftrightarrow H_1:\ \exists i, j \in 1, 2, \cdots, k,\ \ \theta_i \neq \theta_j.$$

样本结构如表4.24所示, 调整秩和检验的主要计算步骤如下.

(1) 对每一个区组$i, i = 1, 2, \cdots, b$来说, 计算其某一位置估计值, 如均值或中位数. 以下计算以均数为例, 即$\bar{X}_{i.} = \dfrac{1}{k}\sum_{j=1}^{k} X_{ij}$.

(2) 每一个区组中的每个观测值减去均值, 即$\mathrm{AX}_{ij} = X_{ij} - \bar{X}_{i.}$, 相减后的值称为调整后的观测值(aligned observation).

(3) 对调整后的观测值, 像Kruskal-Wallis检验中一样, 对全部数据求混合后的秩, 相同的用平均秩, AX_{ij} 的秩仍然记为R_{ij}, 这样编得的秩为调整秩(aligned ranks).

(4) 用$\bar{R}_{\cdot j}$表示第j个处理的平均秩, 即$\bar{R}_{\cdot j}=\dfrac{1}{b}\sum\limits_{i=1}^{b}R_{ij}$. 在零假设之下, $\bar{R}_{\cdot j}$应与$\dfrac{1}{kb}\sum R_{ij}=\dfrac{kb+1}{2}$ 相等. 于是可以使用

$$\tilde{Q}=c\cdot\sum_{j=1}^{k}\left(\bar{R}_{\cdot j}-\frac{kb+1}{2}\right)^2$$

作为检验统计量, 当$\tilde{Q}$取大值时, 考虑拒绝H_0.

(5) Hodges-Lehmmann指出, 当

$$c=\frac{(k-1)b^2}{\sum\limits_{i,j}(R_{ij}-\bar{R}_{i\cdot})^2},$$

这里$R_{i\cdot}=\dfrac{1}{k}\sum\limits_{j=1}^{k}R_{ij}$, 即

$$\begin{aligned}\tilde{Q}&=\frac{(k-1)b^2}{\sum\limits_{i,j}(R_{ij}-\bar{R}_{i\cdot})^2}\sum_{j=1}^{k}\left(\bar{R}_{\cdot j}-\frac{kb+1}{2}\right)^2\\&=\frac{(k-1)\left[\sum\limits_{j=1}^{k}R_{\cdot j}^2-\dfrac{kb^2(kb+1)^2}{4}\right]}{\dfrac{1}{6}kb(kb+1)(2kb+1)-\dfrac{1}{k}\sum\limits_{i=1}^{b}R_{i\cdot}^2},\end{aligned}$$

其中, $R_{\cdot j}=\sum\limits_{i=1}^{b}R_{ij}$, $R_{i\cdot}=\sum\limits_{j=1}^{k}R_{ij}$, 检验统计量的$\tilde{Q}$零假设分布近似于自由度$\nu=k-1$的$\chi^2$分布, 所以结果可以和$\chi^2$分布表进行比较, 这里$k$为处理组数.

当数据中有结点存在时, 用平均秩法定秩, 这时$\tilde{Q}'$统计量为

$$\tilde{Q}'=\frac{(k-1)\left[\sum\limits_{j=1}^{k}R_{\cdot j}^2-\dfrac{kb^2(kb+1)^2}{4}\right]}{\sum\limits_{i,j}R_{ij}^2-\dfrac{1}{k}\sum\limits_{i=1}^{b}R_{i\cdot}^2}.$$

例4.13 现研究一种高血压患者的血压控制效果, 经验表明治疗效果与病人本身的肥胖和身高类型有关的. 现将高血压病人按控制方法分为四类：A, B, C,

D. 从这四种病人中随机抽取8名病人做完全区组设计试验. 进行一段时间的高血压控制治疗后, 测量血压指数(经过一定变化后)如表4.24所示.

表4.24 高血压患者血压控制效果数据表

处理	区 组							
	Ⅰ	Ⅱ	Ⅲ	Ⅳ	Ⅴ	Ⅵ	Ⅶ	Ⅷ
A	23.1	57.6	10.5	23.6	11.9	54.6	21.0	20.3
B	22.7	53.2	9.7	19.6	13.8	47.1	13.6	23.6
C	22.5	53.7	10.8	21.1	13.7	39.2	13.7	16.3
D	22.6	53.1	8.3	21.6	13.3	37.0	14.8	14.8

试问4种血压控制对四种病人降压效果是否相同?

对于这个问题我们先用Friedman检验, 求出秩如下表4.25.

表4.25 Friedman检验区组内秩表

处理	秩								$R_{\cdot j}$
A	4	4	3	4	1	4	4	3	27
B	3	2	2	1	4	3	1	4	20
C	1	3	4	2	3	2	2	2	19
D	2	1	1	3	2	1	3	1	14

由此可计算得Friedman检验统计量$Q = 6.45$, 查表知, 此时的p值为0.091, 如果取$\alpha = 0.05$, 则不能拒绝原假设. 但是从原始数据表中可以看出, 区组间的差异是显然的, 于是使用HL检验如下.

首先计算这8个区组效应的估计值分别为

Ⅰ	Ⅱ	Ⅲ	Ⅳ	Ⅴ	Ⅵ	Ⅶ	Ⅷ
22.735	54.4	9.825	21.475	13.175	44.475	15.775	18.75

由此则可以得到下面全体$X_{ij} - X_{\cdot j}$的秩, 如表4.26所示.

表4.26 Hodges-Lehmmann秩数据表

处理	平均秩								秩和
A	21	29	24	27	10	32	31	26	200
B	18	11	16.5	7	23	28	5	30	138.5
C	15	13	25	14	22	2	6	4	101
D	16.5	9	8	19.5	19.5	1	12	3	88.5

计算HL检验统计量的值为8.53. 由χ^2近似知,其检验的p值为0.036, 对于$\alpha = 0.05$, 拒绝零假设,即认为对病人采取不同的高血压控制,会影响降压效果,这与直观想像是吻合的, 这也表明Friedman 检验与HL 检验是有着显著不同的.

4.8 Cochran检验

一个完全区组设计的特殊情况是观测值只取”是”或”否”、”同意”或”不同意”、”1” 或”0” 等二元定性数据. 这时,由于有太多的重复数据, 秩方法的应用受到限制.Cochran(1950)提出Q检验法, 测量多处理之间的差异是否存在.

假定有k个处理和b个区组, 样本为计数数据,其数据形态如表4.27所示.

表4.27 只取二元数据的完全随机区组数据表

		处理				
		1	2	$\cdots$	k	和
	1	n_{11}	n_{12}	$\cdots$	n_{1k}	$n_{1.}$
区	2	n_{21}	n_{22}	$\cdots$	n_{2k}	$n_{2.}$
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
组	b	n_{b1}	n_{b2}	$\cdots$	n_{bk}	$n_{b.}$
	和	$n_{.1}$	$n_{.2}$	$\cdots$	$n_{.k}$	N

假设检验问题为

$$H_0: k\text{个总体分布相同(或各处理发生的概率相等)},$$
$$H_1: k\text{个总体分布不同(或各处理发生的概率不等)}.$$

统计分析：

以上表观测值$n_{ij} \in \{0,1\}$为计数数据, $n_{.j}$为第j处理中1的个数, 即$n_{.j} = \sum_{i=1}^{b} n_{ij}, j = 1,2,\cdots,k$, 显然各个处理之间的差异可以由$n_{.j}$之间的差异显示出来. n_i为每一区组中1的个数. $\sum_{j=1}^{k} n_{.j} = \sum_{i=1}^{b} n_{i.} = N$, 每格成功概率用$p_{ij}$表示.

当H_0成立时, 每一区组i内的成功概率p_{ij}相等, 对$\forall j = 1,2,\cdots,k, \forall i, p_{i1} = p_{i2} = \cdots = p_{ik} = p_{i.}$, n_{ij}服从两点分布$b(1,p_{i.})$.

$\mathrm{var}(n_{\cdot j})$为$n_{\cdot j}$的方差：

$$\begin{aligned}\mathrm{var}(n_{\cdot j}) &= \mathrm{var}\left(\sum_{i=1}^{b} n_{ij}\right)\\ &= \sum_{i=1}^{b} \mathrm{var}(n_{ij})\\ &= \sum_{i=1}^{b} \hat{p}_{ij}(1-\hat{p}_{ij}).\end{aligned} \tag{4.20}$$

将$\hat{p}_{ij} = \hat{p}_{i\cdot} = n_{i\cdot}\dfrac{1}{k}$代入上式, 得

$$\mathrm{var}(n_{\cdot j}) = \sum_{i=1}^{b} n_{i\cdot}\frac{1}{k}\left(1 - n_{i\cdot}\frac{1}{k}\right)$$

$$= \frac{1}{k^2}\sum_{i=1}^{b}(kn_{i\cdot} - n_{i\cdot}^2).$$

上式的估算值一般都很小, 因而用$k/(k-1)$修正得到下式：

$$\mathrm{var}(n_{ij}) = \frac{n_{i\cdot}(k-n_{i\cdot})}{k(k-1)}. \tag{4.21}$$

将式(4.20)代入式(4.19), 得到估计值为

$$\mathrm{var}(n_{\cdot j}) = \sum_{i=1}^{b} n_{i\cdot}(k-n_{i\cdot})/[k(k-1)]. \tag{4.22}$$

在大样本情况下, $n_{\cdot j}$为近似正态分布, 即

$$\frac{n_{\cdot j} - E(n_{\cdot j})}{\sqrt{\mathrm{var}(n_{\cdot j})}} \overset{L}{\sim} N(0,1). \tag{4.23}$$

式中$E(n_{\cdot j})$为$n_{\cdot j}$的期望值, 一般用样本估计：

$$E(n_{\cdot j}) = \frac{1}{k}\sum n_{\cdot j} = \frac{N}{k}. \tag{4.24}$$

一般$n_{\cdot j}$间并非互相独立, 但当$n_{\cdot j}$足够大时, Tate 和Brown(1970)认为$n_{\cdot j}$近似独立, 故式(4.23)平方后可以累加得自由度$v = k-1$的近似χ^2分布为

$$\sum_{j=1}^{k}\left[\frac{n_{\cdot j} - E(n_{\cdot j})}{\sqrt{\mathrm{var}(n_{\cdot j})}}\right]^2 = \sum_{j=1}^{k}\frac{[n_{\cdot j} - E(n_{\cdot j})]^2}{\mathrm{var}(n_{\cdot j})}. \tag{4.25}$$

将式(4.23)及式(4.21)代入式(4.24)得Cochran Q值为

$$Q=\sum_{j=1}^{k}\frac{\left(n_{\cdot j}-\frac{N}{k}\right)^2}{\sum n_{i\cdot}(k-n_{i\cdot})/[k(k-1)]}$$
$$=\frac{(k-1)\left[\sum n_{\cdot j}^2-\left(\sum n_{\cdot j}\right)^2/k\right]}{\sum n_{i\cdot}-\sum n_{i\cdot}^2/k}. \tag{4.26}$$

结论：当检验统计量的值$Q<\chi^2_{0.05,k-1}$, 不能拒绝H_0, 反之则接受H_1.

例4.14 设有A, B, C三种榨汁机分给10位家庭主妇使用, 用以比较三种榨汁机受喜爱程度是否相同. 对于喜欢的品牌给1分, 否则给0分, 调查结果如表4.28所示.

表4.28 家庭主妇对三种榨汁机喜爱与否统计表

		主妇										
		1	2	3	4	5	6	7	8	9	10	和$n_{\cdot j}$
榨	A	0	0	0	1	0	0	0	0	0	1	2
汁	B	1	1	0	1	0	1	0	0	1	1	6
机	C	1	1	1	1	1	1	1	1	1	0	9
和	$n_{i\cdot}$	2	2	1	3	1	2	1	1	2	2	17

假设检验问题为

H_0 : 三种榨汁机受喜爱程度相同,

H_1 : 三种榨汁机受喜爱程度不同.

统计分析：由于各主妇每人饮食和做家务的习惯不同, 对各榨汁机的功能使用情况也有差异, 故应以主妇为区组. 由式(4.25), 有

$$\sum n_{\cdot j}=\sum R_j=17, k=3,$$
$$\sum n_{i\cdot}^2=2^2+2^2+\cdots+2^2=33,$$
$$\sum n_{\cdot j}^2=2^2+6^2+9^2=121,$$
$$Q=\frac{(3-1)(121-17^2/3)}{17-33/3}=\frac{49.3333}{6}$$
$$=8.2222.$$

结论：现在实际测得$Q = 8.2222 > \chi^2_{0.05,2} = 5.991$, 接受$H_1$, 表示三种榨汁机受喜爱程度不同, 以C榨汁机较受欢迎. 实际上, 从三种榨汁机受喜爱的概率点估计($\hat{p}_{\cdot,1} = 0.12, \hat{p}_{\cdot,2} = 0.35, \hat{p}_{\cdot,3} = 0.53$) 也支持了这一结论.

该题的R程序如下：

```
candid1=c(0,0,0,1,0,0,0,0,0,1)
candid2=c(1,1,0,1,0,1,0,0,1,1)
candid3=c(1,1,1,1,1,1,1,1,1,0)
candid=matrix(c(candid1,candid2,candid3),nrow=10,ncol=3)
nidot.candid=apply(candid,1,sum)
ndotj.candid=apply(candid,2,sum)
k=ncol(candid)
Q=(k-1)*((k*sum(ndotj.candid^2)-(sum(ndotj.candid))^2))/
+(k*sum(nidot.candid)-sum(nidot.candid^2))
pvalue.candid=pchisq(Q,k-1,lower.tail=F)
pvalue.candid
[1] 0.01638955
```

由于p值0.0164远小于0.05, 于是拒绝原假设.

4.9 Durbin不完全区组分析法

由4.1节的预备知识可以知道, 当处理组非常大, 而区组中可允许样本量有限时, 在一个区组中很难包含所有处理, 于是出现了不完全的数据设计结构, 其中较为常见的是均衡不完全区组BIB设计. Durbin于1951年提出一种秩检验, 该检验能用于均衡不完全区组设计中.

采用4.1节的记号, X_{ij}表示第j个处理第i个区组中的观测值, R_{ij}为在第i个区组中第j个处理的秩, 按处理相加得到$R_{i\cdot} = \sum_{j}^{k} R_{ij},\ i = 1, 2, \cdots, b.$

当H_0成立时, 不难得到

$$\mathrm{E}R_{i\cdot} = \frac{r(t+1)}{2},\ i = 1, 2, \cdots, k.$$

k个处理的秩和在H_0下是非常接近的, 秩总平均为$\frac{1}{k}\sum_{i=1}^{k} R_{i\cdot} = \frac{1}{k}\sum R_{ij} = \frac{r(t+1)}{2}$.

当某处理效应大时, 则反映在秩上, 其秩和与总平均之间的差异也较大, 于是可以构造统计量：

$$D=\frac{12(k-1)}{rk(t^2-1)}\sum_{i=1}^{k}\left[R_{i.}-\frac{r(t+1)}{2}\right]^2 \tag{4.27}$$

$$=\frac{12(k-1)}{rk(t^2-1)}\sum_{i=1}^{k}R_{i.}^2-\frac{3r(k-1)(t+1)}{t-1}. \tag{4.28}$$

显然, 在完全区组设计($t=k, r=b$)时, 上面的统计量等同于Friedman 统计量. 对于显著性水平α, 如果D很大, 比如大于或等于$D_{1-\alpha}$, 这里$D_{1-\alpha}$ 为最小的满足$P_{H_0}(D\geqslant D_{1-\alpha})=\alpha$的值, 则可以对于水平$\alpha$拒绝零假设. 零假设下精确分布只对有限的几组$k$和$b$计算过.实践中人们常用大样本近似. 在零假设下, 对于固定的k和t, 当$r\to\infty$时, $D\to\chi^2_{(k-1)}$. 对于小样本,该χ^2 近似不很精确.

此外, 当数据中有结存在时, 实践表明, 只要其长度不大, 结统计量对D统计量的影响不大.

例4.15 设需要对四种饲料(处理)的养猪效果进行试验,用以比较饲料的质量. 选4胎母猪所生的小猪进行试验,每头所生的小猪体重相当, 选择3头进行实验.3个月后测量所有小猪增加的体重(1b) 如表4.29所示,试比较四种饲料品质有无差别.

表4.29 四种饲料的养猪效果数据表

		区组(胎别)				
		Ⅰ	Ⅱ	Ⅲ	Ⅳ	和$n_{.j}$
	A	73(1)	74(1)		71(1)	3
饲	B		75(2.5)	67(1)	72(2)	5.5
料	C	74(2)	75(2.5)	68(2)		6.5
	D	75(3)		72(3)	75(3)	9

注：括号内的数为各区组内按4种处理观测值大小分配的秩.

解 假设检验问题为

H_0:四种饲料质量相同,

H_1:四种饲料质量不同.

统计分析：由式(4.27), $t=4,\ \ k=3,\ \ r=3,\ \ v=4-1=3$, 则

$$\begin{aligned} Q=&\frac{12(4-1)}{3\times 4(3+1)(3-1)}(3^2+5.5^2+6.5^2+9^2)\\ &-\frac{3\times 3(4-1)(3+1)}{3-1}\\ =&60.9375-54\\ =&6.9375. \end{aligned}$$

结论：实测$Q=6.9375<\chi^2_{0.05,3}=7.814$, 不拒绝$H_0$, 没有明显迹象表明四种饲料质量之间存在差异.

习题

4.1 对A, B, C三个灯泡厂生产的灯泡进行寿命测试, 每品牌随机试验不等量灯泡, 结果得到如下列寿命数据(单位：天), 试比较三品牌灯泡寿命是否相同.

A	83	64	67	62	70
B	85	81	80	78	
C	88	89	79	90	95

4.2 在R中编写程序完成例4.7的Dunn检验.

4.3 假设有10个独立的假设检验得到如下有顺序的p 值：

0.00017	0.00448	0.00671	0.00907	0.01220
0.33626	0.39341	0.53882	0.58125	0.98617

在$\alpha=0.05$的显著性水平之下，计算Benferroni检验和BH检验拒绝的原假设的个数。

4.4 针对例4.5的第一组15个检验的数据，编写函数使用IF-PCA方法计算HC值（式(4.6)），对比它拒绝原假设的数量，与例题中的结果一样吗？

4.5 请对第一章的问题2里的Gordon研究，通过4.4编写的程序进行基因有效性的检验，绘制HC图，判断无效基因的数量。

4.6 下表是美国三大汽车公司(A, B, C 三种处理)的五种不同的车型某年产品的油耗, 在R中编写函数完成Hodges-Lehmmann调整秩和检验. 试分析不同公司的油耗是否存在差异，请将Fridman检验与Hodges-Lehmmann调整秩和检验的结果进行比较。

	Ⅰ	Ⅱ	Ⅲ	Ⅳ	Ⅴ
A	20.3	21.2	18.2	18.6	18.5
B	25.6	24.7	19.3	19.3	20.7
C	24.0	23.1	20.6	19.8	21.4

4.7 在一项健康试验中,有三种生活方式, 它们的减肥效果如下表.

生活方式	1	2	3
hline	3.7	7.3	9.0
一个月后	3.7	5.2	4.9
降低的体重	3.0	5.3	7.1
(单位500g)	3.9	5.7	8.7
	2.7	6.5	
$n_i =$	5	5	4

人们想要知道的是从这些数据能否得出它们的减肥效果(位置参数)是一样的. 如果减肥效果不等, 试根据上面这些数据选择方法检验哪一种效果最好, 哪一种最差.

4.8 为考察三位推销员甲、乙、丙的推销能力, 设计实验, 让推销员向指定的12位客户推销商品, 若顾客认为推销员的推销服务满意, 则给1分, 否则给0分, 所得结果如下. 试测验三位推销员的推销效果是否相同.请问该题目可以使用χ^2检验进行分析吗？请讨论比较的结论。

	客 人											
	1	2	3	4	5	6	7	8	9	10	11	12
甲推销员	1	1	1	1	1	1	0	0	1	1	1	0
乙推销员	0	1	0	1	0	0	0	1	0	0	0	0
丙推销员	1	0	1	0	0	1	0	1	0	0	0	1

4.9 现有A, B, C, D四种驱蚊药剂, 在南部四个地区试用,观察实验效果。受试验条件鄠限, 每种药剂只在三个地区试验,每一试验使用400只蚊子, 其死亡数如下. 如何检验四种药剂的药效是否不同?

		地区			
		1	2	3	4
	A	356	320	359	
药	B	338	340		385
剂	C	372		380	390
	D		308	332	348

案例与讨论：薪酬、学历与社会服务之间的关系？

案例背景

哈佛大学图书馆墙上有一条训言是：教育等同收入（The education level represents the income.）俗话说，知识改变命运。知识能带给人基本的生活保障。根据美国人口普查局的统

计，2008年，美国高中学历以下的人每周中位收入是426美元，高中学历的人每周中位收入是591 美元，大专学历的人每周中位收入是736 美元，大学学历的人每周中位收入是978 美元，硕士学历的人每周中位收入是1228 美元，职业性学历（如律师、医生等）的人每周中位收入是1228美元，博士学历的人每周中位收入是1555美元。（中華商報，2009年第38 期）经济合作与发展组织（OECD）发布的2008年度《教育概览》指出，各成员国追求高学历动力依然强劲，在过去10年间，劳动力市场对高学历人才的需求在大幅增加，大多数情况下，随着受教育程度的提高，收入也相应提高。国内领先的网络招聘企业中华英才网2008 年9月发布的《中华英才网2008年度薪酬报告》的人口统计学分析表明，薪酬关于学历（大专以下、大专、本科、硕士（不含MBA）、博士）整体上是呈递增关系的,但是MBA的薪酬是高于硕士（不含MBA）和博士的。

4.9.1 问题提出

企业关心学历对职工收入的影响，收集到某行业某地区20家企业人力资源部的相关数据，想借此来分析不同层次学历、工龄、职位以及工作时间与薪水之间的影响关系。

数据说明与约定

1. 数据来源：某企业人力资源部数据（见光盘）。2. 数据格式：txt纯文本格式。3. 变量说明：五个变量educ(受教育年限)，salary(入职薪水，单位为美元)，salary(目前薪水，单位为美元)，工龄(workyear)，职位TimeforService(与内部其他员工沟通协调工作时间)都是整型变量。

研讨问题

根据案例背景和数据约定，请思考以下几个问题：

(1). 不同教育年限对职员入职薪水有怎样的影响？

(2). 选取一家企业，分析员工工资与和工作时间长短有怎样的关系，如果将服务时间分成三段作为区组，再来看学历对工资的影响，会发现什么规律？

(3). 将20家企业的工资关系全部拿来分析，(2)的规律还一致的成立吗？(2)的规律不成立的企业和(2)的规律成立的企业内部职工人数有怎样的结构性差异吗？请根据数据结合文献《城市劳动力市场中户籍歧视的变化:农民工的就业与工资》给出分析。

(4). 政府每年要修补约两万个城市道路坑洞。道路坑洞积水是发生道路坍塌的一个警示信号，为有效配置资源，政府为以上20 家企业员工配备便携式智能手环。该应用利用智能设备的加速度计和GPS 数据，以非主动方式探测道路坑洞，然后将位置和坑洞数据及时上报给市政府。薪水高的员工有更多的机会接触到更多的员工，被认为有更多的机会反映坑洞问题，请问将智能手环发给薪水高的市民以便及时发现亟需修补的坑洞，有怎样的设计缺陷吗？