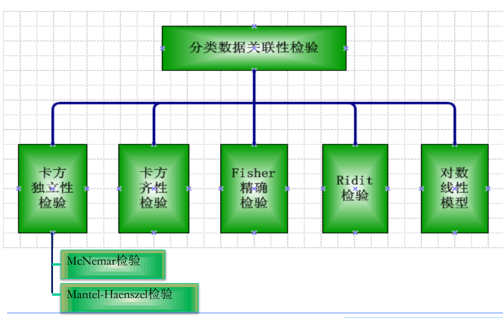
第6章 分类数据关联分析

什么是分类数据

- 统计数据的一种。指反映事物类别的数据。如人按性别分为男、女两类。
- 分类数据（categorical data）是离散数据(discrete data)。分类属性具有有限个（但可能很多）不同值，值之间无序。
- 例子: 200例肿瘤患者中A指标阳性100例，阴性100例；B指标阳性50例，阴性150例。AB都是分类变量。有AB同时阳性的患者20例，想看AB之间是否存在相关。

本章内容



问题：两个分类变量有关系吗？如何度量？

不良习惯 ------------ 健康

	得肺病	没有肺病
吸烟	90	0
不吸烟	0	90

	得肺病	没有肺病
吸烟	80	40
不吸烟	40	20

P(得肺病|吸烟)=?

P(得肺病|不吸烟)=?

$\frac{2}{3}$

明德主楼1019 王星 wangxingscy@gmail.com 82500167

6.1 r x s 列联表和χ^2检验

	B_1	B_2	$\cdots$	B_s	总和
A_1	n_{11}	n_{12}	$\cdots$	n_{1s}	$n_{1\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
A_r	n_{r1}	n_{r2}	$\cdots$	n_{rs}	$n_{r\cdot}$
总和	$n_{\cdot 1}$	$n_{\cdot 2}$	$\cdots$	$n_{\cdot s}$	$n_{\cdot\cdot}$

$n_{i\cdot} = \sum_{j=1}^{s} n_{ij}, i = 1, 2, \cdots, r,$ 表示各行之和;

$n_{\cdot j} = \sum_{i=1}^{r} n_{ij}, j = 1, 2, \cdots, s,$ 表示各列之和;

$n_{\cdot\cdot} = \sum_{j=1}^{s} n_{\cdot j} = \sum_{i=1}^{s} n_{i\cdot}$

χ^2独立性检验

	B_1	B_2	$\cdots$	B_s	总和
A_1	n_{11}	n_{12}	$\cdots$	n_{1s}	$n_{1\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
A_r	n_{r1}	n_{r2}	$\cdots$	n_{rs}	$n_{r\cdot}$
总和	$n_{\cdot 1}$	$n_{\cdot 2}$	$\cdots$	$n_{\cdot s}$	$n_{\cdot\cdot}$

假设检验问题:

$H_0: p_{ij} = p_{i\cdot} p_{\cdot j}$

构造统计量:

$$\chi^2 = \sum_{i,j} \frac{(n_{ij} - e_{ij})^2}{e_{ij}} = \sum_{i,j} \frac{(n_{ij})^2}{e_{ij}} - n_{\cdot\cdot} \qquad e_{ij} = \frac{n_{i\cdot} n_{\cdot j}}{n} \qquad \chi^2 \to \chi^2_{(r-1)(s-1)}$$

当 χ^2 取大值，或者p-值很小的时候，拒绝零假设。

交叉分析

性别 * 可以接受的数码相机的价格 Crosstabulation

			可以接受的数码相机的价格					
			1000元以下	1001-2000元	2001-3000	3000-6000	6001以上	Total
性别	男	Count	31	115	85	64	21	316
		% within 性别	9.8%	36.4%	26.9%	20.3%	6.6%	100.0%
		% within 可以接受的数码相机的价格	11.2%	39.7%	47.5%	83.1%	44.7%	36.4%
		% of Total	3.6%	13.2%	9.8%	7.4%	2.4%	36.4%
	女	Count	245	175	94	13	26	553
		% within 性别	44.3%	31.6%	17.0%	2.4%	4.7%	100.0%
		% within 可以接受的数码相机的价格	88.8%	60.3%	52.5%	16.9%	55.3%	63.6%
		% of Total	28.2%	20.1%	10.8%	1.5%	3.0%	63.6%
Total		Count	276	290	179	77	47	869
		% within 性别	31.8%	33.4%	20.6%	8.9%	5.4%	100.0%
		% within 可以接受的数码相机的价格	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
		% of Total	31.8%	33.4%	20.6%	8.9%	5.4%	100.0%

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	160.399[a]	4	.000
Likelihood Ratio	173.531	4	.000
Linear-by-Linear Association	113.234	1	.000
N of Valid Cases	869		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 17.09.

注意：

1. 交叉列联表中的期望数<5的格点数不超过20%，方可进行Chi square检验。

2. 只有当交叉列联检验通过，才可认为行变量和列变量存在关系，否则只能视为独立。

明德主楼1019

例6.1

例 6.1: 为研究血型与肝病之间的关系，对 295 名肝病患者及 638 名非肝病患者（对照组），调查不同血型的得病情况，如表所示，问血型与肝病之间是否存在关联.

表 6.2. 血型与肝病间的关系

血型	肝炎	肝硬化	对照	合计
O	98	38	289	425
A	67	41	262	370
B	13	8	57	78
AB	18	12	30	60
合计	196	99	638	933

解答

本例中的行和列都是分类变量，因而可用 chisq.test 求出 Pearsonχ^2 值，如下所示：

```
> blood <-read.table("bloodtyp.txt",header=T)
> chisq.test(blood)
    Pearson's chi-square test with Yates' continuity correction
  data:  blood
  X-square = 15.073, df = 6, p-value = 0.020
```

表中输出了 Pearsonχ^2 检验结果，自由度为 $(3-1)(4-1)=6$，χ^2 值为 15.073，p 值为 0.020，由于 p 值小于 0.05，可以拒绝血型与病种独立的假设，认为血型与肝病有一定关联。 C=0.1294459

6.2 齐性检验　例6.2

例 6.2: 对 479 个不同年龄段的人调查他们对种不同类型电视节目的喜爱情况，要求每人只能选出他们最喜欢观看的电视节目类型，结果如下：

表 6.3. 不同年龄层次的人与电视节目类型之间的关系

年龄段	体育类 1	电视剧类 2	综艺类 3	总和
≤ 30	83	70	45	198
$31-50$	91	86	15	192
>50	41	38	10	89
总和	215	194	70	479

问题是想了解不同观众对三类节目的关注率是否一样 .

假设检验问题是：

$$\forall\ i=1,\dots,r, H_0: p_{i1}=\cdots=p_{ir}=p_{i\cdot} \leftrightarrow H_1: \text{等式不全成立}.$$

齐性检验

- H_0： $p_{i1}=p_{i2}=\dots=p_{i\cdot}$
- 在H_0之下，对$p_{i.}$最好的估计是 $\hat{p}_{i.}=n_{i.}/n$
- 对交叉列联表的每个单元格而言，我们希望测量观测频数和期望频数的差异：$(n_{ij} - n_{.j}\hat{p}_{i.})$
- 将上面的结果平方再标准化得到统计量 χ^2_{calc}.
- 注意到$E_{ij}=\hat{p}_{i}.\ n_{.j}=(n_{i}.)(n_{.j})/n$ 这个形式和独立性检验的形式是一致的。

	B_1	B_2	$\cdots$	B_s	总和
A_1	n_{11}	n_{12}	$\cdots$	n_{1s}	$n_{1\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
A_r	n_{r1}	n_{r2}	$\cdots$	n_{rs}	$n_{r\cdot}$
总和	$n_{\cdot 1}$	$n_{\cdot 2}$	$\cdots$	$n_{\cdots}$	$n_{\cdot\cdot}$

χ^2齐性检验

假设检验问题：

$$\forall i=1,\dots,r, H_0: p_{i1}=\dots=p_{ir}=p_{i.} \leftrightarrow H_1: \text{等式不全相等}$$

构造统计量：

$$Q=\sum_{i,j}\frac{(n_{ij}-e_{ij})^2}{e_{ij}}=\sum_{i,j}\frac{(n_{ij})^2}{e_{ij}}-n_{..} \qquad e_{ij}=\frac{n_{i.}n_{.j}}{n_{..}}$$

在零假设下近似有：$\chi^2 \to \chi^2_{(r-1)(s-1)}$

检验方法和独立性检验相同。

解答

例 6.2: 对 479 个不同年龄段的人调查他们对种不同类型电视节目的喜爱情况，要求每人只能选出他们最喜欢观看的电视节目类型，结果如下：

表 6.3. 不同年龄层次的人与电视节目类型之间的关系

年龄段	体育类 1	电视剧类 2	综艺类 3	总和
≤ 30	83	70	45	198
$31-50$	91	86	15	192
>50	41	38	10	89
总和	215	194	70	479

问题是想了解不同观众对三类节目的关注率是否一样 。
假设检验问题是：

$$\forall\ i=1,...,r, H_0: p_{i1}=\cdots=p_{ir}=p_{i\cdot} \leftrightarrow\ H_1: \text{等式不全成立}.$$

χ^2 检验统计量为

$$Q=\sum_{ij}\frac{(n_{ij}-e_{ij})^2}{e_{ij}}=\sum_{i,j}\frac{n_{ij}^2}{e_{ij}}-n_{..};$$

该 χ^2 统计量和独立性检验的统计量形式上完全一致，近似服从自由度为 $(r-1)(c-1)$ 的 χ^2 分布。该例子的 $Q=11.927$，$p-$ 值为 0.00179; 因此可以在水平 $\alpha \ge 0.002$ 时拒绝零假设。

讨论题：

多样本检验和X^2检验相似之处和区别

中国人民大学统计学院

Riddle of Jane Austen

Word	Sense Sbility	Emma	SanditonI	SanditonI
a	147	186	101	83
an	25	26	11	29
this	32	39	15	15
that	94	105	37	22
with	59	74	28	43
without	18	10	10	4

chisq.test(Jane)
Pearson's Chi-squared test

data: Jane
X-squared = 45.5775, df = 15, p-value = 6.205e-05

中国人民大学统计学院

6.3Fisher精确检验

2*2列联表

	B_1	B_2	总和
A_1	n_{11}	n_{12}	$n_{1.}$
A_2	n_{21}	n_{22}	$n_{2.}$
总和	$n_{.1}$	$n_{.2}$	$n_{..}$

在A、B独立时：

$$P\{n_{ij}\}=\frac{n_{1.}!n_{.1}!n_{2.}!n_{.2}!}{n_{..}!n_{11}!n_{12}!n_{21}!n_{22}!}$$

$$P(n_{11})=\frac{\binom{n_{1.}}{n_{11}}\binom{n_{2.}}{n_{21}}}{\binom{n..}{n_{.1}}}$$

中国人民大学统计学院

检验fisher.test

任何一个格子中的的数目都不会过大或者过小，如果过大过者过小就可以考虑拒绝零假设，因而我们考虑 n_{11} 就可以了。当大样本时，可以采用近似正态分布进行检验，即：

$$Z=\frac{\sqrt{n_{..}}(n_{11}n_{22}-n_{12}n_{21})}{\sqrt{n_{1.}n_{2.}n_{.1}n_{.2}}}\rightarrow N(0,1)$$

中国人民大学统计学院

例6.3

例 6.3: 为了解某种药物治疗效果，结果见下表：

表 6.5. 某病两种药物治疗结果

	疗	效	
药　物	有　效	无　效	合　计
A	8	2	10
B	14	18	32
合计	22	20	42

解：统计计算：如果固定边缘值 (10,32,22,20)，那么在零假设条件下出现在四格表中各数值分别为 n_{11},n_{12},n_{21} 及 n_{22} 的概率按超几何分布为：

$$\mathrm{P}\{n_{11}=8\} = \frac{n_{1.}!n_{2.}!n_{.1}!n_{.2}!}{n!n_{11}!n_{12}!n_{21}!n_{22}!} = \frac{10!32!22!20!}{42!8!2!14!18!}=0.0412$$

如果用 fisher.test 可以计算得到 $\mathrm{P}(n_{11}\ge 8)=0.0709$。

中国人民大学统计学院

6.4 Mantel-Haenszel检验

	A	B
存活	42	54
死亡	47	33

当组与组之间常常有不同的背景，而这些背景因子很可能会影响到组与组之间结果存在差异

UU=matrix(c(42,54,47,33),2)

	A	B
存活	20	14
死亡	17	25

	A	B
存活	22	30
死亡	30	8

Pearson's Chi-squared test with continuity correction

data: UU
X-squared = 3.3506, df = 1, p-value = 0.06718

中国人民大学统计学院

h 表示多层四格表的第 h 层, 第 h 层观测病案数为 n_h, $\sum_{h=1}^{k} n_h = n$.

假设检验问题为

H_0 : 试验组与对照组在治疗效果上没有差异;

H_1 : 试验组与对照组在治疗效果上存在差异.

下表是第 h 层四格表的符号表示.

	有效	无效	合计
试验组	n_{h11}	n_{h12}	$n_{h1.}$
对照组	n_{h21}	n_{h22}	$n_{h2.}$
合计	$n_{h.1}$	$n_{h.2}$	n_h

当零假设 H_0 成立时, 先求出第 h 层 n_{h11} 的期望 En_{h11} 和方差 $\mathrm{var}(n_{h11})$:

$$En_{h11} = \frac{n_{h1.}n_{h.1}}{n_h},$$

$$\mathrm{var}(n_{h11}) = \frac{n_{h1.}n_{h2.}n_{h.1}n_{h.2}}{n_h^2(n_h-1)}.$$

$$Q_{MH} = \frac{\left(\sum_{h=1}^{k} n_{h11} - \sum_{h=1}^{k} En_{h11}\right)^2}{\sum_{h=1}^{k} \mathrm{var}(n_{h11})}.$$

表 6.6 不同医院治癌药治癌效果比较

医院	药品	有效	无效	合计
1	A	50	15	65
	B	92	90	182
	合计	142	105	247
医院	药品	有效	无效	合计
2	A	47	135	182
	B	5	60	65
	合计	52	195	247

解 列 R 程序如下:

```
HA=matrix(c(50,92,15,90),2)
HB=matrix(c(47,5,135,60),2)
m=c(HA,HB); x=array(m,c(2,2,2))
mantelhaen.test(x)
```

```
chisq.test(matrix(c(97,150,97,150),2,2))

        Pearson's Chi-squared test

data:  matrix(c(97, 150, 97, 150), 2, 2)
X-squared = 0, df = 1, p-value = 1

>
```

```
Mantel-Haenszel chi-squared test with continuity correction
data:  x Mantel-Haenszel X-squared = 21.9443, df = 1, p-value =
2.807e-06 alternative hypothesis: true common odds ratio is not
```


Simpson悖论(女>男|商，女>男|法，女？男|法+商)

例题：一所美国高校的两个学院，分别是法学院和商学院，新学期招生。人们怀疑这两个学院有性别歧视。

辛普森悖论（Simpson‘s Paradox）亦有人译为辛普森诡论，为英国统计学家E.H.辛普森E.H.Simpson于1951年提出的悖论，即在某个条件下的两组数据，分别讨论时都会满足某种性质，可是一旦合并考虑，却可能导致相反的结论。

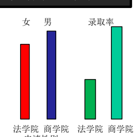
性别	录取	拒收	总数	录取比例
男生	209	95	304	68.8%
女生	143	110	253	56.5%
合计	352	205	557	

法学院

性别	录取	拒收	总数	录取比例
男生	8	45	53	15.1%
女生	51	101	152	33.6%
合计	59	146	205	

商学院

性别	录取	拒收	总数	录取比例
男生	201	50	251	80.1%
女生	92	9	101	91.1%
合计	293	59	352	



配对设计两样本率比较的χ^2检验（mcnemar.test）

方法原理

- 例6.9 用A、B两种方法检查已确诊的乳腺癌患者140名，A法检出91名(65%)，B法检出77名(55%)，A、B两法一致的检出56名(40%)，问哪种方法阳性检出率更高？

A法	B法 +	B法 −	合 计
+	56 (*a*)	35 (*b*)	91
−	21 (c)	28 (*d*)	49
合 计	77	63	140

方法原理

- 显然，本例对同一个个体有两次不同的测量，从设计的角度上讲可以被理解为自身配对设计
- 按照配对设计的思路进行分析，则首先应当求出各对的差值，然后考察样本中差值的分布是否按照H0假设的情况对称分布
- 按此分析思路，最终可整理出如前所列的配对四格表

方法原理

- 注意
  - 主对角线上两种检验方法的结论相同，对问题的解答不会有任何贡献
  - 另两个单元格才代表了检验方法间的差异
- 假设检验步骤如下：
  - *H*0：两法总体阳性检出率无差别，即B = C
  - *H*1：两法总体阳性检出率有差别，即B ≠ C

方法原理

根据 H_0 得 b、c 两格的理论数均为 $T_b = T_c = (b+c)/2$，对应的配对检验统计量为：

$$\chi^2 = \frac{(b-c)^2}{b+c}, \qquad \nu = 1$$

一般在 $b + c < 40$ 时，需用确切概率法进行检验，或者进行校正。

- mcci 56 35 21 28

注意事项

- McNemar检验只会利用非主对角线单元格上的信息，即它只关心两者不一致的评价情况，用于比较两个评价者间存在怎样的倾向。因此，对于一致性较好的大样本数据，McNemar检验可能会失去实用价值。
  - 例如对1万个案例进行一致性评价，9995个都是完全一致的，在主对角线上，另有5个分布在左下的三角区，显然，此时一致性相当的好。但如果使用McNemar检验，此时反而会得出两种评价有差异的结论来。

配对四格表资料的 χ^2 检验

McNemar检验（McNemar's test）

例： 两种血清学检验结果比较

甲 法	乙 法 +	乙 法 −	合计
+	80 (a)	10 (b)	90
−	31 (c)	10 (d)	41
合 计	111	20	131

H_0： B = C H_1： B≠C α=0.05。

$$\chi^2 = \frac{(b-c)^2}{b+c}, \qquad \nu = 1$$

连续性校正： $$\chi^2 = \frac{(|b-c|-1)^2}{b+c}, \qquad \nu = 1$$

当 b+c≥40 时可不校正，而 b+c<40 时则一定要校正。

本例 b+c=10+31=41>40，不需作连续性校正，计算得

$$\chi^2 = \frac{(10-31)^2}{10+31} = 10.76, \qquad \nu = 1$$


```
> ex=matrix(c(80,10,31,10),2,2)
> chisq.test(ex)
        Pearson's Chi-squared test with Yates' continuity correction
data:  ex
X-squared = 2.8817, df = 1, p-value = 0.08959
> mcnemar.test(ex)
        McNemar's Chi-squared test with continuity correction

data:  ex
McNemar's chi-squared = 9.7561, df = 1, p-value = 0.001787
```


序和分布的识别

- 从一般意义上，社会生活不能没有秩序；
  - 公务卡购票
  - 安检
  - 登机
  - 享受飞翔的自由
- 稳定与秩序的辨别：
  - 稳定是被动的，秩序是主动的；
  - 稳定是静态的，秩序是动态的；
  - 稳定是不主张激活的，秩序则是与活力兼容的。

秩序是什么？

- 回想一个随机变量的秩是怎样定义的？
- 秩是独立随机变量向量的一个特征，与**样本量*n***有关
  - 秩与分布分位数的对应关系：

$$q = r/(n+1)$$

$$q = \int_{-\infty}^{m_q} p(x)dx$$

- 如果数据的分布不知道怎么办？
- 一个随机变量的秩是怎样定义的？

问题：A和B两组病人治疗效果是否相同

例子：

组别	疗效 痊愈	显效	好转	无效	合计
治疗组(中医)	68	26	15	3	112
对照组(西医)	737	388	25	5	1155

两组病人：治疗组与对照组；疗效：痊愈、显效、好转、无效

特点：有一个分类是按等级分组的。

两分类数据，其中一个分类变量是分类变量；另一个变量是顺序变量；称为单向有序分组数据

关心的问题：各等级的比例是否相同，各不同的组是否存在整体的优劣之分

6.6 Ridit检验

对于有序分类变量，采用卡方检验方法不能考虑数据的有序性质。为此，对于单向有序问题可考虑Ridit分析.

Ridit检验法的原理：取一个样本数较多的组或者将几组数据汇总成为参照组，根据参照组的样本结构将原来各组响应数变换为参照得分：Ridit得分，利用变换以后的Ridit得分进行处理之间的强弱比较。标准组$\bar{R}$=0.5，其他各对比组均按标准组的各等级R值计算其平均$\bar{R}$，对比组$\bar{R}$在0-1之间波动，最后通过假设检验做出结论

行向量A表示不同比较组，列向量B为顺序尺度变量，假设 $B_1 < \cdots < B_s$，O_{ij} 表示对应格子的响应频数。

假设检验问题：

$H_0: A_1, \cdots, A_r$之间没有强弱顺序 $\leftrightarrow H_1$: 至少一对$A_i \neq A_j$

顺序强度计算步骤

各顺序级别 R_j 计算表

步骤		B_1	B_2	$\cdots$	B_s	合计
	A_1	O_{11}	O_{12}	$\cdots$	O_{1s}	$O_{1\cdot}$
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	A_r	O_{r1}	n_{r2}	$\cdots$	O_{rs}	$O_{r\cdot}$
	总和	$O_{\cdot 1}$	$O_{\cdot 2}$	$\cdots$	$O_{\cdot s}$	$O_{\cdot\cdot}$
(1)		$H_1=\frac{1}{2}O_{\cdot 1}$	$H_2=\frac{1}{2}O_{\cdot 2}$	$H_j=\frac{1}{2}O_{\cdot j}$	$H_s=\frac{1}{2}O_{\cdot s}$	
(2)		0	$C_2=\sum_{k=1}^{1}O_{\cdot k}$	$C_j=\sum_{k=1}^{j-1}O_{\cdot k}$	$C_s=\sum_{k=1}^{s-1}O_{\cdot k}$	
(3)		N_1	N_2	$N_j=H_j+C_j$	N_s	
(4)		R_1	R_2	$R_j=\frac{N_j}{O_{\cdot\cdot}}$	R_s	

(5) 将 R_j 的值按照 O_{ij} 占 $O_{\cdot j}$ 的权重重新配置第 i,j 位置的 Ridit 得分：$R_{ij}=\frac{O_{ij}}{O_{i\cdot}}R_{\cdot j}$.

(6) 计算第 i 处理（类）的 Ridit 得分：$R_i=\sum_{j=1}^{s}R_{ij}$ 这些 Ridit 得分的期望为 0.5 。

两组检验

- 检验问题：

H0:对照组的$\bar{R}$=0.5

H1:对照组的$\bar{R}$≠0.5,α=0.05.

计算标准组Ridit值$\bar{R}$及标准差

R值服从[0,1]的均匀分布,概率密度函数为f(x)=1,由均匀分布的理论可知,R值的总体均数和方差分别是:

$$\mu_R = 0.5, \sigma_R^2 = 1/12$$

$$\bar{R} = 1/n\sum O_jR_j$$

在n足够大时，$\bar{R}$近似服从正态分布,于是有:

$$u = \frac{\bar{R}-\mu_R}{1/\sqrt{12n}} \sim N(0,1)$$

等级 (1)	频数 (2)	累计频数 下移一行 (3)	频数 / 2 (4)	(5) =(3)+(4)	R 值 (5)/总例数
无效	760	0	380	380	0.114
好转	1870	760	935	1695	0.509
显效	670	2630	335	2965	0.890
控制	30	3300	15	3315	0.995
合计	3330	—	—	—	—

$\bar{R}$ 及标准差计算公式：

$$\bar{R} = \frac{\sum O_jR_j}{n} \qquad S_R = \sqrt{\frac{\sum O_jR_j^{\ 2} - \frac{(\sum O_jR_j)^2}{n}}{n-1}}$$

表3：Ridit均值和标准差计算表

等级	R 值	O_j	O_jR_j	$O_jR_j^{\ 2}$
无效	0.114	760	86.64	9.87696
好转	0.509	1870	951.83	484.48147
显效	0.890	670	596.30	530.70700
控制	0.995	30	29.85	29.70075
合计	—	3330	1664.62	1054.76618

本例：

$$\bar{R}_{标准} = \frac{166462}{3330} = 0.5 \qquad S_R = \sqrt{\frac{105476618 - \frac{(166462)^2}{3330}}{3330-1}} = 0.2586$$

- 对照组平均Ridit值计算:
  - 以对照组各级频数与相应的Ridit得分进行加权平均，得到平均R

$$\bar{R}_{中医}=\frac{9\times0.114+51\times0.509+21\times0.890+13\times0.995}{94}=0.624$$

- 计算标准误：样本标准误以标准组的标准差除以样本例数的平方根 $S_{\bar{R}}=\frac{S_R}{\sqrt{n}}=\frac{0.2586}{\sqrt{94}}=0.027$
- 计算$\bar{R}$的可信区间 $\bar{R}\pm u_\alpha S_{\bar{R}}$ $0.624\pm1.96\times0.027$
- 95%置信区间是（0.571，0.677），不包括0.5，结论为差异显著，因此可以认为两种治疗疗效不同，中医疗效优于西医疗效

Ridit得分定义

假设顺序类别B中第j类的边缘分布是$p_{\cdot j}$， j=1,…,s，那么第j类的顺序强度（Ridit得分）定义如下：

$$r_j = \sum_{k=1}^{j-1} p_{\cdot j} + \frac{1}{2}p_{\cdot j},\ j=2,\cdots,s$$
$$= \frac{F_{j-1}^B + F_j^B}{2}$$

其中

$$F_j^B = \sum_{k=1}^{j} p_{\cdot j},\ j=2,\cdots,s.$$

在实际计算中用样本估计 .

计算实例

```
ori=c(20,30,25,44,24,26,16,18)
ori1=ori/2
orisum=sum(ori)
ori3=c(0,ori)
ori4=cumsum(ori3)[1:length(ori)]
ori5=(ori1+ori4)/orisum
ori6=ori*ori5
ori6

now1=c(4,8,3,4,1,0,0,0)
sumnow=sum(now1)
sum(now1*ori5)/sum(now1)
cor(now1,ori)

now2=c(0,1,1,4,7,8,3,15)
sumnow=sum(now2)
sum(now2*ori5)/sum(now2)
cor(now2,ori)

now3=floor(ori*4)
#now3=c(0,1,1,4,7,8,3,15)
sumnow=sum(now3)
sum(now3*ori5)/sum(now3)
```


多组检验

H_0 ： $A_1,\cdots,A_r$之间没有强弱顺序

$\leftrightarrow H_1$ ： 至少存在一对A_i, A_j，使得$A_i\neq A_j$成立.

根据计算的R构造检验统计量：

$$W = \frac{12O_{\cdot\cdot}}{(O_{\cdot\cdot}+1)T}\sum_{i=1}^{r} O_{i\cdot}(R_i-0.5)^2$$

其中T为打结校正因子

当大样本时，T值接近于1，从而检验统计量简化为：

$$W = 12\sum_{i=1}^{k} O_{i\cdot}(R_{i\cdot}-0.5)^2$$

在零假设情况下，W近似服从$\chi^2_{(k-1)}$分布，当W过大或者过小时，可考虑拒绝零假设。近似的置信区间$\bar{R}_i\pm1/\sqrt{3O_{i\cdot}}$

例6.4

例 6.4: 表 6.8 是用头针治疗瘫痪 800 例的疗效分析，不同病因的疗效可以不一样，究竟哪一种病因所引起的瘫痪用头针的治疗效果最佳，哪些次之，哪些最差，是医务人员希望数据回答的问题.

表 6.8. 头针治疗瘫痪 800 例的疗效分析

组别	总数	基本痊愈	显效	有效	无效	恶化	死亡
1 、脑血栓形成及后遗症	500	190	123	162	24	1	0
2 、脑出血及后遗症	132	9	38	73	11	0	1
3 、脑栓塞及后遗症	59	20	13	20	6	0	0
4 、颅内损失及后遗症	54	4	12	33	5	0	0
5 、急性感染性多发性神经炎	10	4	2	3	1	0	0
6 、脊髓疾病	6	1	3	0	2	0	0
总病例数	800	232	202	311	53	1	1

表 6.9. 头针治疗瘫痪 800 例疗效的 Ridit 计算步骤

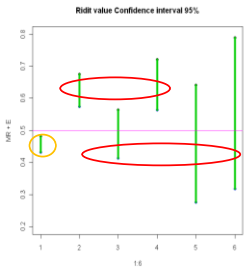
步骤 \ 级别	基本痊愈	显效	有效	无效	恶化	死亡
(I)（病例数总计）	232	202	311	53	1	1
(II)（病例数 ×1/2 ）	116	101	155.5	26.5	0.5	0.5
(III) 累积	0	232	434	745	798	799
(II)+(III)	116	333	589.5	771.5	798.5	799.5
$R=\frac{\text{II+III}}{800}$	0.145	0.416	0.737	0.964	0.998	0.999
合计	33.64	84.082	229.168	51.11	0.998	0.999

解 答

脑血栓形成及后遗症疗效结果的

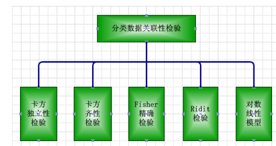
等级	(1)	(2)	(3)
I	194	0.145	28.130
II	134	0.416	55.774
III	182	0.737	134.11
IV	28	0.964	26.992
V	1	0.998	0.998
VI	0	0.999	0
合计	500		246.10

得出 95% 可信限为 0.4564367 +0.02 ，由于组的治疗效果对总数 800 例的效果来讲较好。



本章要求

- 掌握分类数据的独立性研究方法；
- 区分分类数据的独立性和齐性检验的异同；
- 掌握Fisher检验与卡方检验的应用条件的异同；
- 了解Ridit方法和应用；
- 了解对数线性模型和卡方检验的异同；
- 熟练应用R中的相关命令学习如上方法。



中国人民大学统计学院