第六章　列联分析

第一节　列联表

第二节　χ^2 分布与 χ^2 检验

第三节　列联表中的相关测量

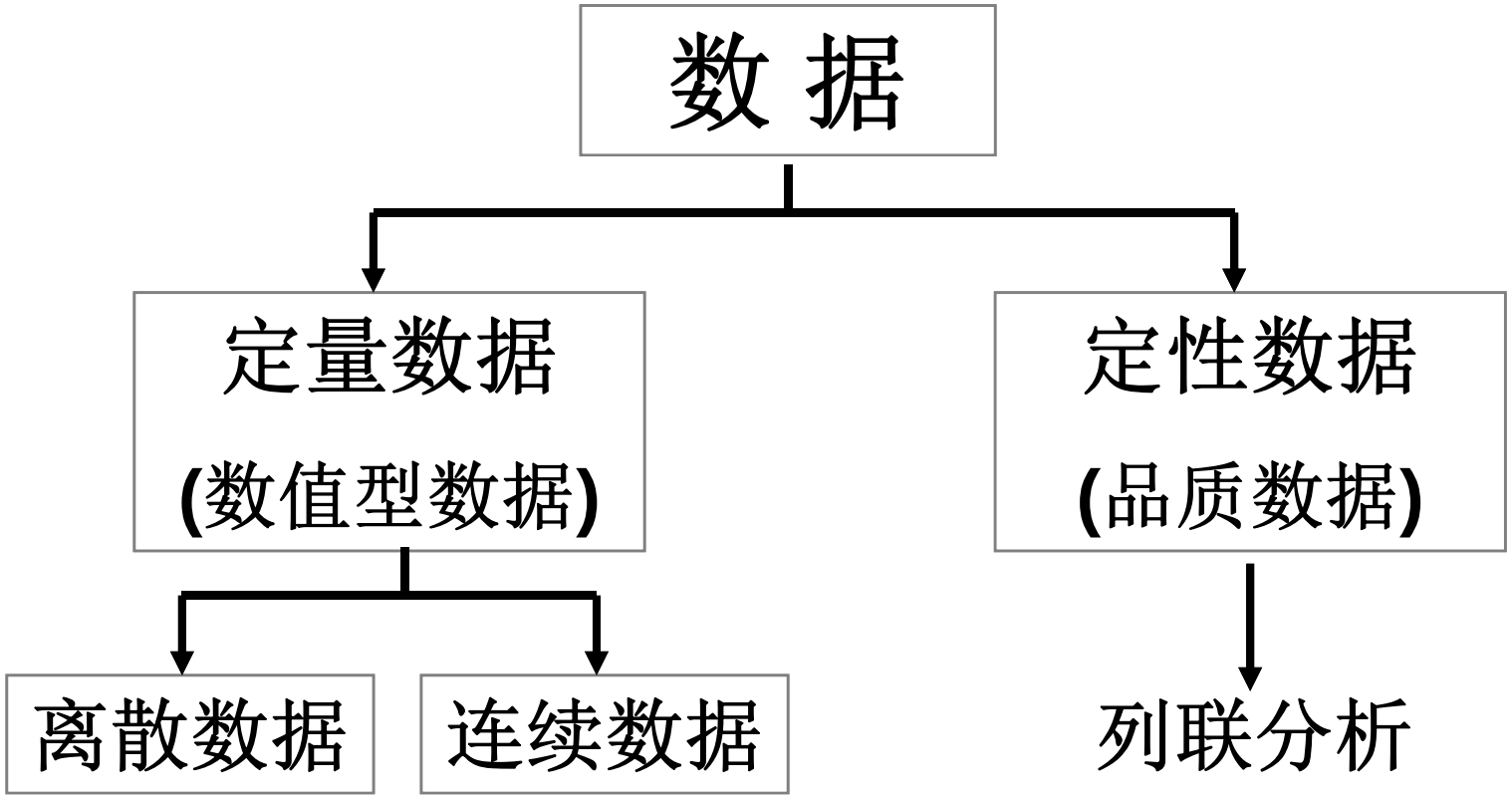
第四节　精确概率检验法

Spss中的操作

学习目标

1. 列联表的构造
2. 进行 χ^2 检验
   - 一致性检验
   - 独立性检验
3. 测度列联表中的相关性
4. 实际应用

数据的类型与列联分析



列联表实例

	老年	中年	青年
戏曲	*20*	*10*	2
歌舞	*5*	*20*	35
球赛	2	10	20

第一节　列联表

一. 列联表的构造

二. 列联表的分布

列联表的构造

列联表的结构

r 行 c 列的列联表

行(r_i) \ 列(c_j)	列(c_j)			合计
	j =1	j = 2	…	
i =1	f_{11}	f_{12}	…	r_1
i = 2	f_{21}	f_{22}	…	r_2
:	:	:	:	:
合计	c_1	c_2	…	n

f_{ij} 表示第 i 行第 j 列的观察频数

列联表实例2

【例】一个集团公司在四个不同的地区设有分公司，现该集团公司欲进行一项改革，此项改革可能涉及到各分公司的利益，故采用抽样调查方式，从四个分公司共抽取420个样本单位(人)，了解职工对此项改革的看法，调查结果如下表

	一分公司	二分公司	三分公司	四分公司	合计
赞成该方案	68	75	57	79	279
反对该方案	32	75	33	31	141
合计	100	120	90	110	420

列联表的分布

观察值的分布

条件频数

行边缘分布

	一分公司	二分公司	三分公司	四分公司	合计
赞成该方案	68	75	57	79	279
反对该方案	32	75	33	31	141
合计	100	120	90	110	420

列边缘分布

百分比分布

行百分比　　列百分比　　总百分比

	一分公司	二分公司	三分公司	四分公司	合计
赞成该方案	16.2%	17.8%	13.6%	18.8%	66.4%
	24.4%	26.9%	20.4%	28.3%	—
	68.0%	62.5%	63.35	71.8%	—
反对该方案	7.6%	10.7%	7.9%	7.4%	33.6%
	22.7%	31.9%	23.4%	22.0%	—
	32.0%	37.5%	36.7%	28.2%	—
合计	23.8%	28.6%	21.4%	26.2%	100%

期望频数的分布

1. 假定行变量和列变量是独立的
2. 一个实际频数 f_{ij} 的期望频数 e_{ij}，是总频数的个数 n 乘以该实际频数 f_{ij} 落入第 i 行和第j列的概率，即

$$e_{ij} = n \cdot \left(\frac{r_i}{n}\right) \cdot \left(\frac{c_i}{n}\right) = \frac{r_i c_j}{n}$$

期望频数的分布

	一分公司	二分公司	三分公司	四分公司	合计
赞成该方案	16.2%	17.8%	13.6%	18.8%	66.4%
	66				—
反对该方案	7.6%	10.7%	7.9%	7.4%	33.6%
					—
合计	23.8%	28.6%	21.4%	26.2%	100%

$$e_{11} = n \cdot \left(\frac{r_1}{n}\right) \cdot \left(\frac{c_1}{n}\right) = \frac{r_1 c_1}{n} = \frac{279 \times 100}{420} = 66.43 \approx 66$$

期望频数的分布

➡根据上述公式计算的前例的期望频数

		一分公司	二分公司	三分公司	四分公司
赞成该方案	实际频数	68	75	57	79
	期望频数	66	80	60	73
反对该方案	实际频数	32	75	33	31
	期望频数	34	40	30	37

第二节　χ^2 分布与 χ^2 检验

一. χ^2 统计量

二. χ^2 检验

χ^2 分布

表5-1 某动物育种试验F_2代资料

	观测值 ***f***	理论值 ***e***	***f - e***
试验一	204	200	4
试验二	24	28	-4

$$\chi^2 = \sum_i \frac{(f_i - e_i)^2}{e_i}$$

χ^2 统计量

实际频数 (f_{ij})	期望频数 (e_{ij})	$f_{ij}-e_{ij}$	$(f_{ij}-e_{ij})^2$	$\frac{(f_{ij}-e_{ij})^2}{e_{ij}}$
68	66	2	4	0.0606
75	80	-5	25	0.3125
57	60	-3	9	0.1500
79	73	6	36	0.4932
32	34	-2	4	0.1176
45	40	5	25	0.6250
33	30	3	9	0.3000
31	37	-6	36	0.9730

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(f_{ij}-e_{ij})^2}{e_{ij}} = 3.0319$$

χ^2 检验

- 拟合优度检验
- 独立性检验
- 配对卡方检验

χ^2 统计量

1. 用于检验列联表中变量之间是否存在显著性差异，或者用于检验变量之间是否独立
2. 计算公式为

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(f_{ij} - e_{ij})^2}{e_{ij}}$$

其自由度为 $(r-1)(c-1)$

式中f_{ij}为实际频数，e_{ij}为期望频数。

拟合优度检验

- 检验观测值与按照理论分布计算出来的理论值是否吻合。

  例：纯合的黄圆豌豆与绿皱豌豆杂交，F_1代自交，F_2代分离数目如下：

Y_ R_（黄圆）	**Y_ rr**（黄皱）	**yyR_**（绿圆）	**yyrr**（绿皱）	合计
315	101	108	32	556

  问是否符合自由组合（独立分配）定律？

解：当性状间相互独立时，根据孟德尔独立分配定律，F_2代的表型可由二项分布给出。显性性状出现的概率为$\varphi = 3/4$。两对基因的自由组合，根据二项展开式

$$\left(\frac{3}{4}+\frac{1}{4}\right)^2 = \frac{9}{16}+\frac{3}{16}+\frac{3}{16}+\frac{1}{16}$$

可以得出理论分离比为：

$$Y\_R\_ : Y\_\mathrm{rr} : \mathrm{yy}R\_ : yyrr = \frac{9}{16} : \frac{3}{16} : \frac{3}{16} : \frac{1}{16}$$

χ^2计算表

	Y_ R_ （黄圆）	Y_ rr （黄皱）	yyR_ （绿圆）	yyrr （绿皱）	合计
观测值	315	101	108	32	556
理论频率	9/16	3/16	3/16	1/16	
理论值	312.75	104.25	104.25	34.75	
$f-e$	2.25	-3.25	3.75	-2.75	
$(f-e)^2/e$	0.016	0.101	0.135	0.218	0.470

H_0: 豌豆F2代分离比符合9:3:3:1的自由组合规律

H_1: 豌豆F2代分离比不符合9:3:3:1的自由组合规律

$$\chi^2 = \sum_{i=1}^{4} \frac{(f_i - e_i)^2}{e_i} = 0.470 < \chi^2_{0.05}(3) = 7.815$$

结论：接受 H_0 ，试验结果符合9:3:3:1的分离比。

$$\chi^2 = \sum_{i=1}^{k} \frac{(f_i - e_i)^2}{e_i}$$

$$= \sum_{i=1}^{k} \frac{f_i^2}{np_i} - 2\sum_{i=1}^{k} f_i + \sum_{i=1}^{k} e_i$$

$$= \frac{1}{n}\sum_{i=1}^{k} \frac{f_i^2}{p_i} - n$$

- 注意：

（1）任何一组的理论值e都不小于5，如果e_i小于5，应将相邻组合并。

（2）当df=1时，χ^2统计量应做连续性矫正：

$$\chi^2 = \sum_{i=1}^{k} \frac{(|f_i - e_i| - 0.5)^2}{e_i}$$

（3）如果总体参数未知，需由样本数据做参数估计，此时的自由度应再减去需要进行估计的参数个数。

女性家长分布类型的推断

女性人数	观测数	理论频率	理论频数
0	0	0.0001	0.01
1	0	0.0019	0.19
2	0	0.0125	1.25
3	4	0.048	4.8
4	14	0.1209	12.09
5	22	0.2087	20.87
6	27	0.2503	25.03
7	19	0.2058	20.58
8	9	0.1111	11.11
9	5	0.0355	3.55
10	0	0.0051	0.51
和	100	0.9999	99.99

独立性检验

1. 检验列联表中的行变量与列变量之间是否独立
2. 检验的步骤为
   - 提出假设
     - H_0：行变量与列变量独立
     - H_1：行变量与列变量不独立
   - 计算检验的统计量

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(f_{ij}-e_{ij})^2}{e_{ij}}$$

   - 进行决策
     - 根据显著性水平α和自由度$(r-1)(c-1)$查出临界值χ_α^2
     - 若$\chi^2 \geq \chi_\alpha^2$，拒绝$H_0$；若$\chi^2 < \chi_\alpha^2$，接受$H_0$

2×2列联表的独立性检验

A(i)	B(j)		总和
	c_1	c_2	
r_1	O_{11}	O_{12}	$R_1=O_{12}{}^{+}O_{22}$
r_2	O_{21}	O_{22}	$R_2=O_{12}{}^{+}O_{22}$
总和	$C_1=O_{11}{}^{+}O_{21}$	$C_2=O_{12}{}^{+}O_{22}$	T

$$E_{ij} = P_{ij}T = \frac{R_i}{T}\frac{C_j}{T}\cdot T = \frac{R_i C_j}{T}$$

例：现对吸烟人群和不吸烟人群是否患有气管炎病进行了随机抽样调查，结果如下表所示，试检验吸烟与患气管炎病有无关联？

$(\alpha = 0.01)$

不同人群	患病	不患病	总和	患病率 /%
吸烟人群	50	250	300	16.67
不吸烟人群	5	195	200	2.5
总和	55	445	500	

H_0: 吸烟与患气管炎病无关

H_1: 吸烟与患气管炎病有关

$$e_{11} = \frac{R_1 C_1}{T} = \frac{300 \times 55}{500} = 33$$

$$e_{12} = \frac{R_1 C_2}{T} = 267 \qquad e_{21} = 22 \qquad e_{22} = 178$$

$$df = (r-1)(c-1) = 1 \qquad \chi^2_{0.01}(1) = 6.63$$

$$\chi^2 = \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{(|f_{ij} - e_{ij}| - 0.5)^2}{e_{ij}} = 23.174$$

结论：决绝H_0，吸烟与患气管炎病相关。

配对卡方检验（McNemer检验）

A	B		合计
	+	-	
+	O_{11}	O_{12}	$R_1=O_{12}{}^{+}O_{22}$
-	O_{21}	O_{22}	$R_2=O_{12}{}^{+}O_{22}$
合计	$C_1=O_{11}{}^{+}O_{21}$	$C_2=O_{12}{}^{+}O_{22}$	T

$$\frac{R_1}{T}=\frac{C_1}{T} \Rightarrow O_{12}=O_{21}$$

$$H_0:\ O_{12} = O_{21} \qquad H_1:\ O_{12} \neq O_{21}$$

$$\chi^2 = \frac{(f_{12} - e_{12})^2}{e_{12}} + \frac{(f_{21} - e_{21})^2}{e_{21}}$$

$$e_{12} = e_{21} \approx \frac{f_{12} + f_{21}}{2}$$

$$\chi^2 = \frac{(f_{12} - f_{21})^2}{f_{12} + f_{21}} \qquad df = r - 1 = 1$$

连续性校正：

$$\chi^2 = \frac{(|f_{12} - f_{21}| - 1)^2}{f_{12} + f_{21}}$$

$r \times c$列联表的独立性检验

【例】某医院用碘剂治疗地方性甲状腺肿，不同年龄的治疗效果如下表。试检验不同年龄的治疗效果有无差异？（α =0.01)

年龄/岁	治愈	显效	好转	无效	合计
11~30	67(45.29)	9(17.87)	10(20.02)	5(5.82)	91
31~50	32(39.32)	23(15.51)	20(19.12)	4(5.05)	79
50以上	10(24.39)	11(9.62)	23(11.86)	5(3.13)	49
合计	109	43	53	14	219

注：括号内数据为理论值

1. 提出假设
   - H_0：治疗效果与年龄无关
   - H_1：治疗效果与年龄有关
2. 计算检验的统计量

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(f_{ij}-e_{ij})^2}{e_{ij}} = 45.48$$

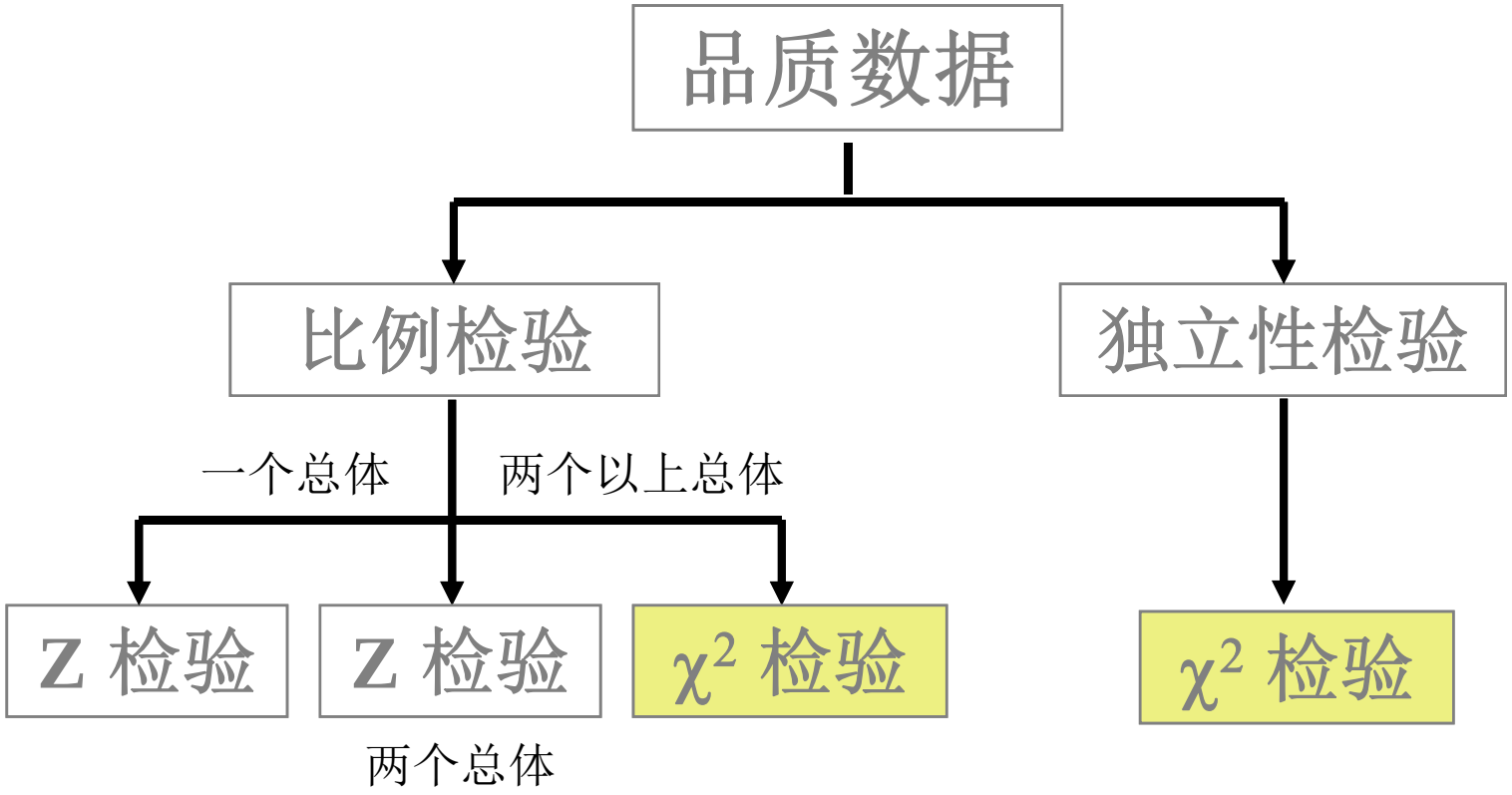
3. 当自由度(4-1)(3-1)=6时，查出相应的临界值$\chi^2_{0.01} = 16.81$。由于χ^2=45.48>χ^2_α，拒绝H_0。说明治疗效果与年龄有关。

年龄/岁	治愈	显效	好转	无效	合计
11~30	67	9	10	5	91
31~50	32	23	20	4	79
合计	99	32	30	9	170

年龄/岁	治愈	显效	好转	无效	合计
11~30	67	9	10	5	91
50以上	10	11	23	5	49
合计	77	20	33	10	140

年龄/岁	治愈	显效	好转	无效	合计
31~50	32	23	20	4	79
50以上	10	11	23	5	49
合计	43	34	43	9	128

品质数据的假设检验



有效性条件

- 设计条件：样本内的观察值相互独立
  - 含有多个随机样本时，观察的是一个分类变量
  - 只有一个随机样本时，观察的是两个分类变量

	红色植株			白色植株		
	花的数量	坐果数量	坐果百分比/%	花的数量	坐果数量	坐果百分比/%
	140	26	19	125	21	17
	116	11	9	134	17	13
	34	0	0	273	81	30
	79	9	11	146	38	26
	185	28	15	103	17	17
	106	11	10	82	24	29
总和	660	85		863	198	

有效性条件

- 样本容量：
  - 每个单元格的理论频数不小于5时，检验有效
  - 如果r和k很大，且平均期望频数至少等于5时，可以接受少量空格内的值很小

- 卡方检验的功效较弱

第三节　列联表中的相关测量

一.　φ 相关系数

二.　列联相关系数

三.　V 相关系数

列联表中的相关测量

1. 品质相关
   - 对品质数据(定类和定序数据)之间相关程度的测度
2. 列联表变量的相关属于品质相关
3. 列联表相关测量的指标主要有
   - φ 相关系数
   - 列联相关系数
   - V 相关系数

φ 相关系数

1. 测度 2×2列联表中数据相关程度的一个量
2. 对于2×2 列联表，φ系数的值在0～1之间
3. φ相关系数计算公式为

$$\varphi = \sqrt{\frac{\chi^2}{n}}$$

φ 相关系数
（原理分析）

一个简化的 2×2 列联表

因素 Y	因素 X		合计
	x_1	x_2	
y_1	a	b	$a+b$
y_2	c	d	$c+d$
合计	$a+c$	$b+d$	n

φ 相关系数
（原理分析）

1. 列联表中每个单元格的期望频数分别为

$$e_{11}=\frac{(a+b)(a+c)}{n} \qquad e_{21}=\frac{(a+c)(c+d)}{n}$$

$$e_{12}=\frac{(a+b)(b+d)}{n} \qquad e_{22}=\frac{(b+d)(c+d)}{n}$$

2. 将各期望频数代入 χ^2 的计算公式得

$$\chi^2=\frac{(a-e_{11})^2}{e_{11}}+\frac{(b-e_{12})^2}{e_{12}}+\frac{(c-e_{21})^2}{e_{21}}+\frac{(d-e_{22})^2}{e_{22}}$$

$$=\frac{n(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)}$$

φ 相关系数
（原理分析）

3. 将χ^2入φ相关系数的计算公式得

$$\varphi = \sqrt{\frac{\chi^2}{n}} = \frac{|ad - bc|}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}$$

- ad 等于 bc ， $\varphi = 0$，表明变量X 与 Y 之间独立
- 若 b=0 ， c=0，或a=0 ， d=0，意味着各观察频数全部落在对角线上，此时φ =1, 表明变量X 与 Y 之间完全相关

Pearson列联相关系数

1. 用于测度大于2×2列联表中数据的相关程度
2. 计算公式为

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}}$$

- C 的取值范围是 $0 \le C < 1$
- $C = 0$表明列联表中的两个变量独立
- C 的数值大小受列联表行数和列数的影响，根据不同行和列的列联表计算的列联相关系数不便于比较

V 相关系数

1. 计算公式为

$$V = \sqrt{\frac{\chi^2}{n \cdot \min[r-1, c-1]}}$$

2. V 的取值范围是 $0 \le V \le 1$
3. $V = 0$表明列联表中的两个变量独立
4. $V = 1$表明列联表中的两个变量完全相关
5. 不同行和列的列联表计算的列联系数不便于比较
6. 当列联表中有一维为2，$\min[(r\text{-}1),(c\text{-}1)]=1$,此时 $V=\varphi$

列联表中的相关测量

【例】一种原料来自三个不同地区，原料质量被分成三个不同等级。从这批原料中随机抽取500件进行检验，结果如下表。分别计算C系数和V系数，并分析相关程度。

地区	一级	二级	三级	合计
甲地区	52	64	24	140
乙地区	60	59	52	171
丙地区	50	65	74	189
合计	162	188	150	500

列联表中的相关测量

解：已知n=500，根据公式计算χ^2＝19.82，列联表为3×3

$$C=\sqrt{\frac{\chi^2}{\chi^2+n}}=\sqrt{\frac{19.82}{19.82+500}}=0.195$$

$$V=\sqrt{\frac{\chi^2}{n\cdot\min[r-1,c-1]}}=\sqrt{\frac{19.82}{500\times 2}}=0.141$$

结论：两个系数均不高，表明产地和原料等级之间的相关程度不高

第四节 Fisher精确概率检验

☆ 总样本数比较小时，可用Fisher精确概率检验

例：现有12例栓塞性脉管炎患者，随机分成新药组与对照组，治疗结果如下表所示，试问两治疗组的治愈率差别是否有统计学意义。

组别	治愈人数	未愈人数	合计
新药组	***6(a)***	***1(b)***	$7(n_{r1})$
对照组	***1(c)***	***4(d)***	$5(n_{r2})$
合计	$7(n_{c1})$	$5(n_{c2})$	***n***

用 P_i 表示第i组的治愈率

H_0: $P_1 = P_2$　　H_1: $P_1 \neq P_2$

新药组的n_{r1}个患者恰好有a例治愈b例未治愈的概率为：

$$P(a|n_{r1}, n_{r2}, n_{c1}, n_{c2}) = \frac{C_{n_{c1}}^{a} C_{n_{c2}}^{b}}{C_{n}^{n_{r1}}}$$

$$= \frac{\frac{n_{c1}!}{a!\,c!} \cdot \frac{n_{c2}!}{b!\,d!}}{n!/(n_{r1}!\,n_{r2}!)} = \frac{n_{r1}!\,n_{r2}!\,n_{c1}!\,n_{c2}!}{n!\,a!\,b!\,c!\,d!}$$

$$P(6|n_{r1} = 7, n_{r2} = 5, n_{c1} = 7, n_{c2} = 5)$$

$$= \frac{7!\,5!\,7!\,5!}{12!\,6!\,1!\,1!\,4!} = 0.0442$$

2	5	$a = 2$
5	0	$P = 0.0265$

3	4	$a = 3$
4	1	$P = 0.2210$

4	3	$a = 4$
3	2	$P = 0.4419$

5	2	$a = 5$
2	3	$P = 0.2852$

6	1	$a = 6$
1	4	$P = 0.0442$

7	5	$a = 2$
5	0	$P = 0.0265$

$(1) H_1: P_1 > P_2 \quad P = P(a = 6) + P(a = 7)$

$(2) H_1: P_1 \neq P_2 \quad P = P(a = 6) + P(a = 7) + P(a = 2)$

操作

- 菜单式操作

  Analyze------Descriptive Statistics------crosstabs
- 由原始数据形成交互表
- 由交互表进行分析时，需要对数据加权。
- 进行卡方检验

结　束