第九章 分类数据分析

9.1 分类数据与 χ^2 统计量

分类数据的结果是频数，使用 χ^2 检验对分类数据的频数进行分析。用 f_0 和 f_e 分别代表观察频数和期望频数。（例题P190）

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e} \tag{1}$$

χ^2 分布的自由度 $df = R - 1$，R 是分类变量类型的个数。自由度越小，分布越左倾；自由度增加，偏斜程度越小，越接近正态分布。

9.3 列联分布

列联表就是将两种以上的变量进行交叉分类得到的频数分布表。（$R \times C$）

独立性检验用于分析i列联表中的行变量和列变量是否独立，即各个变量是否存在依赖关系。在计算独立性的过程中，需要列出列联表里的每个单元格中的 f_e，RT 和 CT 表示该变量所在的行、列的合计。

$$f_e = \frac{RT}{n} \times \frac{CT}{n} \times n = \frac{RT \times CT}{n} \tag{2}$$

根据这个结果按照行、列进行列表，然后得到 χ^2 统计量。χ^2 的自由度 $df = (R-1)(C-1)$。

自由度分析

	C1	C2	C3	Sum
R1	f	f	x	RT_1
R2	f	f	x	RT_2
R3	x	x	0	RT_3

	C1	C2	C3	Sum
Sum	CT_1	CT_2	CT_3	T

9.4 列联表中的相关测量

9.4.1 φ 相关系数

（P196）

φ 相关系数是描述 2×2 列联表数据相关程度的相关系数。$\varphi \in [0, 1]$，越大说明二者相关度越高。

$$\varphi = \sqrt{\chi^2 / n} \tag{3}$$

χ^2 的表达式与§ 9.3 相同，在此有特殊的表达式为（a, b, c, d 的含义是条件频数，也就是观察值。）

$$\chi^2 = \frac{n(ad - bc)}{(a+b)(c+d)(a+c)(b+d)} \tag{4}$$

如果行变量和列变量是相互独立的，则 $ad = bc$（对角线乘积相等）。

	x_1	x_2	Sum
y_1	a	b	a+b
y_2	c	d	c+d
Sum	a+c	b+d	

9.4.2 列联相关系数

列联系数简称 c 系数，用于列联表大于 2×2 的情况。相互独立的时候，$c = 0$；c 永远小于 1。

$$c = \sqrt{\frac{\chi^2}{\chi^2 + n}} \tag{5}$$

9.4.3 V 相关系数

V 相关系数在两个变量相互独立时 $V = 0$；在完全相关时，$V = 1$。当 $\min[(R-1), (C-1)] = 1$ 时 V 退化成 φ。

$$V = \sqrt{\frac{\chi^2}{n \times \min[(R-1), (C-1)]}} \tag{6}$$

9.4.4 数值分析

分析 φ, c, V ，详见 P198。

9.5 列联分析中应注意的问题

9.5.1 百分表的方向

9.5.2 χ^2 分布的期望值准则

如果只有两个单元，则每个单元的期望频数至少是 5.

倘若有两个以上的单元， 20% 的单元期望频数 $f_e < 5$ 则不能使用 χ^2 检验。