独立性检验的强弱分析

报告人　朱建平

厦门大学 经济学院统计系

April 3, 2012

提纲

引言

- $r \times c$ 的二维列联表

在实际中经常要了解两组或多组因素 (或变量) 之间的内在联系.

设有两组因素 A 和 B, 其中因素 A 包含 r 个水平, 即 $A_1, A_2, \ldots, A_r$; 因素 B 包含 c 个水平, 即 $B_1, B_2, \ldots, B_c$. 又设有受制于这两个因素的载体 (或客体) 的集合总体 $\mathcal{N}$. 我们希望通过对总体 $\mathcal{N}$ 关于这两组因素的有关资料 (或抽样资料), 来分析这两组因素的关系.

引言

一般地，设受制于某个载体总体的两个因素为 A 和 B，其中 A 包含 r 个水平.

这里 $A_1, A_2, \ldots, A_r$；B 包含 c 个水平，$B_1, B_2, \ldots, B_c$. 对这两组因素作随机抽样调查，得到一个 $r \times c$ 的二维列联表，记为 $\mathbf{K} = (k_{ij})_{r\times c}$.

这里 $k_{i.} = \sum_{j=1}^{c} k_{ij}$ 表示因素 A 的第 i 个水平的样本个数；$k_{.j} = \sum_{i=1}^{r} k_{ij}$ 表示因素 B 的第 j 个水平的样本个数；$k = k_{..} = \sum k_{ij}$ 表示总的样本个数.

引言

- 问题的提出

我们要通过这一列联表 **K** 来分析两组因素的关联关系. 通常利用独立性检验来推断因素之间是否有联系. 如果两组因素之间不独立, 那么其之间的关联程度有多深, 传统的独立性检验无法描述.

在此, 我们对列联资料的总信息变差进行剖析, 研究独立性检验 χ^2 统计量与总信息变差之间的关系, 通过统计模拟构建独立性检验强弱性分析统计量, 进一步明确独立性检验的内在本质.

提纲

1. 有关记号

对列联表 $\mathbf{K}=(k_{ij})_{r\times c}$ 为一个 $r\times c$，称元素 k_{ij} 为原始频数. 将列联表 $\mathbf{K}$ 转化为频率矩阵，记为 $\mathbf{F}=(f_{ij})_{r\times c}$.

这里 $f_{ij}=k_{ij}/k$ 是样本中属于因素 A 第 i 个水平和因素 B 第 j 个水平的百分比；$f_{i.}=\sum_{j=1}^{c} f_{ij}$，$f_{.j}=\sum_{i=1}^{r} f_{ij}$，
$i=1,2,\ldots,r,j=1,2,\ldots,c$.

列联资料总信息变差的量度

这里我们记

$$\mathbf{f}_r = (f_{1.}, f_{2.}, ..., f_{r.})', \quad \mathbf{f}_c = (f_{.1}, f_{.2}, ..., f_{.c})',$$

$$\mathbf{D}_r = \mathrm{diag}(f_{1.}, ..., f_{i.}, ..., f_{r.}) = \mathrm{diag}(\mathbf{f}_r),$$

$$\mathbf{D}_c = \mathrm{diag}(f_{.1}, ..., f_{.j}, ..., f_{.c}) = \mathrm{diag}(\mathbf{f}_c).$$

那么有，

$$\mathbf{f}_r = \mathbf{F}\mathbf{1}_c, \qquad \mathbf{f}_c = \mathbf{F}'\mathbf{1}_r,$$

$$\mathbf{1}_r'\mathbf{f}_r = \mathbf{1}_c'\mathbf{f}_c = \mathbf{1}_r'\mathbf{F}\mathbf{1}_c = 1.$$

其中 $\mathbf{1}_r = (1, 1, ..., 1)'_{r\times 1}$， $\mathbf{1}_c = (1, 1, ..., 1)'_{c\times 1}$.

列联资料总信息变差的量度

这在此称

$$\mathbf{f}_c^i = (\frac{f_{i1}}{f_{i.}}, \frac{f_{i2}}{f_{i.}}, ..., \frac{f_{ic}}{f_{i.}})' \in \mathbf{R}^c.$$

为因素 A 的第 i 个水平的分布轮廓. 称 $\mathbf{D}_r^{-1}\mathbf{F}$ 为因素 A 的轮廓矩阵. 这里应该注意到, $\mathbf{f}_c^i$, $i = 1, 2, ..., r$ 是超平面 $x_1 + x_2 + ... + x_r = 1$ 的一点集.

同理，因素 B 的第 j 个水平的分布轮廓为

$$\mathbf{f}_r^j = (\frac{f_{1j}}{f_{.j}}, \frac{f_{2j}}{f_{.j}}, \ldots, \frac{f_{rj}}{f_{.j}})' \in \mathbf{R}^r.$$

并称 $\mathbf{D}_c^{-1}\mathbf{F}'$ 为因素 B 的轮廓矩阵，同样 $\mathbf{f}_r^j$， $j = 1, 2, \ldots, c$ 是超平面 $y_1 + y_2 + \ldots + y_c = 1$ 的一点集.

列联资料总信息变差的量度

在此，我们应该明确：

$$\mathbf{D}_r\mathbf{1}_r = \mathbf{F}\mathbf{1}_c, \qquad \mathbf{1}_r'\mathbf{D}_r\mathbf{1}_r = \mathbf{1}_r'\mathbf{F}\mathbf{1}_c = 1,$$

$$\mathbf{D}_c\mathbf{1}_c = \mathbf{F}'\mathbf{1}_r, \qquad \mathbf{1}_c'\mathbf{D}_c\mathbf{1}_c = \mathbf{1}_c'\mathbf{F}'\mathbf{1}_r = 1.$$

从上面的关系式，我们清楚地看到，$\mathbf{D}_r$ 和 $\mathbf{D}_c$ 中的元素起到了权重的作用，称其为权重矩阵.

列联资料总信息变差的量度

2. 总信息变差的量度

针对因素 A 与因素 B 的轮廓矩阵引入卡方 (χ^2) 距离：

$$d^2(i,i') = \sum_{j=1}^{c} \frac{1}{f_{.j}} \left(\frac{f_{ij}}{f_{i.}} - \frac{f_{i'j}}{f_{i'.}}\right)^2$$

$$\text{和} \quad d^2(j,j') = \sum_{i=1}^{r} \frac{1}{f_{i.}} \left(\frac{f_{ij}}{f_{.j}} - \frac{f_{ij'}}{f_{.j'}}\right)^2. \tag{1}$$

这样，根据拟合优度的准则，讨论卡方意义下的总信息变差的量度问题.

列联资料总信息变差的量度

1) 在 χ^2 距离下，以重心计算因素 A 分布轮廓的量度协差阵为

$$\mathbf{S}_r\mathbf{D}_c^{-1} = \mathbf{F}'\mathbf{D}_r^{-1}\mathbf{F}\mathbf{D}_c^{-1} - \mathbf{f}_c\mathbf{f}_c'\mathbf{D}_c^{-1} \triangleq \widetilde{\mathbf{S}}, \tag{2}$$

这里

$$\begin{aligned}\mathbf{S}_r &= \sum_{i=1}^{r} f_{i.}(\mathbf{f}_c^i - \mathbf{f}_c)(\mathbf{f}_c^i - \mathbf{f}_c)' \\ &= \sum_{i=1}^{r} f_{i.}\mathbf{f}_c^i(\mathbf{f}_c^i)' - \mathbf{f}_c\mathbf{f}_c' \\ &= \mathbf{F}'\mathbf{D}_r^{-1}\mathbf{F} - \mathbf{f}_c\mathbf{f}_c',\end{aligned} \tag{3}$$

$$\mathbf{f}_c = \sum_{i=1}^{r} f_{i.}\mathbf{f}_c^i = (f_{.1}, f_{.2}, \ldots, f_{.c})' = \mathbf{1}'\mathbf{D}_c, \tag{4}$$

并且称 $\mathbf{f}_c$ 为关于因素 A 分布轮廓的重心.
在 χ^2 距离下，以原点计算因素 A 分布轮廓的量度协差阵为

$$\mathbf{F}'\mathbf{D}_r^{-1}\mathbf{F}\mathbf{D}_c^{-1} \triangleq \mathbf{S}. \tag{5}$$

列联资料总信息变差的量度

2) 在 χ^2 距离下，以重心计算因素 B 分布轮廓的量度协差阵为

$$\mathbf{S}_c\mathbf{D}_r^{-1} = \mathbf{F}\mathbf{D}_c^{-1}\mathbf{F}'\mathbf{D}_r^{-1} - \mathbf{f}_r\mathbf{f}_r'\mathbf{D}_r^{-1} \triangleq \widetilde{\mathbf{Q}}, \tag{6}$$

$$\begin{aligned}\mathbf{S}_c &= \sum_{j=1}^{c} f_{.j}(\mathbf{f}_r^j - \mathbf{f}_r)(\mathbf{f}_r^j - \mathbf{f}_r)' \\ &= \sum_{i=1}^{r} f_{.j}\mathbf{f}_r^j(\mathbf{f}_r^j)' - \mathbf{f}_r\mathbf{f}_r' \\ &= \mathbf{F}\mathbf{D}_c^{-1}\mathbf{F}' - \mathbf{f}_r\mathbf{f}_r', \end{aligned} \tag{7}$$

$$\mathbf{f}_r = \sum_{j=1}^{c} f_{j.}\mathbf{f}_r^j = (f_{1.}, f_{2.}, \ldots, f_{c.})' = \mathbf{1}'\mathbf{D}_r, \tag{8}$$

并且称 $\mathbf{f}_r$ 为关于因素 B 分布轮廓的重心.
在 χ^2 距离下，以原点计算因素 B 分布轮廓的量度协差阵为

$$\mathbf{F}\mathbf{D}_c^{-1}\mathbf{F}'\mathbf{D}_r^{-1} \triangleq \mathbf{Q}. \tag{9}$$

列联资料总信息变差的量度

那么，以重心量度的总信息变差为 $\mathrm{tr}(\widetilde{\mathbf{S}})$ 和 $\mathrm{tr}(\widetilde{\mathbf{Q}})$；以原点量度的总信息变差为 $\mathrm{tr}(\mathbf{S})$ 和 $\mathrm{tr}(\mathbf{Q})$.

这里应该注意到，$\mathrm{tr}(\widetilde{\mathbf{S}}) = \mathrm{tr}(\widetilde{\mathbf{Q}})$，$\mathrm{tr}(\mathbf{S}) = \mathrm{tr}(\mathbf{Q})$.

提纲

总信息变差与独立性检验统计量的关系

这里主要从两个方面剖析列联表. 一是二维列联表的独立性检验; 二是总信息变差的内涵. 这个问题很少有人在严格的意义上把它们联系起来, 现让我们联系起来分析, 将能深刻地刻划出独立性检验与相应分析的内在关系.

总信息变差与独立性检验统计量的关系

1. 二维列联表的独立性检验

我们知道，频率矩阵 **F** 相应的经验联合抽样分布可以表示为：

$$P\{\xi=i,\eta=j\}=P\{\xi=i\}P\{\eta=j\},\quad i=1,2,...,r,j=1,2,...,c,$$

这里的 ξ 和 η 表示因素 A 和 B 的随机变量. 则根据数理统计理论检验两个变量的独立性用如下统计量

$$\begin{aligned} W_0 &= \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(kf_{ij}-kf_{i.}f_{.j})^2}{kf_{i.}f_{.j}} \\ &= k\sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(f_{ij}-f_{i.}f_{.j})^2}{f_{i.}f_{.j}} \\ &\triangleq k\sum_{i=1}^{r}\sum_{j=1}^{c}(z_{ij})^2, \end{aligned} \tag{10}$$

其中 $z_{ij}=(f_{ij}-f_{i.}f_{.j})/\sqrt{f_{i.}f_{.j}}$. 当假设 H_0：两变量 ξ 和 η 独立成立时，随着 $k\longrightarrow\infty$ 时，统计量 W_0 服从自由度为 $(n-1)(p-1)$ 的 χ^2 分布.

总信息变差与独立性检验统计量的关系

由上面分析知，从因素 A 和因素 B 出发量度总信息变差是一样的，为了叙述方便我们就因素 A 的分布轮廓展开讨论.

定理 1 $\mathbf{f}_c$ 是 $\widetilde{\mathbf{S}} = \mathbf{S}_r\mathbf{D}_c^{-1}$ 的特征值等于 0 时相应的特征向量；$\mathbf{f}_c$ 是 $\mathbf{S} = \mathbf{F}'\mathbf{D}_r^{-1}\mathbf{F}\mathbf{D}_c^{-1}$ 的特征值等于 1 时相应的特征向量.

证： 由于 $(\mathbf{f}_c^i - \mathbf{f}_c)'\mathbf{D}_c^{-1}\mathbf{f}_c = 0$，由 (3) 式知

$$\widetilde{\mathbf{S}}\mathbf{f}_c = \mathbf{S}_r\mathbf{D}_c^{-1}\mathbf{f}_c = \sum_{i=1}^{r} f_{i.}(\mathbf{f}_c^i - \mathbf{f}_c)(\mathbf{f}_c^i - \mathbf{f}_c)'\mathbf{D}_c^{-1}\mathbf{f}_c = 0. \tag{11}$$

即说明，$\mathbf{f}_c$ 是 $\widetilde{\mathbf{S}}$ 的特征值等于 0 时相应的特征向量.
在再根据 (1)、(5) 及 (11) 式

$$\begin{aligned} 0 = \widetilde{\mathbf{S}}\mathbf{f}_c &= \mathbf{F}'\mathbf{D}_r^{-1}\mathbf{F}\mathbf{D}_c^{-1}\mathbf{f}_c - \mathbf{f}_c\mathbf{f}_c'\mathbf{D}_c^{-1}\mathbf{f}_c \\ &= \mathbf{S}\mathbf{f}_c - \mathbf{f}_c, \end{aligned} \tag{12}$$

即

$$\mathbf{S}\mathbf{f}_c = \mathbf{f}_c.$$

从而，$\mathbf{f}_c$ 是 $\mathbf{S}$ 的特征值等于 1 时相应的特征向量. 定理 (1) 得证. ♯

定理 2 除 $\mathbf{f}_c$ 以外，原点协差矩阵 $\mathbf{S}=\mathbf{F}'\mathbf{D}_r^{-1}\mathbf{F}\mathbf{D}_c^{-1}$ 的特征向量 u_k 及其所对应的特征根与重心协差矩阵 $\widetilde{\mathbf{S}}=\mathbf{S}_r\mathbf{D}_c^{-1}$ 是完全一致的.

总信息变差与独立性检验统计量的关系

证: 取重心协差阵 $\widetilde{\mathbf{S}}$ 任一特征向量 $u_k(u_k \neq \mathbf{f}_c)$, 根据 (2) 和 (5) 有

$$\widetilde{\mathbf{S}}u_k = \mathbf{S}u_k - \mathbf{f}_c\mathbf{f}_c'\mathbf{D}_c^{-1}u_k = 0.$$

由定理 1 知, $\mathbf{f}_c$ 与 u_k 均为 $\widetilde{\mathbf{S}}$ 的特征向量, 那么, $\mathbf{f}_c'\mathbf{D}_c^{-1}u_k = 0$, 则

$$\widetilde{\mathbf{S}}u_k = \mathbf{S}u_k,$$

令 u_k 对应的特征值为 β_k, 所以

$$\widetilde{\mathbf{S}}u_k = \beta_k u_k,$$

相应地, 亦有

$$\mathbf{S}u_k = \beta_k u_k,$$

从而, 定理 (2) 得证. $\sharp$

定理 3 在 χ^2 距离意义下，以重心距离反映 $\mathbf{F}$ 的总信息变差与以原点距离反映的总信息变差之间相差单位 1. 即 $\mathrm{tr}(\mathbf{S})-\mathrm{tr}(\widetilde{\mathbf{S}})=1$.

总信息变差与独立性检验统计量的关系

证：由于 $\mathtt{tr}(\mathbf{f}_c\mathbf{f}_c'\mathbf{D}_c^{-1}) = \mathtt{tr}(\mathbf{f}_c'\mathbf{D}_c^{-1}\mathbf{f}_c) = \mathtt{tr}(1) = 1$，再由 (2) 和 (5) 得到

$$\begin{aligned}\mathtt{tr}(\widetilde{\mathbf{S}}) &= \mathtt{tr}(\mathbf{F}'\mathbf{D}_r^{-1}\mathbf{F}\mathbf{D}_c^{-1}) - \mathtt{tr}(\mathbf{f}_c\mathbf{f}_c'\mathbf{D}_c^{-1}) \\ &= \mathtt{tr}(\mathbf{S}) - 1.\end{aligned}$$

从而 $\mathtt{tr}(\mathbf{S}) - \mathtt{tr}(\widetilde{\mathbf{S}}) = 1$. 这样，定理 (3) 得证. $\sharp$

这里我们应该注意到，在 χ^2 距离意义下，以原点距离反映 $\mathbf{F}$ 的总信息变差为 $\sum_{i=1}^r f_{i.} d^2(\mathbf{f}_c^i, 0) = \mathtt{tr}(\mathbf{F}'\mathbf{D}_r^{-1}\mathbf{F}\mathbf{D}_c^{-1})$，而以重心距离反映的总信息变差为 $\sum_{i=1}^r f_{i.}\, d^2(\mathbf{f}_c^i, \mathbf{f}_c) = \mathtt{tr}(\mathbf{S}_r\mathbf{D}_c^{-1})$.

总信息变差与独立性检验统计量的关系

定理 4 设二维列联表的频率矩阵为 $\mathbf{F}=(f_{ij})_{r\times c}$，样本容量为 k. 检验两因素独立性的 χ^2 统计量为 W_0，以重心和原点计算因素 A 分布轮廓的度量协差阵分别为 $\mathbf{S}_r\mathbf{D}_c^{-1}$ 和 $\mathbf{F}'\mathbf{D}_r^{-1}\mathbf{F}\mathbf{D}_c^{-1}$，则

$$k\,\mathrm{tr}(\mathbf{S}_r\mathbf{D}_c^{-1})=W_0 \quad \text{或者} \quad k(\mathrm{tr}(\mathbf{F}'\mathbf{D}_r^{-1}\mathbf{F}\mathbf{D}_c^{-1})-1)=W_0.$$

总信息变差与独立性检验统计量的关系

证： 在此我们对用重心距离的表示详细证明，这一距离相应的总信息变差为

$$
\begin{aligned}
\mathtt{tr}(\mathbf{S}_r\mathbf{D}_c^{-1}) &= \sum_{i=1}^{r} f_{i.}\, d^2(\mathbf{f}_J^i, \mathbf{f}_J) \\
&= \sum_{i=1}^{r} f_{i.} \sum_{j=1}^{c} \frac{1}{f_{.j}} \left(\frac{f_{ij}}{f_{i.}} - f_{.j}\right)^2 \\
&= \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(f_{ij} - f_{i.}f_{.j})^2}{f_{i.}f_{.j}} \\
&= \sum_{i=1}^{r} \sum_{j=1}^{c} (z_{ij})^2 \\
&= W_0/k.
\end{aligned}
$$

即

$$
k\mathtt{tr}(\mathbf{S}_r\mathbf{D}_c^{-1}) = W_0. \tag{13}
$$

根据上面结论，由定理 3 容易得到 $\mathtt{tr}(\mathbf{F}'\mathbf{D}_r^{-1}\mathbf{F}\mathbf{D}_c^{-1}) = W_0/k + 1$. 从而

$$
k(\mathtt{tr}(\mathbf{F}'\mathbf{D}_r^{-1}\mathbf{F}\mathbf{D}_c^{-1}) - 1) = W_0. \tag{14}
$$

这样，定理 (4) 得证. $\sharp$

总信息变差与独立性检验统计量的关系

这里我们需要说明的是，如果以重心和原点计算因素 B 分布轮廓的量度协差阵分别为 $\mathbf{S}_c\mathbf{D}_r^{-1}$ 和 $\mathbf{F}'\mathbf{D}_c^{-1}\mathbf{F}\mathbf{D}_r^{-1}$，同样亦有：

$$\begin{aligned} k\,\mathrm{tr}(\mathbf{S}_c\mathbf{D}_r^{-1}) &= W_0, \\ k(\mathrm{tr}(\mathbf{F}\mathbf{D}_c^{-1}\mathbf{F}'\mathbf{D}_r^{-1}) - 1) &= W_0. \end{aligned} \tag{15}$$

提纲

1. 基本思想

根据数理统计理论，检验两个变量的独立性用统计量 (10)，即为

$$W_0 = k\left(\sum_{i=1}^{r}\sum_{j=1}^{c}\frac{f_{ij}^2}{f_{i.}f_{.j}} - 1\right), \tag{16}$$

另外，我们注意到总信息变差

$$\mathrm{tr}(\mathbf{F}'\mathbf{D}_r^{-1}\mathbf{F}\mathbf{D}_c^{-1}) = \mathrm{tr}(\mathbf{D}^{-\frac{1}{2}}\mathbf{F}'\mathbf{D}_r^{-1}\mathbf{F}\mathbf{D}_c^{-\frac{1}{2}}) = 1 - \sum_{i=1}^{l_0}\beta_i, \tag{17}$$

其中，$\beta_i, i = 1, 2, ..., l_0$ 均为以原点量度的协差阵 $\mathbf{S} = \mathbf{F}'\mathbf{D}_r^{-1}\mathbf{F}\mathbf{D}_c^{-1}$ 非零特征值，且要求 $1 > \beta_1 \geq \cdots \geq \beta_{l_0} > 0$

由定理 (4) 知

$$W_0 = k(\mathrm{tr}(\mathbf{F}'\mathbf{D}_r^{-1}\mathbf{FD}_c^{-1}) - 1) = k\sum_{i=1}^{l_0}\beta_i. \tag{18}$$

即上述的统计量 W_0 就是以原点量度的协差阵 $\mathbf{S}$ 中的小于 1 的特征值和的 k 倍. 因此, 检验零假设 H_0: 两变量 (即两因素) 独立, 完全取决于抽样大小 k 和小于 1 的特征值和的大小. 当给定显著水平 α, 如果 $k\sum_{i=1}^{l_0}\beta_i < \chi^2_{(r-1)(c-1),\alpha}$, 则认为在水平 α 下两组因素是独立的. 这说明所得到的列联表数据仅仅是反映随机误差的, 而没有包含两组因素的关联信息, 这时如果仍然进行两因素关系进行分析的话, 所得的结果只能是虚假的.

如果拒绝了零假设, 则认为适合两组因素之间有一定的关联关系. 那么, 人们会进一步问, 在有关联关系的情形下, 该用分析中的多少个特征值或在多少维投影子空间才能反映两组因素的关联关系, 而其余的则不是呢? 这就需要讨论独立性检验的强弱性.

2. 独立性强弱分析及统计模拟

独立性检验统计量的构造得知统计量 W_0 就是下列矩阵的迹.

$$\mathbf{T}'\mathbf{T},$$

其中

$$\mathbf{T} = \left(\frac{f_{ij}}{\sqrt{f_{i.}f_{.j}}} - \frac{f_{i.}f_{.j}}{\sqrt{f_{i.}f_{.j}}}\right)_{r\times c} = \mathbf{D}_r^{-\frac{1}{2}}\mathbf{F}\mathbf{D}_c^{-\frac{1}{2}} - \mathbf{D}_r^{\frac{1}{2}}\mathbf{1}_r\mathbf{1}_c'\mathbf{D}_c^{\frac{1}{2}}, \tag{19}$$

在此我们引入定义.

定义 设二维列联表的频率矩阵为 $\mathbf{F}$, 相对于因素 A 与因素 B 的权重矩阵为 $\mathbf{D}_r$ 和 $\mathbf{D}_c$. 则称 $\mathbf{D}_r^{-\frac{1}{2}}\mathbf{F}\mathbf{D}_c^{-\frac{1}{2}}$ 为卡方标准化频率矩阵.

定理 5 独立性检验的 χ^2 统计量 W_0 是卡方标准化频率矩阵在正交于矩阵 $\mathbf{S}$ 或 $\mathbf{Q}$ 的最大特征值为 1 时对应的平凡子空间的空间的 k 倍变差.

证： 对卡方标准化频率矩阵 $\mathbf{D}_r^{-\frac{1}{2}}\mathbf{F}\mathbf{D}_c^{-\frac{1}{2}}$ 进行奇异值分解：

$$\mathbf{D}_r^{-\frac{1}{2}}\mathbf{F}\mathbf{D}_c^{-\frac{1}{2}} = \mathbf{D}_r^{\frac{1}{2}}\mathbf{1}_r\mathbf{1}_c'\mathbf{D}_c^{\frac{1}{2}} + \sum_{i=1}^{l_0}\sqrt{\beta_i}\mathbf{D}_r^{-\frac{1}{2}}v_i u_i'\mathbf{D}_c^{-\frac{1}{2}}, \tag{20}$$

这里 u_i 和 v_i 分别是矩阵 $\mathbf{S}$ 和 $\mathbf{Q}$ 对应于小于 1 的第 i 大特征值 β_i 对应的特征向量，且满足

$$u_i'\mathbf{D}_c^{-1}u_j = \begin{cases}0, & i\neq j,\\ 1, & i=j,\end{cases} \quad \text{和} \quad v_i'\mathbf{D}_r^{-1}v_j = \begin{cases}0, & i\neq j,\\ 1, & i=j.\end{cases} \tag{21}$$

由于 $\mathbf{D}_c^{-\frac{1}{2}} u_i$ 和 $\mathbf{D}_r^{-\frac{1}{2}} v_i$ 分别为矩阵 $\mathbf{S}^* = \mathbf{D}_c^{-\frac{1}{2}} \mathbf{F}' \mathbf{D}_r^{-1} \mathbf{F} \mathbf{D}_c^{-\frac{1}{2}}$ 和 $\mathbf{Q}^* = \mathbf{D}_r^{-\frac{1}{2}} \mathbf{F} \mathbf{D}_c^{-1} \mathbf{F}' \mathbf{D}_r^{-\frac{1}{2}}$ 对应于小于 1 的第 i 大特征值的标准特征向量. 特别地, $\mathbf{D}_c^{\frac{1}{2}} \mathbf{1}_c$ 和 $\mathbf{D}_r^{\frac{1}{2}} \mathbf{1}_r$ 分别为 $\mathbf{S}^*$ 和 $\mathbf{Q}^*$ 对应于最大特征值 1 的标准化特征向量.

从而可知 (20) 式就是卡方标准化频率矩阵在依特征值大小的正交特征子空间的奇异分解.

又由于 (20) 式中第一项是在最大特征值 1 对应的子空间的投影, 具有变差 1, 显然这一项是平凡的. 由定理 (4) 知, 独立性检验的 χ^2 统计量 W_0 正是卡方标准化频率矩阵在正交于这一平凡子空间的空间的 k 倍变差. 从而, 定理 (5) 得证. $\sharp$

独立性检验的强弱性分析

针对独立性检验，在两组因素独立的零假设下，即假设总体分布 $\mathbf{F} = (f_{ij})_{r\times c}$ 时，则根据拟合优度检验的有关理论，统计量 W_0 有渐近的自由度为 $(r-1)(c-1)$ 的 χ^2 分布. 由定理 (5) 知，独立检验的零假设 H_0 可表达为总体的卡方标准化分布，有分解：

$$\mathbf{D}_r^{-\frac{1}{2}}\mathbf{F}\mathbf{D}_c^{-\frac{1}{2}} = \mathbf{D}_r^{\frac{1}{2}}\mathbf{1}_r\mathbf{1}_c'\mathbf{D}_c^{\frac{1}{2}}. \tag{22}$$

如果假设被拒绝，则认为两组因素有一定的关联关系，即认为至少有变差 β_1 是反映两组因素有关联关系的.

如果仅有这一个变差是反映这两组因素有关联关系 (记为零假设 H_{10}), 即是假设总体卡方标准化分布矩阵有分解

$$\mathbf{D}_r^{-\frac{1}{2}}\mathbf{F}\mathbf{D}_c^{-\frac{1}{2}} = \mathbf{D}_r^{\frac{1}{2}}\mathbf{1}_r\mathbf{1}_c'\mathbf{D}_c^{\frac{1}{2}} + \sqrt{\beta_1}\mathbf{D}_r^{-\frac{1}{2}}v_1u_1'\mathbf{D}_c^{-\frac{1}{2}}. \tag{23}$$

这一假设 H_{10} 可表达为：总体分布矩阵 $\mathbf{F} = (f_{i.}f_{.j} + \sqrt{\beta_1}v_{1i}u_{1j})_{r\times c}$，其中 $v_1 = (v_{11},...,v_{1r})'$ 和 $u_1 = (u_{11},...,u_{1c})'$ 分别为分布总体下对应矩阵 $\mathbf{S}$ 和 $\mathbf{Q}$ 小于 1 的最大特征值 β_1 的特征向量，且满足 (21) 式，其它参数也为总体参数.

为检验该假设，取统计量

$$W_1 = W_0 - k\beta_1.$$

记 (20) 式两边的样本之差为

$$\mathbf{T}_1 = \mathbf{D}_r^{-\frac{1}{2}}\mathbf{F}\mathbf{D}_c^{-\frac{1}{2}} - \mathbf{D}_r^{\frac{1}{2}}\mathbf{1}_r\mathbf{1}_c'\mathbf{D}_c^{\frac{1}{2}} - \sqrt{\beta_i}\mathbf{D}_r^{-\frac{1}{2}}v_1u_1'\mathbf{D}_c^{-\frac{1}{2}},$$

则有

$$W_1 = k\mathrm{tr}(\mathbf{T}_1'\mathbf{T}_1) = k\sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(f_{ij} - f_{i.}f_{.j} - \sqrt{\beta}_1 v_{1i}u_{1j})^2}{f_{i.}f_{.j}}.$$

通过统计模拟，可知统计量 W_1 有渐近服从自由度为 $(r-2)(c-2)$ 的 χ^2 分布.

独立性检验的强弱性分析

为一般地，考虑零假设 H_{l0}：有且仅有前 $l(\leq l_0)$ 个变差 $\sum_{i=1}^{l}\beta_i$ 反映两组因素关联关系，即其总体的分布矩阵满足：

$$\mathbf{F} = (f_{i.}f_{.j} + \sum_{m=1}^{l}\sqrt{\beta_m}v_{mi}u_{mj})_{r\times c}, \tag{24}$$

或者说总体卡方标准化分布矩阵有分解

$$\mathbf{D}_r^{-\frac{1}{2}}\mathbf{F}\mathbf{D}_c^{-\frac{1}{2}} = \mathbf{D}_r^{\frac{1}{2}}\mathbf{1}_r\mathbf{1}_c'\mathbf{D}_c^{\frac{1}{2}} + \sum_{i=1}^{l}\sqrt{\beta_i}\mathbf{D}_r^{-\frac{1}{2}}v_iu_i'\mathbf{D}_c^{-\frac{1}{2}}. \tag{25}$$

独立性检验的强弱性分析

为检验该假设，取统计量

$$W_l = W_0 - k\sum_{m=1}^{l}\beta_m, \tag{26}$$

即则有

$$W_l = k\sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(f_{ij} - f_{i.}f_{.j} - \sum_{m=1}^{l}\sqrt{\beta_m}v_{mi}u_{mj})^2}{f_{i.}f_{.j}}. \tag{27}$$

同样根据统计模拟，可得知统计量 W_l 有渐近服从自由度为 $(r-l-1)(c-l-1)$ 的 χ^2 分布.

独立性检验的强弱性分析

综上所述，这样就得到了相应分析的依次检验程序：对于给定的显著性水平 α, 首先对零假设 H_0 检验，计算统计量 W_0，判断 W_0 是否大于临界值 $\chi^2_{(r-1)(c-1),\alpha}$，如果否，则检验结束. 认为两因素之间不存在关联关系，并称因素 A 和因素 B 具有零度关联性；如果是，则对零假设 H_{10} 进行检验，计算统计量 W_1，判断 W_1 是否大于临界值 $\chi^2_{(r-2)(c-2),\alpha}$，如果否，则检验结束，并称两因素具有一度关联性；重复上述检验和相应分析，直到对某个 l，如果算得统计量 W_l 对检验假设 H_{l0} 被拒绝，而算得统计量 W_{l+1} 对检验假设 $H_{(l+1)0}$ 被接受，则结束检验，称两因素有 l 度关联性，这时认为两因素的关联程度较强.

独立性检验的强弱性分析

模拟实例. 给定两组因素 A 和 B, 分别含 5 个变量和 4 个变量. 表 a 的数据反映的是两个因素基本独立; 表 b 的数据反映两因素有些相关, 即偏离独立, 但并不严重; 表 c 的数据是反映两因素几乎完全独立, 是用来作比较的, 三个表有相同的行列边际 $(300, 160, 400, 140)'$ 和 $(120, 100, 130, 250, 400)'$.

现在我们对表 a 和表 b 的数据作检验和分析.

独立性检验的强弱性分析

表 1：两因素模拟数据

表 a. 两因素数据 (1)

	B_1	B_2	B_3	B_4
A_1	36	19	48	17
A_2	33	13	40	14
A_3	39	21	52	18
A_4	75	45	100	30
A_5	117	62	160	61

表 b. 两因素数据 (2)

	B_1	B_2	B_3	B_4
A_1	36	19	56	9
A_2	36	10	40	14
A_3	39	21	44	26
A_4	75	48	105	22
A_5	114	62	155	69

表 c. 两因素数据 (3)

	B_1	B_2	B_3	B_4
A_1	36	19	48	17
A_2	30	16	40	14
A_3	39	21	52	18
A_4	75	40	100	35
A_5	120	64	160	56

例中的参数为 $k=1000$，$n=5$，$p=4$. 首先对表 a 的数据作分析. 算得 S 阵的四个非零特征值是 $1, 0.0021, 0.0007, 0$. 检验 H_0(即两组因素独立) 的统计量 W_0 的值为 $W_0=2.7943$，与自由度为 12 的 χ^2 在显著水平 $\alpha=0.05$ 下的临界值 21.03 比较，有 $2.7943<21.03$，故接受零假设. 事实上，该数据与完全独立的表 c 的数据只有稍许差别，可以认为只是随机误差所至. 可见和直观分析一致.

独立性检验的强弱性分析

看表 b 的数据. 算得 **S** 阵的四个非零特征值是 1, 0.0180, 0.0043, 0.0009. 依次检验的统计量数据和结果见表 2.

表 2：对两因素数据 (2) 的检验分析

假设检验	统计量值	χ^2 自由度	显著水平 α	临界值	检验结果
H_0	$W_0 = 23.1976$	12	0.05	21.03	拒绝
H_{10}	$W_1 = 5.1976$	6	0.05	12.59	接受
H_{20}	$W_2 = 0.8976$	2	0.05	5.99	接受

独立性检验的强弱性分析

从表 2 的数据结果判断提示，对表 b 数据有必要作相应分析，而且只需要作降至一维的相应分析就足够了，如果进行降至更高维的相应分析，可能是虚假的，即可能将随机误差当成关联关系. 将表 b 数据和表 c 数据比较，分析有系统差别但并不十分严重，因而可能只存在轻微的关联关系. 可见这里的检验结果和分析是合理的.

提纲

参考文献

[1]. Agrawal, R., Imielinski, T. and Swami, A. (1993), Mining Association Rules Between Sets of Items in Large Database , *Proc of ACM SIGMOD Intl Conf on Management of Data*, 207-216
[2]. Brin, S., Motwani, R. and Silverstein, C. (1997), Beyond Market Basket: Generalizing Association Rules to Correlations ,*1997 Int. Conf. Management of Data*, 265-276
[3]. Benzécri, J. P. (1992), *Correspondence Analysis Handbook* , Marcel Dekker, Inc., New York
[4]. 陈希孺, 倪国熙编著 (1988), *数理统计学教程*, 上海科学技术出版社
[5]. Cramor, H. (1946), *Mathematical Methods for Statistics*, Prineeton Univ. Press
[6]. Everitt, B. S. (1977), *The analysis Contingency Tables* , Chapman and Hall , New York
[7]. 胡国定, 张润楚著 (1989), *多元数据分析方法——纯代数理论*, 南开大学出版社
[8]. Kendall, M. and Stuart, A. (1979), The Advanced Theory of Ststisties, *Charles Griffin & Conpany Limiled*, London, **Vol. 2.Ch**. 33
[9]. Pearson, K. *On the Theory of Contingency and Its Relation to Association and Normal Correlation*, Drapers' Co Memoirs, Biometrie Scries No. 1 London
[10]. Silverstein, C.,Brin, S., Motwani, R. and Ullman, J. (1998), Scalable Techniques for Mining Causal Structures, 1998 Int. Conf. Very Large Data Bases, 594-605
[11]. van de Velden, M. and Neudecker, H. (2000), On an Eigenvalue Property in Correspondence Analysis and Related Methods, *Liner Algebra and its Applications*, **321**, 347-364
[12]. 张尧庭, 谢邦昌, 朱世武 (2001), *数据挖掘入门及应用*, 中国统计出版社, 34-36
[13]. 张润楚, 朱建平 (2002), 相应分析的适应性检验, *第七届全国概率统计学术会议报告*
[14]. 朱建平 (2003), Data Mining 中的统计方法及其应用, *博士论文* 71-93
[15]. 朱建平 (2004.1), 数据挖掘中事务性数据库的压缩及其应用, *统计研究*, **147**, 38-43

Thank you!

Author:	Zhu Jianping
Address:	Dept. of Statistics
	Xiamen University
	FJ, P.R.CHINA, 361005
Phone:	0592-2186371
Email:	xmjpzhu@xmu.edu.cn