目录

1 皮尔森检验及其系数

对于两个分类变量的相关性检验和列联系数计算。

1.1 皮尔森卡方检验（Pearson chi-square test）

在皮尔森检验的独立性检验中，虚无假设（null hypothesis）为：两个变量呈统计独立性。首先，每个观察值会被重新编排到一个“列联表”的二维表格里。如果列联表共有R行C列，那么在独立事件的假设下，每个字段的期望次数计算为：

$$\mathrm{E}_{i,j} = \frac{\left(\sum_{n_C=1}^{C} O_{i,n_C}\right) \cdot \left(\sum_{n_R=1}^{R} O_{n_R,j}\right)}{N}$$

也可以理解为：行合计*列合计/总例数。其中N是样本总量，O为对应的观测值，C为总列数，R为总行数。χ^2统计值的公式为：

$$\chi^2 = \sum_{i=1}^{R} \sum_{j=1}^{C} \frac{\left(O_{i,j} - E_{i,j}\right)^2}{E_{i,j}}$$

自由度计算公式为：$\mathrm{df} = (\mathrm{R}-1)(\mathrm{C}-1)$。

根据设定的置信水准，查出自由度为 df 的卡方分配临界值，比较它与之前得出的卡方统计值，然后推论能否拒绝虚无假设。或查找右尾几率，如果这个值小于临界值则可拒绝虚无假设，既两个变量不呈统计独立性。

卡方检验的限制如下：

1. 如果个别字段的期望次数太低，会使机率分配无法近似于卡方分配。一般要求：自由度$\mathrm{df} > 1$时，期望次数小于 5 的字段不多于总字段的 20%。
2. 若自由度$\mathrm{df} = 1$，且若期望次数<10，则近似于卡方分配的假设不可信。此时可以将每个观察值的离差减去 0.5 之后再做平方，这就是叶氏连续性修正。

1.2 似然比卡方检验（Likelihood-Ratio chi-square test）

似然比卡方检验又称为 G 检验（G-test）。检验的独立性分析是卡方检验的另一种方法，两种检验几乎会给出同样的结果，不过 G 检验需要总的样本数量大于 1000。

G 检验的虚无假设是：一个变量的相对比例独立于另一个变量。G^2统计值的公式为：

$$\mathrm{G}^2 = 2\sum_{i=1}^{R} \sum_{j=1}^{C} O_{i,j} \ln\left(\frac{O_{i,j}}{E_{i,j}}\right)$$

G 检验的自由度计算公式和检验方法同皮尔森卡方检验。

通过以上两种检验方法检验出关联性后可以计算相关列联系数。既当我们拒绝虚无假设时或者当$p < 0.05$时，可以计算列联系数。

1.3 皮尔森列联系数（Pearson Contingency Coefficient）

皮尔森列联系数是从皮尔森卡方衍生而来用于度量关联性的指标。r的取值范围为 0 到$\sqrt{\frac{\min(R,C)-1}{\min(R,C)}}$，计算公式为：

$$r = \sqrt{\frac{\chi^2}{\chi^2 + N}}$$

皮尔森列联系数最好用于 5*5 及以上的表格，得出的值越趋近于 1 代表变量间关联性越大。

1.4 Phi 系数（Phi Coefficient）

Phi 系数是从皮尔森卡方衍生而来主要用于度量两个二元变量间的相关程度。对于一个两行两列的列联表（2*2），ϕ的取值范围为-1 到 1。计算公式为：

$$\phi = \frac{n_{11}n_{22} - n_{12}n_{21}}{\sqrt{n_{1.}n_{.1}n_{2.}n_{.2}}}$$

其中n_{11}、n_{21}、n_{12}、n_{22}为对应列联表的值，$n_{.1}$、$n_{1.}$、$n_{2.}$、$n_{.2}$为对应行列的总和。当取值为-1 时表示完全负相关，为 1 时表示完全正相关，为 0 时表示两者没有关系。

对于其他的列联表，ϕ的取值范围为 0 到$\min(\sqrt{R-1}, \sqrt{C-1})$。

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

1.5 Cramer 氏 V 系数（Cramer's V Coefficient）

Cramer 氏 V 系数是从皮尔森卡方衍生而来主要用于度量关联性的指标。当初设计时，它的最大可达的上限是 1。对于两行两列的列联表（2*2），V的取值范围为-1 到 1。计算公式和 Phi 系数相同。对于其他的列联表，V的取值范围为 0 到 1，计算公式为：

$$V = \sqrt{\frac{\chi^2/n}{\min(R-1, C-1)}}$$

当取值为趋近于 0 时表示两个变量不相关，趋近于 1 时表示两个变量相关。当两个变量完全相同时取值为 1。

2 其他相关性检验系数

2.1 最大信息系数（MIC）

MIC（Maximal Information Coefficient）指标可以用来衡量两变量之间的相似程度，相似程度越高，值越大。传统的衡量指标只能度量两变量之间少量的关系，大都只能度量线性关系，像x，x^2 这样的两个变量之间的关系如果用 Pearson 相关系数就没法度量，但很明显两者之间有某种完全的相似程度，这时如果用 MIC 指标就能得出二者的 MIC 指标值为 1，也就是表示两者完全相似。

下面简单介绍下 MIC 指标具体计首先介绍 Mutual information（MI），互信息。它是衡量两信息的相似程度的指标，相似程度越高值越大。连续情况的计算公式如下：

$$I(x;y) = \int_Y \int_X p(x,y)\log\left(\frac{p(x,y)}{p(x)p(y)}\right) dxdy$$

其中$p(x,y)$表示两个变量的联合分布。MIC 指标在基于 MI 的指标下计算公式为：

$$MIC(x;y) = \max_{XY<B} \frac{I(x;y)}{\log(\min(X,Y))}$$

公式的解释和计算 MIC 的具体三个步骤如下：

1. 给定i、j，对XY构成的散点图进行i列j行网格化，并求出最大的互信息（MI）也就是分子$I(x;y)$。这里给定i、j后，可以得出多种不同的网格化方案。需要从不同的方案中找到互信息的最大值。

2. 对最大的互信息进行归一化，也就是除以分母的$\log(\min(X,Y))$。

3. 选择不同尺度下互信息的最大值作为 MIC 值。1.中我们只给定了一组i、j，实际中我们会有多组不同的i、j，找出所有不同i、j中的最大归一化互信息就是我们需要的 MIC。

3 皮尔森检验和系数在 R 中的实现

在 R 中可以直接使用 *vcd* 包进行皮尔森检验和计算。

假设有如下两个列联表：

	A	B
a	100	100
b	100	100

	A	B
a	100	0
b	0	100

通过 *vcd* 中的 assocstats 函数可得出结果如下，对于第一个列联表：

```
                     X^2 df P(> X^2)
Likelihood Ratio      0  1        1
Pearson               0  1        1

Phi-Coefficient   : 0
Contingency Coeff.: 0
Cramer's V        : 0
```

对于第二个列联表：

```
                     X^2 df P(> X^2)
Likelihood Ratio  277.26  1        0
Pearson           200.00  1        0

Phi-Coefficient   : 1
Contingency Coeff.: 0.707
Cramer's V        : 1
```

具体结果的含义可参见第一节。

4 MIC 在 R 中的计算

可使用 *minerva* 包下的 *mine* 函数来计算 MIC。具体事例如下：

```
> x <- runif(10); y <- 3*x+2
> mine(x,y)
$MIC
    [1] 1
```