阮敬 博士

首都经济贸易大学研究生院 副院长
首都经济贸易大学统计学院 教授

© ruanjing@msn.com

# 列联分析与对应分析

- 人们在研究某一个事物或现象的过程中，有些时候不仅只会考察单独某一个方面的信息，也可以把几个方面的信息联合起来一并考察。如考察某项政策实施之后广大市民对该政策的民意反映，可以用单独一个民意指标“满意状况”来考察。如果把性别指标一并联合起来，考察不同性别人群对该项政策的满意状况，这就是用两个指标来衡量同一个事物，这两个指标的不同表现可以通过交叉的方式形成若干种状况，如男性对该项政策的满意状况、女性对该项政策的不满意状况等，把性别和满意状况这两个变量交叉联合起来，共同对所研究的问题展开研究，这个过程就叫做“交叉分析”。本章将要讲述的列联分析和对应分析就是交叉分析的两种典型形式。

# 列联分析—列联表

- 对于定类或定序等定性数据的描述和分析， 通常可使用列联表进行分析， 本节主要介绍基于列联表$\chi^2$检验的列联分析。
- 两个或两个以上变量交叉形成的二维频数分布表格，称之为“列联表” 。如本章引言部分的不同性别对政策实施满意状况的交叉频数分布，设“性别”变量有“男” 、 “女”2 种属性、 “满意状况”变量有“满意” 、“不满意”2 种属性，得到的列联表形如表 15-1 所示。

表 15-1 二维列联表的一般形式

| 人数 | | 满意状况 | | 合计 |
|---|---|---|---|---|
| | | 满意 | 不满意 | |
| 性别 | 男 | 128 | 117 | 245 |
| | 女 | 109 | 96 | 205 |
| 合计 | | 237 | 213 | 450 |

- 表 15-1 所示的列联表形式非常简单，只是把两个变量的不同属性进行交叉，计算出各种属性组合的频数，作为表格中的主要数据。
- 从列联表中， 可以清楚看到所有人和不同性别的人对该项政策的不同观点分布状况； 同时也可以看到所有满意状况及其两种属性表现的性别分布状况。
- 列联表中变量的属性或取值通常也叫做“水平” ，如性别变量有“男” 、 “女”两个水平，“满意状况”变量有“满意”和“不满意”两个水平。

# 列联分析—列联表

- 列联表行变量的水平个数一般用 R 表示，列变量水平的个数一般用 C 表示，那么一个R 行 C 列的频数分布表叫做 R×C 列联表，如表 15-2 所示。

表 15-2 R ×C列联表

| 频数 | | 列变量 | | | | 行合计 |
|---|---|---|---|---|---|---|
| | | 水平 1 | 水平 2 | … | 水平 c | |
| 行变量 | 水平 1 | $f_{11}$ | $f_{12}$ | … | $f_{1c}$ | $\sum_{j=1}^{c} f_{1j}$ |
| | 水平 2 | $f_{21}$ | $f_{22}$ | … | $f_{2c}$ | $\sum_{j=1}^{c} f_{2j}$ |
| | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| | 水平 r | $f_{r1}$ | $f_{r2}$ | … | $f_{rc}$ | $\sum_{j=1}^{c} f_{rj}$ |
| 列合计 | | $\sum_{i=1}^{r} f_{i1}$ | $\sum_{i=1}^{r} f_{i2}$ | … | $\sum_{i=1}^{r} f_{ic}$ | $\sum_{i=1}^{r}\sum_{j=1}^{c} f_{ij}$ |

- R×C 列联表中各元素 f ij 就是行列变量进行交叉分类得到的观测值个数所形成的频数分布， 行合计表示行变量每个水平在列变量不同水平交叉分类的观测值总数； 列合计表示列变量每个水平在行变量不同水平交叉分类的观测值总数； 行合计加总应当等于列合计加总， 记为总计频数。

# 列联分析—列联表

- 例15-1：某单位欲推行一套新的工资改革方案， 为了考查该方案的合理性， 提高改革方案在公司各部门推行之后的实际效果，特抽查了市场部、客户服务部、发展战略部、综合部、 研发中心等 5 个部门共 220 名员工了解对该套工资改革方案的态度， 以该例数据编制的列联表如表 15-3 所示。

表 15-3　各部门员工对工资改革的态度

| 人数 | | 部门 | | | | | 合计 |
|---|---|---|---|---|---|---|---|
| | | 发展战略部 | 客户服务部 | 市场部 | 研发中心 | 综合部 | |
| 态度 | 反对 | 25 | 15 | 20 | 27 | 29 | 116 |
| | 支持 | 16 | 21 | 23 | 22 | 22 | 104 |
| 合计 | | 41 | 36 | 43 | 49 | 51 | 220 |

- SAS 系统中有两种数据预处理方式可以输出列联表：
- 第 1 种数据预处理方式就是以原始调查数据作为数据集，然后利用前面章节介绍过的FREQ 过程（详见第 6.2.2 节）制表得到列联表；第 2 种数据预处理方式是输入形如表 15-3所示的交叉分组数据，仍然利用 FREQ 过程，并在 FREQ 过程中通过 WEIGHT 语句指定交叉分组频数作为权数，也可得到列联表。

# 列联分析—列联表

- 第 1 种数据预处理方式的具体数据格式如图 15-1 所示。
- 本例所使用的数据值标签如下：

```
proc format;
  value department_fmt  1='发展战略部'
                        2='客户服务部'
                        3='市场部'
                        4='研发中心'
                        5='综合部';
  value attitude_fmt 1='支持'
                     2='反对';
run;
```

Dataset of SASUSER.SALARY_REFORM

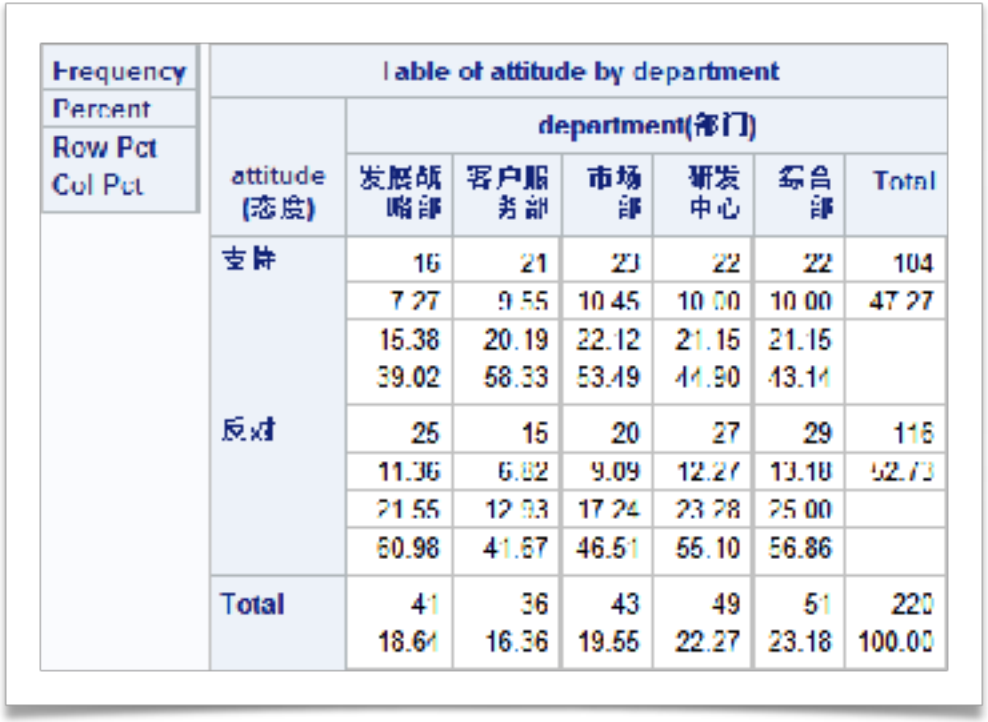
| Obs | ID 编号 | Department 部门 | Attitude 态度 |
|---|---|---|---|
| 1 | 1 | 3 | 2 |
| 2 | 2 | 5 | 2 |
| 3 | 3 | 5 | 2 |
| 4 | 4 | 4 | 2 |
| 5 | 5 | 1 | 1 |
| ... | ... | ... | ... |
| 217 | 217 | 2 | 1 |
| 218 | 218 | 3 | 2 |
| 219 | 219 | 1 | 1 |
| 220 | 220 | 1 | 2 |

# 列联分析—列联表

- 根据第 6 章中介绍过的内容，使用 FREQ 过程编制最为常见的二维列联表的程序如下：

```
proc freq data=sasuser.salary_reform;
  table attitude*department;
  format attitude attitude_fmt. department department_fmt.;
run;
```

- 程序运行之后，可得到如图 15-2 所示的结果。
- 图 15-2 中的表格一共有四行数字，表格的左上角标注了每行数字所代表的意思。第 1行表示交叉分类的频数（Frequency），依次往下分别是百分比（Percent，单位 100%）、行百分比（Row Pct）、列百分比（Col Pct）。



| Frequency<br>Percent<br>Row Pct<br>Col Pct | Table of attitude by department | | | | | |
|---|---|---|---|---|---|---|
| | department(部门) | | | | | |
| attitude(态度) | 发展战略部 | 客户服务部 | 市场部 | 研发中心 | 综合部 | Total |
| 支持 | 16 | 21 | 23 | 22 | 22 | 104 |
| | 7.27 | 9.55 | 10.45 | 10.00 | 10.00 | 47.27 |
| | 15.38 | 20.19 | 22.12 | 21.15 | 21.15 | |
| | 39.02 | 58.33 | 53.49 | 44.90 | 43.14 | |
| 反对 | 25 | 15 | 20 | 27 | 29 | 116 |
| | 11.36 | 6.82 | 9.09 | 12.27 | 13.18 | 52.73 |
| | 21.55 | 12.93 | 17.24 | 23.28 | 25.00 | |
| | 60.98 | 41.67 | 46.51 | 55.10 | 56.86 | |
| Total | 41 | 36 | 43 | 49 | 51 | 220 |
| | 18.64 | 16.36 | 19.55 | 22.27 | 23.18 | 100.00 |

# 列联分析—列联表

- 第 2 种数据预处理格式如图 15-3 所示。
- 该种数据预处理方式应用了交叉分类汇总的形式进行数据输入，即在向 SAS 系统录入数据之前，已经依据行列变量的水平，统计出各种水平交叉分类出现的人数或次数，以数据汇总的方式录入到 SAS 数据集中，这样作为出现人数或次数的那个变量便衡量了其对应交叉情形出现的情况，因此成为各种交叉情形的权重。所以，在第 2 种数据预处理方式下绘制列联表要注意权数的应用。

Dataset of SASUSER.SALARY_REFORM_W

| Obs | Department 部门 | Attitude 态度 | Counts 人数 |
|---|---|---|---|
| 1 | 1 | 2 | 25 |
| 2 | 1 | 1 | 16 |
| 3 | 2 | 2 | 15 |
| 4 | 2 | 1 | 21 |
| 5 | 3 | 2 | 20 |
| 6 | 3 | 1 | 23 |
| 7 | 4 | 2 | 27 |
| 8 | 4 | 1 | 22 |
| 9 | 5 | 2 | 29 |
| 10 | 5 | 1 | 22 |

# 列联分析—列联表

- 程序运行后得到的结果与图 15-2 完全相同。此外，FREQ 过程的 TABLE 语句也提供了用于调整列联表输出内容的语句选项关键字：
  - EXPECTED：输出理论期望频数；
  - DEVIATION：输出观测值与期望值之差；
  - NOFREQ：不输出交叉分类频数；
  - NOPERCENT：不输出百分比；
  - NOROW：不输出行百分比；
  - NOCOL：不输出列百分比；

# 列联分析—列联表的分布

- 列联表中的分布有两种：一种是如表 15-3 或图 15-2 那样，能够直接从样本数据中获得的交叉分类分布，可以直接观测得到。其行、列合计分别称为行边缘分布和列边缘分布；另一种是期望值的分布，是不能直接观测出来的，可以通过样本数据和相关理论进行计算。
- 以例 15-1 为例，如果要想了解不同部门的员工对工资改革方案的态度是否存在显著差异，在没有显著差异的假定条件下，各部门员工不同态度的分布即为列联表的理论分布。据此可以计算出各部门态度人数的理论期望频数值
- 在本例中，持“反对”态度的员工总人数为 116 人，持“支持”态度的员工总人数为104 人。因此对整个单位而言， 对工资改革方案的反对率应当为 116/220=0.5273， 即 52.73%；对工资改革方案的支持率应当为 104/220=0.4727，即 47.27%。

# 列联分析—列联表的分布

- 现假定各部门对工资改革的态度没有差异， 故各部门反对该项政策的人数应当为该部门被调查人数乘以反对率，支持该项政策的人数应当为该部门人数乘以支持率。 如发展战略部一共有员工 41 人，持支持态度的理论人数应当为 41*47.27%=19.38 人。由此计算出来的人数便是列联表的期望值，计算过程及期望值分布如表 15-4 所示。

**表 15-4　列联表的期望分布**

| 期望人数 | | 部门 | | | | | 合计 |
|---|---|---|---|---|---|---|---|
| | | 发展战略部 | 客户服务部 | 市场部 | 研发中心 | 综合部 | |
| 态度 | 反对 | 41*0.5273=21.62 | 36*0.5273=18.99 | 43*0.5273=22.67 | 49*0.5273=25.84 | 51*0.5273=26.89 | 116 |
| | 支持 | 41*0.4727=19.38 | 36*0.4727=17.02 | 43*0.4727=20.33 | 49*0.4727=23.16 | 51*0.4727=24.11 | 104 |
| 合计 | | 41 | 36 | 43 | 49 | 51 | 220 |

- 表 15-4 所示的期望分布可由通过设置 FREQ 过程的 TABLE 语句选项来自动进行计算，程序如下：

```
proc freq data=sasuser.salary_reform_w;
  table attitude*department /expected nopercent norow nocol deviation;
  weight counts;
  format attitude attitude_fmt. department department_fmt.;
run;
```

# 列联分析—列联表的分布

- 本例控制列联表中内容仅为观测频数、 期望频数以及二者之差。 程序运行结果如图 15-4所示。

Frequency
Expected
Deviation

Table of attitude by department

| attitude (态度) | department(部门) 发展战略部 | 客户服务部 | 市场部 | 研发中心 | 综合部 | Total |
|---|---|---|---|---|---|---|
| 支持 | 16 | 21 | 23 | 22 | 22 | 104 |
| | 19.382 | 17.018 | 20.327 | 23.164 | 24.109 | |
| | -3.382 | 3.9818 | 2.6727 | -1.164 | -2.109 | |
| 反对 | 25 | 15 | 20 | 27 | 29 | 116 |
| | 21.618 | 18.982 | 22.673 | 25.836 | 26.891 | |
| | 3.3818 | -3.982 | -2.673 | 1.1636 | 2.1091 | |
| Total | 41 | 36 | 43 | 49 | 51 | 220 |

# 列联分析—$\chi^2$分布与$\chi^2$检验

- 从第 15.1.2 节中可以得知列联表的分布主要有观测值分布和期望值分布， 同时也计算了观测值与期望值之间的偏差。设$f_{ij}^o$表示各交叉分类频数的观测值，$f_{ij}^e$表示各交叉分类频数的期望值，则各交叉分类频数观测值与期望值的偏差为$f_{ij}^o - f_{ij}^e$，则$\chi^2$统计量为：

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

- 当样本量较大时，$\chi^2$（卡方）统计量近似服从自由度为（R-1）（C-1）的$\chi^2$分布，$\chi^2$值与期望值、观测值和期望值之差均有关，值越大表明观测值与期望值的差异越大。因此，可以由此对第 15.1.2 节中计算期望值的假设进行$\chi^2$检验。

- 上一节在计算期望值分布时， 假定各部门对工资改革的态度没有差异， 即各部门对该项改革方案的支持率或反对率均相等， 即员工对该项改革方案的态度与其所在部门无关， 行列变量之间是独立的。据此可以提出原假设和备择假设：

$H_0$:部门与对改革方案态度独立; $H_1$:部门与对改革方案态度不独立

# 列联分析—$\chi^2$分布与$\chi^2$检验

- 对于该假设所进行的检验，除了可用上述介绍的近似$\chi^2$检验之外，在 SAS 系统中还可以进行 FISHER 精确检验， 只需要在 FREQ 过程的 TABLE 语句选项中加入关键字 CHISQ （进行$\chi^2$检验）或关键字 EXACT（进行 FISHER 精确检验） ，利用例 15-1 的数据进行分析的具体程序如下：

```
proc freq data=sasuser.salary_reform_w;
  table attitude*department /chisq exact expected;
  weight counts;
  format attitude attitude_fmt. department
department_fmt.;
run;
```

- 程序运行之后，可得到图 15-2 所示的列联表结果，以及 x 2 检验和 FISHER 精确检验的结果，如图 15-5 所示。

Statistics for Table of attitude by department

| Statistic | DF | Value | Prob |
|---|---|---|---|
| Chi-Square | 4 | 4.0133 | 0.4042 |
| Likelihood Ratio Chi-Square | 4 | 4.0260 | 0.4025 |
| Mantel-Haenszel Chi-Square | 1 | 0.0603 | 0.8061 |
| Phi Coefficient | | 0.1351 | |
| Contingency Coefficient | | 0.1338 | |
| Cramer's V | | 0.1351 | |

| Fisher's Exact Test | |
|---|---|
| Table Probability (P) | 6.543E-05 |
| Pr <= P | 0.4105 |

Sample Size = 220

# 列联分析—$\chi^2$分布与$\chi^2$检验

- 图 15-5 显示，$\chi^2$统计量的值为 4.0133，其对应的 P 值（Prob）为 0.4042，非常不显著。因此，没有充分理由拒绝行列变量即部门与态度之间独立的原假设。
- FREQ 过程还提供了其他几种形式的$\chi^2$统计量进行检验，如似然比$\chi^2$统计量、Mantel-Haenszel$\chi^2$统计量等， 其检验结果均表明在一定的显著性水平条件下没有充分理由拒绝原假设。
- 由于$\chi^2$统计量是一个近似的统计量，该过程的输出结果还提供了 FISHER 精确检验的过程。FISHER 精确检验的统计量服从超几何分布，在样本量较大的时候运算量非常大。
- FISHER 精确检验结果也表明没有充分理由拒绝行列变量，即部门与态度之间独立的原假设。

# 列联分析—列联表中的关联度分析

- 通过$\chi^2$检验，如果得到行列变量之间不是独立的结论，则列联分析中还可以对行列变量之间的相关性进行测量。
- 例15-2：某公司推销一种业务， 为考察该项业务在不同收入消费人群的使用购买意向， 进行了调查，调查结果如图 15-6 所示，试分析收入和购买意向之间的关系。
- 本例使用的数据值标签如下：

```
proc format;
  value income_fmt  1='低收入'
                    2='中收入'
                    3='高收入';
  value propensity_fm 1='不愿购买'
                      2='暂无打算'
                      3='愿意购买';
run;
```

Dataset of SASUSER.PURCHASE

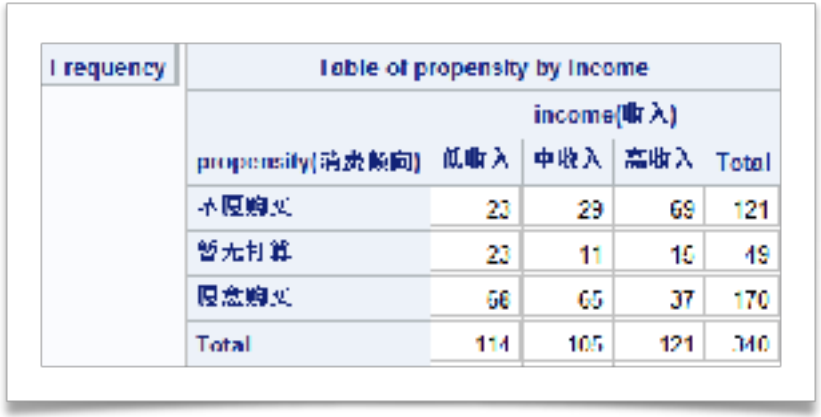
| Obs | Income 收入 | Propensity 消费倾向 | Counts 人数 |
|---|---|---|---|
| 1 | 1 | 3 | 50 |
| 2 | 1 | 1 | 23 |
| 3 | 1 | 2 | 23 |
| 4 | 2 | 3 | 55 |
| 5 | 2 | 1 | 29 |
| 6 | 2 | 2 | 11 |
| 7 | 3 | 3 | 37 |
| 8 | 3 | 1 | 59 |
| 9 | 3 | 2 | 15 |

# 列联分析—列联表中的关联度分析

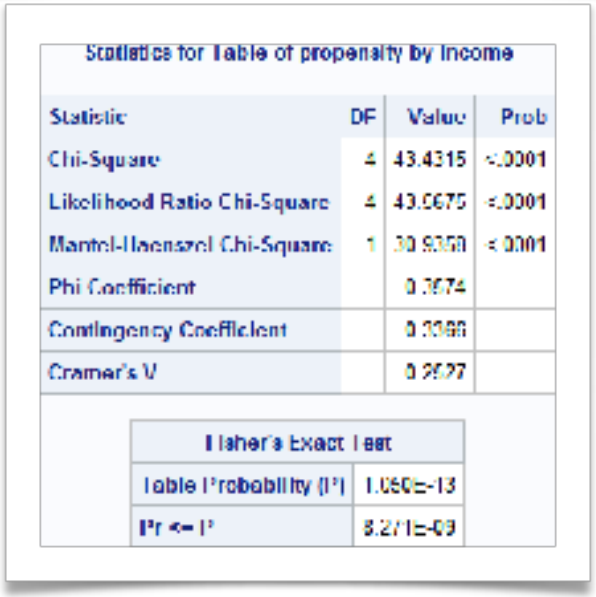
- 在FREQ过程中可以使用TABLE语句的MEASURES选项进行关联度分析， 程序如下：

```
proc freq data=sasuser.purchase;
  table propensity*income /nopercent nocol norow chisq
exact measures;
  weight counts;
  format income income_fmt. propensity propensity_fmt.;
run;
```

- 程序运行之后，首先可得到如图 15-7 所示的列联表。
- 对上表行列变量独立性的检验结果如图 15-8 所示。



| Frequency | Table of propensity by income | | | |
|---|---|---|---|---|
| | | income(收入) | | |
| | propensity(消费倾向) | 低收入 | 中收入 | 高收入 | Total |
| | 不愿购买 | 23 | 29 | 69 | 121 |
| | 暂无打算 | 23 | 11 | 15 | 49 |
| | 愿意购买 | 68 | 65 | 37 | 170 |
| | Total | 114 | 105 | 121 | 340 |



Statistics for Table of propensity by income

| Statistic | DF | Value | Prob |
|---|---|---|---|
| Chi-Square | 4 | 43.4315 | <.0001 |
| Likelihood Ratio Chi-Square | 4 | 43.0675 | <.0001 |
| Mantel-Haenszel Chi-Square | 1 | 30.9350 | <.0001 |
| Phi Coefficient | | 0.3574 | |
| Contingency Coefficient | | 0.3366 | |
| Cramer's V | | 0.2527 | |

| Fisher's Exact Test | |
|---|---|
| Table Probability (P) | 1.0[illegible]E-13 |
| Pr <= P | 8.27[illegible]E-09 |

# 列联分析—列联表中的关联度分析

- 多种$\chi^2$统计量检验和 FISHER 精确检验的结果表明，如果给定 α=0.05 的显著性水平，拒绝原假设，即不认为行变量“购买意愿”与列变量“收入”之间是独立的，即行列变量存在关系。二者的相关程度同样可以在列联表分析中计算出来，如图 15-9 所示。
- 因本例中的行列变量均为定性变量， 为了更好衡量行列变量之间相互影响的关系， 本例已经把收入按照低、中、高收入使用码表的形式用 1、2 和 3 顺序替代，把购买意向按照购买意愿强烈程度也用 1、2 和 3 的顺序替代（如果不按照顺序进行编码，则会在关联度分析中造成混乱和困扰） 。
- 图 15-9 列示了很多种常用的相关系数，如本书第 11 章介绍过的 Pearson 相关系数、Spearman 相关系数、Kendall’s Tau-b 系数等。除此之外，还有 Gamma 系数、Stuart’s Tau-c系数等均在一定程度上衡量了行列变量之间的相关性。图 15-9 中的 Value 列表示对应系数的统计量值，ASE 列表示渐进标准误差，依据给定的理论显著性水平和 ASE 可以计算出对应系数的置信区间。 本例中， 如相关系数 Kendall’s Tau-b 的值为-0.2650， 表明 “收入” 和 “购买意愿”之间存在低度负相关性，即收入越高，可能购买意愿越不强烈。

| Statistic | Value | ASE |
|---|---|---|
| Gamma | -0.4023 | 0.0559 |
| Kendall's Tau-b | -0.2650 | 0.0459 |
| Stuart's Tau-c | -0.2517 | 0.0435 |
| Somers' D C\|R | -0.2701 | 0.0464 |
| Somers' D R\|C | -0.2522 | 0.0438 |
| Pearson Correlation | -0.3021 | 0.0503 |
| Spearman Correlation | -0.2977 | 0.0509 |
| Lambda Asymmetric C\|R | 0.1781 | 0.0495 |
| Lambda Asymmetric R\|C | 0.1882 | 0.0546 |
| Lambda Symmetric | 0.1825 | 0.0490 |
| Uncertainty Coefficient C\|R | 0.0584 | 0.0172 |
| Uncertainty Coefficient R\|C | 0.0645 | 0.0191 |
| Uncertainty Coefficient Symmetric | 0.0613 | 0.0181 |
| Sample Size = 340 | | |

# 列联分析—$\chi^2$分布的期望值准则

- 从 15.1.3 节介绍的内容中可知$\chi^2$检验是一种近似检验，依据观测值和期望值计算出来的统计量在大样本的情况下近似服从$\chi^2$分布。因此，要求在进行列联表检验过程中，样本量应当足够大，而且每个交叉分类的期望频数不能偏小，否则进行$\chi^2$检验可能会得出错误的结论。
- 进行$\chi^2$检验时，$\chi^2$分布的期望值准则主要有两条：
  - 当交叉分类为两类时，要求每一类别的期望值不少于 5；
  - 当交叉分类为两个以上类别时，期望值小于 5 的比例不应超过 20%，否则应把期望值小于 5 的类别与相邻的类别合并。
- 如表 15-5 所示的列联表中，有一个期望频数值为 4，小于 5，依据期望值准则，则不能够进行$\chi^2$检验。

**表 15-5　两个类别的列联表分布**

| 产品合格情况 | 观测值（$f^o$） | 期望值（$f^e$） |
|---|---|---|
| 合格 | 123 | 115 |
| 不合格 | 6 | 4 |

# 列联分析—$\chi^2$分布的期望值准则

- 当数据交叉分类为两个以上类别时， 期望值小于 5 的比例超过 20%时， 如表 15-6 所示。

表 15-6　两个以上类别的列联表分布

| 产品质量分类 | 观测值（$f^o$） | 期望值（$f^e$） |
|---|---|---|
| A | 123 | 115 |
| B | 120 | 132 |
| C | 78 | 87 |
| D | 23 | 45 |
| E | 8 | 4 |
| F | 7 | 3 |

- 表中一共有 6 个分类，其 20%为 6*20%=1.2，但其期望值小于 5 的分类个数为 2，超过了 20%的数目。