论列联表的几种抽样模型及相关性检验

□禹建奇

【内容摘要】本文讨论二维列联表数据的三种常见抽样模型，说明它们之间的联系，以及三种模型下独立性检验的一致性。

【关键词】列联表；抽样模型；独立性检验

【作者简介】禹建奇（1970～），男，湖南邵阳人；桂林理工大学理学院教师，博士；研究方向：数理统计

列联表分析在医学、药学、生物学、社会学、经济学等各领域有着重要的应用作用，而二维列联表则是最基础的。同一个二维列联表，其后的数据来源或者说抽样模型可以是多样的，但是大多数教材，如阿兰·阿格莱斯蒂的《分类数据分析》，均没有详细论述这个问题，本文讨论最常见的三种数据模型：泊松数据模型，多项数据模型，乘积多项数据模型。

一、抽样模型

首先我们来看一个列联表（见《非参数统计》第八章）。

例 1　某疾病有三种处理方法，其结果分“改善”和“没有改善”，数据如表 1。

表 1　某疾病处理结果

改善	没有改善	合计	
处理 A	10	12	22
处理 B	7	8	15
处理 C	6	13	19
合计	23	33	56

问：病情有没改善与处理有关吗？

通常有三种方法取得数据：一是确定各个处理的病人数：22，15，19，再作相应处理，得到表中数据。二是限定总人数为 56，随机抽选处理过的病人，记录他们的情况。三是选取该院得到处理的所有病人，记录他们的情况。

该例数据有如表 2 的一般形式。这里，每个格子的频数 n_{ij} 为随机变量，行频数总和 $n_{i.} = \sum_j n_{ij}$，列频数总和 $n_{.i} = \sum_i n_{ij}$，总频数 $n_{..} = \sum_i n_{i.} = \sum_i n_{.i}$，行因子与列因子的水平分别为 $A_1, A_2, \cdots, A_r$ 及 $B_1, B_2, \cdots, B_c$ · p_{ij}. 表示第 ij 个格子频数占总频数的理论比例（概率）·显然，$P_{ij} = E(n_{ij})/n_{..}$，这里 $E(n_{ij})$ 为 n_{ij} 的期望，而第 i 行的理论比例（概率）$p_{.i}$ 及第 j 列的理论比例（概率）$p_{.j}$ 分别为 $p_{i.} = \sum_j p_{ij}$，$p_{.j} = \sum_i p_{ij}$。

表 2

	B_1	B_2	…	B_C	总和
A_1	n_{11}	n_{12}	…	n_{1C}	$n_{1.}$
A_r	n_{r1}	n_{r2}	…	n_{rC}	$n_{r.}$
总和	$n_{.1}$	$n_{.2}$	…	$n_{.C}$	$n_{..}$

（一）乘积多项分布模型。对于 1），要检验分布的齐性。齐性就是指对任一行，固定行和的条件概率相等。记固定第 i 行后第 j 列的条件概率为 $p_{j/i} = \frac{p_{ij}}{P_i}$，则原假设为：

$H_0: p_{j/i} = p_{j/i^*}, \forall j, i \neq i^*$.

对立假设为 H_1：“零假设中的等式至少有一个不成立”。零假设下，$p_{j/i}$ 与 t 无关，记该概率为 p^j，则 $p \cdot j = \sum_i p_{i.} p_{j/i} = \sum_i p_{i.} p^j = p^j = p^j \sum_i p_{i.} = p^j$，零假设即为：

$H_0: p_{j/i} = p_{.j}, \forall j, i$

具体到例 1，则原假设为“改善的比例与处理无关”。容易看出，“改善的比例与处理无关”意味着“没有改善的比例

行满频带信号加载（50～1000MHz 加满信号）使用，并且可以在 1，550nm 上进行放大以便于将信号馈送给庞大数量的光点。

如图 3 所示，首先通过前置接收机（RX1000 模块）将前端总公司机房所发射的 1，550nm 直播电视信号进行一次光电转换，然后通过混合设备，将广播站机房 IPQAM 输出的本地 VOD、省网 VOD 信号进行电混合，规避了光插播中所存在的调试复杂、插播频点限制等一系列问题。之后通过直调发射模块（DM2000 模块）进行光信号传输，这种传输结构与“1，550nm 城区骨干 + 1，310nm 光分配 + RF 放大/分配”的传统结构将为相似，但是它区别于 1，310nm 发射机的一大特点就是可以像外调 1，550nm 发射机一样进行 EDFA 放大，从而衍生了覆盖面，提高了灵活性，但它的价格又远远低于外调 1，550nm 发射机。

目前吴江有限已经通过此技术在全区 10 个分机房，部署了 90 组互动分组，覆盖双向互动用户 9 万多户，此方案在保障了信号质量的同时，大大降低了部署成本。

【参考文献】

[1] 谷德露. 外调制光发射机关键技术研究［D］. 电子科技大学，2009

[2] 王飚，冯金林. 主路信号 OMI 与光差之间关系的分析与研究［J］. 有线电视技术，2015

[3] 卢剑平，刘玉玲. 1550nm 直调光发送机及其系统应用［J］. 有线电视技术，2015

也与处理无关”。一般而言，为检验数据的齐性，我们通常是预先确定每行的样本数目（$n_{i.}$），再进行抽样得到样本，对这些样本作相应处理，然后记录不同处理下的相应频数。

考虑原假设成立，则 $E_{ij}=E(n_{ij})$ 应该等于 $n_i\cdot p_{.j}$，但 $p_{.j}$ 未知，零假设下，可以用其估计 $\hat{p}\cdot j=\frac{n_{\cdot j}}{n..}$ 代替。这样期望值的估计值为：

$\hat{E}_{ij}=n_i.\hat{p}\cdot j=n_i\cdot n_{.j}/n..$

而第个格子的实际频数为，故 $PearsonX^2$ 统计量为：

$$Q=\sum_i\sum_j\frac{(nij-E_{ij})^2}{E_{ij}}=\sum_i\sum_j\frac{\left(n_{ij}-\frac{n_i\cdot n_{.j}}{n..}\right)^2}{n_i.n_{.j}/n..}$$

它在样本量较大时（$E_{ij}\geqslant5,\forall i,j$）近似地服从自由度为的分布。

推广来看，对一般的列联表，预先确定每一行的行和，对应于每一行，独立进行一次抽样，经过不同处理后，根据处理结果，记录样本落在各个格子的频数，因而，每一行为多项分布，且不同行之间独立，由于，

$$P(n_{ij}=o_{ij},j=1.2.\cdots,c)=\frac{n_i.!}{n_{i1}!\ n_{i2}!\ \cdots n_{ic}!}p_{1/i}{}^{n_{i1}}\cdots p_{c1i}{}^{n_{ic}}$$

其中，o_{ij} 为 n_{ij} 的观测值，$p_{i1j},\cdots.p_{i1c}$ 分别为固定行和下的条件概率。

从而由独立性有：

$$P(n_{ij}=o_{ij},i=1.2.\cdots,r)=\frac{\prod_{i=1}^{r}ni.!}{\prod_{i-1}^{r}\prod_{j-1}^{c}n_{ij}!}\prod_{i-1}^{r}\prod_{i-1}^{c}p_{j/i}{}^{n}_{ij}$$

即该列联表为乘积多项分布数据模型。

（二）整体多项分布数据模型。对应于抽样方案2），我们感兴趣的是行变量和列变量的独立性（INDEPENDENCE）。此时，对应的行列两个概率之积 $p_i.p_{.i}$ 就是第 ij 个格子的理论概率 p_{ij}，故原假设为：

$H_o:p_{ij}=p_{ij}=p_i.p_{.j}\ \forall_i,j$

而零假设下，其估计值为：

$$\hat{P}_{ij}=\hat{P}_i.\hat{P}_{.j}=\frac{n_i.}{n..}\frac{n_{.j}}{n..},$$

而第 ij 个格子的期望值估计为：

$\hat{E}_{ij}=n..\hat{P}_{ij}=n_i.n_{.j}/n..$

容易看到，该过程与方法1）一样，故检验统计量 Q 以及其分布也相同，即为 χ^2 分布。不同于齐性问题，独立性的问题的数据抽样，并不事先固定行和，而是固定总和，然后选取总和数目的样本，经过不同处理后，记录各个格子的频数。

这种抽样方法，预先确定整个列联表的总频数 $n..$，再进行抽样，处理后记录 $n..$ 这个个体落在各个格子的频数，此时，整个数据为一多项分布：

$$p(n_{ij}=o_{ij},j=1.2.\cdots,r)=\frac{\prod_{i=1}^{r}n..!}{\prod_{i=1}^{r}\prod_{j=1}^{c}n_{ij}!}\prod_{i=1}^{r}\prod_{j=1}^{c}p_{ij}{}^{n_{ij}}$$

称整体多项分布数据模型。

（三）泊松分布模型。对于方法3），该模型没有限定样本总数 $n..$，各个格子的频数 n_{ij} 均为独立随机变量，且服从泊松分布 $p(n_{ij}=o_{ij})=\frac{\lambda_{ij}^{o_{ij}}}{o_{ij}!}e\lambda_{ij}$，由泊松分布的性质，各个格子的频数期望值 E_{ij} 就是 λ_{ij}，其估计值 $\hat{E}=\hat{\lambda}=n_{ij}$，故而 $\hat{p}_i.=\frac{n_i.}{n..}$，$\hat{p}_{.j}=\frac{n_{.j}}{n..}$。

我们要检验的问题亦是行和列变量的独立性，即零假设为：

$H_0:p_{ij}=pi\cdot p\cdot j,\forall i,j$

在零假设下，$\hat{p}_{ij}$ 的估计值为 $\hat{p}\,ij=\hat{p}_i.\hat{p}_{.j}=\frac{n_i.\ n_{.j}}{n..n..}$

可见，这和前面两模型一样，由此可以得到一样的统计量 Q，当然也有同样的渐近 χ^2 分布。这类关于独立性的问题的数据获取，通常是观测特定的个体总频数，然后记录这些个体分配到各个格子的频数。它不事先固定样本总频数，因而，样本总频数是随机的。

这种抽样方法，其总频数 $n..$ 也是随机的，记录这 $n..$ 个个体经过处理后落在各个格子的频数，所以，整个数据为乘积泊松分布，满足

$$p(n_{ij}=o_{ij},j=1.2.\cdots,r)=\prod_{i=1}^{r}\prod_{j=1}^{c}\frac{\lambda o_{ij}}{o!\ \ ij!}e^{\lambda ij}$$

该数据模型称为列联表的泊松分布数据模型。

二、三种模型的联系

《列联表的两种抽样模型以及齐性和独立性的检验问题》一文详细阐述了前两种抽样模型的联系，以下说明第三种模型与它们的联系。

定理：若考虑固定总频数前提下的条件概率，则泊松分布数据模型成为整体多项分布数据模型。

证明：泊松分布模型即：

$$P(n_{ij}=o_{ij},j=1,2,\cdots,r)=\prod_{i=1}^{r}\prod_{j=1}^{c}\frac{\lambda o_{ij}}{o_{ij}!}e\lambda_{ij}$$

而 $n_{*}..=\sum n_{ij}$，可见，$n_{*}.$ 亦为泊松分布

$p(n..=o..)=\frac{\lambda^{oi..}}{O..!}e^{\lambda..}$，这里 $\lambda..\sum\lambda_{ij}$

从而，固定总频数的条件概率为：

$$p(n_{ij}=o_{ij},j=1.2.\cdots,c,i=1.2.\cdots,r|n..=o..)$$
$$=\frac{p(n_{ij}=o_{ij},j=1.2\cdots,r)}{p(n..=o..)}$$
$$=\prod_{i=1}^{r}\prod_{j=1}^{c}!\ \frac{\lambda_{ij}^{o_{ij}}}{o_{ij}!}e^{\lambda ij}/\lambda o_i....e^{\lambda..}$$
$$=\frac{o\cdot\cdot!}{\prod_{i=1}^{r}\prod_{j=1}^{c}o_{ij}!}\prod_{i=1}^{r}\prod_{j=1}^{c}p_{ij}{}^{o_{ij}}$$

这里，$p_{ij}=\lambda_{ij}/\lambda..$

即一整体多项分布模型。

三、结语

本文以二维表为例，说明了同一列联表的数据，其来源可能有三种：乘积多项分布模型，整体多项分布模型，泊松分布模型。而整体多项分布模型在限定行和的条件下，就是乘积多项分布模型；泊松分布模型在限定列链表的总频数下的条件分布模型则是整体多项分布模型，而且变量独立性检验，不论在何种模型下进行，都会得到同样的结论。

【参考文献】

[1] 吴喜之，赵博娟. 非参数统计［M］. 北京：中国统计出版社，2013

[2] 阿兰·阿格莱斯蒂. 分类数据分析［M］. 重庆：重庆大学出版社，2012

[3] 禹建奇. 列联表的两种抽样模型以及齐性和独立性的检验问题［J］. 教育教学论坛，2015