中国社会科学院创新工程学术出版资助项目

赵 平 著

定性数据的统计分析

STATISTICAL ANALYSIS OF QUALITATIVE DATA

社会科学文献出版社
SOCIAL SCIENCES ACADEMIC PRESS (CHINA)

中国社会科学院创新工程学术出版资助项目

定性数据的统计分析

赵 平 著

社会科学文献出版社
SOCIAL SCIENCES ACADEMIC PRESS (CHINA)

图书在版编目(CIP)数据

定性数据的统计分析/赵平著. —北京：社会科学文献出版社，2014.4

ISBN 978-7-5097-5134-3

Ⅰ.①定… Ⅱ.①赵… Ⅲ.①定性分析-统计分析 Ⅳ.①C813

中国版本图书馆CIP数据核字（2013）第234933号

定性数据的统计分析

著　　者／赵　平

出 版 人／谢寿光
出 版 者／社会科学文献出版社
地　　址／北京市西城区北三环中路甲29号院3号楼华龙大厦
邮政编码／100029

责任部门／社会政法分社（010）59367156　　责任编辑／杨桂凤
电子信箱／shekebu@ssap.cn　　责任校对／姜夕芬
项目统筹／杨桂凤　　责任印制／岳　阳
经　　销／社会科学文献出版社市场营销中心（010）59367081　59367089
读者服务／读者服务中心（010）59367028

印　　装／三河市尚艺印装有限公司
开　　本／787mm×1092mm　1/16　　印　　张／10.75
版　　次／2014年4月第1版　　字　　数／160千字
印　　次／2014年4月第1次印刷
书　　号／ISBN 978-7-5097-5134-3
定　　价／45.00元

STATISTICAL ANALYSIS OF QUALITATIVE DATA

序

定性变量包含定类变量和定序变量。顾名思义，定类变量包含不同的类别，例如性别、党派；而定序变量是连续的，如情感、意识等。但是定类变量和定序变量不能量化。例如，比较两个家庭的夫妻关系，我们只能说一个家庭的夫妻关系比另一个家庭的夫妻关系好，而不能说“好多少”。在社会科学的一些领域，定类变量和定序变量占有很大的比重，一向是统计学要处理的。

20 世纪 40 年代，数理统计学家和应用统计学家经过不懈的努力，撰写了大量的论文，至 1980 年完善了数学思想，创立了很多关于定序变量和定类变量的统计与检验方法。例如，通过定序变量的层次赋值方法，学者应用尼尔森相关系数，量度列联表的定序变量的关联，并将适用于定序变量和定类变量的对数概率比回归模型和对数线性回归模型引入社会统计学，取代了有问题的概率线性回归模型。时至今日，30 多年过去了，当时针对列联表的量度和检测方法以及对数概率比和对数线性回归已经成为社会统计学的基础，不掌握这方面的知识，不仅不能正确地分析和量度定性变量，而且会阻碍我们学习更先进的统计方法。

本书的重点是介绍定类和定序变量的统计方法。一般来讲，当一个变

量既可以作为定类变量也可以作为定序变量处理时，应该选择后者，因为应用定序变量的连续性质及其层次赋值的方法，具有很多特殊的优点：

- 对于同一变量，定序方法比定类方法更灵敏，可以检测出定类量度否定的变量之间的关联；
- 定序数据和定距数据具有相同的连续性，因此适用于定距数据的统计量，如相关、均值和斜率也适用于定序数据；
- 有很多简单的模型适用于定序变量，并且很容易对参数进行解释，这是定类变量所不及的；
- 在量度含有相同数目变量的模型的拟合优度方面，定序变量的参数往往少于定类变量的参数，所以更简单，易于计算。

从另一方面讲，在社会调查数据中，既有定序数据，也有定类数据。例如在列联表中，应变量是定序变量，解释变量是定类变量。更有甚者，研究人员为了方便或其他目的会将定序变量改成定类变量。因此，在介绍处理定序数据的统计方法时，本书也要涉及处理定类数据的统计方法。

本书的宗旨是尽可能地应用简单的数学知识解释相关的统计学原理，以利于从事社会科学研究的读者掌握并应用定性数据的量度方法。尽管如此，书中仍有大量的数学内容。实际上，这些数学内容并不高深，只要具有高中数学水平就能看懂。在阅读过程中，读者要注意以下几点：首先，数学符号本身不是数学，但是在学习的过程中，常常不是数学知识而是纷杂的符号及其角标令读者望而却步。本书中也有很多数学符号，这是无法回避的。读者只有在理解和记住这些符号后，才能学习那些并不深奥的数学知识。其次，概率比（θ）不仅是对二维列联表和多维列联表的变量关联的量度，而且是对对数概率比模型和对数线性回归模型的关联参数的量度。可以说，概率比是贯穿全书的最重要的概念和量度。最后，列联表和模型对定性数据进行统计量度的目的在于测量定序数据的“线性趋势”，例如列联表的累积概率和概率比描述的概率分布就是为此服务的。在学习对数概率比和对数线性模型时，读者要具备普通线

性回归和方差分析的知识，并结合列联表的分析方法，才能理解模型的设定及其参数的意义。

在撰写本书的过程中，陈婴婴、夏传玲、赵峰和张亮杰同志为本书提供了丰富的社会调查数据。此外，社会科学文献出版社的同志为本书的出版付出了大量精力。对以上诸位同志，我要表示衷心的感谢。

书中如有不当之处，敬请读者批评指正。

目录
CONTENTS

第 1 章

二维列联表

列联表也称交叉分类表，具有直观性和实用性，是分析统计数据，特别是定类和定序数据的重要工具，也是学习和掌握对数概率模型的助手。多维列联表虽然复杂，但是，通过对变量进行层层控制，最终还是要落实到对二元变量的分析。所以，最简单的二维列联表是学习复杂表格的基础。在学习各种统计技术之前，我们先要掌握有关列联表的术语、符号和概念，以及一些基本的统计量度。

1 列联表的频次和概率结构

1.1 观测频次分布

设 X 和 Y 是定类或定序变量，分别有 I 个和 J 个类别或层次，列联表的行表示 X 的类别或层次，列表示 Y 的类别或层次。这样，表格具有 IJ 个分类组合（单元格），可以写作 $I \times J$ 列联表。

列联表的观测频次：单元格频次 n_{ij}，边际频次 n_{i+} 和 n_{+j}，样本量 n。

以上符号的角标 i 表示行的类别或层次，j 表示列的类别或层次。行和列的总数被称为边际频次，各单元格频次的总和就是样本量 n。表 1 - 1 是观测频次分布的一个例子。

表 1-1 是否接受高等教育和性别的交互分类

性 别	是否接受高等教育		
	是	否	总数
男	471	1793	2264
女	410	2078	2488
总 数	881	3871	4752

资料来源：中国社会科学院社会学研究所的社会调查。

$$n_{11} = 471 \quad n_{12} = 1793 \quad n_{1+} = 2264$$
$$n_{21} = 410 \quad n_{22} = 2078 \quad n_{2+} = 2488$$
$$n_{+1} = 881 \quad n_{+2} = 3871 \quad n = 4752$$

1.2 联合概率和条件概率分布

列联表的数据除了可以用频数表示外，还可用概率来表示。当调查数据是总体的时候，例如全国人口普查，列联表的概率分布用 π 表示。概率有不同的类型，例如表 1-2 中有联合概率、条件概率和边际概率。在 $I \times J$ 列联表中有 IJ 个概率，用 π_{ij} 表示行为 I、列为 J 的单元格概率，IJ 个概率用 $\{\pi_{ij}\}$ 表示。

表 1-2 联合、条件和边际概率分布

行	列		总数
	1	2	
1	π_{11} ($\pi_{1(1)}$)	π_{12} ($\pi_{2(1)}$)	π_{1+} (1.0)
2	π_{21} ($\pi_{1(2)}$)	π_{22} ($\pi_{2(2)}$)	π_{2+} (1.0)
总数	π_{+1}	π_{+2}	

资料来源：中国社会科学院社会学研究所。

当变量 X 和 Y 是对称的（也就是变量没有应变量和解释变量之分）时，概率分布 $\{\pi_{ij}\}$ 被称为 X 和 Y 的联合分布。观测频次和联合概率的关系为：

$$n_{ij}/n = \pi_{ij} \quad n_{i+}/n = \pi_{i+} \quad n_{+j}/n = \pi_{+j}$$

行或列的边际概率分布是该行或该列的联合概率分布之和。$\{\pi_{i+}\}$ 表示行（变量 X）的边际概率分布，即每一行的单元格概率之和；$\{\pi_{+j}\}$ 表示列（变量 Y）的边际概率分布，即每一列的单元格概率之和：

$$\pi_{i+} = \sum_j \pi_{ij} \quad 和 \quad \pi_{+j} = \sum_i \pi_{ij}$$

行和列的边际概率分布之和有如下关系：

$$\sum\nolimits_i \pi_{i+} = \sum\nolimits_j \pi_{+j} = \sum\nolimits_i \sum\nolimits_j \pi_{ij} = 1.0$$

即：联合概率分布的行边际概率分布之和等于列边际概率分布之和，等于列联表所有单元格概率之和。

当列联表的变量是不对称的时，Y 是应变量，X 是解释变量。当 X 被确定后（不再是随机的），X 和 Y 的联合分布不再适用。在 X 确定的类别或层次，例如 X 的 i 类别，Y 有条件概率分布，其概率表示为 $\pi_{j(i)}$，$j=1$，…，J。条件概率和观测频次的关系为：

$$n_{ij}/n_{i+} = \pi_{j(i)}, \quad 边际概率分布为 \sum\nolimits_j \pi_{j(i)} = 1, \quad j = 1,\cdots,J$$

如果 X 是应变量，Y 是解释变量，则条件概率和观测频次的关系为：

$$n_{ij}/n_{+j} = \pi_{i(j)}, \quad 边际概率分布为 \sum\nolimits_i \pi_{i(j)} = 1, \quad i = 1,\cdots,I$$

当两个变量是对称的时，可以应用联合概率分布、Y 的条件概率分布（X 确定）〔或 X 的条件概率分布（Y 确定）〕描述它们的关系。当行变量是解释变量、列是应变量时，单元格联合概率除以同一行的边际联合概率就是该单元格的条件概率，这一关系可以由观测频次导出：

$$n_{ij}/n_{i+} = (n_{ij}/n) \div (n_{i+}/n) \quad 等于$$
$$\pi_{j/i} = \pi_{ij}/\pi_{i+}$$

如果解释变量和应变量位置互换，单元格联合概率除以同一列的边际联合概率就是该单元格的条件概率：

$$\pi_{i/j} = \pi_{ij}/\pi_{+j}$$

在一些研究中需要比较和解释定序变量各个层次的条件概率分布。列变量必须是定序应变量 Y，层次用 j 表示；行变量可以是定类或者是定序解释变量 X，类别或层次用 i 表示。

1.3 变量独立的概率表述

当两个变量在统计上独立时，列联表各单元格的联合概率等于对应的边际联合概率之积：

$$\pi_{ij} = \pi_{i+}\pi_{+j} \qquad i = 1,\cdots,I \text{ 和 } j = 1,\cdots,J$$

根据上式，又可推导出下式：

$$\pi_{j/i} = \pi_{ij}/\pi_{i+} = (\pi_{i+}\pi_{+j})/\pi_{i+} = \pi_{+j} \qquad i = 1,\cdots,I$$

公式说明，当变量 X 和 Y 独立时，Y 的第 j 列各单元格的条件概率等于该列的边际联合概率。换言之，当 X 为解释变量、Y 为应变量时，如果对于所有的 $i=1,\cdots, I$,条件概率满足 $\{\pi_{j/1} = \cdots = \pi_{j/I} = \pi_{+j}\}$，$X$ 和 Y 相互独立。

对于从总体随机抽取的样本，我们用 p 替代 π，$\{p_{ij}\}$表示列联表的样本联合概率，$\{n_{ij}\}$表示单元格频次，$n = \sum_i \sum_j n_{ij}$表示样本总量。因此，我们有：

$$p_{ij} = n_{ij}/n$$

和

$$p_{j(i)} = p_{ij}/p_{i+} = n_{ij}/n_{i+} \quad (n_{i+} = np_{i+} = \sum\nolimits_j n_{ij})$$

我们要利用随机样本的概率推论总体的数值，该值是概率的值域。

1.4 概率之差

在 $I\times2$ 列联表中，$\pi_{1(i)}$（$i=1, 2, \cdots, I$）是行为 i、列为 1 的单元格的条件概率。因为边际条件概率是 1，所以（$\pi_{1(i)}$，$\pi_{2(i)}$）=（$\pi_{1(i)}$，$1-\pi_{1(i)}$）被称为二项应变量的条件概率分布。同列不同行的概率差表示为

$\pi_{1(i+1)} - \pi_{1(i)}$，相应地，第 2 列的概率差与第一列的概率差有如下关系：

$$\pi_{2(i+1)} - \pi_{2(i)} = (1 - \pi_{1(i+1)}) - (1 - \pi_{1(i)}) = \pi_{1(i)} - \pi_{1(i+1)}$$

概率差的值域在 -1.0 和 1.0 之间。当行 i 和 $i+1$ 的条件概率分布相同时，概率差为 0。如果各行之间均有 $\pi_{1(i)} - \pi_{1(i+1)} = 0$，应变量 Y 在统计上独立于行的分类或分层，则两个变量独立。

在 $I \times J$ 列联表中，如果 $(I-1)$ $(j-1)$ 个概率差 $\pi_{j(i)} - \pi_{j(I)} = 0$，$i = 1, 2, \cdots, I-1$，$j = 1, 2, \cdots, J-1$，则变量相互独立。

如果将两个变量视为对称的，其概率有联合分布，则行为 i 和 $i+1$ 的条件概率差等于 i 和 $i+1$ 各行的单元格联合概率除以边际概率之差：

$$\pi_{1/(i+1)} - \pi_{1/(i)} = \pi_{(i+1)1}/\pi_{(i+1)+} - \pi_{i1}/\pi_{i+}$$

对于 2×2 列联表，行与行之间和列与列之间的联合概率之差分别为：

$$\pi_{11}/\pi_{1+} - \pi_{21}/\pi_{2+};\ \pi_{11}/\pi_{+1} - \pi_{12}/\pi_{+2}$$

以上两者的值不一定相等。

2　变量独立检验

2.1　卡方检验

卡方检验源于生物学的变异预测，生物学家依据经验提出豌豆种子的变异比例（概率），然后用实验结果验证假设（H_0）。这种方法经英国数学家卡尔·皮尔森发展为卡方检验。

对于二维列联表、多维列联表或定类单变量的一组数据而言，变量独立的卡方检验的 H_0 为：单元格的概率等于一确定的值 $\{\pi_{ij}\}$，这个确定的值等于该单元格的行和列的边际概率之积：

$$\pi_{ij} = \pi_{i+}\pi_{+j}$$

对于总体样本量为 n、单元格观测频次为 $\{n_{ij}\}$ 的样本，期望频次为

$\{m_{ij}=n\pi_{ij}\}$。当 H_0 为真时，期望值等于观测值 $\{m_{ij}=n_{ij}\}$。

假设投掷一枚均匀的硬币，正面朝上的概率是 π，背面朝上的概率就是 $1-\pi$。H_0：硬币正面朝上的概率等于背面朝上的概率，即：$\pi=1-\pi=0.5$。正面朝上的期望频次等于背面朝上的期望频次，同为 $m=n\pi=n/2$。如果 H_0 为真，则 n 次投掷就会得到正面朝上和背面朝上各一半的结果。以上是自然科学的独立检验。

具体到社会统计的二维列联表，我们首先应用列联表的边际频次，算出单元格的期望频次，然后用每一单元格的观测频次 n_{ij} 与该单元格的期望频次 m_{ij} 比较。如果变量独立，即 H_0 为真，则 n_{ij} 应该等于或接近于 m_{ij}；如果 $\{n_{ij}-m_{ij}\}$ 较大，则 H_0 就可能不成立，两个变量会相关。

2.2 皮尔森卡方统计量和卡方分布

检验 H_0 的皮尔森卡方统计量公式为：

$$\chi^2=\sum\frac{(n_{ij}-m_{ij})^2}{m_{ij}} \tag{1-1}$$

当 $n_{ij}=m_{ij}$ 时，χ^2 的值最小，为 0。在样本量确定后，$n_{ij}-m_{ij}$ 越大，χ^2 越大，从而否定 H_0 的可能性越大。

因为 χ^2 越大，否定 H_0 的概率越大，所以检验 H_0 的 p 值是 $\chi^2\geqslant\sum\frac{(n_{ij}-m_{ij})^2}{m_{ij}}$ 的概率。统计量 χ^2 近似于大样本的卡方分布。至于什么是“大样本”，尚没有明确定义。一种观点认为，80% 的 $\{n_{ij}\}\geqslant 5$ 就可以了。p 值是观测值 χ^2 的右尾概率。

卡方分布因自由度（df）的不同而形状各异，其均值等于 df，标准差等于 $\sqrt{2df}$。df 等于备择假设和 H_0 假设的参数数量之差。随着 df 增大，卡方分布趋于钟形，但仍然向右偏斜。因为 $\chi^2\geqslant 0$，所以卡方分布只有非负数值（0 和正数）。图 1-1 是 df 分别等于 1、5、10、20 的卡方密度分布。

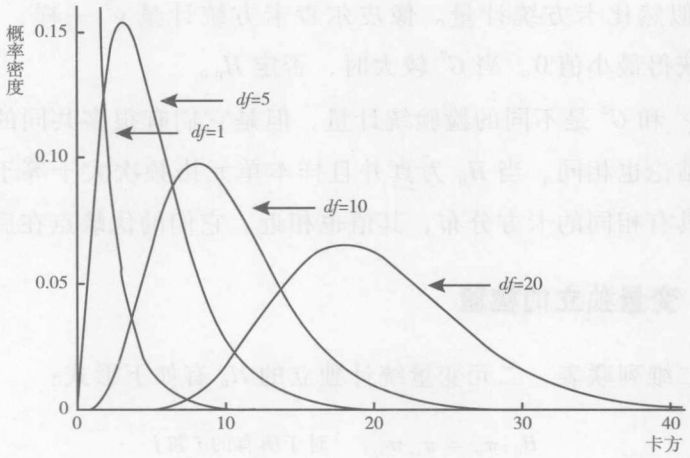

图 1-1　*df* 分别等于 1、5、10、20 的卡方密度分布

资料来源：Agresti，1996。

2.3　似然比卡方统计量

另一种检验 H_0 的统计量是显著性检验的似然比卡方统计量 G^2。显著性检验决定了在 H_0 为真的情况下，使似然函数最大化的参数值，同时也决定了在 H_0 为真的一般情况下，使似然函数最大化的参数值。检验 H_0 的最大似然比公式为：

$$\Lambda = \frac{\text{参数满足 } H_0 \text{ 的最大似然}}{\text{参数不受约束的最大似然}}$$

Λ 的最大值为 1。如果在参数不受 H_0 的约束下最大似然的值很大，则最大似然比 Λ 会大大小于 1，可以肯定地否定 H_0。

似然比检验的统计量等于 $-2\log(\Lambda)$，该值是非负数。Λ 的值越小，$-2\log(\Lambda)$ 的值越大。将统计量取对数的目的在于使统计量趋近卡方抽样分布。对于二维列联表，检验统计量可以简化为：

$$G^2 = 2\sum n_{ij}\log\left(\frac{n_{ij}}{m_{ij}}\right) \tag{1-2}$$

G^2 被称为似然比卡方统计量。像皮尔森卡方统计量 χ^2 一样，G^2 在所有 $n_{ij}=m_{ij}$ 时获得最小值 0。当 G^2 较大时，否定 H_0。

虽然 χ^2 和 G^2 是不同的检验统计量，但是它们有很多共同的特性，往往得出的结论也相同。当 H_0 为真并且样本单元格频次大于等于 5 时，两个统计量具有相同的卡方分布，其值也相近。它们的优缺点在后面讨论。

2.4 变量独立的检验

对于二维列联表，二元变量统计独立的 H_0 有如下形式：

$$H_0:\pi_{ij}=\pi_{i+}\pi_{+j},\quad \text{对于所有的 } i \text{ 和 } j$$

我们一直用 π 表示总体的各种概率，这里的 π_{i+} 和 π_{+j} 是总体的边际联合概率。卡方检验要使用频次而非概率，所以要把概率转换为频次：$m_{ij}=n\pi_{ij}=n\pi_{i+}\pi_{+j}$，$m_{ij}$ 是假设变量独立的期望频次。

因为随机抽样具有省力、省时和节省费用的优点，所以成为社会调查常用的方法。在总体样本很少的情况下，其调查数据 π_{ij}、π_{i+} 和 π_{+j} 也很少。我们可以用随机抽样样本的概率替代总体的概率：

$$p_{ij}=p_{i+}p_{+j}$$

样本的期望频次 $\{m_{ij}\}$ 为：

$$m_{ij}=np_{i+}p_{+j}=n\frac{n_{i+}}{n}\frac{n_{+j}}{n}=\frac{n_{i+}n_{+j}}{n}$$

这是在零假设即变量相互独立的条件下，单元格的期望频次。

对于 $I\times J$ 列联表的独立性检验，皮尔森和似然比卡方统计量等于：

$$\chi^2=\sum\frac{(n_{ij}-m_{ij})^2}{m_{ij}};\quad G^2=2\sum n_{ij}\log\left(\frac{n_{ij}}{m_{ij}}\right)$$

以上两个统计量的自由度为 $df=(I-1)(J-1)$。根据 H_0，单元格的期望概率是由边际联合概率 $\{\pi_{i+}\}$ 和 $\{\pi_{+j}\}$ 确定的，而行的边际联合概率之和与列的边际联合概率之和分别为 1，因此最后一个边际联合概率分别