定性资料的统计分析

张尧庭等编著

广西师范大学出版社

定性资料的统计分析

张尧庭等 编著

广西师范大学出版社

(桂)新登字 04 号

定性资料的统计分析

张尧庭等　编著

☆

广西师范大学出版社出版

(广西桂林市育才路 3 号)

广西新华书店发行

湖南省地质测绘印刷厂印刷

*

开本 850×1168　1/32 印张 6.75 字数 179 千字

1991 年 11 月第 1 版　1991 年 11 月第 1 次印刷

印数:0001—2500

ISBN 7-5633-1206-4/G・992

定价:(平)2.80 元　(精)5.50 元

内 容 简 介

本书叙述定性资料的统计分析，既有理论探讨，又有实际应用，是国内第一本专述这一类方法的专著。全书共四章，第一章介绍定性资料的各种不同的模型；第二、三章分别系统地论述对数线性模型和 logistic 回归的理论、方法和应用；第四章专门论述作者近几年在这方面的研究成果，有些是首次发表的，把多元分析的方法引入定性资料的研究处理中去，获得一些新的方法。每章之后附有参考文献、练习题和附录。

读者对象是统计的理论工作者和实际工作者，大学统计专业的教师、研究生和高年级学生。

序　言

定性资料的统计分析，国内还没有一本专述这一类方法的书。可是随着社会科学内各种研究工作的开展，需要进行定量分析时，定性资料的统计分析方法就显得十分重要。在一些专业杂志上也有介绍这些分析方法的文章，但都着重于方法的介绍和应用的效果，几乎都不谈它们的依据和有关的理论分析。本书是想弥补这一方面的缺陷，提供一本既有理论探讨，又有实际应用的教材。对有兴趣搞理论研究的，可以从中发现不少值得深入研究的问题，可以感受到实际工作会提出丰富的材料，供理论研究的人去加工。对于搞实际工作的，从本书可以了解到处理定性资料的几种比较有代表性的、成熟的方法，它们的理论依据和适用的范围，也可以知道一点新的动向。希望本书在以上这两方面能满足读者的要求，是否满足还请读者批评和建议，以便有机会再版时改进。

全书共四章，第一章介绍了定性资料的各种不同的模型，它们的基本出发点是不将它数量化，就这一类资料本身就可以进行分析。第二、三章分别系统论述了对数线性模型和 logistic 回归的理论、方法和应用。第四章专门论述编者近几年这方面的研究成果，有些是首次发表的，把多元分析的方法引入定性资料的研究处理中去，获得一些新的方法。凡是涉及数学、理论较多的章节，我们都赋以 * 号，搞实际工作的人可以跳过不看。较长推理的数学论证放入每一章的附录。我们以章为独立的单位，公式（1.6）指 §1 的第六个标记的公式，引用另一章的公式和例题时，都说明是第几章的。全书假定读者已学过一般的概率统计课，是适宜高年级大学生或研究生的参考材料。对搞实际工作的读者，本书可

以分章处理，只看第一、二章也行，只看第一、三章也行，不必从头到尾。

这本书断断续续写了好几年，最初是想根据夏立显、安希忠同志翻译的 Agresti，A. 著的书《有序分类资料的分析》加以改写，补充一些国内的实例，后来感到这一内容不能完全代表定性资料分析的方法，logistic 的模型在国内医学界正在流行，需要从理论上给以阐明，所以又重新组织。后来自己和研究生胡飞芳同志在处理定性资料上有一些新的想法，于是又拖了一些时间，重新再组织材料编写。这样就拖了近二年，给出版社带来很大的困难。

全书是由我执笔写的，编者应该是夏立显、安希忠、奚李群、胡飞芳和我五个人，因为材料都是他们提供的。

必须感谢广西师范大学出版社和王炜炘同志，没有出版社的支持和编辑同志的辛勤劳动，本书不可能这么快和读者见面。

张尧庭

1991 年 1 月 31 日

目　录

第一章 绪 论

§1 引 言

定性资料的统计分析，随着统计方法应用范围的扩展和深入，日益受到重视。首先我们对定性资料的概念作一较为详细的说明。

数理统计中经常遇到的资料可以分为以下四类：

(1) 计量的。例如人的身高、体重、血压、……，气象上的温度、相对湿度、……等等。这一类资料的特点是：原则上它的取值可以是在某一区间内的任一实数，通常称这类资料是连续的，或考察的指标是连续的。它的统计分析是与具有密度的连续随机变量的分布有关，在一般的统计教科书中，这是主要的、甚至是全部的论述对象。

(2) 计数的。例如一个平方米内某一种害虫的个数、一个居民区内拥有的电视机的台数、一个单位内职工的总人数、……等等。这一类资料的特点是：它们的取值范围是整数，大部分还只在非负整数范围内取值，通常称这一类资料是计数的，或考察的指标是计数的。它的统计分析是与离散的随机变量的分布有关，在一般的统计教科书中，这一内容涉及不多，有些教材几乎不讨论。

这两类资料原则上是可以分清的，但实际上有时也困难。例如一个人的年龄，按理可以认为是连续的，然而实际上只能按年或月或日计算，是计数的。从这里可以看到，有时为了方便，连续的资料是可以离散化的。

(3) 有序的。有些资料，既不能计量，也不能计数。例如这一块布的颜色比那一块深，但无法量化；又如评定毛料的手感程度，感觉这一种比那一种丰满；评定酒或茶的好坏时，只能评出一个顺序，而无法量化。这一类指标和资料，我们称它为有序的。

要注意的是，人们为了标记方便，往往把有序的量用等级 1，2，3，……等表示，比如砂眼的程度用“＋”号、“＋＋”号等表示，这里的 1，2 计算“＋”的个数，实际上是反映患病的深浅程度，已不是计数的意义了，因而这两类资料不能混淆，在统计分析方法上也是不同的。在一般的统计教科书中，几乎不讲有序资料的统计分析方法，偶而也有涉及一些的。

(4) 名义的。有些资料不是计量的、计数的，也不是有序的，它仅仅是一个名义值。例如给各种不同的颜色赋以代号，给不同的书籍赋以代码，……等等，这些赋值只起一个名义的作用，它的值的顺序和大小并无统计意义。这一类资料，我们称它为名义的。

以上这四类指标或四类资料，也可以粗分为两类：定量的(计量的和计数的)、定性的(有序的和名义的)。本书主要的内容是对名义资料的统计分析，然而这一方法也可以用来处理有序资料，有时例子中也反映这一内容。

实际问题中，往往是定量和定性的资料同时出现的，所以把定性资料数量化以后，就可以全部作为定量的资料来统一处理，这一类的方法已有专门的书籍介绍。然而定性资料的数量化方法可以很不相同，这样所得的结论是否会与数量化的方法有关，就使人产生疑虑，因此，寻求一种与数量化不相干的分析方法就显得十分重要，这一课题这些年来，日益受到统计家的关注，论文和专著也出现了不少，本书着重介绍这一方面的理论和方法。

如果我们把实际问题的范围略加一些限制，问题的分类就比较清楚。通常都是从资料来分析变量之间的关系，例如探讨两个

量之间是否独立？不独立时有什么形式的函数关系？如何去进一步估计函数的形式和函数中的参数？……等等。为了便于说明，我们把一部分变量称为因变量，另一部分变量称为自变量，于是按变量是定性或定量的情况来分类，就可以列出如下的表：

因 变 量	自 变 量	统计问题归类
(*a*)定量	定量	回归(或线性模型)
(*b*)定量	定性	方差分析
(*c*)定量	定性、定量	协方差分析(或线性模型)
(*d*)定性	定性	列联表
(*e*)定性	定量	判别分析、聚类分析
(*f*)定性	定性、定量	对数线性模型等
(*g*)定性、定量	定性、定量	?

上述的表是一个大致的分类。本书的重点是(*d*)和(*f*)这一类的问题。

第一章我们着重介绍问题和数学模型，对一个典型的例子进行剖析，展示各种不同的观点怎样导出各自的处理方法，其目的是帮助读者了解这些观点的特点，以便在以下各章分别展开，讨论更复杂的问题时能把握住基本的思想，不致被繁琐的推导和形式复杂的公式所迷惑。

以下我们所用的记号与一般教科书是相同的，为了便于实际工作者阅读，每个符号第一次使用时都会给以说明。有些内容在一般教科书上不易见到的，我们都在相应各章的附录中给一简明、适用的介绍。

§2 问题和模型

我们先看两个实际的例子，然后讨论它的模型。

例 2.1(见 Cornfield 1956[1]) 对肺癌患者与对照组的居民，分别调查吸烟与不吸烟的人数，得到下表：

	对照组	肺癌患者
吸 烟	32	60
不吸烟	11	3

表 2.1 肺癌患者与对照组中吸烟与不吸烟人数

问题是患肺癌是否与吸烟有关？

例 2.2(见 Agresti 1984[2]) 对 1976—1977 年佛罗里达州 20 个地区杀人案件中被告和被判处死刑与否的 326 个对象得到下表：

被告 \ 判刑	死	否
白人	19	141
黑人	17	149

表 2.2 判死刑与被告肤色的分类表

问题是法院判处死刑是否与被告的肤色有关？

这两例的问题是很相似的，一个是医学上的，一个是社会法律制度上的，但从资料的统计分析来看，它们可以概括为同一个类型的统计问题。下面我们以例 2.1 为讨论对象，引入数学模型。

是否吸烟与是否患肺癌均可由取值为 0 或 1 的指示变量来描述，即定义

(2.1)
$$x=\begin{cases}1 & \text{该人吸烟}\\ 0 & \text{该人不吸烟}\end{cases}$$

(2.2)
$$y=\begin{cases}1 & \text{该人患肺癌}\\ 0 & \text{该人不患肺癌}\end{cases}$$

x 这个变量指示一个人是否吸烟，y 这个变量指示一个人是否患肺癌。假定我们考察的对象是人数相当多的一个人群，被调查的对象是随机抽取的，则每一个被抽到的人可以确定他的(x, y)的取值，因而 x，y 是两个随机变量，描述 x，y 取值情况的是它们的联合分布。记

$$\pi_{ij}=P(x=i, y=j)\ (x=i, y=j \text{ 的概率})$$

于是有：

$\pi_{00}=$人群中不吸烟、不患肺癌的比例

$\pi_{10}=$人群中吸烟，但不患肺癌的比例

$\pi_{01}=$人群中不吸烟，但患肺癌的比例

$\pi_{11}=$人群中吸烟又患肺癌的比例

$\pi_{\cdot 0}=\pi_{00}+\pi_{10}=$人群中不患肺癌的比例

$\pi_{\cdot 1}=\pi_{01}+\pi_{11}=1-\pi_{\cdot 0}=$人群中患肺癌的比例

$\pi_{0\cdot}=\pi_{00}+\pi_{01}=$人群中不吸烟的比例

$\pi_{1\cdot}=\pi_{10}+\pi_{11}=1-\pi_{0\cdot}=$人群中吸烟的比例

有了这些记号后，可以把问题用数学的形式表示出来。

从例 2.1 的问题来看，就是要问吸烟的人与不吸烟的人患肺癌的概率是否相同？实际上是问两个条件概率是否相等？今条件概率

$$P(y=1\mid x=0)=\frac{P(x=0, y=1)}{P(x=0)}=\frac{\pi_{01}}{\pi_{0\cdot}}$$

$$P(y=1 \mid x=1)=\frac{P(x=1,\ y=1)}{P(x=1)}=\frac{\pi_{11}}{\pi_{1\cdot}}$$

于是要检验的假设是

$$H_0:\ \frac{\pi_{01}}{\pi_{0\cdot}}=\frac{\pi_{11}}{\pi_{1\cdot}} \tag{2.3}$$

是否成立。注意到 $\pi_{0\cdot}$ 与 $\pi_{1\cdot}$ 是 x 的边缘分布，因而有 $\pi_{0\cdot}+\pi_{1\cdot}=1$，于是当(2.3)式成立时，有

$$\frac{\pi_{01}}{\pi_{0\cdot}}=\frac{\pi_{11}}{\pi_{1\cdot}}=\frac{\pi_{01}+\pi_{11}}{\pi_{0\cdot}+\pi_{1\cdot}}=\pi_{\cdot 1}$$

于是从上式就得到 $\pi_{01}=\pi_{0\cdot}\pi_{\cdot 1}$，$\pi_{11}=\pi_{1\cdot}\pi_{\cdot 1}$，而

$$\pi_{10}=\pi_{1\cdot}-\pi_{11}=\pi_{1\cdot}(1-\pi_{\cdot 1})=\pi_{1\cdot}\pi_{\cdot 0}$$

$$\pi_{00}=\pi_{0\cdot}-\pi_{01}=\pi_{0\cdot}(1-\pi_{\cdot 1})=\pi_{0\cdot}\pi_{\cdot 0}$$

因此(2.3)式等价于要检验 x，y 相互独立，即

$$H_0:\ \pi_{ij}=\pi_{i\cdot}\pi_{\cdot j}\qquad \forall\ i,j=0,1\ \text{成立} \tag{2.4}$$

从例 2.2 的问题来看，如用 x 表示被告是黑人(取 1)还是白人(取 0)，y 表示判死刑(取 1)还是不判死刑(取 0)，则也是问 x，y 是否独立？

为了方便，今后我们引入记号≙，$A\triangleq B$ 表示“用 A 表示 B”，也可以表示“将 A 记为 B”，从上下文很容易看出它是哪一种含意，不会发生混淆。从例 2.1 和例 2.2 就可引出一般的 2×2 的列联表，通常称为四格表。四格表相应的数学模型是：假定有 A，B 两种属性，用 1 表示具有某种属性，用 0 表示不具有这种属性，在抽取样本后，我们得到的调查表格是下表：

A \ B	0	1
0	n_{00}	n_{01}
1	n_{10}	n_{11}

表 2.3　A，B 属性的频数分布表

表中 n_{ij}≙具有“A 为 i，B 为 j”的样品数，称为第(i, j)格的频数。为什么调查所得的频数 n_{ij}是这样分配的呢？这与具有属性 A，B 的概率分布有关。如用 π_{ij}≙具有“A 为 i，B 为 j”的比例，则随机抽取一个，出现“A 为 i，B 为 j”的样品的概率就是 π_{ij}，因而可以设想一张 2×2 的概率分布表如下：

A \ B	0	1
0	π_{00}	π_{01}
1	π_{10}	π_{11}

表 2.4　A，B 属性的概率分布表

很明显，频数表是实际的、已知的，概率表是客观存在的，然而往往是未知的，统计问题就是想利用频数表推断概率表的一些性质。例如要问 A，B 这两个属性是否相互独立？就是要问是否表 2.4 具有如下的结构：

(2.5)　$\pi_{ij}=p_iq_j$，p_i，q_j 均非负，$p_0+p_1=1$，$q_0+q_1=1$

容易看出，如关系式(2.5)成立，实际只有两个未知参数 p_0，q_0(或 p_1，q_1)，如(2.5)不成立，就有三个参数(因 $\pi_{00}+\pi_{01}+\pi_{10}+\pi_{11}=1$)。一般说来，结构简单的模型相应的参数就少，结构复杂的模型，相应的参数就多。

在进行数据分析之前，有两点应该指出：(1)对于例 2.1 和2.2 而言，把实际资料看成是随机抽样取得的，那就与真实情况相差太远。例 2.1 是一种回顾性的调查资料，例 2.2 是一种综合性的调查资料，随着介绍方法的逐步深入，我们要注意这一点。如果把它们看成是随机样本，也只着重于说明方法，而不在于问题本身。(2) 2×2的四格表虽然是常见的、比较典型的列联表，然而这种分类过于简单，往往不易得到正确的结论，随着内容的不断深化，我们将会明显地感到这一点，这也是希望读者注意的。

§3 相合系数 (Coefficient of association)

历史上很早就注意到四格表的分析了。为了便于推广到一般情形，四格表中的频数用 n_{11}，n_{12}，n_{21}，n_{22}表示，即四格表为：

A \ B	有 B	无 B
有 A	n_{11}	n_{12}
无 A	n_{21}	n_{22}

如果属性 A 不依赖于属性 B，则在具有属性 B 的类中和不具 B 的类中，有 A 与无 A 的频数之比“应该相同”，即有

$$\frac{n_{11}}{n_{11}+n_{21}}=\frac{n_{12}}{n_{12}+n_{22}}=\frac{n_{11}+n_{12}}{n} \quad (n=n_{11}+n_{12}+n_{21}+n_{22}) \tag{3.1}$$

于是有(见练习 3)

$$\begin{cases} \dfrac{n_{21}}{n_{11}+n_{21}}=\dfrac{n_{22}}{n_{12}+n_{22}}=\dfrac{n_{21}+n_{22}}{n}, \\ \dfrac{n_{11}}{n_{11}+n_{12}}=\dfrac{n_{21}}{n_{21}+n_{22}}=\dfrac{n_{11}+n_{21}}{n}, \\ \dfrac{n_{12}}{n_{11}+n_{12}}=\dfrac{n_{22}}{n_{21}+n_{22}}=\dfrac{n_{12}+n_{22}}{n} \end{cases} \tag{3.2}$$

从(3.1)就可得

$$n_{11}=\frac{(n_{11}+n_{12})(n_{11}+n_{21})}{n}$$

如果有

$$n_{11}>\frac{(n_{11}+n_{12})(n_{11}+n_{21})}{n} \tag{3.3}$$

这就表明具有 B 的属性中具有 A 的，比不具 B 的而且有 A 的“频数相对地多”，也即具有属性 B 的“容易”有属性 A，这反映这两种属性具有正的相合性，或者说，A 与 B 是相合的。如果

$$n_{11}<\frac{(n_{11}+n_{12})(n_{11}+n_{21})}{n}$$

则表明 A 与 B 是不相合的(disassociated)。

因此(3.3)中左、右两端的差距记为 D，就可以衡量相合性，今

$$D=n_{11}-\frac{(n_{11}+n_{12})(n_{11}+n_{21})}{n}=\frac{n_{11}n_{22}-n_{12}n_{21}}{n}$$

于是相合系数应是 D 的增函数。Yule(1900,1912)引入了相合系数

(3.4) $$Q=\frac{nD}{n_{11}n_{22}+n_{12}n_{21}}=\frac{n_{11}n_{22}-n_{12}n_{21}}{n_{11}n_{22}+n_{12}n_{21}}$$

应怎样来理解 Q 呢？用 x 和 y 分别表示属性 A 与 B 的指示变量，则有

$$x=\begin{cases}1 & \text{个体具有 } A\\ 0 & \text{个体不具 } A\end{cases}\qquad y=\begin{cases}1 & \text{个体具有 } B\\ 0 & \text{个体不具 } B\end{cases}$$

假定从全部考察对象中随机抽取了两个个体，它们相应的指示变量分别为(x_1,y_1)，(x_2,y_2)，考虑

$$\pi_s=P((x_1-x_2)(y_1-y_2)=1)\quad\text{（属性变化一致的概率）}$$

$$\pi_d=P((x_1-x_2)(y_1-y_2)=-1)\quad\text{（属性变化相反的概率）}$$

$$\pi_0=P((x_1-x_2)(y_1-y_2)=0)$$

于是考虑系数

(3.5) $$\gamma=\frac{\pi_s}{1-\pi_0}-\frac{\pi_d}{1-\pi_0}=\frac{\pi_s-\pi_d}{1-\pi_0}=\frac{\pi_s-\pi_d}{\pi_s+\pi_d}$$

它反映出属性变化相同的概率比不相同的高还是低。如果用频率代替概率，γ 用频数来表示时，可以看到有 $n_{11}n_{22}$ 对是变化一致的，有 $n_{12}n_{21}$ 对是变化相反的，因而 γ 相应的值就是 Q，这样对 Q 给出了一个理论上的解释。

很明显，我们可以看出 Q 有如下的性质：