〔英〕B. S. 艾沃日特 著

列联表分析

科学出版社

列联表分析

〔英〕B. S. 艾沃日特 著

刘韵源　周家丽 译

科学出版社

1980

内 容 简 介

属性统计在医学、生物学、工农业和社会科学中都有着广泛的应用。本书深入浅出地系统介绍了目前列联表分析中的常用方法，特别是多维列联表分析方法，并列举了不少生物医学、社会科学的实际例子。书中避开了繁冗的数学计算和公式推导，着重阐述概念及方法的应用。是有关属性统计的一本较好的入门书。可供上述学科中的专业工作者参考，亦适合医学院校和应用统计学专业的学生阅读。

B. S. Everitt
THE ANALYSIS OF CONTINGENCY TABLES
Chapman and Hall Ltd.
London 1977

列 联 表 分 析

〔英〕B. S. 艾沃日特 著
刘韵源 周家丽 译

*

科学出版社出版
北京朝阳门内大街 137 号
中国科学院印刷厂印刷
新华书店北京发行所发行 各地新华书店经售

*

1980 年 1 月第 一 版 开本：787×1092 1/32
1980 年 1 月第一次印刷 印张：4 1/4
印数：0001—4,380 字数：92,000

统一书号：13031·1224
本社书号：1704·13—6

定价：0.68 元

译 者 的 话

本书深入浅出地系统介绍了目前列联表分析中常用的方法，是一本有关属性统计的较好入门书。并且在介绍各种分析方法的同时，还推荐了有关文献，以供读者深入研究参考。

全书共分六章，前三章为基础部分，后三章讨论多维列联表的分析。第一章叙述基本知识；第二章详细叙述了 2×2 表的分析方法；第三章研究一般 $r\times c$ 列联表的分析；第四章介绍了多个属性变量所产生的列联表，为下一章打基础；第五章详细讨论了列联表数据的 log-线性模型拟合和参数估计，为全书的精彩部分；第六章简述了特殊类型列联表的分析。作者仔细避开了繁冗的数学计算和公式推导，而把重点集中在概念的阐述和方法的应用上。书中列举了大量实际例子，这对生物学、医学、农业和社会科学工作者尤其适宜。本书也适合医学院校和学习应用统计学的学生阅读。

限于译者水平，误译和不当之处在所难免，请读者批评指正。

译 者

1978 年 12 月于北京

代　　序

几年来我的《定性数据分析》(Analysing Qualitative Data)一书就待修订。自 1961 年出版以来，或许部分由于它的缘故，论述列联表分析的许多重要新著作已经发表。B. S. Everitt 先生欣然应承作此修订，但当他回顾了新近的文献后，明显感到仅仅将原书作些修补已不够，新著作十分广泛，使许多老方法变得陈旧。因此，征得出版者同意后，决定修订本实际上将是一本新书。

现在这本书的前两章并不如此明显，作为导引，它包括了老底子。但在其后各章中，新方法显著增多。这能举一个例子来说明。复习了直至 1961 年的文献后，会因当时分析多维表方法的贫乏和低效感到失望。不仅在社会科学研究中是这样，而且，普遍感到如此。但此严重不足往后就得到确实满意的弥补，不仅增添了检验更细微假设的手段，而且产生了多维频数数据的 log -线性模型拟合和以参数形式描述这些数据的方法。Everitt 在本书第五章中对后一论题作了透彻描述，这章确实向读者提出了有力的挑战。

十多年来，《定性数据分析》一书在研究工作者中广为流传，但用一本更现代的书来补充它已势为必要。Everitt 写的这本书承担了此任务，我确信本书将证明会与其先驱一样深受欢迎。

A. E. Maxwell

1976 年 1 月 于伦敦大学精神病学学院

序

本书起因于 A. E. Maxwell 教授的名著《定性数据分析》的一个修订计划。在回顾了新近的文献后，明显感到须作重要改写才能使其具备最新内容。于是，在 Maxwell 教授的赞同和鼓励下，决定用现在的书名，由我尝试写一本差不多是新的著作。熟悉原书的读者会发现，《列联表分析》前三章的内容与原书前六章的部分内容很相似。然而，本书后三章讨论多维列联表的分析，是原书中未曾包括的论题。

本书所预期的适合对象与原书相同，即精神病学、社会科学、心理学等学科的研究工作者，也适合医学和应用统计学专业的学生阅读。这意味着本书已仔细考虑了将所要求的数学知识，限制在较低程度上。

B. S. Everitt

1976 年 5 月于伦敦大学精神病学学院

目　　录

第一章　列联表和卡方检验

1.1 引　言

本书主要涉及以交叉分类或列联表形式出现的计数数据的分析方法。在这章中，我们将着手定义各种术语，介绍一些命名法，叙述这样的数据能如何产生。后面各节描述卡方分布，并给出用卡方测验检验列联表独立性的数值例子。

1.2 分　类

能用多种不同方式对一总体（用以表示人或物的任何适当规定的种类的通称）的成员进行分类。例如，人能分为男性和女性，已婚和单身，有选举权和无选举权，等等。这是二级分类的一些例子。多级分类也不寡见，例如人被分为惯用左手的，两手俱利的，惯用右手的；或者（譬如说）为作一次盖洛普民意测验（gallup poll），把打算投票的人分为五类：（a）保守党，（b）工党，（c）自由党，（d）主意未定者，（e）其它。我们的主要兴趣将是种类完尽且互不相容的分类。一种分类是完尽的，是指它提供了足够多的种类，以顺应总体的全部成员；各类互不相容则是指总体的每个成员能被正确地分派到其中一类，且仅仅一类。乍看起来，分类的完尽性要求似乎限制过严。例如，或许我们所关心的不是就其整体说，而仅仅就大学生选民的投票意向，进行一次盖洛普民意测验。如果记起一个总体的定义，则此困难迎刃而解。这个词的统计定义

比它在通常用法中的规定更不固定。确实如此，我们可以把上述测验中的总体规定为“全部有选举权的大学生”。各类可以调整、变更或合并；例如在投票的例子中，将类(d)和类(e)合并，未必就会损失很多信息。

当总体被分为若干类后，我们便可以计数每类中个体的数目。这些计数或频数，是本书将主要涉及的数据类型，即我们将研究定性数据，而非诸如高度、温度等连续变量测量中所得到的定量数据。

诚然，一般说来不能得到来自整个总体的信息，我们仅须研究抽自总体的一个样本。统计科学的主要作用之一，就是论证怎样从样本信息的研究中作出关于某些总体的正确推断。在此过程中，根本的一环是确保所抽的样本具有代表性(无偏的)。这能通过随机抽样实现，在一随机样本中，所研究的总体的每个成员有相等概率被抽上。随机抽样概念将在第二章中详细讨论。

1.3 列 联 表

本书主要分析抽自于某总体的一样本按照两个或多个定性变量分类时所产生的数据。例如表1.1中展示了5375个因结核病死亡的人所组成的样本，他们按照两个定性变量，性别和造成死亡的结核病类型分类(注意：如表所示，这些变量的各类是完尽的且互不相容)。

表1.1 因结核病的死亡

	男 性	女 性	总 计
呼吸系统结核	3534	1319	4853
其他类型结核	270	252	522
总 计	3804	1571	5375

表 1.1 这样的一张表称为列联表，这个 2×2 的例子是其最简单的型式(样本成员用两种不同方式被一分为二)。如果此两变量拥有两个以上的类别，则列联表的方格数将多于所示的 4 个。各方格中的表列值为频数，它们能变换成比例或百分率，重要的是须记住，这些数据最初是频数或计数，而不是连续的测量结果。当然，连续数据常能用连续标尺上的一些区间做成离散形式。例如年龄为一连续变量，但如果将人们分为不同年龄群组，则对应于这些组群的各个区间能看作是些分立单元。

由于表 1.1 仅包含两个变量，称为二维列联表。在以后各章中，我们将研究三维和高维列联表，它们是一样本按照两个以上定性变量分类时产生的。

1.4 命 名 法

在这一节中，我们介绍二维列联表的一些表示方法，以后讨论高维列联表时再加以推广。

表 1.2 给出了二维列联表的一般型式，其中由 N 次观测所组成的样本按照两个定性变量分类，一个变量有 r 类，另一个有 c 类，称为 $r \times c$ 列联表。

表 1.2 二维列联表的一般型式

	列 (变 量 2) 1 2 …… c	总 计
行(变量 1) 1	n_{11} n_{12} …… n_{1c}	$n_{1\cdot}$
行(变量 1) 2	n_{21}	$n_{2\cdot}$
r	n_{r1} n_{rc}	$n_{r\cdot}$
总 计	$n_{\cdot 1}$ $n_{\cdot 2}$ $n_{\cdot c}$	$n_{\cdot\cdot} = N$

行变量第 i 类、列变量第 j 类中的观测频数或实计数，即表中第 ij 格中的频数，用 n_{ij} 表示。行变量第 i 类中的总观测数用 $n_{i\cdot}$ 表示，列变量第 j 类中的总观测数用 $n_{\cdot j}$ 表示，称它们为边缘总计，由方格频数给出：

$$n_{i\cdot} = n_{i1} + n_{i2} + \cdots + n_{ic} = \sum_{j=1}^{c} n_{ij} \tag{1.1}$$

$$n_{\cdot j} = n_{1j} + n_{2j} + \cdots + n_{rj} = \sum_{i=1}^{r} n_{ij} \tag{1.2}$$

同理有：

$$n_{\cdot\cdot} = \sum_{i=1}^{r}\sum_{j=1}^{c} n_{ij} \tag{1.3}$$

$$= \sum_{i=1}^{r} n_{i\cdot} = \sum_{j=1}^{c} n_{\cdot j} \tag{1.4}$$

n 表示样本的总观测数，常用 N 表示。

这个表示法常称为点表示，各点表明对特定下标求和。

对表 1.1 中的数据，我们有：

(I) $r = c = 2$，两个变量各有两类；

(II) $n_{11} = 3534$，$n_{12} = 1319$，$n_{21} = 270$，$n_{22} = 252$，为方格频数；

(III) $n_{1\cdot} = 4853$，$n_{2\cdot} = 522$，为行边缘总计，是两种类型结核病的死亡总数；

(IV) $n_{\cdot 1} = 3804$，$n_{\cdot 2} = 1571$，为列边缘总计，是样本中男女两性的总数；

(V) $N = 5375$，是样本的总观测数。

1.5 独立分类与连带

研究了与我们有关的数据类型后，现在需要考察对这类数据有兴趣的问题。通常最重要的问题是：构成列联表的定

性变量是否为独立的 (independent)。为回答这个问题，必须弄清分类间的独立性条件是什么。对于 2 × 2 表，这些条件比较容易看出。例如回到表 1.1 的数据，显然，如果引起死亡的结核病类型和病人的性别无关，则能期待死于呼吸系统结核的男女两性病人各自的比例数相等。如果这些比例数不同，则表明与其他因素相比，呼吸系统结核所引起的死亡，倾向于与性别有着更为密切的连带 (association) 关系。(当然，由于抽样的偶然因素，以及可能归结为随机原因的其他因素，这两个比例数可以一定程度地不等；需要确定的是，在这两个比例数间所观察到的差异是否太大，以致不能归因于上述理由，为此需要作检验，这将在下节讨论。)

我们已直观看到，2 × 2 表中独立性意味着两个比例数相等，现在稍微更正式地研究一般 $r \times c$ 列联表中这个概念的含义。先假定在总体 (样本抽自于它) 中，属于行变量第 i 类、列变量第 j 类的一次观测的概率用 p_{ij} 表示；因此，由抽样 N 个个体所得表的第 ij 格中的期望频数，或预计数，F_{ij} 用下式给出：

$$F_{ij} = Np_{ij} \tag{1.5}$$

[熟悉数学期望和概率分布的读者知道，在观测频数以概率值 p_{ij} 遵循多项分布的假设下，有 $F_{ij} = E(n_{ij})$；例如可参阅 Mood 和 Greybill, 1963, 第三章。]

现在让 $p_{i\cdot}$ 表示总体中一次观测是属于行变量第 i 类的概率 (此时不涉及列变量)，用 $p_{\cdot j}$ 表示列变量第 j 类的相应概率。则由概率乘法律，总体中两变量间的独立性意指：

$$p_{ij} = p_{i\cdot}p_{\cdot j} \tag{1.6}$$

用列联表中的期望频数，则独立性意味着：

$$F_{ij} = Np_{i\cdot}p_{\cdot j} \tag{1.7}$$

读者也许要问：这能有什么帮助呢？因为两变量的独立性只

是由未知的总体概率值所确定。然而须知这些概率实际上可以简单地从观测频数估计。不难证明，概率 $p_{i\cdot}$ 和 $p_{\cdot j}$ 的"最好"估计 $\hat{p}_{i\cdot}$ 与 $\hat{p}_{\cdot j}$，是以观测值的相应边缘总计为依据的，即有：

$$\hat{p}_{i\cdot} = \frac{n_{i\cdot}}{N} \quad 和 \quad \hat{p}_{\cdot j} = \frac{n_{\cdot j}}{N} \tag{1.8}$$

（这是极大似然估计，见 Mood 和 Greybill，第十二章。）运用方程（1.8）所给出的 $p_{i\cdot}$ 与 $p_{\cdot j}$ 的估计，能进而估计表中第 ij 格的期望频数，如果这两个变量是独立的话。用 E_{ij} 表示此估计，注意到方程（1.7），则有：

$$E_{ij} = N\hat{p}_{i\cdot}\hat{p}_{\cdot j} = N\frac{n_{i\cdot}n_{\cdot j}}{NN} = \frac{n_{i\cdot}n_{\cdot j}}{N} \tag{1.9}$$

当两个变量独立时，用公式（1.9）所估计的频数与观测频数间仅仅相差一个可归因于偶然因素的量；然而如果两个变量是非独立的，则能预期会有一个大的差异出现。因此建立在两组频数 n_{ij} 和 E_{ij} 间差值大小基础上的，对构成二维列联表的两变量之独立性的任何检验，似乎是灵敏的。这样的检验在下节讨论。（本书后面各部分中，当不存在把它们和频数 F_{ij} 搞混淆的危险时，期望频数的估计 E_{ij} 将常常简称为"期望值"。）

1.6 卡方检验

上节讨论了两个变量的独立性概念。为检验独立性，已指出需要研究以下假设的真实性：

$$p_{ij} = p_{i\cdot}p_{\cdot j} \tag{1.10}$$

通常称它为零假设，用符号 H_0 表示。

也指出了：假设应该建立在 H_0 真实时期望频数的估计值（为 E_{ij}）与观测频数（为 n_{ij}）之差的基础上。首先由 Pearson

(1904)所提出的这样一个检验用了统计量 χ^2:

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(n_{ij} - E_{ij})^2}{E_{ij}} \tag{1.11}$$

能看出这个统计量的大小与差值($n_{ij} - E_{ij}$)有关。如果两个变量是独立的,这些差将小于两个变量非独立时的差;从而 H_0 真实时的 χ^2 值将小于 H_0 不真实时的值。此后需要有一个方法来确定导致接受和拒绝 H_0 时的 χ^2 值。这样的一个方法,是以在独立性假设真实的假定下对 χ^2 所导出的概率分布为依据的。假设的接受或拒绝以所得 χ^2 值的概率为基准;有"低"概率的值导致拒绝假设,其余值时则为接受假设。这是统计学中导出显著性检验的标准程序。一般取 0.05 或 0.01 为"低"概率,并称为检验的显著性水平 (significance level)。(显著性水平和有关论题的更详细的讨论可见 Mood 和 Greybill, 第十二章。)

当假定观测频数有一多项分布,且期望频数不是太小时(见本书 3.4 节),统计量 χ^2 近似遵循卡方分布。独立性假设的检验现在可以通过比较所算得的 χ^2 值与卡方分布的查表值而得以实现。

1.7 卡方分布

假定读者已熟悉正态分布(大多数统计教科书中都有叙述,也可参阅 Mood 和 Greybill, 第六章),卡方分布作为若干独立变量平方和的概率分布由它产生,这些独立变量的每个都有一标准正态分布,即均值为零,标准差为 1 的正态分布。分布形状与所包含的独立变量数目有关。例如,由 ν 个独立标准正态变量平方和所构成的卡方变量(χ^2):

$$\chi^2 = z_1^2 + z_2^2 + \cdots + z_\nu^2 \tag{1.12}$$

有一个仅依赖于 ν 的分布。许多教科书中给出了对不同的 ν 值，分布取不同形状的曲线图（Hays, 1973, 第十一章）。通常将构成卡方变量的独立变量数称为自由度（degrees of freedom）；在上面的情况下，则说成自由度（d. f.）为 ν 的卡方。

由于可以利用已经制成表的卡方值（附录 A），这儿无需讨论卡方分布的数学表达式。为确定由某些列联表所算得的 χ^2 值是否应导致我们接受或者拒绝独立性假设的全部必要信息，能从附录 A 的表中得出。然而我们必须先知道 χ^2 的自由度；它与构成列联表的每个变量的类数有关，其值在下节中导出。知道 ν 后，对于一些预先确定的显著性水平 α（常取为 0.05 或 0.01），查看自由度为 ν 的卡方分布表，找出所需要的卡方值。如果 χ^2 值大于在 α 水平上的表查值 χ^2_ν，则表明所得结果极少可能是偶然出现的（小于 100 α% 次）；从而表现出与独立性的真实偏离，导致拒绝零假设 H_0。

1.8 二维列联表的自由度

用以检验构成一列联表的两个变量独立性的统计量已取为：

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(n_{ij} - E_{ij})^2}{E_{ij}} \qquad (1.13)$$

当独立性假设真实时，近似 χ^2 的分布的卡方分布之自由度，在给定列联表行列边缘总计的情况下，简单地为 (1.13) 中独立项的数目。(1.13) 中总项数为 $r \times c$，也就是表中方格数。但它们当中的若干项能由行列总计的知识确定。例如知道了 r 行总计便能每行一个地决定 r 个频数 n_{ij}，并因此确定 (1.13) 中的 r 项。(1.13) 中的独立项数于是减到 $(rc - r)$。