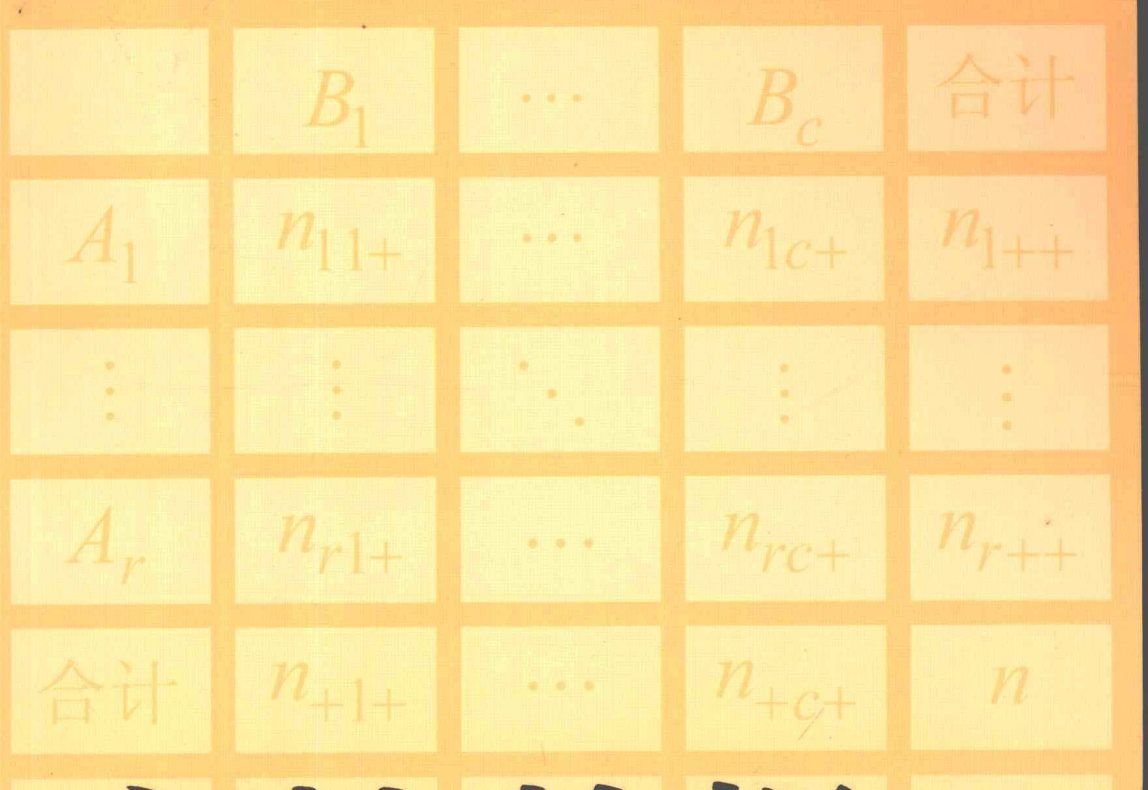

定性数据统计分析

◉ 王静龙 梁小筠 编著

中国统计出版社
China Statistics Press

定性数据统计分析

◉ 王静龙 梁小筠 编著

中国统计出版社
China Statistics Press

(京)新登字 041 号

图书在版编目(CIP)数据

定性数据统计分析/王静龙,梁小筠编著.
-北京:中国统计出版社,2008.7
ISBN 978-7-5037-5496-8

Ⅰ.定…
Ⅱ.①王… ②梁…
Ⅲ.定性分析-统计分析(数学)
Ⅳ.0212

中国版本图书馆 CIP 数据核字(2008)第 079845 号

定性数据统计分析

作　　者/王静龙　梁小筠
责任编辑/陈悟朝
装帧设计/艺编广告
出版发行/中国统计出版社
通信地址/北京市西城区月坛南街 57 号　邮政编码/100826
办公地址/北京市丰台区西三环南路甲 6 号
网　　址/www.stats.gov.cn/tjshujia
电　　话/邮购(010)63376907　书店(010)68783172
印　　刷/利兴印刷有限公司
经　　销/新华书店
开　　本/710×1000mm　1/16
字　　数/290 千字
印　　张/16.25
印　　数/1-2000 册
版　　别/2008 年 7 月第 1 版
版　　次/2008 年 7 月第 1 次印刷
书　　号/ISBN 978-7-5037-5496-8/O·65
定　　价/35.00 元

前言

定性数据统计分析是统计分析的一个重要内容，它在实践中有着广泛的应用。华东师范大学统计系早在上世纪90年代就开设了这一门课。张尧庭教授所写的书《定性资料的统计分析》，以及他后来翻译的《离散多元分析：理论与实践》一书，是我们教学的参考用书。在教学的过程中，我们陆续编写了一份份各个章节的讲义。修改多次的讲义经整理加工而成的《定性数据分析》2005年在华东师范大学出版社出版。尽管那本书内容不够充实，叙述不够深刻，但大家给于我们极大的鼓励，并提出了很多的修改建议。现在这本《定性数据统计分析》较那本书有很大的不同。首先本书增加了对数线性模型与对应分析等两方面的内容。对数线性模型是处理定性数据的一个非常有效的模型。我们还结合对数线性模型与逻辑斯蒂线性回归模型，简要介绍广义线性模型。对应分析是利用多元统计分析中的主成分分析和典型相关分析的方法，寻找列联表各个属性间的对应关系。我们统计系毕业的在市场咨询与社会调查工作的同学都说，对应分析在他们的工作中应用非常广泛。他们还讲到列联表的某些格上的数是0时的处理问题，为此本书增加了不完备列联表统计分析的内容。列联表的独立性除了用概率描述外，本书还用期望频数描述独立性，并由此引出了列联表的相关模型。用期望频数描述独立与相关性，我们就可以很自然地引入列联表的对数线性模型。在逻辑斯蒂线性回归模型这一部分，我们还增加了逻辑斯蒂判别分析和多项逻辑斯蒂回归模型的内容。

本书共分八章，第一章介绍定性数据的描述性统计分析方法。第二章介绍分类数据的统计推断方法，第三、四和五章介绍交叉分类数据，即列联表的统计推断方法。第六章介绍逻辑斯蒂线性回

归模型。第七章介绍对数线性回归模型。第八章介绍对应分析。本书在选材时，注意到应用统计软件，例如 EXCEL、MINITAB、SPSS 和 SAS 等的需要。书中收集、编写了大量的例子，它们反映了定性数据应用的很多方面的问题，也是各种统计方法如何运用的示范。本书将有关的理论证明放在附录中，由于教学时间紧，或急于了解统计方法应用的读者可以跳过去。

本书除了作为大学统计专业的教学用书外，还可以作为从事理论研究和应用的统计工作者、教师和学生的参考用书，此外，本书也适宜于在社会学、心理学、人口学、市场学和医学等领域进行应用研究，以及从事市场咨询与社会调查的人士阅读，也可以作为这些学科的教学用书。

感谢张尧庭教授，他写的和翻译的上面所述的两本书使得我们对定性数据的统计分析产生了浓厚的兴趣。本书的完稿得益于他的教诲。如今，张尧庭教授已过世，仅以此书寄托我们对他的怀念与哀思。我们也要感谢华东师范大学统计系的同事、历届本科生和研究生。感谢茆诗松教授的推荐，感谢中国统计出版社严建辉社长和陈悟朝先生的辛勤劳动，感谢华东师范大学出版社朱建宝先生的支持。

王静龙　梁小筠

2008 年 4 月

目 录

第一章　定性数据

§1.1　定性数据

数据按其取值来分有以下四种类型：

(1) 计量数据　如人的身高、体重、……，产品的长度、直径、重量、……，股票的价格、市盈率、……。它们的取值可以是某个区间内的任意一个实数。

(2) 计数数据　如企业职工人数、成交股票股数、单位时间内通过某交叉路口的汽车数等。它们在整数范围内取值，大部分还仅在非负整数范围内取值。

(3) 有的时候，观察值不是数，而是事物的属性，如人的性别(男、女)，婚姻状况(未婚、有配偶、丧偶、离婚等)，物体的颜色、形状。我们常用数来表示属性的分类，例如用数“1”和“2”分别表示男和女。这些数只起一个名义的作用，只是一个代码，没有大小关系，也不能进行运算。在这里，“2”与“1”不能比较大小，“1 + 2”也没有意义。这一类数据称为名义定性数据，简称名义数据。

(4) 有些事物的属性有一个顺序关系，如人的文化程度由低到高可分为文盲，小学，初中，高中、中专和大专、大学等 5 类。用数 0，1，2，3 和 4 分别表示文盲，小学，初中，高中、中专和大专、大学。又如顾客对某商场营业员服务态度的评价分为“满意”、“一般”、“不满意”三类，可分别用“3”、“2”、“1”表示。这些数只起一个顺序作用，类与类之间的

差别是不能运算的。例如,“满意”比“一般”好,但“好多少”是不能计算的,即这里的“3 - 2”是没有意义的。这一类数据称为有序定性数据,简称有序数据。

计量数据和计数数据称为定量数据。名义数据和有序数据称为定性数据。

$$\text{数据}\begin{cases}\text{定量数据}\begin{cases}\text{计量数据}\\ \text{计数数据}\end{cases}\\ \text{定性数据}\begin{cases}\text{名义数据}\\ \text{有序数据}\end{cases}\end{cases}$$

类似地,有定量变量和定性变量。定量变量中有计量定量变量和计数定量变量。定性变量中有名义定性变量和有序定性变量。

实际问题中,有时所有的数据都是定性数据或定量数据,有时既有定性数据又有定量数据。本书讨论含有定性数据统计问题的分析方法。

§1.2 定性数据的描述性统计

得到一批定性数据后,要进行整理,从中提取有用的统计信息。整理定性数据常用的方法有表格法、图示法和数值法。

§1.2.1 表格法

例 1.1 向 50 个被访者调查“在下列 5 种饮料中,您最喜欢喝的是哪一种饮料?”

可口可乐、苹果汁、橘子汁、百事可乐、杏仁露

得到结果见表 1.1。

表 1.1 中的数据使人看了眼花缭乱,不得要领。如果统计一下每一种饮料出现的次数(频数),可以看到“可口可乐”出现了 17 次,“苹果汁”出现了 8 次,“橘子汁”出现了 7 次,“百事可乐”出现了 7 次,“杏仁露”出现了 11 次。这些结果汇总在下面的频数频率分布表 1.2 中。

表 1.1 被访者最喜欢的饮料

橘子汁	可口可乐	杏仁露	可口可乐	可口可乐
苹果汁	苹果汁	橘子汁	杏仁露	苹果汁
可口可乐	杏仁露	可口可乐	杏仁露	百事可乐
苹果汁	橘子汁	杏仁露	可口可乐	苹果汁
可口可乐	可口可乐	百事可乐	可口可乐	可口可乐
百事可乐	百事可乐	橘子汁	杏仁露	可口可乐
可口可乐	苹果汁	橘子汁	可口可乐	杏仁露
苹果汁	橘子汁	可口可乐	杏仁露	杏仁露
可口可乐	杏仁露	苹果汁	百事可乐	可口可乐
可口可乐	百事可乐	杏仁露	橘子汁	百事可乐

表 1.2 最喜欢的饮料的频数频率分布表

饮料名称	频数	频率(%)
可口可乐	17	34
苹果汁	8	16
橘子汁	7	14
百事可乐	7	14
杏仁露	11	22
合 计	50	100

从表 1.2 中可以看出:喜欢“可口可乐”的频数最高,“杏仁露”其次,接下来的“苹果汁”,“橘子汁”和“百事可乐”受欢迎的程度差不多。这样的信息单凭观察表 1.1 的原始数据是不容易得出的。

频数分布表是表明几个不相重叠的类中每一类的频数的表格。表 1.2 是名义数据的频数频率分布表。对于有序数据,在制作频数频率分布表时还可以统计累计频率。

例 1.2 某班有 55 名学生,数学课程考试的成绩为:优 4 人,良 11 人,中 23 人,及格 14 人,不及格 3 人。频数分布表见表 1.3。表 1.3 的“累计频率”这一栏告诉我们:成绩优良的学生占 27%,95% 的学生达到要求。

表 1.3 某班学生数学成绩的频数频率分布表

成绩	人数	频率(%)	累计频率(%)
优	4	7	7
良	11	20	27
中	23	42	69
及格	14	26	95
不及格	3	5	100
合 计	55	100	

表 1.2 按饮料名称分组,如果我们还想考察这些饮料受欢迎的程度与性别是否有关,那就需要制作饮料名称和性别的交叉分组列表,见表 1.4。

表 1.4 饮料名称和性别的交叉分组列表

		性别		合 计
		男	女	
饮料名称	可口可乐	13	4	17
	苹果汁	2	6	8
	橘子汁	3	4	7
	百事可乐	5	2	7
	杏仁露	2	9	11
合 计		25	25	50

表 1.4 告诉我们 50 个被访者中男性和女性各有 25 人。这些饮料受欢迎的程度与性别是有关系的。男性被访者最喜欢可口可乐,其次是百事可乐,而女性最喜欢杏仁露,其次是苹果汁。表 1.4 就是所谓的两种方式分组的交叉表,类似地有三,或更多种方式分组的交叉表。交叉表又称列联表(contingency table)。列联表的统计分析见本书第三、四和五各章。

§1.2.2 图示法

§1.2.2.1 条形图

条形图是用宽度相同的长方形的高低或长短来表示数据变动特征

的图形。长方形可以竖放也可以横放。竖放时，常在横轴上标记定性数据的每一类别，在纵轴上表示频数或频率。每一类都对应一个长方形，这个长方形的高度表示这一类的频数或频率。

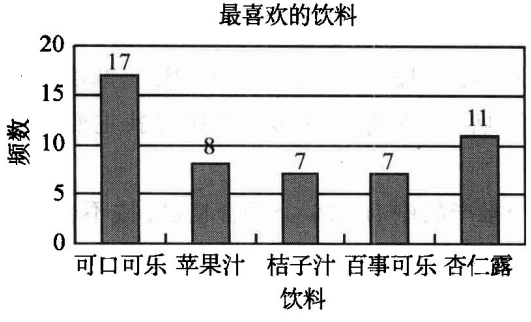

图 1.1 “最喜欢的饮料”的条形图

图 1.1 是“最喜欢的饮料”的条形图，它是利用 Excel 软件画出来的。图中横轴表示五种饮料，每一种饮料对应一个长方形，长方形的高度表示相应的频数。

§1.2.2.2 圆形图

圆形图用一个圆及圆内几个扇形的面积来表示数据的频数(频率)分布。定性数据的每一类对应一个扇形，它的中心角等于 360°乘以该类出现的频率。

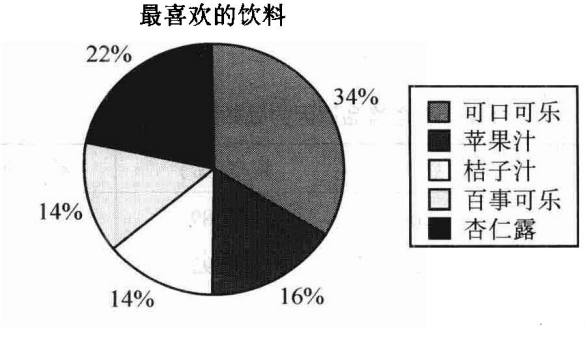

图 1.2 “最喜欢的饮料”的圆形图

图 1.2 是“最喜欢的饮料”的圆形图，它也是利用 Excel 软件画出来的。每一种饮料对应一个扇形，扇形的中心角与该饮料出现的频率

成比例。例如,“可口可乐”出现的频率为34%,它对应的扇形的中心角就等于360°×0.34=122.4°。

§1.2.2.3 排列图

排列图,又称帕累托(Pareto)图。它的全称是“主次因素排列图”,在质量管理中很有用。人们通过生产实践发现,大部分的质量问题往往只由少数几个原因引起,找出这几个原因,是解决质量问题的关键。排列图可以在影响产品质量的众多因素中寻找主要因素,以明确改进质量的方向。

例1.3 一批产品中有976个不合格品,按不合格品产生的原因分类,得表1.5。

表1.5 不合格品原因频数分布表

原因	频数
操作	22
设备	526
工具	292
工艺	89
材料	47
合计	976

把表1.5按频数从大到小重新排列,计算频率和累积频率,得表1.6。

表1.6 不合格品原因的频数频率分布表

原因	频数	频率(%)	累积频率(%)
设备	526	53.89	53.89
工具	292	29.92	83.81
工艺	89	9.12	92.93
材料	47	4.82	97.75
操作	22	2.25	100
合计	976	100	

图 1.3 是根据表 1.6 画出来的排列图。它右面的图标分别表示频数和累积频率。左面的图下方的数字 1、2、3、4 和 5 分别表示设备、工具、工艺、材料和操作。这个图像条形图，但在各个条形或其上方又画了以小方块表示的累积频率。这些小方块连成一条折线，这条折线称为累积频率折线，也称为帕累托折线。

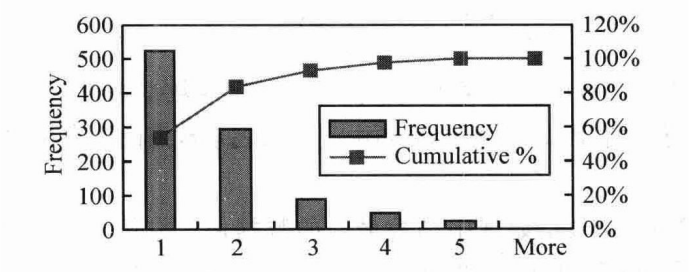

图 1.3 排列图

根据累积频率在 0～80％之间的因素为主要因素的原则，可以在累积频率为 80％处画一条水平线，在该水平线以下的折线部分对应的原因项便是主要因素。

从图 1.3 可知，造成不合格品的主要原因是设备与工具，要减少不合格品首先应该从以上两方面着手。在这两个问题解决之后，再次对不合格品产生的原因画排列图进行分析，这时工艺很可能成了造成不合格品的主要原因，成了我们为减少不合格品的首要抓手。工艺问题解决之后，再画排列图，依此类推。这样我们就有可能达到质量管理工程 6 西格玛(sigma)的要求，使得不合格品率不超过 3.4/1000000。排列图是著名的 6 西格玛管理的一个重要工具。

排列图是意大利经济学家帕累托发现的帕累托原则在质量管理工程方面的应用。对帕累托原则有兴趣的读者可参阅本书附录 1。

使用 Minitab 制作帕累托图(pareto chart)的步骤：

输入 Data -> Stat -> Quality Tools -> Pareto Chat

§1.2.3 数值法

表格法和图示法描述了定性数据大致的分布形状，数值法是用代表性的数值描述定性数据的统计分布的特征。代表性的数值有两类，

一类描述定性数据的中心位置，另一类描述定性数据的离散程度。由于定性数据值仅仅是表示事物属性的代码，所以很显然地，描述定性数据统计分布特征的这些代表性的数值应与定性数据的取值没有关系。

§1.2.3.1 中心位置的描述

设有一批数据 $x_1, x_2, \cdots, x_n$。在一般的统计教材中，样本均值

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{1.2.1}$$

是数据中心位置的最主要的代表值。但对定性数据来说，数据的加法是没有意义的，因此，常用众数和中位数来表示数据的中心位置。

① 众数

众数(mode)是数据中频数最高的数据值。在例 1.1 中“可口可乐”的频数最高，因而，它也就是“可口可乐”是众数。在这里，众数提供了被调查者偏好的信息。对名义数据来说，众数是描述数据中心位置的量度。众数记为 m_o。至此，读者可能会有个疑问，顾名思义众数应该是个数，“可口可乐”怎么能说它是众数。事实上，我们这里所说的数据有广泛的含意，人的性别，饮料的名称等也都可以看成是数据。英文名词“data”有多种含意，它可理解为数据，也可理解为资料。为此有的书说饮料的名称是定性资料，而表示饮料名称的数是定性数据。为方便起见，本书将定性资料和定性数据统称为定性数据。不难设想，把定性资料与定性数据区分开来，既不方便，也没有必要。

有时，频数最高的数据值可能不止一个，这时，就存在不止一个众数。如果在数据中有两个众数，则称此数据为双众数的。如果有三个或三个以上的众数，则称数据为多众数的。在多众数的情况下，众数对于描述数据的中心位置，已经没有多大意义了。

众数的应用很广泛，且不难理解。例如农贸市场卖蔬菜的小贩，一天下来他最关心的是今天哪一个蔬菜卖得最多。为此农贸市场上某类商品的价格常以众数值为代表。由交通事故驾驶过程因素分析的表 1.7 知，“察觉得晚”是事故起因内在因素的众数。交通管理部门最关心的就是如何使得驾驶员不会察觉得晚。一方面教育驾驶员不要疲劳驾驶、醉酒驾驶，驾车时注意力集中，另一方面研究分析道路的管理，例如信号控制系统。

表 1.7 交通事故驾驶过程因素分析

内在因素	事故起数	构成率(%)
察觉得晚	1192	59.6
判断错误	696	34.8
驾驶错误	96	4.8
其它	16	0.8
合计	2000	

1995 年根据天津市统计局提供的最近汇总的全国出生人口按月分布状况的抽样调查资料表明,我国新生婴儿出生最多的月份是 10 月,而出生婴儿最少的月份是 6 月。20590 名新生婴儿中,出生于 10 月的多达 2076 人,占首位,10 月是众数;而出生于 6 月份的最少,只有 1477 人。2007 年公安部对我国 13 亿多户籍人口的一项统计分析显示,“王”姓是第一大姓,有 9288.1 万人,“李”姓其次,有 9207.4 万人,第三是“张”姓,有 8750.2 万人。他们分别占全国人口总数的 7.25%、7.19%和 6.83%。然而 2006 年初,国家自然科学基金委、中国科学院遗传与发育生物学研究所历时两年的百家姓统计研究的结果是李、王、张分列前三位。看来 2007 年的版本与 2006 年的版本在谁是中国第一大姓的问题上有所差别。差别的原因很可能是在于,2006 年版的是抽样调查,调查的范围还包括港、澳、台地区,共有 3 亿人口的数据。2007 年版是公安部将全部户籍人口的姓排序生成的,但港、澳、台地区没有纳入公安部的户籍登记。2006 年版的数据表明,王姓的分布北少南多,港、澳、台都是王姓少而李姓多,台湾地区李姓超过王姓几十万人。看来,包括港、澳、台地区,在全世界的范围内,“李”姓是第一大姓。

② 中位数

中位数(median)是将数据按由递增或递减的顺序排列后位于中间的数值。如果数据的个数为奇数,中间的数就是中位数;如果个数为偶数,中间两个数的平均值就是中位数。中位数记为 m_e。

例如,有 5 个数:2,3,5,7,10,中间的数 5 就是中位数 m_e,有两个数比它小,两个数比它大。如果有 6 个数:2,3,5,7,10,14,中间的两个数 5 和 7 的平均值 6 就是中位数 m_e,有三个数比它小,三个数比它大。