定性数据分析

王静龙 梁小筠* 编著

DingXingShuJuFenXi

华东师范大学出版社

定性数据分析

王静龙　梁小筠　编著

华东师范大学出版社

图书在版编目(CIP)数据

定性数据分析/王静龙，梁小筠编著. —上海：华东师范大学出版社，2005.7
ISBN 7-5617-4349-1

Ⅰ.定… Ⅱ.①王…②梁… Ⅲ.定性分析：统计分析 Ⅳ.C813

中国版本图书馆 CIP 数据核字(2005)第 086164 号

华东师范大学教材出版基金资助出版

定性数据分析

编　　著　王静龙　梁小筠
特约编辑　陈信漪
封面设计　卢晓红
版式设计　蒋　克

出版发行　华东师范大学出版社
　　市场部 电话 021-62865537
　　门市(邮购)电话 021-62869887
　　门市地址　华东师大校内先锋路口
业务电话　上海地区 021-62232873
　　华东 中南地区 021-62458734
　　华北 东北地区 021-62571961
　　西南 西北地区 021-62232893
业务传真　021-62860410　62602316
http://www.ecnupress.com.cn
社　　址　上海市中山北路 3663 号
　　邮编 200062

印 刷 者　商務印書館上海印刷股份有限公司印刷
开　　本　890×1240　32 开
印　　张　6
字　　数　166 千字
版　　次　2005 年 9 月第一版
印　　次　2005 年 9 月第一次
印　　数　3 100
书　　号　ISBN 7-5617-4349-1/O·151
定　　价　9.00 元

出 版 人　朱杰人

前　言

定性数据分析是数据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九十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》，是我们教学的参考用书. 原来的教学安排为每周 3 课时，共 18 周. 后来改为每周 2 课时，共 18 周. 在教学的过程中，我们陆续编写了各个章节的讲义. 本书就是将修改多次的讲义整理加工而成的.

本书共分六章，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 Logistic 线性回归模型. 很遗憾，由于学时数有限，本书没有介绍对数线性回归模型. 本书在选材时，注意到应用统计软件，例如 SPSS 和 SAS 等的需要.

我们建议，教学时数的分配计划如下表：每周 2 学时，教学周 18 周，含复习考试 1 周.

章　次	一	二	三	四	五	六
学　时	2	4	8	8	8	4

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. 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. 此外，§ 3.2 中的 3.2.4，χ^2 检验和似然比检验的比较，也可以跳过去.

对理论研究有兴趣的读者不妨阅读一下.

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,本书也适宜于进行社会学、心理学、人口学、医学等学科的研究及从事抽样调查的人士阅读,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.我们也要感谢华东师范大学统计系的历届学生,因为有他们的参与,我们在教学中对所涉及的内容越来越有体会,享受着极大的乐趣.如果没有他们的参与,本书难以成稿.还要感谢茆诗松教授,他在百忙之中审阅了书稿,提出了很多宝贵的意见,并推荐书稿早日出版.感谢华东师范大学教材出版基金的资助.最后,要感谢华东师范大学出版社,没有他们的辛勤劳动,本书不可能很快出版.

王静龙、梁小筠

2004 年 12 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) 名义数据 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) 有序数据 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、高中或中专和大专或大学等 5 类. 用数 0、1、2、3 和 4 分别表示文盲、小学、初中、高中或中专和大专或大学. 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**.

数据
- 定量数据
  - 计量数据
  - 计数数据
- 定性数据
  - 名义数据
  - 有序数据

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的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“在可口可乐、苹果汁、橘子汁、百事可乐、杏仁露等 5 种饮料中，您最喜欢喝的是哪一种饮料?”得到的结果如表 1.1.

表 1.1 被访者最喜欢的饮料

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

（续 表）

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

上面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中.

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积频率.

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　某班学生数学成绩的频数频率分布表

成　绩	人　数	频率(%)	累积频率(%)
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 1.3 的“累积频率”这一栏告诉我们：成绩优良的学生占 27%，95%的学生达到要求.

1.2.2　图　示　法

(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.

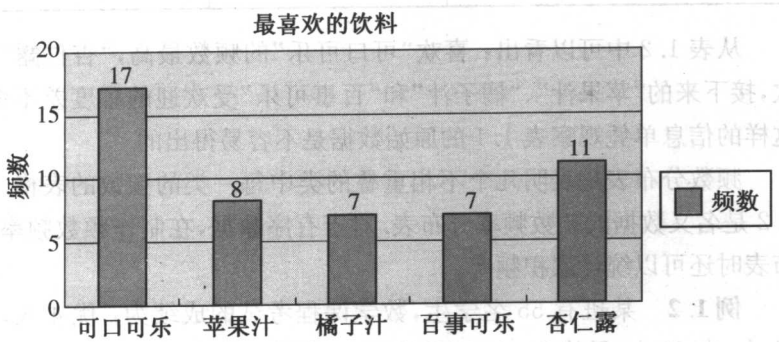

图 1.1　“最喜欢的饮料”的条形图

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.

(2) 饼图

饼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.

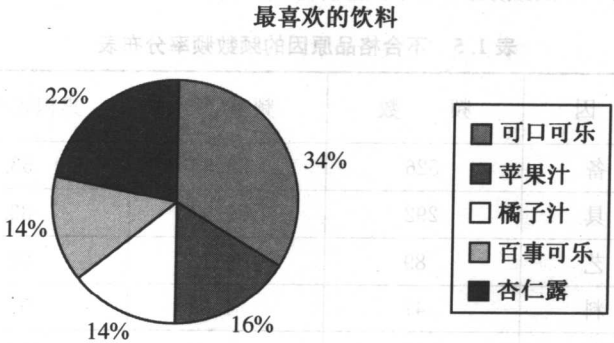

图 1.2　“最喜欢的饮料”的饼图

图 1.2 是“最喜欢的饮料”的饼图，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 $360^\circ \times 0.34 = 122.4^\circ$.

(3) 排列图

排列图(*Pareto* 图)在质量管理中很有用，它的全称是“主次因素排列图”. 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 1.4.

表 1.4　不合格品原因频数分布表

原　因	频　数	原　因	频　数
操　作	22	工　艺	89
设　备	526	材　料	47
工　具	292	合　计	976

把表 1.4 按频数从大到小重新排列，计算频率和累积频率，得表 1.5.

表 1.5　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	

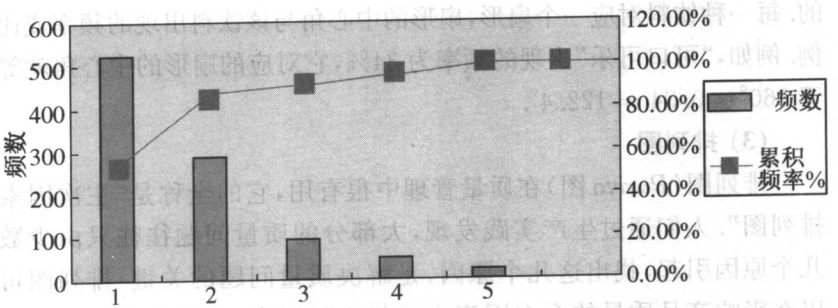

图 1.3　排列图

图 1.3 是利用 Excel 软件画出来的排列图. 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 *Pareto* 折线.

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.

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精确地描述定性数据的统计分布. 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.

1.2.3.1 中心位置的描述

设有一批数据 x_1, x_2, …, 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

众数是在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,它就是众数. 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 mo.

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. 在这种情况下，众数对于描述数据的中心位置，已经没有多大意义了.

抽样调查资料显示，我国婴儿出生于十月份的最多. 在 20 590 名新生婴儿中，出生于十月的有 2 076 人，占首位，十月是众数；而出生于六月份的最少，只有 1 477 人. 我国 56 个民族的姓氏多达 11 969 个. 世界第一大姓是人口超过 8 700 万的“李”姓，“李”姓是众数. 李、王、张、刘、陈 5 大姓氏总人数达 3.5 亿.

(2) 中位数

中位数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 me.

例如，有 5 个数：2、3、5、7、10，中间的数 5 就是中位数 me，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 me，有三个数比它小，三个数比它大.

如果数据中有极端值，中位数不受极端值的影响，能够很好地反映数据的中心位置. 例如，某地区某年农民户均年纯收入增长了 2.9%. 但是该地区农民减收户却多达 60%，平收、增收的农户只占$\frac{1}{3}$强. 所以户均年纯收入的增长是由并不占多数的农户的收入增长拉动的. 相对平均数而言，若计算农户年纯收入的中位数，就能较好地反映农户收入的真实情况.

对定性有序数据来说，中位数是描述数据中心位置的量度.

例 1.4 某儿童活动中心对游客进行问卷调查，其中的两个问题为：

(1) 您对活动设施满意吗？

(2) 您对工作人员的服务态度满意吗？

调查结果如表 1.6 所示.

表 1.6 游客对活动设施及服务态度的评价(%)

指 标	很满意 1	满意 2	一般 3	不满意 4	很不满意 5
活动设施	19.7	39.9	30.3	8.2	1.9
累积百分比	19.7	59.6	89.9	98.1	100
服务态度	8.4	36.5	42.2	11.7	1.2
累积百分比	8.4	44.9	87.1	98.8	100

从累积百分比可以看出：对活动设施评价的中位数(50%处)位于“满意”这一类；对服务态度评价的中位数(50%处)位于“一般”这一类. 也就是说游客对儿童活动中心的活动设施和服务态度评价的代表值分别为“满意”和“一般”.

一般来说，定性有序数据的中位数可以按以下的步骤计算：

a) 将原数据按递增(由小到大)的顺序排列；

b) 计算 $i=\frac{n+1}{2}$，其中 n 是数据的个数；

c) 若 i 是整数，则定性有序数据的中位数是将数据按递增的顺序排列后位于第 i 位次的数；若 i 不是整数，将 i 向上取整，即取 k 为大于 i 的毗邻整数，则定性有序数据的中位数是将数据按递增的顺序排列后位于第 k 位次的数.

(3) 百分位数

百分位数用于衡量数据的位置的量度. 但它所衡量的，不一定是中心位置. 儿童的身高、体重常以百分位数的形式报告. 例如，某个 5 岁的男童身高 114.9 厘米，单凭这个数字，还不能说该男童身高究竟如何. 如果同时报告这个高度达到 80% 分位数，就可以说大约有 80%的同龄男孩比他矮，20%的同龄男孩比他高. 一般，第 p 百分位数是这样的一个值，它使得至少有 p%的数据项小于或等于这个值，且至少有(100−p)%的数据项大于或等于这个值. 第 50 百分位数就是中位数. 第 25 百分位数称为**下四分位数**，记为 Q_L. 第 75 百分位数称为**上四分位数**，记为 Q_U.