高等学校现代统计学系列教材

属性数据分析

王静龙　梁小筠　王黎明　编著

Analysis of Categorical Data

高等教育出版社
HIGHER EDUCATION PRESS

高等学校现代统计学系列教

Analysis of Categorical Data

属性数据分析

Shuxing Shuju Fenxi

王静龙　梁小筠　王黎明　编著

高等教育出版社·北京
HIGHER EDUCATION PRESS　BEIJING

内容提要

本书共九章，主要内容包括属性数据，单一属性分类数据，四格表，二维列联表，高维列联表，逻辑斯谛回归模型，对数线性模型，列联表的对应分析，属性数据的贝叶斯统计推断。附录中对教材的部分理论证明做了补充。全书结合统计软件 Excel、Minitab、SPSS 和 SAS，注重统计方法的应用。本书还配有大量的例题，有助于学生巩固所学的属性数据统计分析的方法及应用。

本书可作为高等学校统计学专业本科生和研究生的教学用书，也可作为社会学、心理学、人口学、市场学和医学等领域从事理论研究和应用的统计工作者的参考用书。

图书在版编目(CIP)数据

属性数据分析/王静龙，梁小筠，王黎明编著. --北京：高等教育出版社，2013.7

ISBN 978-7-04-037621-0

Ⅰ.①属… Ⅱ.①王…②梁…③王… Ⅲ.①社会调查-统计分析-统计程序-高等学校-教材 Ⅳ.①C915-39

中国版本图书馆 CIP 数据核字(2013)第 134631 号

策划编辑	张晓丽	责任编辑	徐 可	封面设计	赵 阳	版式设计	余 杨
插图绘制	尹 莉	责任校对	孟 玲	责任印制	尤 静		

出版发行	高等教育出版社	咨询电话	400-810-0598
社 址	北京市西城区德外大街 4 号	网 址	http://www.hep.edu.cn
邮政编码	100120		http://www.hep.com.cn
印 刷	北京宏信印刷厂	网上订购	http://www.landraco.com
开 本	787mm×960mm 1/16		http://www.landraco.com.cn
印 张	16.75	版 次	2013 年 7 月第 1 版
字 数	310 千字	印 次	2013 年 7 月第 1 次印刷
购书热线	010-58581118	定 价	26.50 元

物 料 号 37621-00

高等学校现代统计学系列教材编委会

（按姓氏笔画排序）

总　序

统计学是一门收集、整理和分析数据的科学和艺术。这里的“数据”通指“信息的载体”，涵盖了大千世界中的文本、图像、视频、时空数据、基因数据等。统计学是一个独立的学科，在历史上曾隶属于数学，但统计学与数学有着本质的区别，因此统计学教育有其自身的特点和要求，这些特点表现为：(1) 统计学研究的是随机现象，而数学研究的是确定性的规律；(2) 统计学是一门应用性很强的学科，许多概念和原理来自于实际的需要，不是数理逻辑的产物；(3) 数据在统计学中扮演了重要的角色。目前，统计学已被列为一级学科。

在过去的30年中，随着生命科学、信息科学、物质科学、资源环境、认知科学、工程技术、经济金融和人文科学等众多学科的发展，产生了许多新的统计学分支，如风险管理、数据挖掘、基因芯片分析等。此外，计算机及其有关软件在统计教育和应用中扮演了越来越重要的角色，它们提供了越来越多的图形表达和分析的方法，使得许多原来教科书中重要的内容，现在已变得无足轻重。统计教育必须要改革才能适应高速发展的形势。

大学的统计教育可分为两大类，一类是非统计学专业的课程，另一类是统计学专业的教学设计。非统计学专业的学生学习统计的目的是为了应用，在大学阶段，课程不多，主要是学习基础的统计概念和方法，学会使用统计软件，培养其解决实际问题的能力。统计学专业的课程设置十分重要，应向国际靠拢，对教师队伍的要求也较高。虽然这两类学生的教育有很多共同点，但在课程设置中必须加以区分。

我国的统计教育在过去受苏联的影响很深，把统计学作为数学的一个分支，在内容上偏理论，少应用，过于强调概率论在统计中的作用。统计学是一门应用性很强的学科，应从实际问题、从数据出发，通过统计的工具来揭示数据内部的规律。用“建模”的思路来教统计，使学生能更加容易理解统计的概念和方法，知道如何将实际问题抽象为统计模型，反过来又指导实践。对非统计学专业的学生，要强调统计的应用。学生要能熟练地使用至少一个统计软件包。对于统计学专业的学生，要培养学生对实际问题的建模能力。有些实际问题可直接应用现有的统计方法来解决，如问卷调查的统计分析。有些问题在初次接触时并不像一个统计问题，必须有坚实的统计基础和对实际问题的洞察力，才能从中发掘出统计模型。要培养学生的这种能力及统计思想(统计思想是统计文化的一

部分，是用统计学的逻辑思考问题）。教师在授课中要结合较多的应用例子，要求学生做案例研究，鼓励学生参加建模比赛，参加企业的实际项目。

为满足我国统计教育发展的需要，我们计划编写一套面向高校本科生、特别是一般院校，适用于统计学专业和非统计学专业的系列教材。系列教材的编写宗旨是：突出教学内容的现代化，重视统计思想的介绍，适应现代统计教育的特点及时代发展的新要求；以统计软件为支撑，注重统计知识的应用；内容简明扼要，生动活泼，通俗易懂。编写原则为：(1) 从数据出发，不是从假设、定理出发；(2) 从归纳出发，不是从演绎出发；(3) 强调案例分析；(4) 重统计思想的阐述，弱化数学证明的推导。系列教材分为两个方向，一个面对统计学专业，另一个面对非统计学专业和应用统计工作者。

高等学校现代统计学系列教材是适应形势的要求，由高等教育出版社邀请专家组成“高等学校现代统计学系列教材编委会”负责选题、审稿，由高等教育出版社出版。

以上是我们编写这套教材的背景和理念，希望得到读者的支持，特别是高校领导和教学一线的教师的支持。我们希望使用这套教材的师生和读者多提宝贵意见，使教材不断完善。

高等学校现代统计学系列教材编委会

前　言

属性数据(又称定性数据)分析是统计分析的重要内容,它在实践中有着广泛的应用。华东师范大学金融与统计学院早在20世纪90年代就开设了这门课程。张尧庭教授编写的《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。在教学过程中,我们陆续编写了各个章节的讲义,经过多次修改整理,于2005年在华东师范大学出版社出版了教材《定性数据分析》。尽管那本书内容不够充实,叙述不够深刻,但大家给予我们极大的鼓励,并提出了很多修改建议。之后,《定性数据统计分析》于2008年由中国统计出版社出版,它在之前《定性数据分析》的基础上,增加了对数线性模型与对应分析两方面的内容。这两本书没有讲述贝叶斯统计推断,我们始终感到遗憾。现在这本《属性数据分析》增加了由上海财经大学王黎明教授撰写的属性数据的贝叶斯统计推断的内容。

本书共分九章,第一章介绍属性数据的描述性统计分析方法。第二章介绍单一属性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。第六章介绍逻辑斯谛线性回归模型。第七章介绍对数线性回归模型。第八章介绍对应分析。第九章介绍属性数据的贝叶斯统计推断。本书在选材时,注重统计软件的应用,例如Excel、Minitab、SPSS和SAS等。书中收集了大量可反映属性数据应用问题的例题,也可作为各种统计方法如何运用的示范。本书将正文中的部分理论证明放在附录中,教学时间紧,或只求了解统计方法应用的读者可以跳过去。本书标有星号"*"的某些章节,也可以跳过去。

感谢张尧庭教授,他写的和翻译的上述两本书使得我们对属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金融与统计学院茆诗松教授、诸位同事、历届本科生和研究生。感谢方开泰教授的推荐。感谢高等教育出版社李蕊与张晓丽女士以及徐可先生的关心、支持和辛勤劳动,感谢华东师范大学出版社朱建宝先生与中国统计出版社陈悟朝先生的支持。

本书内容、讲授方法的不当之处恳望大家批评指正。

王静龙　梁小筠

2012年6月

目　录

第一章

属性数据

§1.1 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属性数据，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属性数据，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**属性数据**，或称为定性数据.

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

实际问题中，有时所有的数据都是属性数据或定量数据，有时既有属性数据

又有定量数据. 本书讨论含有属性数据统计问题的分析方法.

类似地,有**定量变量**和**属性变量**. 定量变量中有**计量定量变量**和**计数定量变量**. 属性变量中有**名义属性变量**和**有序属性变量**.

§1.2 属性数据的描述性统计

统计大致可分为两类:描述统计(Descriptive Statistics)与推断统计(Inference Statistics). 本小节着重讲解属性数据的描述统计,也就是讲解如何对属性数据进行整理分析,从中提取有用的信息. 整理分析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 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 1.1 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 1.2 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.

启动 Excel 中文版“数据”菜单上的“数据透视表和数据透视图(P)”命令就可以制作频数频率分布表 1.2,其步骤如下:

1) 建立数据文件,例如将原始数据表 1.1 放在 A 列的第 2 至第 51 个单元格,且在 A 列的第 1 个单元格上输入项目名称“饮料”;

2) 选择“数据”下拉菜单;

3) 选择“数据透视表和数据透视图(P)”选项;

4) 选择:“Microsoft Office Excel 数据列表或数据库(M)”,选择“数据透视表(T)”,选择“下一步”;

5) 在选定区域栏中键入“A1:A51”,选择“下一步”;

6) 选择“现有工作表”,键入“D1”,选择“完成”;

7) 首先将项目“饮料”拖入行字段,然后将“饮料”拖入中间部分.

注:① 在 A 列的第 1 个单元格上必须输入项目名称“饮料”.② 在步骤 5)键入“D1”,意思是说,让输出的频数分布表的左上角在 D 列的第 1 个单元格.

为输入方便,可以用 1、2、3、4 和 5 分别作为可口可乐、苹果汁、橘子汁、百事可乐和杏仁露的代码,将表 1.1 记录的 50 个被访者最喜欢喝的饮料名称转化为 50 个数.这就是名义属性数据.有了名义属性数据,启动 Excel 制作频数频率分布表 1.2 有下面两种方法:

方法一 在将表 1.1 记录的 50 个被访者最喜欢喝的饮料名称转化为名义属性数据之后,仍可使用“数据”菜单上的“数据透视表和数据透视图(P)”命令制作频数频率分布表 1.2,步骤 1 到 7 与前面所述的完全相同,只是多了一个步骤:

8) 在输出的交叉分组列表的左上角上右击鼠标,选择(Field Settings)

字段设置(N),然后在数据透视表字段的对话框的汇总方式(S)的菜单中选择计数.

方法二 启动 Excel 中文版“工具”菜单上的“数据分析”命令也可以制作频数频率分布表 1.2,其步骤如下:

一、建立数据文件.用 1、2、3、4 和 5 分别作为可口可乐、苹果汁、橘子汁、百事可乐和杏仁露的代码.例如将原始数据放在 A 列的第 1 至第 50 个单元格.

二、Excel 要求输入数据分组的情况.因而在例如 B 列的第 1 至第 5 个单元格上依次输入饮料的代码:1、2、3、4 和 5.

三、按下面的顺序制作频数分布表:

1. 选择工具下拉菜单;

2. 选择数据分析选项;

3. 在分析工具框中选择直方图;

4. 在直方图对话框中:

1)在输入区域栏中键入 A1:A50;

2)在接收区域栏中键入 B1:B5;

3)选择输出区域,并在输出区域栏中键入 D1;

4)单击确定.

注:选择了数据分析选项之后,除直方图之外,分析工具框中还有很多的选项,每一个选项都对应着某个数据分析功能.Excel 的数据分析模块有很多的数据分析功能.

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	频数	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,那就需要在调查的时候记录下被调查者的性别.调查得到结果见表 1.4,它是表 1.1 的拓展.

表 1.4 被访者的性别与最喜欢的饮料

男, 可口可乐	男, 百事可乐	女, 杏仁露	女, 杏仁露	男, 可口可乐
男, 杏仁露	女, 苹果汁	男, 可口可乐	女, 可口可乐	男, 可口可乐
男, 可口可乐	男, 橘子汁	男, 可口可乐	女, 可口可乐	男, 苹果汁
女, 可口可乐	女, 百事可乐	女, 苹果汁	男, 百事可乐	男, 可口可乐
男, 橘子汁	男, 杏仁露	男, 可口可乐	男, 可口可乐	女, 杏仁露
女, 橘子汁	女, 橘子汁	女, 杏仁露	女, 橘子汁	男, 百事可乐
女, 百事可乐	男, 百事可乐	男, 橘子汁	女, 杏仁露	女, 杏仁露
女, 可口可乐	女, 杏仁露	男, 可口可乐	女, 橘子汁	女, 苹果汁
男, 苹果汁	女, 苹果汁	男, 百事可乐	女, 苹果汁	女, 杏仁露
男, 可口可乐	男, 可口可乐	男, 可口可乐	女, 苹果汁	女, 杏仁露

根据表 1.4 可制作饮料名称和性别的交叉分组列表,见表 1.5.表 1.5 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 1.5 就是所谓的两种方式分组的交叉表,类似地有三种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.

表 1.5 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

启动 Excel 中文版“数据”菜单上的“数据透视表和数据透视图(P)”命令,除了可以制作频数频率分布表(见表 1.2),还可以制作两种方式分组的交叉表 1.5. 制作交叉表的步骤与频数频率分布表基本相同,只是有下面一些的不同:

① 步骤 1)建立数据文件时,除了在 A 列的第 1 个单元格上输入项目名称“饮料”,并且将饮料名称放在 A 列的第 2 至第 51 个单元格之外,还需在 B 列的第 1 个单元格上输入另一个项目名称“性别”,并且将性别男或女放在 B 列的第 2 至第 51 个单元格;

② 步骤 5)在选定区域栏中键入“a1:b51”;

③ 步骤 7)除了首先将项目“饮料”拖入行字段,还需将将项目“性别”拖入列字段,最后将“饮料”,或“性别”拖入中间部分.

④ 建立数据文件时,如果输入的是代码:1、2、3、4 和 5 分别作为可口可乐、苹果汁、橘子汁、百事可乐和杏仁露的代码;0 和 1 分别作为男性和女性的代码,则多了一个步骤:在输出的交叉分组列表的左上角上右击鼠标,选择字段设置(N),然后在数据透视表字段的对话框的汇总方式(S)的菜单中选择计数.

§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,常在横轴上标记属性数据的每一类别,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 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.

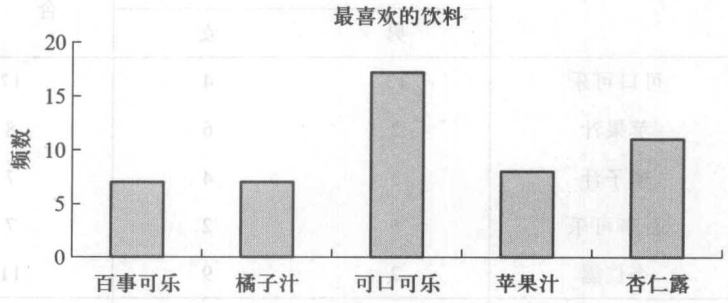

图 1.1 “最喜欢的饮料”的条形图