普通高等教育“十一五”国家级规划教材

数理统计讲义

郑　明　陈子毅　汪嘉冈　编著

博学·数学系列

复旦大学出版社
www.fudanpress.com.cn

普通高等教育“十一五”国家级规划教材

数理统计讲义

郑 明 陈子毅 汪嘉冈 编著

博学·数学系列

复旦大学出版社
www.fudanpress.com.cn

图书在版编目(CIP)数据

数理统计讲义/郑明,陈子毅,汪嘉冈编著. —上海:复旦大学出版社,2006.2
(博学·数学系列)
ISBN 978-7-309-04836-0

Ⅰ.数… Ⅱ.①郑…②陈…③汪… Ⅲ.数理统计-高等学校-教材 Ⅳ.0212

中国版本图书馆 CIP 数据核字(2005)第 140207 号

数理统计讲义

郑 明 陈子毅 汪嘉冈 编著

出版发行 复旦大学出版社 上海市国权路 579 号 邮编 200433
86-21-65642857(门市零售)
86-21-65118853(团体订购) 86-21-65109143(外埠邮购)
fupnet@fudanpress.com http://www.fudanpress.com

责任编辑 范仁梅
总 编 辑 高若海
出 品 人 贺圣遂

印 刷 浙江临安市曙光印务有限公司
开 本 787×960 1/16
印 张 17.25
字 数 290 千
版 次 2007 年 2 月第一版第二次印刷
印 数 4 001—7 100

书 号 ISBN 978-7-309-04836-0/O·352
定 价 25.00 元

前　言

自复旦大学统计学专业建立以来, 数理统计学就始终是本专业的一门基础课程. 随着专业培养目标和教学计划的不断完善, 该课程的内容也在不断地探索、变化和充实. 笔者自 20 世纪 90 年代后期开始在复旦大学管理学院统计学专业讲授这门课程. 在参考使用兄弟院校的相关教材的基础上逐步形成了适应自己专业培养要求的内容. 本书就是根据这些年共同的教学实践而形成的实际使用的教材.

本书假定读者具备高等数学(或数学分析)、线性代数和概率论的知识. 本书先简单介绍了数据描述; 在此基础上介绍了总体、样本和统计量等统计的基本概念, 并将这些概念与概率论的基础知识联系起来给出统计量与抽样分布的概念和实例; 然后叙述了数理统计的基础部分——统计推断(即参数估计和假设检验). 至于统计中的回归分析、线性模型、抽样理论、多元分析等内容将会在专门的课程中讲授.

在本教材的编写和教学中, 我们力求做到以下几点:

(1) 这是统计学专业的入门课程, 所以本书注意问题的实际由来, 既强调统计的基本思想, 同时又介绍实际的做法.

(2) 充分利用已学过的数学基础知识, 阐明统计的概念和原理, 这样会有利于以后进一步学习其他统计课程.

(3) 通过必要的练习帮助读者掌握原理和方法. 所以教材的每一章都配有较多的习题.

本书只是我们过去几年工作的一个汇总. 如何更好地在统计学专业的这一基础课程中反映统计学的不断发展与进步, 正是我们需要努力的. 随着统计学的广泛应

用, 已有越来越多的院系和专业开设这方面的课程, 这就需要有多种不同层次和不同风格的教材来满足广泛的需求. 本书至多只是满足某一需要的一种尝试.

本书初稿完成后, 承蒙应坚刚教授细心审阅, 提出了宝贵的意见. 在此表示衷心的感谢. 最后, 编者诚挚地感谢复旦大学出版社范仁梅女士, 是她的帮助完成了本书的出版.

由于水平有限, 本书虽经不断修改, 但仍会有不少缺点和问题, 编者欢迎读者的批评指正.

编 者

2005 年 8 月

内容提要

本书是一本理学类统计学专业的基础课程教材，书中介绍了数理统计的基本知识和基本理论：首先，简单介绍了数据描述；在此基础上介绍了总体、样本和统计量等统计的基本概念，并将这些概念与概率论的基础知识联系起来，给出了统计量与抽样分布的概念和实例；最后叙述了数理统计的基础部分——数理推断（即参数估计和假设检验）．为帮助读者掌握数理统计的原理和方法，本书的每一章中都配有较多的习题．书后还附有有关表格．

本书可作为统计学专业或相关专业数理统计课程的教材和统计类课程的教学参考书，亦可供上述有关专业的研究生、教师和科研人员阅读参考．

目　录

第一章　基本知识

在中国大百科全书中, 关于数理统计学的描述是

> 数学的一个分科. 研究怎样去有效地收集、整理和分析带有随机性的数据, 以对所考察的问题作出推断或预测, 直至为采取一定的决策和行动提供依据和建议.

联系到本课程对这个描述需要说明的是:

- 数理统计首先是面向数据的. 它不能离开对数据的采集、加工和理解.
- 在数理统计中, 讨论如何收集数据一般是试验设计和抽样调查这两个分支学科的内容, 本课程并不涉及这方面的内容.
- 如何有效地汇总和描述数据, 这是认识数据的第一步, 也是从数据作出推断的基础.
- 数理统计不仅仅限于描述数据, 它要从数据中提取有用的信息, 据此作出推断并为决策和行动提供依据和建议.
- 数理统计依托数据作出推断是基于对数据一定的假定之下进行的, 即数据的来源满足一定的模型: 数据是来自满足一定条件的总体的随机样本.
- 统计推断的任务就是要在关于数据的这些假定之下寻求由样本推断总体的最优方法. 对数据满足不同的模型时可以有不同的最优解. 对不同的推断问题也有不同的解决方法.

§ 1.1　数据描述

一、数据表

由于计算技术的发展, 现今的数据采集、存储、加工和分析都离不开计算机的应用. 所以下面对数据描述的介绍也侧重于计算机对数据表的处理.

1.1.1 在统计分析中, 所面向的数据往往是关于若干个个体 (对象) 的描述, 而对每个个体的描述又是所关心的若干个 (属性) 指标. 例如对学生一般状况的调查结果的数据就是一个班级或不同班级若干个学生的记录, 而每个学生可以记录他的姓名、性别、年龄、身高和体重等等指标. 抽象地看可以将第 i 个对象的 n 项指标记为

$$x_{i1},\ x_{i2},\ \cdots,\ x_{in}.$$

而总共 m 个学生的记录有 $m \times n$ 个数据, 可记为表格的形式, 如表 1.1-1 所示.

表 1.1-1

记录序号	变量 1	变量 2	$\cdots$	变量 n
1	x_{11}	x_{12}	$\cdots$	x_{1n}
2	x_{21}	x_{22}	$\cdots$	x_{2n}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
m	x_{m1}	x_{m2}	$\cdots$	x_{mn}

表 1.1-2 所示是一个班级中 40 个学生的姓名、身高、年龄、性别和体重的信息.

表 1.1-2

姓 名	年龄	身高 (厘米)	性别	体重 (千克)	姓 名	年龄	身高 (厘米)	性别	体重 (千克)
LAWRENCE	17	172	男	78.1	JOE	13	154	男	47.7
JEFFERY	14	169	男	51.3	MARY	15	152	女	41.8
EDWARD	14	167	男	50.8	LINDA	17	152	女	52.7
PHILLIP	16	167	男	58.1	MARK	15	152	男	47.2
KIRK	17	167	男	60.8	PATTY	14	152	女	38.6
ROBERT	15	164	男	58.1	ELIZABET	14	152	女	41.3
JACLYN	12	162	女	65.8	JUDY	14	149	女	36.8
DANNY	15	162	男	48.1	LOUISE	12	149	女	55.8
CLAY	15	162	男	47.7	ALICE	13	149	女	48.6
HENRY	14	159	男	54	JAMES	12	149	男	58.1
LESLIE	14	159	女	64.5	MARIAN	16	147	女	52.2
JOHN	13	159	男	44.5	TIM	12	147	男	38.1
WILLIAM	15	159	男	50.4	BARBARA	13	147	女	50.8
MARTHA	16	159	女	50.8	DAVID	13	145	男	35.9
LEWIS	14	157	男	41.8	KATIE	12	145	女	43.1
AMY	15	157	女	50.8	MICHAEL	13	142	男	43.1
ALFRED	14	157	男	44.9	SUSAN	13	137	女	30.4
CHRIS	14	157	男	44.9	JANE	12	135	女	33.6
FREDRICK	14	154	男	42.2	LILLIE	12	127	女	29.1
CAROL	14	154	女	38.1	ROBERT	12	125	男	35.9

表 1.1-3 所示是在计算机中显示的类似的数据.

表 1.1-3

5	Nom	Int	Int	Nom	Int
40	NAME 姓名	AGE 年龄	HEIGHT 身高（厘米）	SEX 性别	WEIGHT 体重（公斤）
1	LAWRENCE	17	172	男	78.1
2	JEFFERY	14	169	男	51.3
3	EDWARD	14	167	男	50.8
4	PHILLIP	16	167	男	58.1
5	KIRK	17	167	男	60.8
6	ROBERT	15	164	男	58.1
7	JACLYN	12	162	女	65.8
8	DANNY	15	162	男	48.1
9	CLAY	15	162	男	47.7
10	HENRY	14	159	男	54

1.1.2 数据文件　记录各个观测的每项属性的数据在计算机中以特定格式的数据文件存放, 通常可以直观地将它当作一个矩形数据表. 在不同的场合这些数据文件有不同的名称. 在数学上, 它就是一个矩阵, 在数据库的术语中称为表, 在 SAS 软件中称为数据集, 在统计中它就是对样本观测结果的记录. 这个矩阵的每一行, 也称为一条记录或观测, 它记录了一个观测对象的各项特性的指标值. 例如, 在上述关于学生信息的数据表 1.1-2 中的第二行就记录了名为 Jeffery 的学生的姓名、年龄、身高、性别和体重的数据. 矩阵的每一列, 也称为字段或变量, 它记录了所有观测对象某一项特性的指标值. 例如, 表 1.1-2 中的第三列就依次记录了所有学生的身高数据.

在计算机存储的数据表中, 一般要求同一列的数据有相同的属性, 例如表示性别的一列单元格中的数据都只能用字符 “男” 或 “女” 记录, 表示身高的一列数据都只能是数.

1.1.3 表的拼接　在生产和社会实践中, 为了便于快速收集和存储数据, 同一对象的不同特性往往存放在不同的数据表中. 而统计分析常常需要将同一对象的各方面的特性汇集在同一张数据表中. 例如, 一个单位员工的履历、健康状况、工作业绩和工资收入可能存放在不同部门的不同数据表中, 若需要利用这些资料进行统计分析, 就需要将来自不同部门的资料整合到同一张数据表中. 这往往是十分耗时的任务, 但却是必须的任务.

1.1.4 缺失值　在数据的收集过程中, 有时无法得到所关心对象的所有项目的记录.

例如, 在问卷调查中, 涉及个人收入和财产的信息往往无法获得. 这时在记录所有结果的数据表中有些记录的个别字段无法填入. 通常用约定的特殊符号表示. 也称为缺失值(missing value). 在数据处理和分析中给以特殊的处理.

1.1.5 变量的类型 数据表的变量按其本身的含义及计量的不同, 可分为定量的和定性的, 如图 1.1-1 所示.

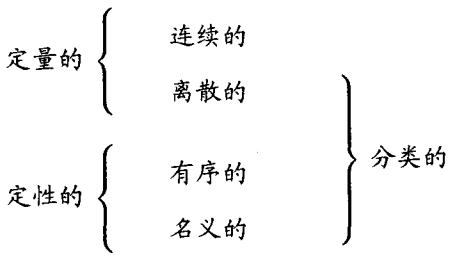


图 1.1-1 变量的类型

• 定量的(quantitative): 例如身高、体重、产量、产值、年龄等. 这类变量的取值是一个数量, 并且对它的数值进行算术运算有实际的含义. 定量变量在计算机中必须以数值存放.

对定量的变量还可进一步分为:

◇连续的(continuous). 这类变量可能取的值可以是充满整个区间, 例如身高、体重、产值等. 也称它为区间的(interval) 或实型的(real).

◇离散的(discrete). 这类变量只能取到有限个数值, 例如年龄、事故次数、每周的第几个交易日等.

在一个具体问题中, 有时一个定量的变量既可确定为连续的也可确定为离散的, 例如, 年龄就是这样. 这往往依赖于一个变量在分析中的作用. 一般地, 在分析中将一个数值型变量按其值进行分类时就应该是离散的. 而将一个变量的取值作为分析对象并对其进行运算时就应该是连续的.

• 定性的(qualitative): 例如, 性别、省份、品种、产品型号等. 这类变量的不同取值有不同含义, 对其不同的值无法进行算术运算或没有实际含义. 在计算机中这类变量的值可以直接用表示其含义的字符存放, 也可经过编码后以数值存放.

对定性的变量还可进一步分为:

◇名义的(nominal). 例如, 性别、省份、品种等. 这些变量所取的值之间没

有自然的次序关系.

◇有序的(ordinal). 例如, 型号 (特大、大、中、小), 疗效 (恶化、无效、稍有改善、大有改善、痊愈) 等.

对变量的上述分类还和数据分析中采用的方法密切相关. 从数据统计分析采用的方法来看, 将离散的、名义的和有序的 3 类合称为分类的(categorical) 或属性的.

许多常用的方法都是适用于连续型的变量. 但当因变量为分类变量时这些常用的基于对变量取值进行算术运算的统计方法就不适用, 于是必须使用分类变量的统计方法. 对分类型变量, 不论它是离散的或有序的, 在分析中若不利用变量取值的次序信息, 就可将其作为名义型的, 若利用了次序信息, 就可将它作为有序型的.

二、频数统计

1.1.6 频数统计 数据表中包含了多个变量 (属性) 的观测结果. 首先从单个变量的描述分析开始. 设取出某一个变量的 n 次观测结果:

$$x_1,\ x_2,\ \cdots,\ x_n.$$

为了描述这些数据, 首先需要回答的是它们取了什么值和取各个不同值的比例——这两者合起来也称为这些观测值的分布. 表 1.1-4 就是表 1.1-2 中 40 个学生的性别和年龄数据取值的分布.

表 1.1-4

性 别				
SEX	Frequency	Percent	Cumulative Frequency	Cumulative Percent
男	22	55.00	22	55.00
女	18	45.00	40	100.00
年 龄				
AGE	Frequency	Percent	Cumulative Frequency	Cumulative Percent
12	8	20.00	8	20.00
13	7	17.50	15	37.50
14	12	30.00	27	67.50
15	7	17.50	34	85.00
16	3	7.50	37	92.50
17	3	7.50	40	100.00

在上述第二张关于年龄分布的表中, 其第二列频数 (frequency) 就表示整个 40 个学生的数据中, 12 岁的学生有 8 个, 13 岁的有 7 个等等.

第三列百分数 (percent) 由第二列的频数除以数据总数 40 并化为百分数而得到: $8/40 \times 100 = 20$, $7/40 \times 100 = 17.5$ 等等.

第四列累计频数 (cumulative frequency) 由第二列频数累加得到: $8 + 7 = 15$, $8 + 7 + 12 = 27$ 等等. 由累计频数马上可以看出不超过 13 岁的学生有 15 名, 不超过 14 岁的学生有 27 名等等.

第五列累计百分数 (cumulative percent) 由累计频数除以样本容量 40 并化为百分数而得到: $8/40 \times 100 = 20$, $15/40 \times 100 = 37.5$ 等等. 由累计频数可以看出不超过 13 岁的学生占 37.5%, 不超过 14 岁的学生占 67.5% 等等.

虽然频数统计是一个简单的计数过程, 但通过它可以了解观测数据中变量取值的分布, 并为进一步的分析提供依据. 此外, 从样本取值的分布中也可发现一些明显不合理的例外记录. 因此, 进行频数统计也是数据处理中重要的一步.

变量取值的信息除了上面提到的各个变量取值频数的统计外, 还可以进行多个变量取值组合的频数统计. 例如, 对班级信息的数据可同时考虑男生和女生不同年龄的分布, 即构成按两种或多种方式分类的频数表.

1.1.7 柱状图 对于频数统计, 除了用列表记录变量取不同数值时的频数、百分数外, 还常用柱状图或饼图等图形工具来表示. 柱状图也称条形图. 在垂直柱状图里,

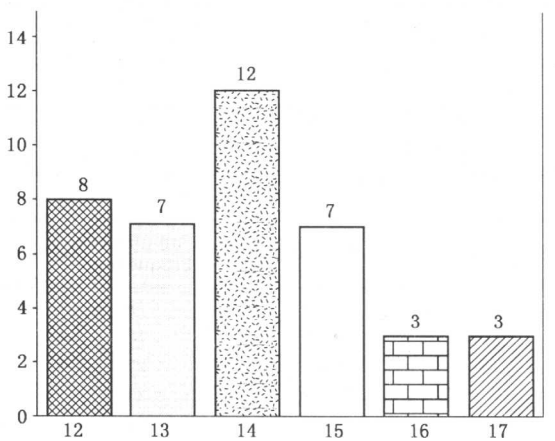
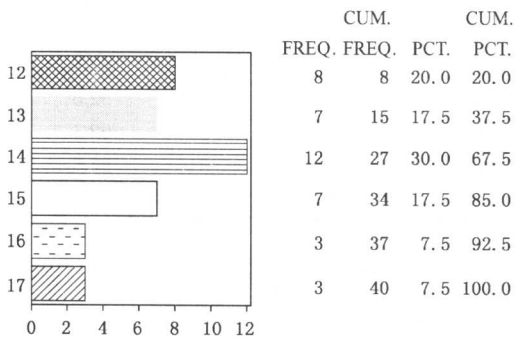


图 1.1-2 年龄分布的垂直柱状图

有多个宽度相同的柱并列, 对变量取到的每个值, 都用一个柱描绘. 柱的高度表示频数 (或百分数). 所以从柱的不同高度可以对变量取值的频数分布有一个整体的印象. 由于不同类别的频数和百分数只差一个公共的因子, 所以用频数或百分数作柱的高度, 两种柱状图的形态是相似的. 图 1.1-2 就是相应班级年龄数据的柱状图 (以频数为柱的高度).

与垂直柱状图类似的还有水平柱状图和饼图, 如图 1.1-3 所示.



(a) 年龄分布水平柱状图　　(b) 年龄分布饼图

图 1.1-3　年龄分布的水平柱状图和饼图

1.1.8 直方图　对于区间型变量, 尤其是它可能取到很多不同的数值时, 它取到特定数值的频数就不很多. 例如, 对表 1.1-2 中体重计算它的频数分布时, 可得到如表 1.1-5 所示的体重频数表.

表 1.1-5

体重	频数	百分数	累计频数	累计百分数
29.1	1	2.5	1	2.5
30.4	1	2.5	2	5.0
33.6	1	2.5	3	7.5
35.9	2	5.0	5	12.5
36.8	1	2.5	6	15.0
⋮	⋮	⋮	⋮	⋮
55.8	1	2.5	33	82.5
58.1	3	7.5	36	90.0
60.8	1	2.5	37	92.5
64.5	1	2.5	38	95.0
65.8	1	2.5	39	97.5
78.1	1	2.5	40	100.0