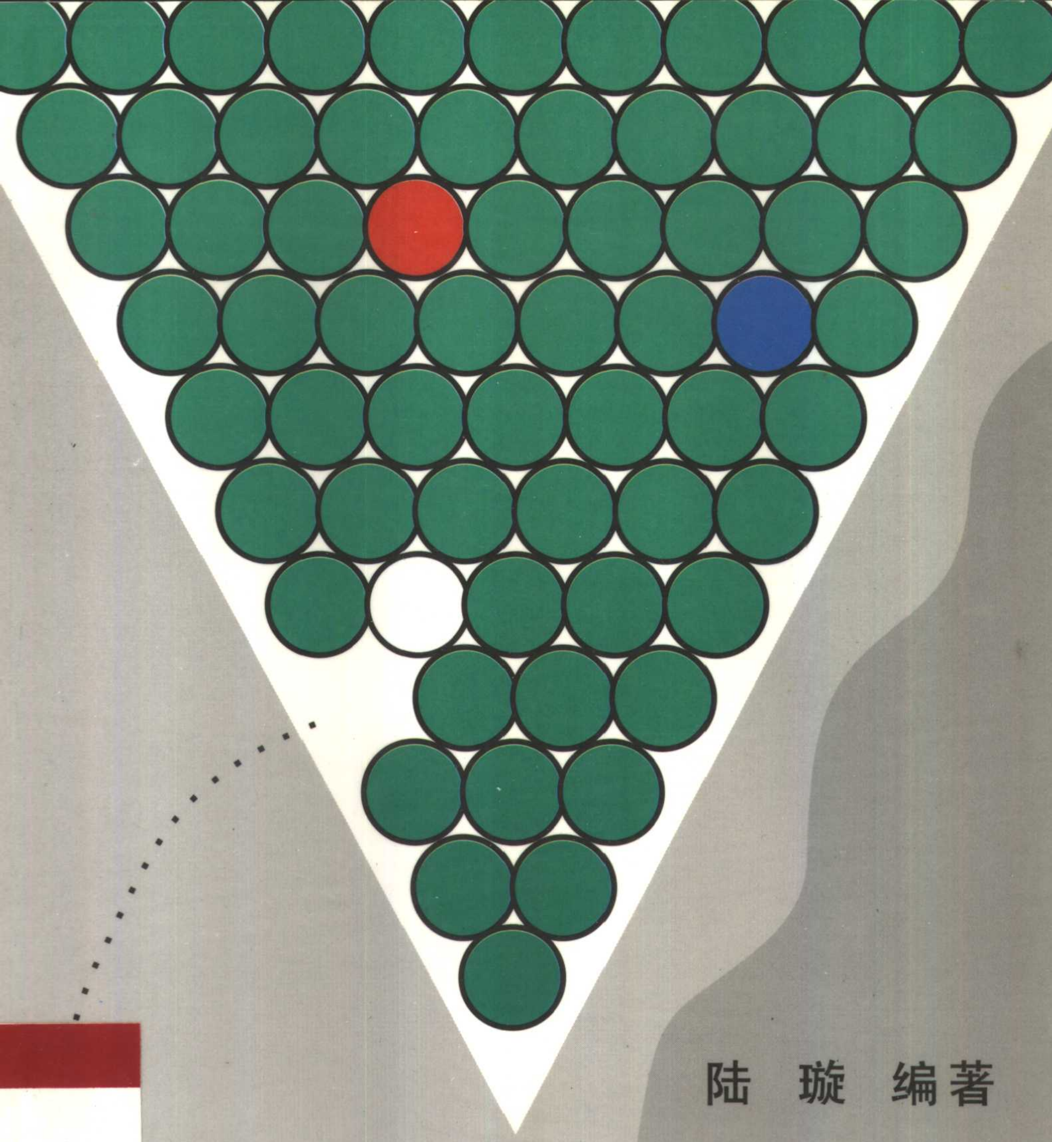

陆璇 编著

# 数理统计基础

清华大学出版社
http://www.tup.tsinghua.edu.cn

(京)新登字 158 号

## 内 容 简 介

本书在撰写中力图兼顾理论和应用两个方面，深入浅出地介绍数理统计学的基本概念和方法。全书内容包括：统计模型与统计量、点估计、区间估计、假设检验、实用线性模型(方差分析与线性回归分析)。在每章后安排了一些精选的习题及供阅读的补充材料。最后附了一些常用的统计数表，以备读者查用。

本书适合高等学校数学专业和应用数学专业作为本科生教科书，也可作为相关专业的本科生或研究生教材。

**图书在版编目(CIP)数据**

数理统计基础/陆璇编著 —北京：清华大学出版社，1998.10
ISBN 7-302-02990-3

Ⅰ.数… Ⅱ.陆… Ⅲ.数理统计 Ⅳ.O212

中国版本图书馆 CIP 数据核字(98)第 13361 号

出版者：清华大学出版社（北京清华大学学研楼，邮编 100084）
http://www.tup.tsinghua.edu.cn
印刷者：北京市清华园胶印厂
发行者：新华书店总店北京发行所
开　本：850×1168 1/32　印张：9.5　字数：326 千字
版　次：1998 年 9 月 第 1 版　2000 年 3 月 第 2 次印刷
书　号：ISBN 7-302-02990-3/O·194
印　数：4001～6000
定　价：11.00 元

# 序 言

统计学(statistics)是一门研究从随机数据中获取信息、发现规律并指导决策的数学方法的科学。在科学研究中,用统计学方法从数据中所获得的信息和发现的初步规律往往成为重大科学发现的先导。统计学方法在自然科学和工程技术研究的许多领域内都得到广泛的应用,例如生物学、遗传学、医药学、地质学、遥感技术、语音识别、等等,我们可以列出长长的一张表。在社会科学研究领域中,例如在人口调查、社会统计、心理学、教育学、保险科学、以及金融工程等领域中,统计方法的应用也极其广泛。对社会数据进行统计分析后所得到的结论也广泛地被政府部门和大公司作为支持决策的依据。统计学本身的发展也离不开对科学研究的参与。经典统计学的奠基人之一费歇尔(R. A. Fisher),就是在进行农业科学试验时总结出方差分析和试验设计等统计模型与方法,而这些方法又在现代工业的质量控制中有着重要的应用并得到发展。计算机的普遍使用不仅使统计学方法的应用范围日益广泛,而且为统计学本身的发展提供了新的生长点。综上所述,我国现代化建设事业的发展不仅需要大量的统计专门人才,而且要求在各个领域内都有懂得统计学,善于用统计方法来解决实际问题的人才。

为培养各种类型的统计研究型和应用型人才,需要有适合于不同培养目标和不同基础的学生使用的教科书。本书的目标是兼顾使学生打好扎实的理论基础和培养初步应用能力两个方面。它适合于作为数学、应用数学或对统计基础要求较高的其它专业的本科大学生数理统计课的教学用书,也适合于作为有一定数学基础的读者学习统计学的自学用书。学习本书所需要的先修课程为

数学分析(高等数学)、线性代数和初等概率论。

理论与实际相结合是本书写作的基本原则,在这个基本原则下,本书有以下几个特点。首先,在内容取舍上,优先考虑那些既有深刻的理论意义,又有重要实用价值的统计概念和方法。例如,极大似然方法是在参数模型下最重要的一种统计方法,在参数估计(包括点估计和区间估计)和假设检验中有着极其重要的地位。因此,本书中的相当大的篇幅是讨论极大似然方法的。其次,某些基本结论(例如极大似然估计的渐近正态性)由于证明起来比较困难(需要较为高深的数学理论),故而在许多初等的统计学教科书中常加以回避。而在一些高等的统计学教科书中,这些结论的繁复证明又往往使初学者和实际工作者望而生畏。考虑到这些结论的理论重要性和广泛的应用价值,在本书中采用如下的折中方案:在给出结论的同时,尽可能地给出一个虽然不太严格、但有助于学生领会证明思路的形式上的"证明"。其目的在于使学生"知其然",同时又在一定的深度上"知其所以然"。第三,为了促进学生进一步自学、思考,在本书的每章的最后一节,不仅给出基本习题,还加入了一些补充材料。这些补充材料的阅读应该与作习题一起,构成课外练习的有机组成部分。

本书是作者多年在清华大学应用数学系进行数理统计教学的经验和体会的基础上经过充实、取舍而形成的,全书共分五章。前四章分别介绍统计模型的基本概念,及统计推断的三大问题(参数点估计、区间估计和假设检验)。在内容选择上以参数统计为主体,适当地加入一些非参数统计的内容。第五章介绍两个应用范围十分广泛、且理论上有紧密联系的实用统计模型:方差分析与线性回归分析。在使用本书作为教学用书时,要在一个学期的课堂教学中全部讲完全书,内容略显多了一些。但是从保证内容的系统性和完整性的角度来看,作者认为目前这些内容是适宜的、必需的。在实际用本书进行教学时,教师可根据情况作必要的取舍。根据作者的

经验，如果采用教师在课堂上讲重点、难点，其余内容由学生课外阅读的方法进行教学，在一个学期中学完全部内容也是完全可以做到的。

作者在进行本书的撰写时，试图将它写成一本“统计味”较浓的教科书。实际效果究竟如何，有待于教学实践的检验。作者衷心希望专家学者、教师、学生和各界读者对本书提出批评和建议。限于作者的水平，在本书中不可避免地存在疏漏之处，希望读者予以指正，以便将来有可能再版时予以纠正。

# 目　录

# 第1章 统计模型与统计量

统计学(statistics)分析处理的对象是带有随机性的数据(data). 由于来源的广泛性、多样性及获取方法的不同,随机数据的形态是很丰富的. 为对数据进行定性、定量的分析处理,首先要对它们建立一定的数学模型,在此基础上再选用适当的数学方法进行整理加工. 概率论是研究随机现象的数学理论,因此,在统计学中将随机数据看成是有一定分布的随机变量,这样建立起来的数学模型就是统计数学模型,或简称为统计模型.

## 1.1 随机数据

本节介绍几个随机数据的例子,以使读者先获得一些感性认识.

### 1.1.1 随机数据的例子

**例 1.1.1** 为检验某灯泡厂生产的产品的质量,从该厂库存的一大批灯泡中随机地抽取 10 个,检测其寿命(单位 h). 假定测得的寿命为

$$1980,2800,3060,4500,2760,3270,1560,0,3200,1940.$$

这组数据就是随机数据,因为这 10 个供检测的样品灯泡是从一大批灯泡中随机地抽取的. 这里"随机地抽取"的含义,在 1.3 中将予以回答. 这 10 个样品虽然是一大批灯泡中的很小一部分,但它们的寿命长短却可以在一定程度上反映整批灯泡寿命的分布情况.

**例 1.1.2** 假定按照某种规定的标准，灯泡的寿命在 3000 h 以上为正品，而在 3000 h 以下为次品. 同样是上面的 10 个样品，其正品或次品的记录为

次品，次品，正品，正品，次品，正品，次品，次品，正品，次品.

这样一组记录也是一组随机数据. 这组数据与例 1.1.1 中的数据形态不同，所包含的信息也不同.

**例 1.1.3** 从某个城市的居民中随机地选取 500 人，调查他们的年龄、职业、受教育程度及年收入. 由于数据量很大，不可能将这组数据都列出来，只列出前 5 个人的调查结果于表 1.1.1. 从表中可以看到，对每个人调查 4 项指标(变量)，其中两个(年龄和年收入)为数量化指标，另外两个(职业和受教育程度)为非数量化指标. 对非数量化指标如何用数学的方法来分析处理，将在 1.2 节中回答这个问题. 由于这 500 个被调查的居民是从一个城市的居民中随机地选取的，因此这一组数据是随机数据. 通过对这 500 个居民的调查结果，可以对整个城市居民的上述四项指标的分布与关系有一定程度的了解.

**表 1.1.1 城市居民调查表**

| 编号 | 年龄 | 职业 | 教育程度 | 年收入/万元 |
| --- | --- | --- | --- | --- |
| 1 | 42 | 干部 | 高中 | 0.82 |
| 2 | 35 | 工人 | 初中 | 0.63 |
| 3 | 50 | 医生 | 大学 | 0.98 |
| 4 | 47 | 工人 | 小学 | 0.78 |
| 5 | 36 | 教师 | 大学 | 0.85 |

**例 1.1.4** 为研究甲醛与尿素的反应时间对所生成树脂强度的影响，在三个反应时间：80 min、100 min 及 120 min 分别做三次

试验，记录下试验结果，列于表 1.1.2. 假定树脂强度为反应时间的未知函数. 由于在工业生产中进行试验，各种条件(例如试验温度、原料的纯度、原料混合的均匀程度等)不一定能掌握得十分稳定，因此在同样的设计条件(反应时间)下所得到的结果会有围绕着某个期待值(未知的理论值)的随机浮动. 由表 1.1.2 可看到，在同一反应时间下的三次试验的结果都不同. 因此所得到的数据应视为随机数据. 从这些试验数据中可以分析出甲醛与尿素的反应时间与生成树脂的强度之间的关系.

**表 1.1.2 树脂强度试验结果**

| 反应时间/min | 强度/(kg/cm) |
|---|---|
| 80 | 7.8，8.2，7.4 |
| 100 | 10.7，11.1，12.3 |
| 120 | 10.3，9.5，9.8 |

以上几个例子提供了随机数据的一些常见形态. 它们的来源、维数、数量性质(数量、非数量)等都有不同. 在实际问题中，统计学处理的大量数据是多维的，但本书的内容是介绍数理统计理论基础，因此，今后所讨论的数据主要是一维数据. 统计学的第一个任务就是要将各种不同形态的随机数据纳入有一定统一性的数学模型之中，在此基础上才能对随机数据用适当的数学方法进行分析，从中获取信息，进行推断，得出科学的结论或支持决策.

## 1.1.2 数据的简单处理

远在数理统计的理论体系建立起来之前，人们就在广泛地使用一些直观易行的方法来处理数据，从中获得有用的信息. 这里介绍几种常用的对单一来源的一维数据作简单加工处理的方法. 其中有些方法将在后面的内容中作深入的研究；限于篇幅，其它的方

法(虽然也是有理论上的优良性质)将不再涉及.

首先,对一组数据,我们希望有一种直观明了的方式来表示它们的散布状况,常用的方法是画“直方图”(histogram). 设数据为 $x_1,x_2,\cdots,x_n$. 首先,取这组数据的最小值 $x_{n1}=\min\{x_1,\cdots,x_n\}$和最大值 $x_{nn}=\max\{x_1,\cdots,x_n\}$. 于是区间$[x_{n1},x_{nn}]$就界定了这组数据所在的范围. 区间的长度为 $r=x_{nn}-x_{n1}$,称为“极差”(range). 然后,将某个包含$[x_{n1},x_{nn}]$的适当区间,根据数据量的大小划分成若干个等长的小区间. 假定得到 $k$ 个小区间. 对第 $j$ 个小区间,计数其中数据的个数 $n_j$, $j=1,\cdots,k$. 则数据落在 $j$ 个小区间中的“频率”为 $f_j=n_j/n$, $j=1,\cdots,k$. 根据这组频率值,我们可以在数轴上画直方图:在第 $j$ 个小区间上画一矩形,其高度为 $f_j$. 这样一组高低不等,相互联接的矩形就构成直方图. 例如,将例 1.1.1 中的数据在区间[0,5000]上分成四组,分别在四个子区间上:[0,1250],(1250,2500],(2500,3750],(3750,5000]. 数据散布的频率如下表:

| 区间 | [0000,1250] | (1250,2500] | (2500,3750] | (3750,5000] |
|---|---|---|---|---|
| 频率 | 0.1 | 0.3 | 0.5 | 0.1 |

根据上表画出的直方图如图 1.1.1 所示. 在这个例子中,由于数据量较小,因此画出的直方图较为粗糙. 但是可以清楚地由图中看出,这 10 个数据在区间(2500,3750]上最集中,在两端的两个区间上最稀疏,在小于 0 和大于 5000 上则没有分布. 如果数据量大一些的话,我们可以把区间划得更多、更细一些,画出的直方图可以更细致地反映数据散布的特点.

进一步,我们还可以构造刻画数据“中心位置”的量. 首先是数据的“均值”(mean) $\bar{x}=n^{-1}\sum_i x_i$ .

另一个刻画数据中心位置的量是“中位数”(median):将数据 $x_1,\cdots,x_n$ 依从小到大的顺序重排后,记为 $x_{n1}\leqslant x_{n2}\leqslant\cdots\leqslant x_{nn}$,则数

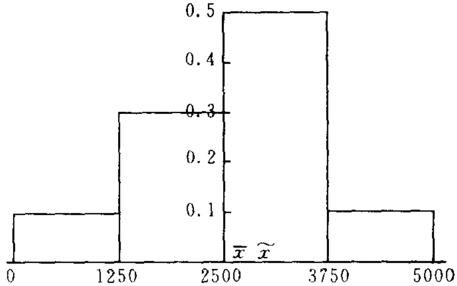


图 1.1.1　直方图

据的中位数定义为

$$\tilde{x}=\begin{cases}\{x_{ni},\ i=1,\cdots,n\}\ \text{的中间一个}, & \text{当 } n \text{ 为奇数};\\ \{x_{ni},\ i=1,\cdots,n\}\ \text{的中间两个的算术平均}, & \text{当 } n \text{ 为偶数}.\end{cases}$$

对例 1.1.1 中的数据，$\bar{x}=2507$. 为计算 $\tilde{x}$，我们将这组数据重排如下：

0，1560，1940，1980，2760，2800，3060，3200，3270，4500

现在 $n=10$，位于中间的两个数据为 2760 和 2800，因此 $\tilde{x}=(2760+2800)/2=2780$. 由图 1.1.1 可以看出，$\bar{x}$ 和 $\tilde{x}$ 确实处于数据分布的中心位置. 除了刻画数据的中心位置的量之外，还有用来刻画数据的散布度的量. 数据的极差 $r$ 可以刻画数据散布范围的大小，但它不能刻画数据在这个范围内散布的集中或离散的程度. 常用的刻画数据的散布度的量为

$$\hat{\sigma}^2=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2=\frac{1}{n}\sum_{i=1}^{n}x_i^2-\bar{x}^2,$$

或

$$s^2=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2=\frac{n}{n-1}\hat{\sigma}^2$$

$\hat{\sigma}^2$ 和 $s^2$ 都称为数据的“方差”(variance)，二者之间只有微小区别. $\hat{\sigma}^2$ 的含义容易解释. $(x_i-\bar{x})^2$ 可解释为单个数据 $x_i$ 对于整组数据

的中心位置 $\bar{x}$ 的偏离度，这个量越大则表示 $x_i$ 对于 $\bar{x}$ 的偏离度越大，反之，这个量越小则表示 $x_i$ 对于 $\bar{x}$ 的偏离度越小. $\hat{\sigma}^2$ 就是所有数据对于 $\bar{x}$ 的偏离度的算术平均，代表一组数据整体上的离散度，或平均离散度. 至于 $s^2$ 与 $\hat{\sigma}^2$ 的微小区别在于其平均加权系数由 $n^{-1}$ 修正为 $(n-1)^{-1}$. 这样作的理由，读者通过学习以后的内容就会知道. 在实际使用时，我们用 $\hat{\sigma}=\sqrt{\hat{\sigma}^2}$ 或 $s=\sqrt{s^2}$（都称为数据的“标准差”(standard deviation)）来反映数据的离散度. 在例1.1.1的数据中，$s=384$，$\hat{\sigma}=364$. 现在我们来看另一组数据（读者可以把这组数据想象成另一组灯泡的寿命）：

3030，2760，2840，2990，3210，2950，2530，3080，2970，3030.

直观地可以看出，这组数据的离散度比第一组数据的小，或者反过来说，它的集中程度比第一组数据的大. 实际算出这组数据的 $s=190$，$\hat{\sigma}=180$. 显然，这两个值比第一组数据的相应值要小很多.

另外一种刻画数据的离散程度的方法是利用“四分位数”(quartile). 将数据依从小到大的顺序重排之后，分成数目相等的四份（每份各占四分之一的数据量）. 从左向右，依次处于第 1、第 2 第 3 个分位点上的数据（或处于分位点两侧的数据的算术平均）称为第 1、第 2、和第 3 四分位数，通常记为 $q_1$，$q_2$，$q_3$. 显然，$q_2=\tilde{x}$. 比如，在 例 1.1.1 中共有 10 个数据，因此第 1 四分位数为数据中第 3 小的，而第 3 四分位数为数据中第 8 小的（第 3 大的），即 $q_1=1940$，$q_3=3200$. 在区间 $[q_1, q_3]$ 内集中了 50% 的数据. 因此，$q_3-q_1$ 的大小可以用来度量数据集中、离散的程度. 这个值大，说明数据的离散度大；反之，这个值小，说明数据的离散度小.

以上我们介绍了几种常用的对单一来源的一维数据作简单处理的方法. 其中的手法有重新排序，作直方图，计算一些简单的、刻画数据的某一方面特性的量，等等. 这些方法所产生的结果统称为“统计量”(statistics). 到目前为止，从统计量中所获得的各种信息

还只看成是反映数据本身的一些数学特征.但是,更重要的是,由于数据本身在一定程度上包含了它的来源中的信息(被测试的灯泡的寿命中包含了一大批灯泡的寿命分布状况的信息,被调查的市民的数据中包含了一个城市中居民全体的信息,等等),我们用上述方法从数据中所获得(集中、浓缩)的信息也在一定程度上反映了数据来源的某一或某些方面的数学特征.由于搜集数据、分析数据的目的根本在于要研究产生数据的客体,因此,将由数据中获得的信息转化、解释为产生数据的客体的统计数学特征,其意义远远超过单纯解释数据本身.为此就需要对数据建立一定的统计数学模型,这些模型是用来刻画产生数据的客体的统计数学特征的.模型中存在着一些未知因素.通过对数据的分析来对这些未知因素进行推断,从而达到对模型中未知因素的(一定程度的)认识,最终解释所研究的客体,这就是统计分析的最主要目的.

## 1.2 样本与样本分布

### 1.2.1 总体与总体分布

随机数据的来源是在统计分析问题中的研究对象的全体所构成的集合.这样的集合就称为总体(population).总体中的每个元素称为个体,是数据的载体.当总体中个体的数目有限时,该总体称为有限总体;否则就称为无限总体.对总体中的每个个体,有一个或多个刻画其特性的(数量的或非数量的)变量,是统计学真正研究的对象.以后,我们用 $X$ 记这个(或这些)变量.对应总体中的不同的个体,$X$ 的值是不同的,因此 $X$ 是定义在总体上的函数.以后,我们就称 $X$ 为总体变量,或就简称为总体.

总体变量 $X$ 具有一定的分布,称为总体分布.原则上,这个分布应由对应 $X$ 在一定范围内取值的个体数占总体中个体总数的