高等统计学

郑忠国 童行伟 赵慧 编著

北京大学出版社
PEKING UNIVERSITY PRESS

高 等 统 计 学

郑忠国　童行伟　赵慧　编著

北京大学出版社
PEKING UNIVERSITY PRESS

图书在版编目(CIP)数据

高等统计学/郑忠国，童行伟，赵慧编著．—北京：北京大学出版社，2012.2

ISBN 978-7-301-16096-1

Ⅰ．①高… Ⅱ．①郑… ②童… ③赵… Ⅲ．①统计学–研究生–教材 Ⅳ．①C8

中国版本图书馆 CIP 数据核字(2012)第 016397 号

书　　名：高等统计学
著作责任者：郑忠国　童行伟　赵慧　编著
责 任 编 辑：潘丽娜
标 准 书 号：ISBN 978-7-301-16096-1/O·0805
出 版 发 行：北京大学出版社
地　　址：北京市海淀区成府路 205 号　100871
网　　址：http://www.pup.cn　电子邮箱：zpup@pup.pku.edu.cn
电　　话：邮购部 62752015　发行部 62750672　编辑部 62752021
出版部 62754962
印 刷 者：北京世知印务有限公司
经 销 者：新华书店
890 毫米×1240 毫米　A5　13.5 印张　530 千字
2012 年 2 月第 1 版　2012 年 2 月第 1 次印刷
定　　价：32.00 元

未经许可，不得以任何方式复制或抄袭本书之部分或全部内容.
版权所有，侵权必究
举报电话：010-62752024　电子邮箱：fd@pup.pku.edu.cn

内 容 简 介

统计学是研究数据处理方法的学科，本书系统介绍了统计学中的基本概念和理论. 首先介绍如何将社会各行各业中出现的数据系统化成统计模型，成为统计学处理的对象；然后介绍统计学中常用的统计方法和相关理论，为学生将来进行统计研究或实际工作打下一个坚实的基础.

由于统计学发展迅速，内容广泛，难以全面介绍，本书突出重点，着重介绍经典数理统计学中的主要内容，全书分成五章：第一章介绍数据的统计模型以及相应统计量的性质和指数分布族；第二章介绍参数估计方法；第三章介绍大样本方法；第四章介绍统计检验方法，第五章介绍贝叶斯分析方法以及相关的统计决策模型.

本书可作为高等院校概率统计学专业的硕士学位基础课教材，也可作为医学、经济、金融、社会学等相关专业的博士生教材. 此外，其他从事科学研究和实际数据处理的人员，比如，各科研单位的科研工作者，工程、医药卫生、商业、银行、保险业等行业的从业人员也可将本书作为参考资料.

作 者 简 介

郑忠国　北京大学数学科学学院教授、博士生导师, 1962 年毕业于北京大学数学力学系, 1965 年北京大学研究生毕业. 长期从事数理统计的教学和科研工作, 研究方向是非参数统计、可靠性统计以及统计计算. 发表论文近百篇. 主持完成国家自然科学基金项目“不完全数据统计理论及其应用(1999—2001)”, 教育部博士点基金项目“应用统计方法研究”和“工业与医学中的应用统计研究”等. 研究项目“随机加权法”获国家教委科技进步二等奖. 出版的教材有《高等统计学》(北京大学出版社, 1998)和《概率与统计》(北京大学出版社, 2007)等.

童行伟　北京师范大学数学科学学院副教授、硕士生导师, 2003 年毕业于北京大学数学科学学院, 获博士学位. 研究方向是生物统计和金融统计. 发表论文二十余篇. 主持国家自然科学基金项目“面板技术数据的统计推断以及应用(2010—2012)”, 教育部科技研究重大项目“多元经济数据中的因果分析, 经验似然和 Copula 方法(2009—2011)”.

赵慧　华中师范大学数学与统计学学院副教授、硕士生导师, 2005 年北京大学数学科学学院博士毕业, 2007 年中国科学院数学与系统科学研究院博士后出站. 近年来从事数理统计的教学与科研工作, 研究方向为图模型与因果推断、生存分析与可靠性. 在国内外期刊上已经发表论文十余篇.

试读结束：需要全本请在线购买： www.ertongbook.com

前　言

自 20 世纪 80 年代以来, 我国统计学进入蓬勃发展时期. 在此形势之下, 北京大学为概率统计专业硕士研究生开设了硕士学位课程. 高等统计学就是学位课程中三门必修课之一. 在这门课程的讲稿的基础上, 于 1998 年出版了《高等统计学》. 经过 10 年的教学实践, 我们在此基础上, 扩充成本书.

高等统计学是概率统计方向硕士研究生的一门专业基础课, 因此本书在编写中特别强调了统计学中的基本概念和基本理论, 为学生将来从事科学研究和实际工作打下良好的基础. 同时, 由于近年我国经济迅速发展, 各行各业对统计学的需求越来越迫切, 高校的其他专业对这种需求也愈加重视, 比如, 经济学, 社会科学, 金融学, 医学等专业都开设相关统计学课程. 面对越来越广阔的读者面, 在编写本书的时候, 特别注意材料的选编. 力求内容精练、难度深度适中, 语言通俗易懂, 并且着重让学生把握统计思想的内涵 (这也是统计学与数学的一个主要区别). 在编写本书的时候, 还注意增加一些贴近现代生产生活实际的例子, 以便学生加深理解所学的统计学基础知识. 由于统计学的飞速发展, 我们还注意增加一些近年发展起来的新方法. 例如, 在估计部分增加了 EM 算法的介绍和稳健估计的一些基本知识, 如影响函数、样本崩溃点等, 同时介绍一类特殊的稳健估计: M- 估计; 在检验部分增添了广义似然比检验、χ^2 检验、非参数假设检验, 多重假设检验等内容, 在统计决策和贝叶斯分析部分介绍了先验知识在统计分析中的重要性, 以及先验信息提取中的技巧和各种不同看法.

本书的内容比一般硕士课程 (一学期) 教学内容多一些, 大约多出 20% 左右. 由于统计学以介绍方法为主, 教材内容可以分成若干块, 讲

课教师可以有充分剪裁内容的自由，可根据进度或学生接受程度对内容作选择和增删. 每章后面，配有一定数量的习题，协助学生加深理解课堂讲授的内容. 书末还对这些习题给出了答案与提示，方便学生学习. 本书的书写中若有通顺性不尽如人意之处或内容疏漏之处，诚望读者提出批评指正.

作 者

2012 年 1 月于北京

目　　录

表格目录

插 图 目 录

试读结束：需要全本请在线购买： www.ertongbook.com

第一章　统计的基本概念

§1.1　数据、统计模型和分布族

在学习数理统计以前, 我们必须弄明白什么是数理统计学, 数理统计学的研究对象是什么. 要回答这些问题, 我们要引用《中国大百科全书 • 数学》(中国大百科全书出版社, 1992) 中关于数理统计学的定义: "数理统计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." 由这个定义可以看出统计学是在应用科学中十分重要的研究方法. 无论什么科学和事业, 社会科学也好, 自然科学也好, 工业也好, 农业也好, 人们都需要了解存在于事物中的规律. 为了了解这些规律, 我们需要收集一些数据, 而这些数据绝大多数是带有随机性的. 统计学的任务恰恰是研究和分析这些带有随机性的数据, 从中找到事物内在的规律, 而这些规律正好是实际工作者所需要的结论. 我们只举一个例子说明统计学与实际问题结合的密切程度.

例 1.1 (仪器设备的使用寿命问题)　任何设备, 大到火箭、嫦娥月球卫星上的仪器, 小到家用电器中的一台电视机, 一节电池, 都有使用寿命问题. 无论是仪器研制者、生产者或使用者都十分关心仪器的使用寿命问题. 头一个问题是: 什么叫做仪器设备的使用寿命? 即使是一个没有统计知识的人也会回答这个问题. 一个仪器从开始使用起直到仪器失效这一段时期称为仪器的使用寿命. 但是, 若进一步问, 一个仪器, 或一节电池的使用寿命有多长? 对这个问题就不好回答, 一般人只能说, 品牌好的电池寿命长一些. 从统计学的角度来看这个问题, 由于电池的寿命是带有随机性的, 应该由随机变量来刻画电池的寿命. 现在的问题是: 应该怎样研究仪器的寿命? 这就需要做试验, 收集数据. 若讨论的问题是电池的寿命, 这个问题相对简单. 只要找几节电池

进行试验, 观察它们的使用寿命, 就可以大致知道电池的使用寿命的状况. 但是像嫦娥月球卫星上的仪器, 情况就不是那样简单. 若没有现成的数据, 还需要设计试验, 并收集数据. 由于仪器是贵重的设备, 同时又是高可靠性的仪器, 人们不可能对仪器的寿命值进行观测, 只可能进行删失寿命试验, 即有些仪器没有到寿终就停止试验. 这些数据也称不完全数据. 收集数据以后, 还需要对数据进行分析, 找到有关仪器寿命的信息. 找到信息以后, 还要对该仪器的使用提出一些建议, 例如, 预测这个仪器能不能顺利完成任务. 若不能完成任务, 可提出一些建议, 改进产品质量, 以提高仪器的使用寿命. 上述问题的陈述, 充分说明我们所做的工作就是数理统计学所规定的任务. 但是要准确地陈述使用寿命问题, 还需要介绍数理统计的一些基本概念.

首先我们的研究对象是数据. 什么是数据呢? 广义地讲, 数据就是我们在实际工作中的记录. 例如, 某工厂为考察某些电子产品的使用寿命, 随机地抽取了 18 台产品做试验, 测得寿命数据如下 (见表 1.1.1):

表 1.1.1　电子仪器使用寿命 (单位: 小时)

17	29	50	68	100	130	140	270	280
340	410	450	520	620	190	210	800	1100

这 18 个寿命数据就是我们的研究对象. 又例如, 某社会工作者调查某城市中成年吸烟者占成年人口的比例, 共调查了 339 人, 其中 205 人为吸烟者, 134 人为不吸烟者. 在数据处理时我们通常用 x 表示数据, 此处 x 既可以是一个数, 也可以是一个向量或其他的量. 当明确表示向量时, 或者在论证中, 既有向量, 又有这个向量的分量的时候, 我们用 $\boldsymbol{x}$ 表示向量, 用 x_i 表示相应的分量. 数据的主要形式是 $\boldsymbol{x}=(x_1,\cdots,x_n)$. 在实际问题中, 有时候光一个字母是不够用于表达数据的. 例如, 在连续 10 天的气象记录中, 得到 $m_1,\cdots,m_{10},M_1,\cdots,M_{10}$, 其中 $m_i(i=1,\cdots,10)$ 是每天的最低气温, $M_i(i=1,\cdots,10)$ 是每天的最高气温. 此时的数据为 $\boldsymbol{x}=\{(m_i,M_i),\ i=1,\cdots,10\}$. 但是在学习统计学的时候, 用 $\boldsymbol{x}=(x_1,\cdots,x_n)$ 表示数据是最方便并且能够抓住数据本质的一种方法. 利用这种表达方法, 表 1.1.1 中寿命数据可表达

成 $\boldsymbol{x}=(x_1,\cdots,x_{18})$ 或 $\boldsymbol{x}=(x_1,\cdots,x_n), n=18$.

引入数据的概念以后, 我们要记住统计学的任务之一是对数据进行分析, 并且对所考察的问题作出推断. 在寿命数据的问题中我们的任务是考察该厂生产的电子产品的使用寿命. 但是我们必须弄明白, 什么是工厂生产的电子产品的使用寿命, 而且还要弄清楚工厂的这 18 台产品的寿命与工厂生产的产品的使用寿命之间的联系. 由经验知, 一个工厂生产的产品的使用寿命是带随机性的. 因此, 我们把一个工厂所生产的电子产品的使用寿命 X 看成一个随机变量是最合适不过了. 做这样的抽象以后, 可以把人们思维中直观的概念精确化成为一个数学对象. 若没有这种抽象, 就不可能对工厂生产的产品的使用寿命进行精确的研究.

什么是我们所需要的信息? 随机变量的某些特征是我们最关心的. 例如, X 的期望 $\mathbf{E}X$, $\mathbf{E}X$ 大, 说明产品的使用寿命长. $\mathbf{E}X$ 的大小说明了该工厂生产的产品质量的好坏. 除了 $\mathbf{E}X$, X 的标准方差 $\sigma X=\sqrt{\mathrm{var}(X)}$ 也是一个很重要的指标. 当然 X 的分布体现了工厂产品的使用寿命的全部信息. 因此若要了解工厂生产的电子产品的使用寿命, 我们只需了解随机变量 X 的分布.

而数据又是什么? 它与随机变量 X 的关系是什么? 从数据形成的过程可知, 电子产品的寿命 x_1 是工厂生产的一台产品的寿命, 即 X_1 的一个观察值 $X_1=x_1$. 同样 x_n 是 X_n 的观察值. 这样, $\boldsymbol{x}=(x_1,\cdots,x_n)$ 是 $\boldsymbol{X}=(X_1,\cdots,X_n)$ 的观察值. 有经验的实际工作者一定会明白, 我们收集数据的目的是为了了解工厂生产的产品的质量. 因此在采样时一定不会为某种利益去故意选择好的产品或坏的产品进行检查, 所选的产品一定是代表工厂产品质量的随机变量 X 的观察值, 而且这些产品也是相互独立地采样而得到的. 用数学的语言来描述: $X_1,\cdots,X_n$ 为相互独立且同分布的随机变量序列, 其公共分布与 X 的分布相同. 由于数据 $\boldsymbol{x}$ 与 X 有这样一层关系, 我们就指望从 $\boldsymbol{x}$ 得到 X 的分布的信息.

下面以电子产品的使用寿命为例引入统计学的基本概念. 我们将

工厂所生产的某种电子产品的总和称为**总体**. 总体是统计学最基本的概念. 所谓总体, 并不是仅仅表示工厂已经生产的全体产品. 总体是工厂所能生产的产品的总和, 因而包括还没有生产出来的产品. 除了产品的总和的概念外, 我们还需要在更深层次上理解总体的概念, 因为工厂生产的全体产品, 就是由工厂一个一个地生产出来的. 这样, 总体可以理解为工厂生产产品的过程或能力. 现在假定我们关心的是工厂所生产的产品的使用寿命. 这样总体又可以简化为工厂所生产的电子产品的寿命的总和, 而这些寿命值是工厂一个一个生产出来的产品的使用寿命值, 可以把生产一个产品的使用寿命看成一次随机试验. 因此我们将总体归结为一个随机变量. 根据不同的情况, 有时候, 也可把总体归结为随机向量、随机函数等随机量. 不过随机变量是最简洁的概念, 而且本书中所讨论的总体以随机变量为主.

定义 1.1　我们所考察的对象的总和称为**总体**. 在统计学中它可以归结为随机变量或其他形式的随机量. 代表总体的随机变量的分布也称为**总体分布**.

我们还以电子产品的使用寿命为例, 解释有关统计学的概念. 总体是寿命问题的信息之载体, 是信息存在的一种形式. 作为试验者, 并不确知总体的分布. 因此只能认为总体分布属于某个分布族. 用记号 X 表示我们研究的总体, X 是一个随机变量. 记 F 为 X 的分布, 我们只能假定 F 属于某个分布族 $\mathcal{F}$. 确定了 $\mathcal{F}$ 之后, 对所研究的总体形成了一个统计模型.

定义 1.2　设所考察的总体为 X, X 的分布 $F \in \mathcal{F}$, 则称 $\{X \sim F, F \in \mathcal{F}\}$ 为**总体模型**.

为所考察的问题建立总体模型是十分重要的事情, 若建立了一个正确的总体模型, 也就是给了实际问题一个严格的陈述. 建立总体模型最重要的任务是给出 X 的分布族 $\mathcal{F}$, 通常需要一些经验知识, 或先验信息. 在建立模型或确定 $\mathcal{F}$ 的过程中, 所确定的分布族范围越小, 说明利用的先验信息或经验越多, 将来作统计分析所得到的结论越具体. 若对 X 的分布没有任何知识, 那么只有假定 $\mathcal{F} = \{F : F$ 为 X 的任意一个分布$\}$.

现在讨论表 1.1.1 中寿命数据所对应的总体. 根据经验, 电子产品的寿命的分布只是所有分布组成的大类中的一个子类. 例如可以假定寿命分布为**指数分布**, 即其分布密度具有下列形式:

$$p(x,\theta)=\frac{1}{\theta}\exp\{-x/\theta\},\quad x>0,\theta>0, \tag{1.1}$$

上式中 $\theta>0$ 是分布的参数. 由 (1.1) 所刻画的总体模型称为参数模型, 因为这个模型中的分布可由参数来刻画. 我们可将这个模型写成 $\{F_\theta:\theta\in(0,\infty)\}$ 或 $\{p(x,\theta):\theta\in(0,\infty)\}$, 此处 F_θ 是总体分布, $p(x,\theta)$ 是总体分布的密度函数, θ 是模型的参数, $\Theta=(0,\infty)$ 称为参数空间. 我们做这种假定也不是毫无根据的. 我们根据前人的经验, 或者分析以前的数据, 得到如下的认识: 用指数分布去刻画电子产品的寿命是足够精确的. 当然, 人们也可以不相信这种假定, 对数据进行分析以后, 可以进一步假定电子产品的寿命分布应该属于更大的一个分布类, 例如**威布尔 (Weibull) 分布族**.

观察数据 $\boldsymbol{x}=(x_1,\cdots,x_n)$ 可以看成来自总体 F_θ 的一组观察值, 或者是 $\boldsymbol{X}=(X_1,\cdots,X_n)$ 的观察值, 其中 $X_i\sim\text{iid }F_\theta$ (此处记号 $X_i\sim\text{iid }F_\theta$ 表示 $X_1,\cdots,X_n$ 为独立同分布随机变量序列, 其公共分布为 F_θ), 而 θ 为参数的真值. 在统计中我们称 $\boldsymbol{X}=(X_1,\cdots,X_n)$ 为**样本**, 称 n **为样本量或样本的大小**, 称 $\boldsymbol{X}$ 的取值 $\boldsymbol{x}=(x_1,\cdots,x_n)$ 为**样本值**, 样本值就是我们观察到的数据. 由于数据 $\boldsymbol{x}=(x_1,\cdots,x_n)$ 被看成随机向量 $\boldsymbol{X}$ 的取值, $\boldsymbol{X}$ 的所有可能的取值的集合称为样本空间, 用 $\mathcal{X}$ 表示之, 即 $\boldsymbol{x}$ 的取值集合为 $\mathcal{X}$, $\boldsymbol{x}\in\mathcal{X}$. 由于 $X_i\sim\text{iid }F_\theta$, 可知 $\boldsymbol{X}$ 在 $\mathcal{X}$ 上产生一个分布, 记这个分布为 $\mathbf{P}_\theta$. 我们称 $(\boldsymbol{X},\mathbf{P}_\theta),\theta\in\Theta$ 为统计模型. 我们将上述的基本概念总结成下面的定义.

定义 1.3 当总体模型 $(X,F_\theta),\theta\in\Theta$ 确定以后, 将数据 $\boldsymbol{x}=(x_1,\cdots,x_n)$ 看成总体随机变量的一组独立观察值, 即 $\boldsymbol{x}=(x_1,\cdots,x_n)$ 是随机向量 $\boldsymbol{X}=(X_1,\cdots,X_n)$ 的观察值, 其中 $X_1,\cdots,X_n\sim\text{iid }F_\theta$. $(X_1,\cdots,X_n)$ 称为**简单随机样本**, 简称**样本**, 样本 $\boldsymbol{X}=(X_1,\cdots,X_n)$ 是一个随机向量. $\boldsymbol{x}$ 称为**样本值**, 即我们得到的数据. $\boldsymbol{X}$ 的取值空间 $\mathcal{X}$ 称为**样本空间**. $\boldsymbol{X}$ 和它相应的分布族 $\{\mathbf{P}_\theta,\ \theta\in\Theta\}$ 称为**统

计模型.

总体模型和统计模型在理论上是没有什么差别的, 数学上都是一个随机量和它的分布族. 不过我们在学习的时候, 要明白: 总体模型的重点是随机变量的分布, 它代表我们所需要的信息. 而统计模型所强调的是观察数据, 我们要从观察数据提取总体的信息. 而总体模型不强调观察数据, 一旦有了数据, 就成为一个统计模型.

对统计模型进行定义的时候, 模型中的参数 θ 是一个抽象的量, 它可以是数值, 也可以是常数向量或是其他的量. 其主要特征是: 一旦参数的值确定以后, 统计模型中的分布就完全确定. 不过常见的模型中, 参数 θ 是以数值或向量的形式出现的. 在某些统计问题中, 我们需要了解与参数有关的量, 即 θ 的函数 $g(\theta)$. $g(\theta)$ 也是人们需要了解的信息, 为了简便, 将 θ 的函数 $g(\theta)$ 也称为参数.

对于实际问题的研究发现, 总体模型中最重要的概念是总体随机变量 X 的分布族 $\mathcal{F}$. 前面已经说过, 若没有任何先验知识, 只有把 $\mathcal{F}$ 定义成一切分布所组成的集合. 由于问题的物理背景特殊, 人们根据问题的特殊背景, 确定其分布族. 这是建立模型之前我们必须考虑的重要问题. 现在我们介绍一些常见的分布族. 我们用例子的方式进行介绍.

例 1.2 (两点分布)　两点分布又称为**伯努利分布**, 具有代表性的实际背景是不合格品率的检查问题. 设工厂生产一批产品, 每生产一件产品为不合格品的概率为 $\theta, \theta \in (0,1)$, 记

$$X = \begin{cases} 1, & \text{产品为不合格}, \\ 0, & \text{其他}. \end{cases}$$

此时, X 为离散随机变量, 它服从两点分布, 其分布列为 $p(x,\theta) = \theta^x(1-\theta)^{1-x}, x = 0,1$, 相应的总体模型为 $\{p(x,\theta) : \theta \in (0,1)\}$. 设 $\boldsymbol{X} = (X_1, \cdots, X_n)$ 为 X 的一个样本, 则 $\mathcal{X} = \prod\limits_{i=1}^{n}\{0,1\}$ 为样本空间, $\mathcal{X}$ 是长度为 n 的 0-1 序列之集合. $\mathbf{P}\{\boldsymbol{x}\} = \prod\limits_{i=1}^{n}\theta^{x_i}(1-\theta)^{1-x_i}, \theta \in (0,1)$ 为相应的统计模型. 以后当我们遇到实际问题, 需要建立模型的时候, 若

试读结束：需要全本请在线购买： www.ertongbook.com