高等统计学

●薛留根　编著

科学出版社

高等统计学

薛留根　编著

科学出版社

北京

内 容 简 介

本书介绍高等统计学的基本概念、方法和理论，其内容包括基本概念、点估计、统计决策与 Bayes 统计、假设检验、区间估计和置信域. 本书着重阐述高等统计学的思想、概念和方法，尽量简化公式推导和理论证明. 此外，每章列举一些典型例题，给出较详细的解题方法和技巧，并有选择地安排一些模拟计算和图示.

本书可以作为本科高年级学生或硕士研究生的教材，也可以作为科技工作者自学或查阅资料的参考书.

图书在版编目(CIP)数据

高等统计学/薛留根编著. —北京：科学出版社, 2015

ISBN 978-7-03-044978-8

Ⅰ. ①高… Ⅱ. ①薛… Ⅲ. ①统计学-高等学校-教材 Ⅳ. ①C8

中国版本图书馆 CIP 数据核字(2015) 第 130102 号

责任编辑：陈玉琢 / 责任校对：张凤琴
责任印制：张 倩 / 封面设计：陈 敬

科学出版社 出版
北京东黄城根北街 16 号
邮政编码：100717
http://www.sciencep.com

三河市骏主印刷有限公司 印刷
科学出版社发行 各地新华书店经销
*
2015 年 8 月第 一 版 开本：720 × 1000 1/16
2015 年 8 月第一次印刷 印张：16 1/4
字数：320 000

定价：98.00 元

(如有印装质量问题，我社负责调换)

前　言

本书是为统计学及相关专业的学生和实际工作者编写的教科书. 阅读本书只需要高等数学和概率统计的基础知识, 读完本书即可进入统计学各相关领域的学习和研究. 因此, 本书可以作为本科高年级学生或硕士研究生的教材, 也可以作为科技工作者自学或查阅资料的参考书.

全书共五章, 依次为基本概念、点估计、统计决策与 Bayes 统计、假设检验、区间估计和置信域. 本书着重阐述高等统计学的思想、概念和方法, 尽量简化公式推导和理论证明. 此外, 每章尽可能多地列举一些典型例子, 给出较详细的解题方法和技巧, 并有选择地安排了一些模拟计算和图示.

本书作为北京工业大学统计学专业的研究生教材使用了 10 多年, 并经过反复修改. 趁这次出版的机会, 我们又对一些章节作了较大的改动和润色, 充实了一些新内容, 添加了若干典型例题, 并配置了一定数量的习题. 本书在取材与写作上, 重视以下三个方面: 一是在内容安排上注意由浅入深, 既考虑广度, 又要有深度, 体现新颖性和应用性; 二是在语言叙述上尽量通俗易懂, 既易于理解, 又不失文字的严谨性; 三是有侧重地安排一些典型例题, 着力说明统计学的方法和应用, 且配置的习题能够使读者得到各种基本训练.

众所周知, 统计学的生命力在于应用. 如果抛开实际问题, 那么统计学就无用武之地. 在实际应用中的确要用到诸多统计方法, 然而这些方法也需要用理论来指导. 高等统计学作为一门专业基础课程, 在介绍方法论的同时, 还需要阐述统计方法的产生根源、发展过程和理论基础. 因此, 作者在这些方面作了努力, 尽量使读者能够了解到诸多统计方法的来龙去脉, 不但使读者知其然, 而且知其所以然. 本书虽然不侧重介绍统计的实际应用案例, 但它所涉及的内容是进行统计理论、方法及应用研究所必备的基础知识. 希望读者读完本书后能够起到夯实基础的作用, 为以后进入理论研究领域和应用研究领域迈出坚实的一步.

本书的出版得到科学出版社陈玉琢编辑的鼓励和帮助, 得到了国家自然科学基金 (11171012)、高等学校博士学科点专项科研基金 (20121103110004)、北京市自然科学基金 (1142003, L140003) 和北京工业大学研究生精品课程建设项目 (CR2014-009) 的资助, 作者谨在此一并表示感谢.

由于作者水平所限, 书中难免有不足之处, 敬请广大读者斧正.

薛留根

2015 年 5 月

目　录

第1章 基本概念

本章主要介绍统计学中的基本概念, 主要内容包括统计模型与常用分布族、统计量及其分布、充分统计量、完备统计量、指数型分布族. 这些内容将为后面各章的学习提供帮助.

1.1 统计模型与常用分布族

统计学方法和理论的研究是基于某个统计模型展开的, 而对统计模型的讨论涉及分布族. 因此, 统计模型与分布族在统计学中扮演着重要角色. 本节首先引入统计模型的概念, 然后介绍一些常用的分布族.

1.1.1 统计模型

在开始学习统计学之前, 首先要明白什么是统计学. 《大英百科全书》的解释是: 统计学 (Statistics) 是一门收集与分析数据, 并且根据数据进行推断的艺术与科学. 按照上述对统计学的解释, 我们可以看出统计学有两个主要任务: 一是收集数据; 二是分析数据. 第一个任务的内容属于统计学中的两门课程 ——"抽样调查" 和 "试验设计"; 第二个任务需要利用各种统计方法来完成. 本书仅考虑第二个任务, 即讨论如何对已有的数据进行统计分析的问题. 由于数据来源于自然和社会的各个方面, "应用" 是统计学的一个十分重要的特征, 但实际应用更需要理论为基础. 因此, 本书不但介绍统计学中的基本概念和方法, 而且也涉及主要的统计理论.

在统计学中, 数据是样本的观测值, 数据分析的目的是利用样本来对事物的某些未知方面进行统计推断或预测. 假定样本 X 的一切可能取值为 $\mathcal{X}$, 那么通常称 $\mathcal{X}$ 为样本空间, 称 $(\mathcal{X}, \mathcal{B})$ 为可测空间, 其中 $\mathcal{B}$ 是 $\mathcal{X}$ 的某些子集构成的 σ 域. 依 X 的分布而从 $\mathcal{X}$ 中随机抽出的一个元素就是样本. 对一维总体, 容量为 n 的样本 X 记为 $(X_1, \cdots, X_n)^\tau$, 其中 "τ" 表示向量或矩阵的转置, 此时样本空间 $\mathcal{X}$ 是 n 维欧氏空间 $\mathbf{R}^n$ 或 $\mathbf{R}^n$ 的某个 Borel 子集, 而取 $\mathcal{X}$ 的一切 Borel 子集作为 $\mathcal{B}$. 这样的样本空间称为欧氏样本空间. 对于 k 维总体, 也可以作类似理解. 有了这个约定, 我们就不必在每个场合下对样本空间进行说明了.

随机变量 X 有一定的概率分布 F. 大家知道, 在概率论中 F 是给定的, 概率和数字特征的计算是在 F 已知的情况下进行的. 对统计学中的问题, F 总是未知的, 或仅知道其形式而其中含有未知参数. 因此, 我们可以把这个意思说成: F 属

于某个分布族 $\mathcal{F}$. 它在特定的统计问题中有具体的含义. 当 F 是样本分布时, $\mathcal{F}$ 称为样本分布族; 而当 F 是总体分布时, $\mathcal{F}$ 则称为总体分布族. 二者统称为分布族, 但其含义有些差别. 例如, 如果总体 $\widetilde{X}$ 有分布 $\widetilde{F}$, 从 $\widetilde{X}$ 中抽取独立同分布 (iid) 样本 $X_1,\cdots,X_n$, 则 $X=(X_1,\cdots,X_n)^\tau$ 有分布 $F=\widetilde{F}\times\cdots\times\widetilde{F}$, 它完全由 $\widetilde{F}$ 所决定. 我们可以把样本 $X_1,\cdots,X_n$ 视为在完全同等的条件下对 $\widetilde{X}$ 所作的 n 次独立观测值. 此时通常把由 $\widetilde{X}$ 的分布 $\widetilde{F}$ 所构成的集合称为总体分布族, 它决定了样本 X 的分布族 —— 样本分布族. 因此在这个特例下, 总体分布族与样本分布族有不同的含义.

样本空间 $\mathcal{X}$、σ 域 $\mathcal{B}$ 和样本分布族 $\mathcal{F}$ 构成了一个统计问题的三个基本要素. 我们称三元组 $(\mathcal{X},\mathcal{B},\mathcal{F})$ 为统计模型. 如果分布族 $\mathcal{F}$ 仅依赖于某一个参数 (或参数向量)θ, 则称该模型为参数 (统计) 模型, 并称 $\mathcal{F}$ 为参数分布族. 如果 $\mathcal{F}$ 中的分布不能用有限个参数来刻画, 则称该模型为非参数 (统计) 模型, 并称 $\mathcal{F}$ 为非参数分布族. 例如, 设 $\mathcal{F}_1=\{F_\theta:\theta\in\Theta\}$, 其中 θ 为参数, Θ 为参数空间, 那么 $(\mathcal{X},\mathcal{B},\mathcal{F}_1)$ 为参数模型, 其中 $\mathcal{F}_1$ 为参数分布族. 又如, 设 $\mathcal{F}_2=\{F:F$为实数集 $\mathbf{R}$上的对称分布$\}$, 那么 $(\mathcal{X},\mathcal{B},\mathcal{F}_2)$ 为非参数模型, 其中 $\mathcal{F}_2$ 为非参数分布族.

在实践中, 对具体问题可以借助于专业知识和经验积累来确定统计模型. 人们通常希望从参数模型出发来研究统计学中的问题, 因为参数模型含有较多的信息, 由此出发可以获得精度较高的参数估计. 但这样做要承担一定的风险, 这是因为当参数模型不真时, 统计推断结果可能会偏离实际, 甚至与实际相背离. 如果选用非参数模型, 所冒风险就会很小, 因为非参数模型适应面广, 但它所含的信息较少, 统计推断结果的精度一般不会很高. 在这两类模型下所用的统计推断方法有很大差别, 这就形成了统计学中的两类方法 —— 参数统计方法和非参数统计方法.

在 20 世纪 80 年代, 人们提出了另一类模型 —— 半参数模型. 部分线性模型就是其中的一种, 即有形式

$$E(Y|X=x,U=u)=\beta^\tau x+g(u), \tag{1.1.1}$$

其中 $\beta=(\beta_1,\cdots,\beta_p)^\tau$ 为 p 维未知参数向量, $g(u)$ 为定义在某区间上的未知函数. 模型 (1.1.1) 由两部分构成: 第一部分 $\beta^\tau x$ 为 $x=(x_1,\cdots,x_p)^\tau$ 的线性组合; 第二部分 $g(u)$ 为 u 的非线性函数. 因此称它为部分线性模型. 该模型不能作为参数模型, 因为 (X,U,Y) 的分布族不能通过有限个参数来刻画. 由于模型 (1.1.1) 的第一部分是参数性的, 而第二部分是非参数性的, 因此它应归入半参数模型. 按照这一思想, 可以举出其他一些半参数模型的例子. 例如, 单指标模型、部分线性单指标模型、部分线性变系数模型、可加部分线性模型等. 对半参数模型的讨论超出了本书的范围, 这里不再赘述.

本书主要讨论参数模型及参数统计方法, 但也涉及非参数统计方法. 关于非参

数模型及非参数统计方法的详细讨论, 可以阅读相关的非参数统计书籍, 例如, 陈希孺和柴根象 (1993), 孙山泽 (2000), 王静龙和梁小筠 (2006), 李竹渝与鲁万波和龚金国 (2007), 薛留根 (2013, 2015) 等. 对于半参数模型的讨论, 可参阅柴根象和洪圣岩 (1995)、薛留根 (2012) 等人的著作.

下面引入可控分布族和可控模型的概念. 为此, 我们从测度的绝对连续性谈起. 设 μ 和 ν 是可测空间 $(\mathcal{X},\mathcal{B})$ 上两个 σ 有限测度. 如果对任何 $N\in\mathcal{B}$, 由 $\mu(N)=0$ 可以推出 $\nu(N)=0$, 则称 ν 相对于 μ 绝对连续, 或者称 ν 被 μ 所控制, 记为 $\nu\ll\mu$. 根据 Radon-Nikodym 定理, 在 $\nu\ll\mu$ 条件下, 一定存在 $\mathcal{X}$ 上的一个 $\mathcal{B}$ 可测函数 $f(x)$, 使得

$$\nu(B)=\int_B f(x)\mathrm{d}\mu(x),\quad \forall B\in\mathcal{B}.$$

函数 $f(x)$ 称为 ν 对 μ 的 Radon-Nikodym 导数, 记为

$$f(x)=\frac{\mathrm{d}\nu}{\mathrm{d}\mu},\ \text{a.e. }\mu.$$

并且这个函数 $f(x)$ 在 a.e. μ 意义下是唯一的, 其中 a.e. μ 表示 "在 $\mathcal{X}$ 中去掉一个测度 μ 为 0 的集合后对 x 都成立", 常称为 "关于 μ 几乎处处成立". 利用绝对连续的概念, 可以给出可控分布族和可控模型的定义.

定义 1.1.1 设 $(\mathcal{X},\mathcal{B},\mathcal{F})$ 为一统计模型. 如果在可测空间 $(\mathcal{X},\mathcal{B})$ 上存在这样一个 σ 有限测度 μ, 使得 $\mathcal{F}$ 中每一个概率分布 F 对 μ 都是绝对连续的, 即对任意 $F\in\mathcal{F}$, 都有 $F\ll\mu$, 则称 $\mathcal{F}$ 为可控分布族, 称 $(\mathcal{X},\mathcal{B},\mathcal{F})$ 为可控模型, 并称 μ 为控制测度, 相应的 Radon-Nikodym 导数 $\mathrm{d}F/\mathrm{d}\mu$ 称为密度函数, 简称为密度.

对控制测度 μ, 如无特殊声明, 均指非负测度. 统计学中常用来作控制的 σ 有限测度有两种: 计数测度和 Lebesgue 测度. 下面举例加以说明.

例 1.1.1(计数测度) 设 $\mathcal{X}=\mathbf{R}$, $\mathcal{B}$ 是直线上一切 Borel 集组成的 σ 域, 在 $(\mathcal{X},\mathcal{B})$ 上定义如下测度:

$$\mu(B)=B\text{中非负整数的个数},\quad \forall B\in\mathcal{B}.$$

容易验证, 测度 μ 是 σ 有限测度, 并称为计数测度. 它可以用来控制任一个定义在非负整数集合 N(或其子集) 上的概率分布族, 其 Radon-Nikodym 导数就是通常的概率分布列. 如对 Poisson 分布族来说, 任一个不含非负整数的 Borel 集 A 的计数测度 $\mu(A)$ 为零, 而在这样的集合上 Poisson 概率 $P(A)$ 必为零. 对任一个 Borel 集 B, Poisson 概率 $P(B)$ 可表示为

$$P(B)=\int_B\frac{\lambda^x}{x!}\mathrm{e}^{-\lambda}\mathrm{d}\mu(x)=\sum_{x\in B\bigcap N}\frac{\lambda^x}{x!}\mathrm{e}^{-\lambda},$$

其中 x 仅取非负整数. 因此, Poisson 分布对计数测度的密度函数为

$$f(x;\lambda)=\frac{\lambda^x}{x!}\mathrm{e}^{-\lambda},\quad x=0,1,\cdots,$$

类似地, 可对二项分布族、负二项分布族作出解释.

今后对离散型随机变量的分布所谈论的密度函数, 就是指该分布对计数测度的 Radon-Nikodym 导数. 下面给出 Lebesgue 测度的定义.

例 1.1.2(Lebesgue 测度)　设 $\mathcal{X}=\mathbf{R}$, $\mathcal{B}$ 是直线上的一切 Borel 集组成的 σ 域, 在 $(\mathcal{X},\mathcal{B})$ 上基于区间长度定义 Lebesgue 测度

$$\mu(B)=B\text{中不相交区间的长度之和或其极限},\quad \forall B\in\mathcal{B}.$$

容易验证, Lebesgue 测度是 σ 有限测度, 它可以控制任一个定义在实数集 $\mathbf{R}$ 上的连续分布 F; 其 Radon-Nikodym 导数就是通常的密度函数 $f(x)$. 如对正态分布而言, 任一个 Borel 集 B 的概率总可表示为

$$P(B)=\int_B\frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}\mathrm{d}x,$$

其中 $\mathrm{d}x=\mathrm{d}\mu(x)$ 是 Lebesgue 测度.

一般来说, 对于一个参数模型 $(\mathcal{X},\mathcal{B},\mathcal{F})$, 如果分布族 $\mathcal{F}=\{F_\theta,\theta\in\Theta\}$ 是可控的, 其控制测度为 μ, 则相应的密度函数也依赖于参数 θ, 即

$$\frac{\mathrm{d}F_\theta(x)}{\mathrm{d}\mu}=f(x;\theta),\quad \theta\in\Theta.$$

此时, 可控分布族也可以用密度函数 $f(x;\theta)$ 表示, 即

$$(\mathcal{X},\mathcal{B},\{f(x;\theta):\theta\in\Theta\}).$$

显然, 控制一个分布族的测度并不唯一, 比如当 $F\ll\mu\ll\mu'$ 时, 则有 $F\ll\mu'$, 这时有

$$\frac{\mathrm{d}F}{\mathrm{d}\mu'}=\frac{\mathrm{d}F}{\mathrm{d}\mu}\frac{\mathrm{d}\mu}{\mathrm{d}\mu'}.$$

存在既不被计数测度控制, 又不被 Lebesgue 测度控制的分布族. 一个特殊的例子是 Marshall-Olkin 的二元指数族. 对该分布族的详细讨论可参阅茆诗松等 (2006) 的著作, 这里不再赘述.

1.1.2　常用分布族

在统计模型 $(\mathcal{X},\mathcal{B},\mathcal{F})$ 中, 样本空间 $\mathcal{X}$ 和 σ 域 $\mathcal{B}$ 是不可缺少的, 它指出了样本的取值范围以及应讨论哪一类事件是有意义的. 但分布族 $\mathcal{F}$ 是统计模型的核心,

它在统计推断中起着重要作用. 在概率论与数理统计的教科书中已介绍过一些常用的分布族, 其中包括:

(1) 二项分布族 $\{B(n,\theta):0<\theta<1\}$;

(2) Poisson 分布族 $\{P(\lambda):\lambda>0\}$;

(3) 正态分布族 $\{N(\mu,\sigma^2):(\mu,\sigma^2)\in\mathbf{R}\times\mathbf{R}^+\}$, 其中 $\mathbf{R}^+$ 是正实数集;

(4) 均匀分布族 $\{U(a,b):-\infty<a<b<\infty\}$.

这些分布族及其性质都是大家所熟悉的, 这里不再一一赘述. 此外, 在统计学中还经常涉及另外的一些分布族, 它们是: Gamma 分布族、Beta 分布族、t 分布族、F 分布族等. 下面逐个介绍这些分布族.

(5) Gamma 分布族. Gamma 分布的密度函数为

$$f(x;\alpha,\lambda)=\frac{\lambda^\alpha}{\Gamma(\alpha)}x^{\alpha-1}\mathrm{e}^{-\lambda x}I(x>0),$$

记为 $\mathrm{Ga}(\alpha,\lambda)$, 其中 α 和 λ 是两个正的参数, α 称为形状参数, λ 称为尺度参数, $I(A)$ 表示集合 A 的示性函数, $\Gamma(\alpha)$ 为 Gamma 函数, 其表达式为

$$\Gamma(\alpha)=\int_0^\infty x^{\alpha-1}\mathrm{e}^{-x}\mathrm{d}x.$$

Gamma 分布族记为 $\{\mathrm{Ga}(\alpha,\lambda):\alpha>0,\lambda>0\}$. 对 Gamma 分布作如下讨论.

(i) Gamma 分布的密度曲线. 当固定尺度参数 λ 时, 改变 α 的值将导致 Gamma 分布的密度曲线形状的改变. 图 1.1.1 给出了不同 α 值下的 Gamma 分布的密度曲线. 从图中可以得到如下结论: 当 $\alpha\leqslant 1$ 时, $f(x)$ 是严减函数; 当 $1<\alpha\leqslant 2$ 时, $f(x)$ 先凸后凹; 当 $\alpha>2$ 时, $f(x)$ 先凹后凸, 最后又凹, 此时它有两个拐点.

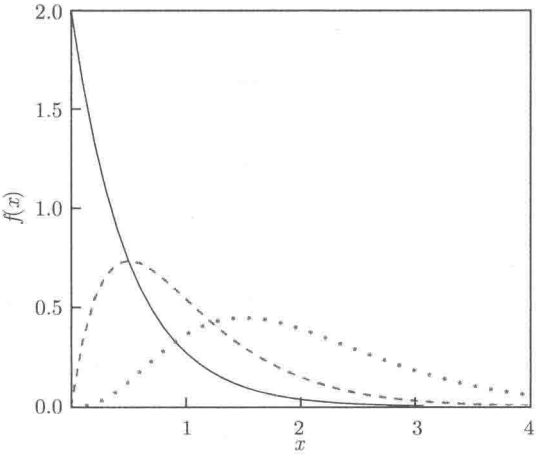

图 1.1.1 几种特殊的 Gamma 分布的密度曲线

实曲线对应 $\alpha=1$, 虚曲线对应 $\alpha=2$, 点曲线对应 $\alpha=4$

(ii) Gamma 变量 X 的数字特征. X 的 k 阶矩为

$$E(X^k)=\frac{\Gamma(\alpha+k)}{\Gamma(\alpha)}\frac{1}{\lambda^k}.$$

它的期望和方差分别为

$$E(X)=\frac{\alpha}{\lambda},\quad \operatorname{var}(X)=\frac{\alpha}{\lambda^2}.$$

(iii) Gamma 分布族的两个重要子族——指数分布族和χ^2分布族. 在 Gamma 分布中令 $\alpha=1$, 即得指数分布, 记为 $\operatorname{Exp}(\lambda)$, 其密度函数为

$$f(x;\lambda)=\lambda \mathrm{e}^{-\lambda x}I(x>0).$$

在 Gamma 分布中令 $\alpha=\dfrac{n}{2}$, $\lambda=\dfrac{1}{2}$, 即得自由度为 n 的 χ^2 分布, 记为 χ^2_n, 其密度函数为

$$f(x;n)=\frac{1}{2^{n/2}\Gamma(n/2)}x^{\frac{n}{2}-1}\mathrm{e}^{-\frac{x}{2}}I(x>0),$$

其中自由度 n 可为任意正实数, 但在实际问题中常用的自由度 n 为自然数, 并编制了 χ^2 分布表.

(6) Beta 分布族. Beta 分布的密度函数为

$$f(x;a,b)=\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}x^{a-1}(1-x)^{b-1}I(0<x<1),$$

记为 $\operatorname{Be}(a,b)$, 其中 a 和 b 是正的参数. Beta 分布族记为 $\{\operatorname{Be}(a,b):a>0,b>0\}$. 对 Beta 分布族作如下解释.

(i) Beta 分布的密度曲线. 参数 a 和 b 的值的改变将导致 Beta 分布的密度曲线形状的改变. 图 1.1.2 在 a 和 b 的不同值下给出了几种特殊的 Beta 分布的密度曲线. 从图中可以得到如下结论: 当 $a<1$ 和 $b<1$ 时, $f(x)$ 的曲线呈 U 型, 在 $(1-a)/(2-a-b)$ 处达到最小值, 特别地, 对 $a=b=0.5$, 该分布为反正弦分布, 对 $a=b=1$, 该分布就是区间 $(0,1)$ 上的均匀分布, 记为 $U(0,1)$; 当 $a>1$ 和 $b>1$ 时, $f(x)$ 的曲线呈单峰状, 在 $(a-1)/(a+b-2)$ 处达到最大值; 当 $a\leqslant 1$ 和 $b>1$ 时, $f(x)$ 是严减函数; 当 $a>1$ 和 $b\leqslant 1$ 时, $f(x)$ 是严增函数.

(ii) Beta 变量 X 的数字特征. X 的 k 阶矩为

$$E(X^k)=\frac{\Gamma(a+b)\Gamma(a+k)}{\Gamma(a)\Gamma(a+b+k)}.$$

它的期望和方差分别为

$$E(X)=\frac{a}{a+b},\quad \operatorname{var}(X)=\frac{ab}{(a+b)^2(a+b+1)}.$$

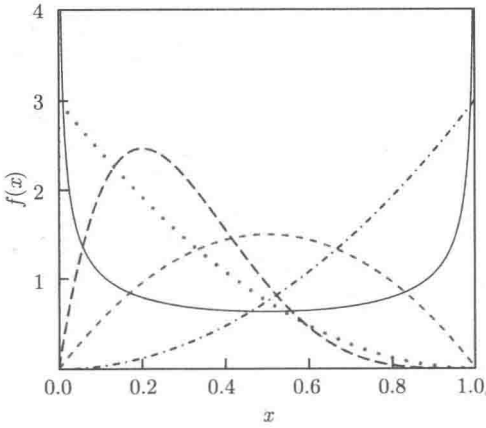

图 1.1.2 几种特殊的 Beta 分布的密度曲线

实曲线对应 $a=b=0.5$, 虚曲线对应 $a=b=2$, 点曲线对应 $a=1$ 和 $b=3$, 点虚曲线对应 $a=3$ 和 $b=1$, 长虚曲线对应 $a=2$ 和 $b=5$

(7) t 分布族. t 分布的密度函数为

$$f(x;\alpha)=\frac{\Gamma\left(\frac{\alpha+1}{2}\right)}{\sqrt{\alpha\pi}\Gamma\left(\frac{\alpha}{2}\right)}\left(1+\frac{x^2}{\alpha}\right)^{-\frac{\alpha+1}{2}},\quad x\in\mathbf{R},$$

记为 t_α, 其中参数 α 是正的实数. t 分布族记为 $\{t_\alpha:\alpha>0\}$. 在实际问题中常用的自由度 α 为自然数, 并编制了 t 分布表. 对 t 分布族作如下解释.

(i) t 分布的密度曲线. t 分布的密度曲线形状呈现 "中间高, 两边低, 左右对称", 很像标准正态分布 $N(0,1)$ 的密度曲线, 如图 1.1.3 所示.

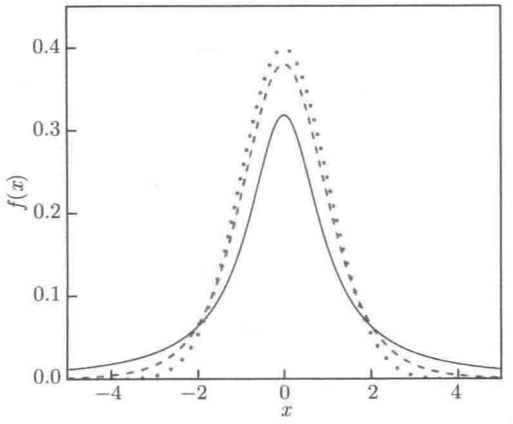

图 1.1.3 t 分布的密度曲线

实曲线对应 $\alpha=1$, 虚曲线对应 $\alpha=5$, 点曲线对应标准正态分布 $N(0,1)$

可以证明: 当 $\alpha \to \infty$ 时, t 分布的密度函数趋于标准正态分布的密度函数, 即

$$\lim_{\alpha\to\infty} f(x;\alpha) = \frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{x^2}{2}}. \tag{1.1.2}$$

事实上, 利用 Stirling 公式可以算得

$$\lim_{\alpha\to\infty} \frac{\Gamma\left(\dfrac{\alpha+1}{2}\right)}{\sqrt{\alpha}\Gamma\left(\dfrac{\alpha}{2}\right)} = \frac{1}{\sqrt{2}}.$$

此外, 由数学分析中的极限公式可得

$$\lim_{\alpha\to\infty}\left(1+\frac{x^2}{\alpha}\right)^{-\frac{\alpha+1}{2}} = \lim_{\alpha\to\infty}\left[\left(1+\frac{x^2}{\alpha}\right)^{\frac{\alpha}{x^2}}\right]^{\frac{\alpha+1}{\alpha}\cdot\left(-\frac{x^2}{2}\right)} = \mathrm{e}^{-\frac{x^2}{2}}.$$

综上所述, 即证明了式 (1.1.2). 在实际应用中, 当 $\alpha > 30$ 时常用标准正态分布来代替 t 分布, 其近似精度很高.

(ii) t 变量 X 的数字特征. 由于 t 分布的密度函数是幂函数, 因此比较它的幂次可以看出: 自由度为 α 的 t 分布仅存在低于 α 阶矩. 又由于 t 分布的密度函数是偶函数, 故存在的奇数阶矩为零, 即 $E(X^{2k+1}) = 0$, $2k+1 < \alpha$. 存在的偶数阶矩为

$$E(X^{2k}) = \frac{\alpha^k}{\sqrt{\pi}}\frac{\Gamma\left(\dfrac{\alpha}{2}-k\right)\Gamma\left(k+\dfrac{1}{2}\right)}{\Gamma\left(\dfrac{\alpha}{2}\right)}, \quad 2k < \alpha.$$

它的期望和方差分别为 $E(X) = 0$ 和

$$\mathrm{var}(X) = \frac{\alpha}{\alpha-2}, \quad \alpha > 2.$$

(iii) t 分布的特例. 自由度为 1 的 t 分布为 Cauchy 分布, 其密度函数为

$$f(x) = \frac{1}{\pi(1+x^2)}, \quad x \in \mathbf{R}.$$

Cauchy 分布的一般形式是

$$f(x;a,b) = \frac{b}{\pi[b^2+(x-a)^2]}, \quad x \in \mathbf{R},$$

其中 $(a,b) \in \mathbf{R}\times\mathbf{R}^+$. Cauchy 分布的期望和方差都不存在.

(8) F 分布族. F 分布的密度函数为

$$f(x;n_1,n_2) = \frac{\Gamma\left(\dfrac{n_1+n_2}{2}\right)}{\Gamma\left(\dfrac{n_1}{2}\right)\Gamma\left(\dfrac{n_2}{2}\right)}\left(\frac{n_1}{n_2}\right)^{\frac{n_1}{2}}\cdot\frac{x^{\frac{n_1}{2}-1}}{\left(1+\dfrac{n_1}{n_2}x\right)^{\frac{n_1+n_2}{2}}}I(x>0),$$

记为 F_{n_1,n_2}, 其中 n_1 和 n_2 是正的参数, 称为自由度. F 分布族记为 $\{F_{n_1,n_2}: n_1 > 0, n_2 > 0\}$. 当 n_1 和 n_2 均为自然数时, 人们编制了 F 分布表, 它是常用的统计分布之一. 图 1.1.4 给出了几种 F 分布的密度曲线.

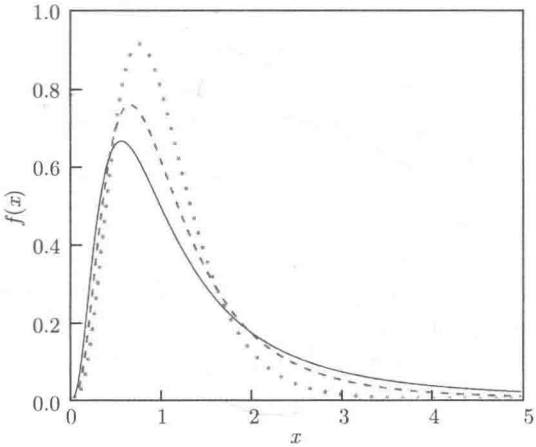

图 1.1.4 F 分布的密度曲线

在 $n_1 = 10$ 时, 实曲线对应 $n_2 = 5$, 虚曲线对应 $n_2 = 10$, 点曲线对应 $n_2 = 50$

除了上面介绍的一些分布族, 还有负二项分布族、倒 Gamma 分布族、多项分布族、多元正态分布族、非中心分布族. 关于这几个分布族的介绍可参阅茆诗松等 (2006) 的著作, 这里不再赘述.

1.2 统计量及其分布

对于某一个统计模型 $(\mathcal{X}, \mathcal{B}, \mathcal{F})$, 人们根据实际问题可以确定样本空间 $\mathcal{X}$ 和 σ 域 $\mathcal{B}$, 也可以知道分布族 $\mathcal{F}$ 的类型, 但不知道 $\mathcal{F}$ 中哪一个分布最适合. 要解决这个问题, 就需要从 $\mathcal{X}$ 中获得样本, 利用样本中的信息对总体分布作出判断. 在对这类问题的研究中, 统计量和它的分布起着非常重要的作用.

1.2.1 统计量

样本是总体的代表, 其中含有总体的信息, 但在一般情况下, 不能直接用样本作统计推断. 人们需要把样本中的信息进行加工处理, 以样本的函数形式把样本中分散的信息集中起来, 然后再作统计推断. 在实践中, 人们往往是针对具体问题构造样本的某种函数, 通过它提取样本中与总体有关的信息, 以便推断总体的分布或其数字特征. 例如, 为了估计总体 X 的期望 $E(X)$, 人们把样本 $X_1, \cdots, X_n$ 加工成