TSINGHUA UNIVERSITY

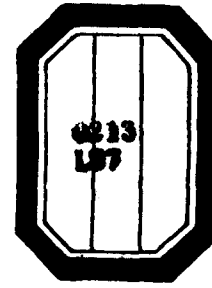
应用统计

陆 璇 编著

清华大学出版社

应用统计

陆 璇 编著



清华大学出版社

(京)新登字 158 号

内 容 简 介

本书较详细地介绍了多种常用的统计模型和分析方法，其中包括：极大似然法、分布的估计与检验、线性回归、方差分析与试验设计、寿命数据的模型与分析以及分类数据的模型与分析等。书中各章还配备了适量的练习和参考文献。在书后附有关于 SAS /STAT 统计软件的简介及常用的统计数表，以备查阅。

本书可作为高等院校非数学专业的硕士和博士研究生和高年级本科生学习应用统计课程的教科书，也可作为具有相当学历的读者自学应用统计的参考书。

图书在版编目(CIP)数据

应用统计/陆璇编著. —北京：清华大学出版社，1999.11
ISBN 7-302-03676-4

Ⅰ.应… Ⅱ.陆… Ⅲ.统计-方法 Ⅳ.C81

中国版本图书馆 CIP 数据核字(1999)第 34940 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：北京人民文学印刷厂
发行者：新华书店总店北京发行所
开 本：850×1168 1/32 **印张**：11.5 **字数**：298 千字
版 次：1999 年 12 月第 1 版 1999 年 12 月第 1 次印刷
书 号：ISBN 7-302-03676-4 /O・219
印 数：0001～4000
定 价：14.50 元

光华基金会为支持学术专著和研究生教材的出版，给予我社资助，本书即为由光华基金会资助出版的专著之一。

前　言

本书是作者在清华大学连续三年讲授“应用统计”课之备课教案的基础上完成的。这是一本供研究生和大学高年级学生学习应用统计的教科书，也可供需要学习一些应用统计知识的读者作为自学参考书。学习本书的预备知识包括：微积分、线性代数、初等概率统计。

考虑到极大似然估计和似然比检验的重要性，本书第 1 章介绍这方面的内容。实际上，后面各章中的估计和检验的方法都或多或少地和极大似然原理有关，其中第 5、6 章有直接联系。第 2 章的内容是关于分布的估计与检验，其中，χ^2 统计量在第 6 章中有重要应用。第 3 章到第 6 章介绍四个有广泛应用的统计模型：回归分析、方差分析与试验设计、寿命数据的模型与分析以及分类数据的模型与分析。每一个模型都有丰富的内容，并有各种专著介绍。在本书中，作者根据自己的理解进行择要介绍，重点放在应用上。

多元分析是统计学中一个重要而有广泛应用的分支，内容十分丰富。但限于篇幅，以及考虑到保持内容的联系与完整性，本书中并没有包含多元分析的内容。这并不意味着作者对统计学的这一重要领域有任何的忽视或轻视。

本书的内容不可能，也没有必要在一个学期的课程内全部讲完。作者认为，对于在专业研究中需要使用较多统计知识的研究生或高年级本科生来说，迫切需要一本有一定深度的综合性参考书，便于选择自己所需要的内容。此外，教师在讲授应用统计课程时也需要有这样一本教材，以便在讲授时可以根据情况对内容做灵活的取舍。无论是教或学，都不能把统计学的模型和方法当成一种固

定不变的僵硬程序，而应该充分发挥和理解统计学的思想，达到能够灵活运用的目的。要深入全面地掌握任何一个统计模型和有关的方法，都需要在学习一定基础知识的前提下进一步研究专著和文献。

限于作者的水平，本书在选材、内容的组织和叙述上肯定会有不少缺陷，作者欢迎使用本书的师生和所有读者及时将意见和建议反馈回来，以便将来有机会再版或重印时加以修改、订正。

作　者

1999 年 4 月于清华大学

目　录

第 1 章　极大似然法及其应用

1.1　极大似然估计

参数估计问题在统计推断中是一个核心的问题. 读者知道，通常的初等统计教科书都会介绍两种广泛使用的参数估计方法：矩方法和极大似然法. 矩方法所得到的估计具有直观和易于计算的优点. 但其缺点是使用范围窄，它要求样本为简单的，即从同一总体中获得的. 极大似然估计则不受此限制，是参数估计问题中最重要的一种估计方法. 它的主要优点为：

(1) 有优良的统计性质和好的近似分布，可用来作参数的区间估计和假设检验；

(2) 有一套完整、成熟的数值算法；

(3) 基于似然原理的似然比检验是一种理论上和应用上都非常重要的检验方法.

由于以上特点，极大似然法的思想和方法渗透到统计学的各个分支中，与各种各样的统计推断方法有直接或间接的联系. 因此，了解极大似然法的特点，并学会在各种不同的条件下灵活地应用它，对一个高素质的应用统计人才来说是非常必要的. 通常初等的统计教科书中虽然都无例外地介绍极大似然估计的基本概念，但对它的性质及应用介绍甚少，无法体现极大似然法的特点和重要性. 本章的目的是对极大似然法做一较为全面的介绍，以满足实际统计数据分析者的需要. 读者将会看到，在以后各章中所介绍的统计方法都与本章的内容有或多或少的联系.

1.1.1 极大似然估计的定义

记样本为 $X_1,\cdots,X_n$，样本的观测值为 $x_1,\cdots,x_n$. 设样本在 $x_1,\cdots,x_n$ 的密度函数值(当 X 的分布连续)或概率值(当 X 的分布离散)为 $f(x_1,\cdots,x_n;\theta)$，其中 θ 为未知参数，在一定的范围内取值. 当给定样本 $X_1=x_1,\cdots,X_n=x_n$ 时，定义似然函数(likelihood function)为

$$L(\theta)=f(x_1,\cdots,x_n;\theta).$$

$L(\theta)$看成为(给定样本时)参数 θ 的函数，它在 θ 的某个值 θ_1 处的值 $L(\theta_1)$越大，就意味着 θ 在 θ_1 处的"似然性"越大；反之，它在 θ 的某个值 θ_2 处的值 $L(\theta_2)$越小，就意味着 θ 在 θ_2 处的"似然性"越小. 一个自然的想法是：用使 $L(\theta)$达到最大值的 θ 值去估计 θ. 这就是极大似然估计. 具体地说，若 $\hat{\theta}$ 满足

$$L(\hat{\theta})=\max_{\theta}L(\theta), \tag{1.1.1}$$

则它称为 θ 的极大似然估计(maximum likelihood estimate). 以下我们简记极大似然估计为 ML 估计. ML 估计 $\hat{\theta}$ 是一个统计量，也就是说，它是样本值 $x_1,\cdots,x_n$ 的函数，因为 $L(\theta)$的极大值和极大值点是由 $x_1,\cdots,x_n$ 所决定的.

例 1.1.1 设样本 $X_1,\cdots,X_n$ 为取自正态总体 $N(\mu,\sigma_0^2)$的简单样本，其中 μ 为未知参数，而 σ_0^2 已知. 当给定 $X_1=x_1,\cdots,X_n=x_n$ 时，似然函数，即样本的联合密度为

$$L(\mu)=(2\pi\sigma_0^2)^{-n/2}\exp\left\{-\frac{1}{2\sigma_0^2}\sum_i(x_i-\mu)^2\right\}.$$

为便于对 $L(\mu)$求极大，取 $L(\mu)$的自然对数(以 e 为底的对数) $l(\mu)=\ln L(\mu)$. 由于对数函数为单调增函数，对 $L(\mu)$求极大等价于对 $l(\mu)$求极大

$$l(\mu) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\sigma_0^2 - \frac{1}{2\sigma_0^2}\sum_i (X_i - \mu)^2.$$

由于 $l(\mu)$ 关于变量 μ 有连续导数，由微积分的知识可知，$l(\mu)$ 在它的极大值点必然满足条件：$\mathrm{d}l(\mu)/\mathrm{d}\mu = 0$. 对上式两端关于 μ 求导数，并令导数等于 0，得到

$$\frac{\mathrm{d}l}{\mathrm{d}\mu} = \frac{1}{\sigma_0^2}\sum_i (x_i - \mu) = 0.$$

不难由上述方程解得 μ 的 ML 估计为

$$\hat{\mu} = \frac{1}{n}\sum_i x_i = \bar{x}.$$

为证明 $\bar{x}$ 确实使 $l(\mu)$ 达到极大，原则上还应该验证 $l(\mu)$ 关于 μ 的二阶导数小于零. 在这个例子中，这个结论不难验证.

当样本分布中的未知参数有多个时，设样本的密度函数值（当样本服从连续分布）或概率值（当样本服从离散分布）为 $f(x_1,\cdots,x_n;\theta_1,\cdots,\theta_k)$，其中 $\theta_1,\cdots,\theta_k$ 为未知参数，在一定的范围内取值. 可以记 $\boldsymbol{\theta} = (\theta_1,\cdots,\theta_k)^{\mathrm{T}}$，则 $\boldsymbol{\theta}$ 为一个 k 维参数. 在这种情况下，似然函数仍然定义为

$$L(\boldsymbol{\theta}) = f(x_1,\cdots,x_n;\boldsymbol{\theta}).$$

这样，ML 估计仍然可以按照(1.1.1)来定义，只不过现在是一个关于 k 个变量求极大的问题.

例 1.1.1 在所有的初等统计教科书中都会找到. 我们要从这个例子中引出若干问题，并在本章后续的内容中逐个予以解答.

问题 1　在例 1.1.1 中，μ 的 ML 估计可以通过求解方程 $\mathrm{d}l(\mu)/\mathrm{d}\mu = 0$ 来得到. 这个方程通常称为“似然方程”. 但并不是在任何情况下都可以通过求解似然方程来得到 ML 估计. 为此要弄清楚能够通过求解似然方程来得到 ML 估计的条件.

问题 2　在例 1.1.1 中，参数 μ 的 ML 估计有直接的函数表达形式 $\hat{\mu} = \bar{x}$. 在这种情况下，称 ML 估计有“解析解”. 但并不是

在任何情况下 ML 估计都有解析解，或者说，ML 估计只能通过似然方程以隐函数的方式表出. 在这种情况下只能求 ML 估计的数值解. 为此要知道如何求 ML 估计的数值解.

问题 3　在例 1.1.1 中，如果用随机变量形式的样本 $X_1, \cdots, X_n$ 代入 μ 的 ML 估计，就得到 $\hat{\mu} = \overline{X}$. $\overline{X}$ 的期望为 μ，因此是 μ 的无偏估计. 此外，估计的精度是由它的方差来决定的，方差越小，精度越高. 由于 $\overline{X}$ 的方差为 σ_0^2/n，与 n 成反比. 因此，当 n 较大时，它的精度是很高的. 事实上可以证明，$\overline{X}$ 是 μ 的“方差一致最小无偏估计”，也就是说，在 μ 的所有无偏估计中，$\overline{X}$ 的方差最小. 此外，$\overline{X}$ 仍然服从正态分布，这一性质保证我们可以计算 $\overline{X}$ 在一定范围内取值的概率. 利用这个性质，可以从 $\overline{X}$ 出发来构造 μ 的区间估计或作关于 μ 的假设检验. 在一般情况下，我们要知道 ML 估计是否有类似的优良性质.

在以下各小节中，将回答上述问题.

1.1.2　似然方程及 ML 估计的数值解法

在例 1.1.1 中的 ML 估计是通过求解似然方程来得到的. 这个方法有相当广泛的一般性. 首先考虑 θ 是一维参数的情况. 假定似然函数 $L(\theta)$ 关于 θ 有连续的一阶导数，$l(\theta) = \ln L(\theta)$. 由于对数函数为单调增函数，则 $l(\theta)$ 与 $L(\theta)$ 有相同的极值点，且 $l(\theta)$ 也有连续的一阶导数. 因此，若 $\hat{\theta}$ 为 θ 的 ML 估计（$L(\theta)$ 和 $l(\theta)$ 的极大值点），则它应该满足方程

$$\frac{\mathrm{d}l(\hat{\theta})}{\mathrm{d}\theta} = 0. \tag{1.1.2}$$

求解方程(1.1.2)，若有唯一解，则此解一般就是极大值点，因而为 ML 估计. 若解不唯一，则原则上可以计算 $l(\theta)$ 在所有解点的值，其中最大值所对应的解就是 ML 估计.

当样本分布中的未知参数为多维参数时，情况是类似的. 设参数有 k 个，记作向量的形式：$\theta=(\theta_1,\cdots,\theta_k)^{\mathrm{T}}$. 设似然函数 $L(\theta)$ 关于 $\theta_j, j=1,\cdots,k$ 有连续一阶偏导数. 在这种情况下，仍然记 $l(\theta)=\ln L(\theta)$，根据同样的理由，求 $L(\theta)$ 的极大值点的问题转化为求 $l(\theta)$ 的极大值点的问题. 根据多元微积分的知识可知，$l(\theta)$ 的极大值点 $\hat{\theta}=(\hat{\theta}_1,\cdots,\hat{\theta}_k)$ 应该满足下列方程组：

$$\frac{\partial l(\hat{\theta})}{\partial \theta_j}=0,\quad j=1,\cdots,k, \tag{1.1.3}$$

这就是在多维情形下的似然方程组. 求解此方程组就可得到 ML 估计 $\hat{\theta}=\hat{\theta}_1,\cdots,\hat{\theta}_k)$.

根据上述讨论，我们知道，ML 估计可以通过求解似然方程(组)来得到的必要条件是 $L(\theta)$ 关于 θ 有连续一阶(偏)导数. 若此条件不满足，则(偏)导数无法求，当然也不存在求解似然方程(组)的问题. 对许多常用的简单参数模型来说，参数的 ML 估计不仅可以通过求解似然方程来得到，而且有解析表达式. 下面我们对简单样本 $X_1=x_1,\cdots,X_n=x_n$ 列出一些重要分布的参数的 ML 估计.

例 1.1.2 总体 X 为 $B(1,p)$ 分布，即 $P(X=1)=p=1-P(X=0)$，则似然函数为 $L(p)=p^{\sum x_i}(1-p)^{n-\sum x_i}$，对数似然函数为

$$l(p)=\ln L(p)=\sum_i x_i\ln p+\left(n-\sum_i x_i\right)\ln(1-p).$$

似然方程为

$$\frac{\mathrm{d}l(p)}{\mathrm{d}p}=\frac{\sum_i x_i}{p}-\frac{n-\sum_i x_i}{1-p}=0.$$

由似然方程可以直接解得 p 的 ML 估计为 $\hat{p}=\frac{1}{n}\sum_i x_i=\bar{x}$.

例 1.1.3 总体 X 为泊松(Poisson)分布 $P(\lambda)$，即