数理统计教程

潘继斌　胡宏昌　编著

湖北科学技术出版社

数理统计教程

潘继斌 胡宏昌 编著

湖北科学技术出版社

图书在版编目(CIP)数据

数理统计教程/潘继斌,胡宏昌主编. -武汉:湖北科学技术出版社,2009.11.

ISBN 978-7-5352-4436-9

Ⅰ.数… Ⅱ.①潘…②胡… Ⅲ.数理统计-教材 Ⅳ.0212

中国版本图书馆 CIP 数据核字(2009)第 184766 号

策　　划:李慎谦

责任编辑:谭学军　王小芳　　封面设计:戴　旻

出版发行:湖北科学技术出版社　　电话:027-87679468

地　　址:武汉市雄楚大街 268 号　　邮编:430070

(湖北出版文化城 B 座 12-13 层)

网　　址:http://www.hbstp.com.cn

印　　刷:湖北鄂东印务有限公司　　邮编:438000

850×1168　1/32　8 印张　1 插页　196 千字

2009 年 11 月第 1 版　　2009 年 11 月第 1 次印刷

定价:20.00 元

序　言

笔者根据十几年从事数理统计教学和科研的经验，尝试编写本教程。本书像大多教材一样包含数理统计学的基本内容：抽样分布、参数估计和假设检验。这样编排内容既有利于学生掌握统计的基本思想，又有利于学生进一步学习其他统计课程。但不同之处在于注重基本概念的理解与扩展、强调统计思想方法的实质、并融进现代统计观点；强化统计的实用性。本教材的习题采用填空、选择、计算与证明体系，使学生能通过从易到难的练习，较好地掌握统计基本知识。将线性回归和方差分析的有关理论与实际应用联系起来是本教程的另一特点，通过这两章的学习，相信学生会用它们来解决相关实际问题。这样做，避免了只重理论轻应用的问题，有利于学生了解统计学应用性强的特点，从而激发学生学习统计的热情，又有利于锻炼和培养学生的动手能力和解决实际问题的能力。

本书第一章至第三章由潘继斌执笔，继承了魏宗舒编《概率论与数理统计教程》的统计部分体系好、理论性强等优点，适当进行了扩展、编排和组织。第四章和第五章由胡宏昌执笔，吸收了一些有关统计学、数据分析及 SAS 软件教材的优点，使理论与应用和谐统一。

本书可作为数学与应用数学专业及信息与计算科学专业的本科生教材。对于数学与应用数学专业，若学时较少，则建议只讲第 1 章至第 3 章中的基本部分，太长的证明删去，重点讲第四章。

编本书的初衷，是想展示统计学的基本理论和体现应用性很强的特点，但由于笔者学识浅薄，不足甚至错误之处在所难免，恳请批评指正！

编　者

2009 年 9 月

目　录

第 1 章　数理统计的基本概念

概率论和数理统计是研究随机现象统计规律的数学学科. 它们之间联系密切,但也有本质的区别:在概率论中研究的出发点是给定描述试验的定量概率模型(Ω,F,P),研究这个概率空间的各种性质,如对某个给定的事件 A,估计 A 在一系列 l 次独立随机实验中发生的频率分布 $F(x;A,l)=P(V_l(A)\leqslant x)$. 由于随机变量及其概率分布能够全面地描述随机现象的统计规律性,故概率论中使用推理方法主要是演绎法;而在数理统计中研究的出发点是给定描述试验的定性模型(Ω,F),并给定根据一个未知的概率测度 P 出现的独立同分布数据 $x_1,x_2,\cdots$,再加上人们对概率测度 P 或这组数据的一些认识(即各种假定),研究这个概率测度或它的某个泛函. 因此,归纳法是数理统计中主要使用的方法.

§1.1　数理统计概述

1.1.1　数理统计研究的基本问题

首先我们通过一些实际例子,说明数理统计研究的基本问题和使用的一般方法.

实例 1　如何估计一个鱼池中鱼的数量 x 和产量 y?

实例 2　怎样研究某厂所生产的一批电视机显像管的平均寿命?

实例 3　设某厂生产一种灯管,其寿命服从正态分布 $N(\mu,40\ 000)$,从过去较长一段时间的生产情况来看,灯管的平均寿命为 $\mu=1\ 500$ 小时. 现采用新工艺后,在所生产的灯管中抽取 25 只,测得平均寿命为 1 675 小时. 问采用新工艺后,灯管的寿命是否有明显提高?

实例 4 一个试验者对未知的物理量 μ 进行研究时，只能测量与之有某种关系的量 x，且测量值一般要受到各种随机因素的影响，如何估计 μ？

对于以上实际例子，我们作以下讨论：

(1)数理统计所要研究的问题是随机现象的分布或分布特征问题. 如关于例 1，我们可以先撒网捞取 M 条鱼，并做上记号后再放回鱼池中，问题即转化为估计鱼池中有记号鱼的比例和鱼池中的鱼的平均重量. 而再从鱼池中任取一条鱼，它可能有记号，也可能没有记号，每次实验结果应服从两点分布. 同时鱼池的鱼的重量也有一定的频率分布，其平均重量即是其平均值.

(2)对研究的随机现象由于不能进行全面观测(如例 1、例 4)或观测具有破坏性(如例 2、例 3)，故对数理统计所要研究的问题采用的一般方法只能是从整体中抽取一部分进行观测，再据此对整体作出推断，即归纳法是数理统计中常采用的方法. 如对于例 1，我们可以在放回标有记号鱼一定时间后，再撒网捞取 N 条鱼，观测其中有记号鱼的个数 P 及这 N 条鱼的平均重量 Q，从而不难想象 $x \approx MN/P, y \approx xQ$.

由于观测和试验是随机现象，依据有限个观测或试验对整体所作出的推断不可能绝对准确，多少总含有一定程度的不确定性，而不确定性用概率的大小来表示是最恰当的. 概率大，推断就比较可靠；概率小，推断就相对不可靠. 数理统计学中，一个基本问题就是依据观测或试验所取得有限的信息对整体如何推断的问题. 每个推断又必定伴随一定的概率以表明推断的可靠程度. 这种伴随有一定概率的推断称为**统计推断**. 因此，数理统计中主要采用**归纳推理方法**，对其研究方法的评价和推理结果的解释必须是针对大量重复实验进行.

(3)在数理统计中，关于总体的分布是未知的，但我们总是从经验出发对其分布或观测的数据作出一些假设，如例 2 一般假设

电视机显像管的寿命服从指数分布，例 3 的正态分布假设，对例 4 可设观测随机因素作用具有“加性”，即 $x = f(\mu) + \varepsilon$. 在随机因素由大量独立而又没有一个因素起主要作用条件下，可进一步假设 $\varepsilon \sim N(0, \sigma^2)$. 对整体分布类型已知，仅有未知参数的抽样推断称为参数统计，否则称为非参数统计.

(4)数理统计学研究的内容十分广泛，本书涉及内容主要是经典统计研究的参数估计问题(如例 1、例 2)、假设检验(如例 3)、回归分析(如例 4)和方差分析.

1.1.2　数理统计中常用的基本概念

在数理统计学中我们把研究对象的全体所构成的一个集合称为**总体**或**母体**，而把组成母体的每一成员称为**个体**. 如例 2 中一批显像管的全体就组成一个母体，其中每一只显像管就是一个个体.

在实际中我们所研究的往往是母体中个体的某些数值指标，如例 2 显像管的寿命指标 ξ，它是一个随机变量. 假设 ξ 的分布函数是 $F(x)$. 如果我们主要关心的只是这个数值指标 ξ，为了方便起见，可以把这个数值指标 ξ 的可能取值的全体看作母体，并且称这一母体为具有分布函数 $F(x)$ 的母体. 这样就把母体和随机变量联系起来了，并且这种联系也可以推广到 $k(k \geqslant 2)$ 维. 例如要研究母体中个体的两个数值指标 ξ 和 η，可以把这两个指标所构成的二维随机向量 (ξ, η) 可能取值的全体看作一个母体，简称二维母体. 这个二维随机向量 (ξ, η) 在母体上有一个联合分布函数 $F(x, y)$，并称这一母体为具有分布函数 $F(x, y)$ 的母体.

前面已经提过，数理统计学中我们总是通过观测和试验以取得信息，对其研究方法的评价和推理结果的解释必须是针对大量重复实验进行. 数理统计学中我们一般总是假定从客观存在的母体中按机会均等的原则随机地抽取一些个体，然后对这些个体进行观测或测试某一指标 ξ 的数值. 这种按机会均等的原则选取一些个体进行观测或测试的过程称为**随机抽样**. 假如我们抽取了 n

个个体，且这 n 个个体的某一指标为 $(\xi_1,\xi_2,\cdots,\xi_n)$，我们称这 n 个个体的指标 $(\xi_1,\xi_2,\cdots,\xi_n)$ 为一个**子样**或**样本**，n 称作子样的**容量**. 在重复抽样中每个 ξ_i 是一个随机变量，从而可以把容量为 n 的子样 $(\xi_1,\xi_2,\cdots,\xi_n)$ 看成一个 n 维随机向量. 在一次抽样以后，观测到 $(\xi_1,\xi_2,\cdots,\xi_n)$ 的一组确定的值 $(x_1,x_2,\cdots,x_n)$ 称作容量为 n 的**子样的观测值**(或数据). 容量为 n 的子样的观测值 $(x_1,x_2,\cdots,x_n)$ 可以看作一个随机实验的一个结果，它的一切可能结果的全体构成一个样本空间，往后我们称为**子样空间**. 它可以是 n 维空间，也可以是其中的一个子集. 而子样的一组观测值 $(x_1,x_2,\cdots,x_n)$ 是子样空间的一个点. 如果要研究母体中个体的两个指标 (ξ,η)，则所抽取的 n 个个体的指标 $(\xi_1,\eta_1),(\xi_2,\eta_2),\cdots,(\xi_n,\eta_n)$ 构成一个容量为 n 的子样. 由此可见，二维母体的容量为 n 的子样由 $2n$ 个随机变量构成，它的一组观测值 $(x_1,y_1,x_2,y_2,\cdots,x_n,y_n)$ 是 $2n$ 空间中一个点. 二维母体的子样空间可以是 $2n$ 维空间，也可以是它的一个子集.

实际上，从母体中抽取子样有很多种方法，为了使子样能够很好地代表母体(只有这样，才能依靠子样对母体作出可靠的推断)，就需要对抽样方法作一些要求：

(1)母体中的每个个体有同等机会被选入子样.

(2)子样的分量 $\xi_1,\xi_2,\cdots,\xi_n$ 是相互独立的随机变量.

满足以上两个要求所取得的子样称作**简单随机子样**. 如有放回随机抽取的子样是简单随机子样；当母体数目很大而容量较小时，不放回随机抽取的子样可认为是简单随机子样.

下文中所述的子样都是指简单随机子样.

设母体 ξ 的分布函数为 $F(x)$，则容量为 n 的子样 $(\xi_1,\xi_2,\cdots,\xi_n)$ 的联合分布函数为

$$F^*(x_1,\cdots,x_n)=\prod_{i=1}^{n}F(x_i).$$

1.1.3　数理统计中的基本定理

下面我们讨论统计学的基本定理，它是统计学中为什么能从子样观测结果推理母体情况的根本原因.

定义 1.1　设 $(\xi_1,\xi_2,\cdots,\xi_n)$ 是从母体中抽取的样本容量为 n 的一个子样，由

$$F_n(x)=\frac{1}{n}\sum_{i=1}^{n}I_{(\xi_i\leqslant x)},\qquad \forall x\in\mathbf{R} \tag{1.1}$$

所确定的分布称为样本分布或经验分布，对每一个样本观测值，$F_n(x)$ 是一个分布函数，称为样本分布函数或经验分布函数，对每一个固定的 $x\in\mathbf{R}$，$F_n(x)$ 又是样本 $(\xi_1,\xi_2,\cdots,\xi_n)$ 的函数，故 $F_n(x)$ 又是一个随机变量.

由于可把示性函数 $I_{(\xi_i\leqslant x)},(i=1,2,\cdots,n)$ 看作独立同分布，仅取 0 或 1 的随机变量，故

$$EF_n(x)=\frac{1}{n}\sum_{i=1}^{n}EI_{(\xi_i\leqslant x)}=\frac{1}{n}\sum_{i=1}^{n}P(\xi_i\leqslant x)=F(x),$$

$$Var[F_n(x)]=\frac{1}{n^2}\sum_{i=1}^{n}VarI_{(\xi_i\leqslant x)}=\frac{1}{n}F(x)[1-F(x)]\leqslant\frac{1}{4n}\leqslant\frac{1}{4}$$

根据贝努利大数定律，对 $\forall\varepsilon>0$，有

$$\lim_{n\to\infty}P(|F_n(x)-F(x)|>\varepsilon)=0 \tag{1.2}$$

式(1.2)说明了当 n 充分大时，经验分布函数是母体分布函数的良好近似. 关于经验分布函数有下面更强结论，它被称为统计学基本定理.

定理 1.1(格里汶科)　对任意给定的自然数 n，设 $x_1,x_2,\cdots,x_n$ 是取自总体分布函数 $F(x)$ 的一个子样观测值，$F_n(x)$ 为其经验分布函数. 记 $D_n(x)=\sup\limits_{-\infty<x<\infty}|F_n(x)-F(x)|$，则有

$$P(\lim_{n\to\infty}D_n=0)=1.$$

该定理表明：当 n 无限大时，对于所有的 x，$F_n(x)$ 与 $F(x)$ 之

差的绝对值是一致的愈来愈小，这个事件发生的概率为 1. 关于其收敛速度的研究也有大量的成果，在此不加叙述.

例 1.1 某单位对 100 名女学生测定血清总蛋白含量 (g/L)，数据如下：

74.3	78.8	68.8	78.0	70.4	80.5	80.5	69.7	71.2	73.5
79.5	75.6	75.0	78.8	72.0	72.0	72.0	74.3	71.2	72.0
75.0	73.5	78.8	74.3	75.8	65.0	74.3	71.2	69.7	68.0
73.5	75.0	72.0	64.3	75.8	80.3	69.7	74.3	73.5	73.5
75.8	75.8	68.8	76.5	70.4	71.2	81.2	75.0	70.4	68.0
70.4	72.0	76.5	74.3	76.5	77.6	67.3	72.0	75.0	74.3
73.5	79.5	73.5	74.7	65.0	76.5	81.6	75.4	72.7	72.7
67.2	76.5	72.7	70.4	77.2	68.8	67.3	67.3	67.3	72.7
75.8	73.5	75.0	73.5	73.5	73.5	72.7	81.6	70.3	74.3
73.5	79.5	70.4	76.5	72.7	77.2	84.3	75.0	76.5	70.4

由 SAS 系统中 proc capability 过程容易得到子样的经验分布函数及其相应的母体分布(正态分布)函数如图 1.1 所示，从该图容易看出经验分布函数能够很好地近似母体分布函数 $N(73.668, 3.9389^2)$.

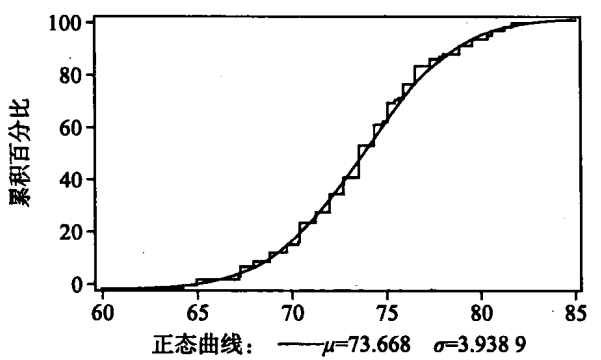

图 1.1 血清总蛋白含量的经验分布函数与母体的正态分布函数

1.1.4　统计学发展简史

了解了统计学中的基本问题、基本概念、基本方法、基本内容和基本原理后，我们来简单回顾一下统计学的发展史.

统计学产生于 17 世纪中叶，是从几个不同的领域开始的. 一个是英国威廉·配第的《政治算术》(1676 年)中对英国、法国、荷兰三国的经济实力比较；第二个是英国的约翰·格朗特的《关于死亡表的自然观察与政治观察》中对人口与社会现象中重要的数量规律的发现；第三个是法国的帕斯卡和费马对古典概率论的研究. 经过几代统计学家的努力，到 19 世纪末形成了古典统计学(主要是描述统计学)的基本框架.

20 世纪初，大工业的发展对产品质量检测问题提出了新的要求，即只抽取少量产品作为样本对全部产品质量做出推断. 因为大批量产品要做全面全部质量检验，既费时、费钱，又费人力，加之有些产品质量的检验要做破坏性检验，全部检验已不可能. 1907 年，英国的戈赛特提出了小样本 t 检验统计量，利用 t 统计量就可以从大批产品中只抽取较少的产品完成对全部产品质量的检验与推断，从而使统计学进入到现代统计学(主要是推断统计学)的新阶段. 以后经过著名统计学家费希尔给出的 F 统计量、最大似然估计、方差分析等思想和方法，奈曼和皮尔逊的置信区间估计和假设检验、沃尔德的序贯抽样和统计决策函数等，到 20 世纪中叶构筑了现代统计学的基本框架.

从 20 世纪 50 年代以来，统计理论、方法和应用进入了一个全面发展的新阶段. 一方面，统计学受计算机科学、信息论、混沌理论、人工智能等现代科学技术的影响，新的研究领域层出不穷，如多元统计分析、现代时间序列分析、贝叶斯统计、非参数统计、线性统计模型、探索性数据分析、数据挖掘等. 另一方面，统计方法的应用不断扩展，不论是自然科学、工程技术、农学、医学、军事科学，还是法律、历史、语言、新闻等人文科学和社会科学都离不开统计

方法.

§1.2 统计量及其分布

子样中含有母体的信息,但是较为分散,子样所含的信息不能直接用于解决我们所要研究的问题,而需要把子样所含的信息进行数学上的加工使其浓缩起来,从而解决我们的问题,这在数理统计中往往通过构造一个合适的依赖于子样的函数——统计量来达到.

1.2.1 统计量及其分布

定义 1.2 一个统计量是子样的一个函数,且这个函数是不依赖于任何未知参数的随机变量.

按照定义,随机变量 $\eta=\sum_{i=1}^{n}\xi_i$ 是一个统计量,前一节中提到的经验分布函数 $F_n(x)$ 也是一个统计量.但当 μ,σ^2 未知时,$\eta=\frac{\xi_i-\mu}{\sigma}$ 就不是一个统计量,这里 $\xi_1,\xi_2,\cdots,\xi_n$ 是一个子样.关于统计量我们作如下讨论:

(1)一个统计量不依赖于任何未知参数的要求是为了得到样本的一组观测值后能立即算出统计量的值,而不受总体分布尚未知的影响.尽管一个统计量不依赖于任何未知参数,但是它的分布可能是依赖于未知参数(如下面的定理 1.2、定理 1.3).

(2)对统计量应要求其具有"可测性",即统计量是样本可测空间 (R^n,B^n) 到其值域空间 (R^k,B') 的可测映射,其中 B 是 F 构成的乘积 σ 域,B' 是统计量诱导的其值域上的 σ 域,$k<n$,这是为了保证遇到与统计量有关的事件时,总是有概率可言.

(3)统计量具有数据的压缩功能,它具体体现在对样本的不同观测值,统计量的值可以有相同的值.

例如,设 $X=(X_1,X_2,X_3)$ 是从二点分布 $b(1,\theta)$ 中抽取的一

个容量为 3 的样本，由于每个 X_i 仅取 0 或 1，其样本空间只含有 8 个点，产生的 σ 域由 256 个事件组成. 这些样本实际上提供了两种信息：一是 3 次实验中成功了多少次；二是成功出现在那些实验点上. 对估计成功概率 θ 而言，有用的是第一种信息，第二种信息对估计 θ 并不重要. 如果采用统计量 $T_1 = X_1 + X_2 + X_3$，则它的值域仅含 4 个点，产生的 σ 域由 16 个事件组成，且将第一种信息充分反映出来了，即 T_1 既压缩了数据，又不损失有关 θ 的信息.

(4)统计量是对样本的信息“加工”，其目的是充分集中总体的有关信息，消去样本本身的信息. 统计量的构造既要压缩数据，又要不损失有关参数的信息，同时数学上要便于处理. 如对(3)中的问题，若取统计量 $T_2 = X_1X_2X_3$ 或 T_3 为任一常数，则它们虽然也压缩了数据，但损失了有关 θ 的信息，在统计学上把不损失信息的统计量称为充分统计量，有关充分统计量的内容在下一章讨论.

定义 1.3　统计量的分布称为**抽样分布**.

抽样分布在研究统计量的性质和评价一个统计推断的优良性等方面十分重要. 统计量是子样的函数，它的抽样分布函数可以从子样的联合分布函数推出. 设母体 ξ 的概率密度函数为 $p(x)$，则容量为 n 的子样 $(\xi_1, \xi_2, \cdots, \xi_n)$ 的联合概率密度函数为

$$p(x_1, \cdots, x_n) = \prod_{i=1}^{n} p(x_i)$$

如果 $T = T(\xi_1, \xi_2, \cdots, \xi_n)$ 是一维统计量，则其分布函数为

$$F_T(x) = P(T(\xi_1, \cdots, \xi_n) \leqslant x) = \int_D \cdots \int p(x_1, x_2, \cdots, x_n)\mathrm{d}x_1 \cdots \mathrm{d}x_n$$

其中积分区域 $D = \{(x_1, \cdots, x_n): T(x_1, \cdots x_n) \leqslant x\}$ 是 n 维欧氏空间的一个子集. 当 T 为可微函数时，对 $F_T(x)$ 求导可得 T 的密度函数；当 T 为 $k(k < n)$ 时，可以类似求其联合分布函数；当总体密度函数含有参数时，T 的分布或密度也含有此参数.