统计计算

高惠璇 编著

北京大学出版社

统 计 计 算

高惠璇 编著

北京大学出版社
北 京

新登字(京)159号

图书在版编目(CIP)数据

统计计算 / 高惠璇编著. —北京:北京大学出版社,
1995.7
ISBN 7-301-02827-X

Ⅰ.统… Ⅱ.高… Ⅲ.概率统计计算法-高等学校-教材
Ⅳ.0242.28

书　　名:统计计算
著作责任者:高惠璇
责任编辑:刘　勇
标准书号:ISBN 7-301-02827-X/O·357
出 版 者:北京大学出版社
地　　址:北京市海淀区中关村北京大学校内 100871
电　　话:出版部 2502015　发行部 2559712　编辑部 2502032
排 印 者:北京大学印刷厂
发 行 者:北京大学出版社
经 售 者:新华书店
850×1168 毫米 32 开本 13.125 印张 338 千字
1995 年 7 月第一版　1995 年 7 月第一次印刷
印　　数:0001—3,000 册
定　　价:15.00 元

内 容 提 要

统计计算是数理统计、计算数学和计算机科学的交叉学科。本书系统地介绍了统计计算的基本方法，并给出各种算法的统计原理和数值计算的步骤，以及部分例子，使读者掌握用统计方法解决具体问题的全过程.

本书内容包括误差与数据处理、分布函数和分位数的计算、随机数的产生与检验、矩阵计算、无约束最优化方法、多元线性和非线性回归的算法及随机模拟方法等.各章内容丰富，并配有适量的习题和上机实习题.

本书可作为理工科院校概率统计、数学、应用数学、计算机科学等系大学生的教材，也可作为教师、研究生以及从事统计、信息处理工作的有关工程技术人员的参考书.

前　言

数理统计方法是以概率论为理论基础，通过样本来了解和推断总体统计特性的科学方法，内容极为丰富．随着计算机使用的日益广泛，为了更好地应用数理统计方法来解决实际问题，从事统计工作或实际工作的人们都很关心如何应用计算机来更快完成各种统计数据的分析处理工作，故而出现了“统计计算”(Statistical Computation)这个方向．统计计算是数理统计、计算数学和计算机科学三者的结合，它是一门综合性学科．

在科学研究和生产实际的各个领域中，普遍地存在着大量数据的分析处理工作．如何应用数理统计学中的回归分析、多元分析、时间序列分析等统计方法来解决实际问题，以及如何解决在应用中出现的计算问题，对实际工作者来说是极需解决的问题．本书的目的力求把统计思想、数值计算步骤及在计算机上的实现结合起来，使读者掌握用统计方法解决实际问题的全过程．

本书是作者在北京大学多年讲授“统计计算”课程的讲义基础上编写的．内容可分为两部分：第一部分(第一、二、三、五章及第七章的前半部分)是基本统计计算方法，包括数据处理、常用分布的分布函数和分位数的计算、随机数的产生和检验、常用的矩阵算法及无约束最优化方法．第二部分(第四、六章及第七章的后半部分)介绍应用最广泛的线性与非线性回归分析的各种算法、随机模拟方法及在各方面的应用．因此，本书包括了统计计算的基本内容，并在每章末配有适量的习题和上机实习题，以利于培养学生应用统计方法解决实际问题的能力．本书适用于数理统计、计算数学和应用数学等专业，课程学时为 60～80 学时的统计计算课程教材．编写此书的过程中，作者力求内容充实，阐述通俗易懂、深入浅出，

并便于自学.本书对于从事统计、信息处理工作等领域的实际工作者也是一本很适用的参考书.

本书由我系汪仁官教授和耿直教授审阅,他们在百忙中抽出宝贵时间认真审阅全书,提出不少修改意见,在此向他们表示衷心的感谢.由于编者水平有限,书中难免存在错误或不足之处,恳请读者批评指正.

编　者

1994年10月于北京大学

目　录

第一章　误差与数据处理

在生产实践和科学实验中，经常会遇到大量的各种不同类型的数据. 这些数据为我们提供了很有用的信息，它可以帮助我们认识事物的内在规律、研究事物之间的关系、预测事物的可能发展，是指导生产实践和科学实验的重要依据.

但是这些有用的信息并非一目了然，而是蕴藏在大量的数据之中. 要想从这大量的数据中找到有用的信息，必须对数据进行分析整理，去伪存真、去粗取精、由表及里、抓住主要矛盾，尽可能充分地、正确地从数据中提取出有用的信息. 数理统计学为我们分析、处理数据提供了许多有用的统计方法.

观测得到的数据难免存在误差，对数据分析处理时，也会产生一些计算误差. 因误差的普遍存在，我们有必要对误差及含有误差的数据处理问题进行研究. 本章介绍误差的类型和基本特点，数据处理的基本方法；同时复习统计学中经常用到的基本概念和术语.

本章的参考文献有[1]，[3]～[5]，[8]，[17]，[18]，[21]，[22]，[27].

§1　误　　差

用统计方法解决实际问题时，首先要建立它的统计数学模型，也就是要把这个具体问题经过抽象化简，建立有关量应满足的统计关系式——即制定描述这些量的统计数学模型. 统计数学模型总是近似的，它包含模型误差，这种误差会影响分析结果. 研究如何提出更为合适的统计数学模型，这是多元统计分析、时间序列分析等其他学科讨论的问题.

在统计数学模型选定之后，接着要利用观测数据估计模型参数. 观测数据总是通过实验，用测量工具观测得到的，它们不可能绝对准确，总存在一定的实验误差，这种误差是不可避免的.

用数学方法计算模型参数的估计值时，也常遇到一些典型的计算误差问题. 如计算一个无穷级数之和时，总是用它前面的若干项之和来近似. 截去了该级数的后段，就产生了误差. 这类误差叫做截断误差.

另外，计算工作都是用计算机实现的，在计算机内存中，最简单的有理数，如$\frac{1}{3}$，$\frac{1}{7}$等都只能用有穷位小数近似；至于无理数，如π，e，$\sqrt{2}$等更是如此. 这都要产生截断误差.

最后在进行乘、除等运算时，得到的结果只能按“四舍五入”原则用有限位数表示，这样产生的误差就是“舍入误差”.

实验误差、计算误差和模型误差对于计算结果的影响，都是统计计算中不能忽视的问题.

（一）实验误差

实际问题中遇到的数据总是通过观测、实验得来的. 例如测量物体的高度，用同一种方法重复测量n次，得n个实验数据：$x_1, x_2, \cdots, x_n$. 虽然物体的高度客观存在，是一个常数l，但每次测量的结果不完全相同，也就是说，实验数据中存在误差. 这种误差常称为实验误差.

记$\varepsilon_i = x_i - l \ (i = 1, 2, \cdots, n)$，$\varepsilon_i$就是第$i$次实验误差. 误差与真值之比称为相对误差，因真值未知，而测量值与真值接近，故也可以把误差与测量值之比作为相对误差的近似值，即

$$\text{相对误差} = \frac{\text{误差}}{\text{真值}} \approx \frac{\text{误差}}{\text{测量值}}.$$

实验误差按其性质可分为三类：

(1) 随机误差(偶然误差)

这是在实验过程中，由一系列随机因素引起的不易控制的误差. 这类误差在实验中是不可避免的. 在一次实验中，误差的取值

可正可负，可大可小，但当重复实验次数 n 充分大时其均值趋于零，具有这种性质的误差称为随机误差. 如果用 ε 表示每次实验的误差，则 ε 是随机变量，且均值为 0，方差为 σ^2. 进一步地可认为 ε 服从正态分布，记为 $\varepsilon \sim N(0,\sigma^2)$.

(2) 系统误差

由于某种人为因素引起实验结果有明显的固定偏差，这种固定偏差称为系统误差. 如由于仪器使用不当，格值不准、观测方法不合理等引起的误差. 如果用 ε 表示因人为因素使得每次实验产生的误差，则 $\varepsilon \sim N(\mu_0,\sigma^2)(\mu_0 \neq 0)$；常数 μ_0 就是系统误差. 这类误差不可能通过增加实验次数来消除；但可以用统计检验的方法进行检查. 当发现有系统误差后，必须找出引起误差的原因. 通过改进仪器性能，测定仪器常数、改善观测条件等措施来加以克服.

(3) 过失误差

把明显歪曲实验结果的误差称为过失误差(也称为异常值). 它是由于实验观测系统测错、传错或记错等不正常原因造成的. 在数据处理中这类误差一定要消除，否则会严重影响计算结果的准确度，甚至给出不正确的结论.

在一组实验数据中，实验误差总是综合性的，即随机误差、系统误差和过失误差同时错综复杂的存在于实验数据中. 我们应通过分析整理数据，把系统误差、过失误差消除. 随机误差虽不可避免，但它有统计规律性，经多次重复观测，可消去随机误差的影响.

(二) 计算误差

用数学方法解决实际问题时，除了因实验误差的存在会影响计算结果外，还有计算误差同样对计算结果有影响. 计算误差包括截断误差和舍入误差.

(1) 截断误差

计算函数 $f(x)$ 的值时，采用某种数值计算方法计算，得到的近似值与准确值之差称为方法误差或截断误差.

例如，计算 $\sin x$ 值时，利用公式：

$$\sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \cdots,$$

当 x 很小时，取第一项 x 作为 $\sin x$ 的近似值，这时截去部分引起的误差就是截断误差. 从理论上可以证明，产生的误差不会超过 $\frac{|x|^3}{6}$.

另外由于计算机的字长有限，原始数据输入计算机内存后产生的误差也是一类截断误差. 例如 $\frac{1}{3}$，$\frac{1}{7}$，$\sqrt{2}$，e，π 等数存入计算机后只能用有限位小数来代替，这都要产生截断误差. 这类截断误差与计算过程无关. 它是客观存在的，在理论上完全确定的误差.

(2) 舍入误差

以上提到的无理数或有些有理数存入计算机后，因只能用有限位数表示而产生截断误差. 计算机处理这类截断误差时，一般是按“四舍五入”的原则截取，故这类误差也可称为舍入误差.

还有一类舍入误差与计算过程及选用的计算公式有关，即在进行乘、除等运算时产生的误差. 这类误差在大量复杂的计算中也是不可忽视的.

(3) 在数值计算中应注意的几条原则

为了减少计算误差的产生，在设计算法及编写程序时，以下几个简单的原则值得引起注意：

① 注意计算顺序；

② 避免相近的大数值相减及相差很大的两数值做加减运算；

③ 简化计算公式，减少计算次数；

④ 注意某些确定的值作为实数在计算机内存中可能是一近似值.

实验误差、计算误差对于计算结果的影响都是统计计算中必须引起重视的问题. 而随机误差不可避免地存在于实验数据中，故在数据分析处理过程中，要消除实验误差对分析结果的影响.

§2 总体的数字特征

(一) 总体和分布

总体：我们所研究的对象的全体叫做总体. 总体中的每一个基本单位称为个体. 例如物体高度的全部可能测量值组成我们研究的总体，而每一次的测量值 x_i 就是一个个体.

样本：从总体中观测得到的部分结果称为样本(或子样). 如测量 n 次，$x_1,\cdots,x_n$ 就是一个样本量(或称容量) 为 n 的样本.

如果用 X 表示物体的测量高度，则 X 是随机变量(简记为 R. V.). X 的所有可能取值组成我们研究的对象全体. 而从数学上说，所谓总体就是一个随机变量 X. 样本 $X_1,\cdots,X_n$ 是 n 个相互独立且和总体 X 有相同统计规律的随机变量.

因实验误差的存在，物体的测量值 X 是随机变量. 重复测量 n 次所得数值不完全一样，即这些数值有波动；另一方面这些数值有统计规律性，即它们的数值虽不等，但多数在常数 l 附近，离 l 值越远的数值越少. 这表明我们考察的总体 X 有一定的概率分布(统计规律性).

对于随机变量 X，如果存在非负函数 $p(x)$，使得对任意 $a<b$，有 $P\{a<X<b\}=\int_a^b p(x)\mathrm{d}x$，则随机变量 X 是连续型的；且称 $p(x)$ 为 X 的概率密度函数.

由样本值 $x_1,\cdots,x_n$(即一批观测值)，利用直方图法可给出分布密度函数 $p(x)$ 的近似图形.

对一般随机变量 X (可以是连续型的，也可以是离散型的，甚至更一般的)，称函数

$$F(x)=P\{X\leqslant x\}$$

为 X 的分布函数.

把样本 $x_1,\cdots,x_n$ 按取值由小到大顺序排列，得次序统计量

$x_{(1)} \leqslant x_{(2)} \leqslant \cdots \leqslant x_{(n)}$. 令

$$F_n(x) = \begin{cases} 0, & \text{当 } x < x_{(1)}, \\ \dfrac{k}{n}, & \text{当 } x_{(k)} \leqslant x < x_{(k+1)}, \\ 1, & \text{当 } x \geqslant x_{(n)}, \end{cases}$$

称 $F_n(x)$ 为 X 的经验分布函数. 它和分布函数 $F(x)$ 有类似性质.

(二) 随机变量的数字特征

在实际问题中,常常需要用几个有代表性的数字来描述总体 X 的基本统计特性,通常称它们为随机变量 X 的数字特征.

(1) 位置的数字特征量

反映 X 取值位置的特征量有均值、中位数和众数等.

定义 2.1 ① 设 X 是离散型随机变量,概率分布是

$$P\{X = x_k\} = p_k \quad (k = 1, 2, \cdots),$$

则称和数 $\sum\limits_k x_k p_k$ 为 X 的**均值**,记作 $\mathrm{E}(X)$;

② 设 X 是连续型随机变量,密度函数为 $p(x)$,则称积分

$$\int_{-\infty}^{\infty} x p(x) \mathrm{d}x$$

为 X 的**均值**,记作 $\mathrm{E}(X)$.

以上定义中,要求无穷级数(或积分)绝对收敛,否则相应的均值不存在.

定义 2.2 对任意随机变量 X,任给 $p(0 \leqslant p \leqslant 1)$,满足:

$$\begin{cases} P\{X \leqslant x_p\} \geqslant p, \\ P\{X \geqslant x_p\} \geqslant 1 - p \end{cases}$$

的数值 x_p 称为随机变量 X 的 p **分位数** ①,当 $p = 1/2$ 时,称 x_p 为**中位数**,记为 m_e 或 $x_{1/2}$.

定义 2.3 ① 若 X 是连续型随机变量,其概率密度函数为

---

① p 分位数也可以定义为:设 X 的分布函数为 $F(x)$,若 $x_p = \inf\{x: F(x) \geqslant p\}$,则称 x_p 为总体 X 的 p 分位数. 这样定义的 p 分位数是唯一的.

$p(x)$，称满足：$p(m_0)=\sup\limits_{\text{一切}x} p(x)$ 的数值 m_0 为 X 的**众数**.

② 若 X 是离散型随机变量，其概率分布为：$P\{X=x_k\}=p_k$ $(k=1,2,\cdots)$，如果数值 m_0 使得 $P\{X=m_0\}=p^*$，而 $p^*=\max\limits_i p_i$，则称 m_0 为随机变量 X 的众数.

均值是反映总体 X 取值的“平均”位置的特征量；中位数是刻画总体 X 取值的“中心”位置的特征量；而众数是刻画总体 X 的最可能取值位置的特征量.

例 2.1 若 $X\sim N(\mu,\sigma^2)$，则 X 的中位数、众数及均值都为 μ.

例 2.2 设离散随机变量 X 的概率分布为 $P\{X=i\}=0.5$ $(i=0,1)$，则均值 $\mathrm{E}(X)=0.5$；中位数 $m_e\in[0,1]$；众数 $m_0=0$ 或 1. 中位数和众数总存在，但可能不唯一.

(2) 离散性的数字特征量

反映随机变量 X 取值分散程度的特征量有方差、极差、四分位极差和变异系数等.

定义 2.4 ① 设 X 是离散型随机变量，概率分布是

$$P\{X=x_k\}=p_k \quad (k=1,2,\cdots),$$

则称和数 $\sum\limits_k (x_k-\mathrm{E}(X))^2 p_k$ 为 X 的**方差**，记作 $\mathrm{Var}(X)$.

② 设 X 是连续型随机变量，密度函数为 $p(x)$，则称积分

$$\int_{-\infty}^{\infty}(x-\mathrm{E}(X))^2 p(x)\mathrm{d}x$$

为 X 的**方差**，记作 $\mathrm{Var}(X)$.

定义 2.5 ① 称随机变量 X 的最大取值与最小取值之差为**极差**，记为 R，即 $R=\max X-\min X$.

② 称随机变量 X 的 3/4 分位数与 1/4 分位数之差为**四分位极差**，记为 Q，即 $Q=x_{0.75}-x_{0.25}$.

③ 设随机变量 X 的均值 $\mathrm{E}(X)$ 记为 μ，方差 $\mathrm{Var}(X)$ 记为 σ^2，则称比值 σ/μ 为 X 的**变异系数**，记为 C_V.

方差是反映随机变量 X 取值分散程度最常见的数字特征，但