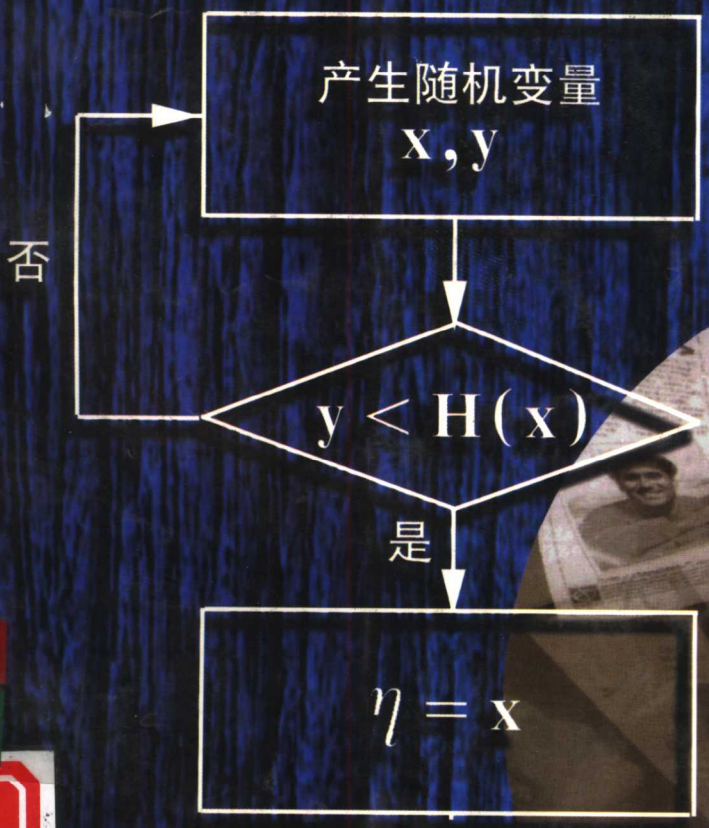
周德才 孙亦鸣 编著

计算机随机模拟原理、方法及计算程序



·华中理工大学出版社·

计算机随机模拟原理、方法及计算程序

周德才　孙亦鸣　编著

华中理工大学出版社

图书在版编目(CIP)数据

计算机随机模拟原理、方法及计算程序/周德才，孙亦鸣编著
武汉：华中理工大学出版社， 1998 年 11 月
ISBN 7-5609-1844-1

Ⅰ. 计…
Ⅱ. ①周… ②孙…
Ⅲ. 计算机模拟
Ⅳ. TP302.1

计算机随机模拟原理、方法及计算程序
周德才 孙亦鸣 编著
责任编辑： 龙纯曼
*
华中理工大学出版社出版发行
（武昌喻家山 邮编：430074）
新华书店湖北发行所经销
华中理工大学出版社照排室排版
华中理工大学出版社印刷厂印刷
*
开本：787×1092 1/16 印张：14.75 字数：360 000
1998 年 11 月第 1 版 1998 年 11 月第 1 次印刷
印数：1—1 500
ISBN 7-5609-1844-1/TP・301
定价：18.80 元

内 容 提 要

本书较全面、系统地介绍了计算机随机模拟的基本原理、方法及一些计算程序. 内容包括：实验数据的预处理与几类基本算法；常用统计分布的数值计算；随机数与 Monte Carlo 方法；多元统计分析和过程统计等及用 Fortran 语言编写的 26 个计算程序.

全书的基本内容只要具备概率论基础知识的读者就可读懂. 本书可作为高等院校教材或教学参考书，也可作为科学工作者及实际工作者的参考书.

前　言

计算机科学和概率统计研究成果的结合，形成了具有独特风格的数值计算方法的一个分支——计算机随机模拟，它既能解决大量的随机性的问题，也能求解各种确定性的数学问题. 随着科学技术发展的需要，使用计算机随机模拟以替代耗资巨大的物理实验成为当代科技发展的必要手段和潮流，而这方面的研究正是计算机随机模拟的基本内容之一，因此，计算机随机模拟的地位越来越重要了.

目前，较实用的计算机随机模拟的书籍比较缺乏，而涉及这方面的资料，或偏重于理论，或偏重于计算程序. 本书的目的在于，从实际应用的角度出发，对计算机随机模拟原理与方法进行较全面、较系统的介绍，在编写过程中，使内容深入浅出，通俗易懂，以适应科研工作者和广大的实际应用工作者的需要.

全书共分五章. 第一章介绍了实验数据的预处理方法和几类基本算法，以提高实验数据的分析效率，获取随机模拟所需的统计信息，为设计最佳随机模型作必要的准备；第二章介绍了计算机随机模拟常用统计分布的数值计算，讨论了离散分布(如二项分布，泊松分布)可归为连续分布的数值计算原理；第三章介绍了独具风格的 Monte Carlo 方法(随机抽样的基本原理与方法)，并给出了相应的基本计算程序；第四章简单介绍了多元统计分析的几种常用方法；第五章介绍了过程统计的数字时间序列分析方法，主要有时域分析、频域分析及 FFT 算法.

本书可作为高等院校的教科书或教学参考书，也可作为计算数学工作者和从事科学计算的科技工作者的参考书.

中国科学院计算中心张建中研究员，杨自强研究员对本书提出了宝贵意见. 华中理工大学随机研究中心黄志远教授，华中理工大学理学院和华中理工大学出版社对本书的出版给予了大力的支持和帮助，在此谨向他们表示衷心的感谢.

由于作者水平有限，书中难免有不足之处，我们衷心希望读者提出宝贵意见.

编者

1998 年元月于华中理工大学

目　录

引　论

计算机随机模拟，是概率论、数理统计、计算数学和计算机科学等学科的一个交叉性、边缘性及应用性很强的学科分支. 计算机随机模拟利用概率论、数理统计中提供的概率统计模型，在数字计算机上进行模拟计算，对试验观测数据或随机模拟数据进行统计分析或处理，给出实际问题性质的统计描述、统计预测和适时控制. 计算机随机模拟在科学技术、国防建设、工农业生产以及社会科学等各方面有着广泛的应用.

从计算方法的角度看，计算机随机模拟可概括为实验数据的预处理与统计描述，统计分布函数和分位数的计算，数据的统计分析，过程统计和蒙特-卡洛(Monte-Carlo)方法等，五类计算问题.

1. 实验数据的预处理与统计描述

为了保证实验数据统计分析的准确性和可靠度，在进行统计分析以前，对实验数据要进行必要的检查修正，对异常数据进行舍弃或加工，对缺失了的数据进行适当的填补，有时还需要将数据按大小排序以及对数据进行适当的变换处理(例如指数变换、取对数变换)等等，这包含计算有关数据的位置特征参数(即代表性参数，例如均值、分位数、极小值和极大值)，散布特征参数(如极差、方差、变异系数)，分布特征参数(如偏度系数、峰度系数)，及几组数据间相关特征参数(如相关系数)，并且给出直方图，以弄清数据的基本形态、散布中心和散布范围，以利计算机随机模拟的进一步处理和统计分析.

2. 统计分布函数和分位数(单侧百分位数)的计算

统计分析结论的可靠程度依赖于假设检验. 有电子计算机进行模拟计算，就没有必要用查表的方式寻找所需要的分布函数及分位数进行假设检验. 因此，分布函数和分位数算法的研究在计算机随机模拟中有着基本的意义.

3. 数据的统计分析

数据的统计分析是计算机随机模拟的核心问题. 数据的统计分析包括数理统计中各种常见的统计分析方法，内容相当丰富，应用极其广泛，如参数统计、统计检验、方差分析、回归分析，多元统计分析中的因子分析、非线性映射、主成分分析(用于简化多元数据结构)、典型相关分析、判别分析和聚类分析，以及多元数据的图表示等.

4. 过程统计

如数字时间序列分析、预报，快速傅氏变换与谱计算. 虽然过程统计涉及更多的概率论的内容，但在计算方法方面，离不开上述的统计分析方法. 因此越来越多的人认为“过程”是统计，其应用也相当广泛.

5. 蒙特-卡洛方法

蒙特-卡洛方法就是统计试验方法. 其基本思想是将计算的问题化成一个概率模型，并使模型的数字特征(如均值)为所需的计算结果，然后用抽样试验和统计方法去求这些数字特征的估计值. 所谓统计试验方法是在计算机上大量产生所需的随机变量的抽样值——随机数，然后进行相应的计算，求得估计值. 这类方法是计算机随机模拟中最具特色的一类方法，可以解决其它方法无法解决的实际问题，在理论上也可以起到补充和辅助作用.

随着电子计算机的迅猛发展，计算机随机模拟的发展速度很快，目前业已成为最活跃的学科之一. 有人根据加拿大计算中心的资料估计，目前世界上所进行的计算中有 60%以上用到计算机随机模拟的方法，在计算机随机模拟计算中有 80%以上用到回归分析. 有了电子计算机，使原来无法进行计算机随机模拟的计算得以实现；随着新技术革命浪潮的兴起，科学技术、工商管理及各行各业将提出更多的计算机随机模拟计算问题，可以预计，计算机随机模拟将踞更重要的地位，发挥更大的作用. 作为一个科学工作者，必须在计算机上熟练地从事计算机随机模拟统计分析计算工作，才能适应我国国民经济发展的需要，有效地学习国外的先进技术，顺利而有效地完成科学计算工作和模拟试验.

为讨论方便，下面引进计算机随机模拟中常见的有关概念.

1. 误差

设 x_0 为某产品参数 X(例如某电子元件的电阻)的标称值(或真值)，不管你事先怎样控制客观条件，按技术规定生产的产品的参数 X 并不是一个常数，而是一个随机变量. 设实际测量的数据为

$$x_1, x_2, \cdots, x_n, \cdots, x_N, \tag{0-1}$$

则差值

$$\varepsilon_n = x_n - x_0, \quad n = 1, 2, \cdots, N \tag{0-2}$$

称为误差(或称为实验误差)，即有

$$\varepsilon = X - x_0. \tag{0-3}$$

显然，ε 为随机变量. 引起误差的原因，按其性质大致分为三类[2][3].

1) 随机误差：由一系列偶然因素引起的一类不易控制的测量误差称为随机误差(或称为偶然误差). 实际中，随机误差是难免的，其取值可大可小，可正可负，具有统计规律性，服从一定的概率分布，大多数随机误差服从正态分布或从正态分布派生出来的分布. 随机误差最明显的特征是随着 N 的增加，随机误差取值的算术平均值愈来愈小，逐渐接近于零，即设 ε_1^* 为随机误差，则有

$$E(\varepsilon_1^*) = 0. \tag{0-4}$$

随机误差的标准差 $\sigma_{\varepsilon_1^*}$ 称为精确度.

2) 系统误差：具有确定性规律的误差称为系统误差. 一般，系统误差是一个非零常数，与随机误差不同，系统误差往往是可控制的，或可识别的. 因而，可以通过一定的手段识别并消除这类误差. 但不能通过增加实验次数 N 通过算术平均进行消除. 系统误差也称为测量设备的准确度.

3) 过失误差：一般把明显歪曲实验结果的误差称为过失误差，把含有过失误差的实验数据称为异常数据(或异常点). 可用统计的方法(如 3σ 法则，Chauvenet 法则，Grubbs 法则等)识别过失误差，对于怀疑为异常的数据，最好能分析出明确的物理或工程技术方面的原因. 在实验数据的整理加工过程中，应首先识别并舍弃异常数据，消除过失误差，以免影响计算结果.

随机误差、系统误差和过失误差往往同时存在于同一实验数据中，即有

$$\varepsilon = \varepsilon_1^* + \varepsilon_2^* + \varepsilon_3^*, \tag{0-5}$$

其中 ε 为实验误差，ε_1^* 为随机误差，ε_2^* 为系统误差，ε_3^* 为过失误差. 计算机随机模拟的基本任务之一是识别并剔除异常点，消除过失误差，找出实验数据的统计规律，估计实验数据的基本统计参数.

2. 总体和子样

计算机随机模拟研究对象的全体称为总体(又称为母体);组成总体的最小研究单位(或单元)称为个体.不过,在实际中,我们关心的常常是研究对象的某个指标 X(如灯泡的寿命,某标准件的直径等),它是一个随机变量.因而,总体通常是指某个随机变量 X 取值的全体,而每次的实验观测数据 x_n 就是一个个体,每一个体都是一个实数.如果要研究的指标不止一个,那么,可分为若干个总体来研究.

一个随机变量 X 取值小于实数 x 的可能性大小是一个在[0,1]上取值的实数,记为

$$P\{X < x\} = F(x). \tag{0-6}$$

我们称总体 X 的分布为 $F(x)$或称 $F(x)$为随机变量 X 的概率分布函数,简称为分布函数.

设 X 为具有分布函数 F 的随机变量,若 $X_1, X_2, \cdots, X_N$ 为具有同一分布 F 的相互独立的随机变量,则 $X_1, X_2, \cdots, X_N$ 为从总体 X(或总体 F)得到的容量为 N 的子样(或样本).子样是总体(全集)的部分结果(子集).如果 $x_1, x_2, \cdots, x_N$ 是实验测量的具体实数值,则 $x_1, x_2, \cdots, x_N$ 称为总体 X 的 N 个独立的观察值(或实现).显然,由于抽样(即具体的测量或观察)的随机性与独立性,每个个体 $x_n(n=1,2,\cdots,N)$都可以看作是某一随机变量 $X_n(n=1,2,\cdots,N)$所取的观察值,$(x_1, x_2, \cdots, x_N)$可以看作是 N 维向量$(X_1, X_2 \cdots, X_N)$的观察值或一个实现.以后,我们在符号上不再严格区分随机变量 X_n 与观察值 x_n,通常用 x_n 表示,根据上下文,并不难区分它们.一般,在理论上进行推导时,x_n 看作是随机变量,在具体计算中,x_n 看作是一个个体(或实现).这样可以避免经常换用符号而引起的更多的麻烦.

3. 统计量及数字特征参数

在实际问题中,往往只需要几个有代表性的数字(称为数字特征)来描述总体 x 的基本统计特征,如总体的数学期望 $E(x)=\mu$,方差 $D(x)=\sigma^2$,在离散型随机变量中,它们唯一地确定了如二项分布、泊松分布的统计规律,在连续型随机变量中,它们唯一地确定了如均匀分布、正态分布的统计规律.因此,在计算机随机模拟计算中,数字特征参数的计算是必不可少的.

用以估计总体数字特征的连续函数 $g(x_1, x_2, \cdots, x_N)$称为统计量.其中,$x_1, x_2, \cdots, x_N$ 为总体 x 的一个样本.g 不包含任何未知参数.一般称 $\hat{\theta}=\hat{\theta}(x_1, x_2, \cdots, x_N)$为总体 x 的待估参数 θ 的估计量.对应于样本的一个实现,$\hat{\theta}$ 称为 θ 的估计值,仍简记为 $\hat{\theta}$.估计量是一个随机变量,估计值是一个具体的数值,$\hat{\theta}$ 是随机变量还是估计值,可由上下文判断.

下面给出的样本均值 $\bar{x}$ 和样本方差 S^2 等都是统计量(统计量都是随机变量).$\bar{x}$ 作为总体数学期望 $E(x)=\mu$ 的无偏估计.S^2 作为总体方差 $D(x)=\sigma^2$ 的极大似然估计.如果 $x_1, x_2, \cdots, x_N$ 相互独立地来自同一总体,可以用以下数字特征参数对实验数据进行预处理和对总体进行初步的统计推断.

(1) 样本均值

样本均值是指

$$\bar{x} = \frac{1}{N}(x_1 + x_2 + \cdots + x_N) = \frac{1}{N}\sum_{n=1}^{N} x_n = \frac{1}{N}\sum_{n} x_n. \tag{0-7}$$

样本均值 $\bar{x}$ 是描述实验数据位置特征的一个参数,它表征了概率分布的中心位置.描述实验数据位置特征的参数还有众数、中位数、分位数、极小值、极大值等.这里不再一一介绍.这些位置特征参数中,样本均值用得最多.

(2) 样本方差和标准差

$$S^2 = \frac{1}{N}\sum_{n}(x_n - \bar{x})^2 = \frac{1}{N}\sum x_n^2 - \bar{x}^2, \tag{0-8}$$

$$S=\sqrt{\frac{1}{N}\sum_{n}(x_n-\bar{x})^2}. \tag{0-9}$$

(3) 标准均差

$$M=\sqrt{\frac{N}{1-\frac{2}{\pi}}}\left(\frac{1}{N}\sum\frac{|x_n-\bar{x}|}{S}-\sqrt{\frac{2}{\pi}}\right). \tag{0-10}$$

(4) 变异系数(或离差系数)

$$\mathrm{CV}=S/\sqrt{\bar{X}}. \tag{0-11}$$

样本方差、标准差、标准均差及变异系数是散布特征参数. 它们表征了实验数据在实数轴上的分散程度. 一般使用样本方差或标准差就够了,但对于不同量纲的总体进行比较时,常常用标准均差或变异系数. 它们都是无量纲的数值(即没有物理单位的无名数). 当考虑无偏性时,样本方差可用

$$S^2=\frac{1}{N-1}\sum_{n=1}^{N}(x_n-\bar{x})^2=\frac{1}{N-1}\left(\sum x_n^2-N\bar{x}^2\right). \tag{0-12}$$

当 N 较大时,样本方差可用(0-8)式,也可用(0-12)式,不会影响结果的分析.

(5) 标准偏度系数(或标准偏倚系数)

$$g_1=\sqrt{\frac{1}{6N}}\sum_{n}\left(\frac{x_n-\bar{x}}{S}\right)^3. \tag{0-13}$$

(6) 标准峰度系数(或标准峰凸系数)

$$g_2=\sqrt{\frac{N}{24}}\left[\frac{1}{N}\sum_{n}\left(\frac{x_n-\bar{x}}{S}\right)^4-3\right]. \tag{0-14}$$

偏度系数和峰度系数是描述总体密度函数 $f(x)$图形特征的分布特征参数.

当总体 x 以均值 $E(x)=\mu$ 为中心对称分布取值时,它的奇数阶中心矩

$$E[(x-\mu)^{2K+1}]=0,\quad K=0,1,2,\cdots \tag{0-15}$$

成立. 除了一阶中心矩外,任一个不为零的奇数阶中心矩都可用来衡量分布的不对称(即偏斜)程度,其中以三阶中心矩最为简单. (0-13)式就是将三阶中心矩加工的无量纲数值. 其直观意义见图 0-1. 一般当 $g_1>0$ 时,称 $f(x)$有正偏度;当 $g_1<0$ 时,称 $f(x)$有负偏度;当 $g_1=0$ 时,称 $f(x)$对称. 值得注意的是[4]:$g_1=0$ 时,$f(x)$可能根本不对称,如图 0-1(d). 稍微改动一下曲线就能随意地给出正的或者负的三阶矩. 因此,实际中不能仅就(0-13)式下结论.

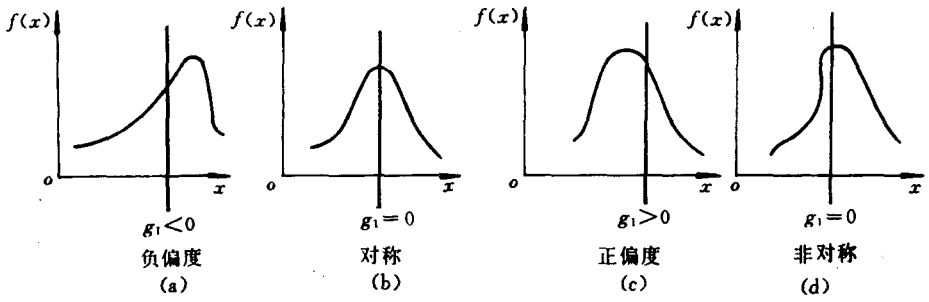


图 0-1

四阶中心矩可以用来测定密度函数 $f(x)$图形顶峰的凸平度;作为分布特征的另一个描述参数,(0-14)式是一个无量纲数值,便于与标准正态分布的峰度进行比较并进行统计检验.

(7) 线性相关系数

$$r(j)=\frac{1}{N-j}\sum_{n=1}^{N-j}\left(\frac{x_n-\bar{x}}{S}\right)\left(\frac{x_{n+j}-\bar{x}}{S}\right),\ j=1,2,\cdots,K;K\ll N \tag{0-16}$$

(0-16)式是总体理论相关系数 $\rho(j)$时滞为 j 时的估计值. $\rho(j)$是实验数据(0-1)各次实验之间相依性的一个定量测度. 若有独立性,则 $\rho(j)=0$,否则有 $0<|\rho(j)|\leqslant 1$,表示实验数据间存在着线性关系. $\rho(j)$愈接近 1,则表示实验数据间的线性关系愈强. $r(j)$给出了由 x_n 线性预报 x_{n+j}可能性的大小.

(8) 线性时关系数

$$r_{xt}=\frac{1}{N}\sum_{n}\left(\frac{x_n-\bar{x}}{S}\right)\sqrt{12}\left(\frac{n}{N}-\frac{1}{2}\right), \tag{0-17}$$

其中 n 视为实验观测时的相对时间. (0-17)式给出了实验数据 x_n 与时间 n 之间的线性相关程度,以判定 x_n 有无系统误差随时间 n 变化.

线性相关系数和线性时关系数为相关特征参数. 这些参数的计算程序见第一章 § 1.6.

5. 实验数据变换式

在实际应用中,常常需要将实验数据经过适当的变换,使之利于统计计算和统计分析. 正态分布在理论上比较完善,实际应用也比较方便. 因而,基于正态分布的合适的数据变换能扩大计算机随机模拟的应用范围和提高计算机随机模拟的实用效果. 例如,回归问题中,非线性的回归问题可以化为线性回归处理.

设

$$y_1,y_2,\cdots,y_n,\cdots,y_N \tag{0-18}$$

为原始实验数据,给出如下变换式:

$$x_n=\begin{cases}y_n, & \delta=1;\\ \alpha y_n+\beta, & \delta=2;\\ \ln(\alpha y_n+\beta), & \delta=3;\\ (\alpha y_n+\beta)^{\gamma}, & \delta=4;\\ \gamma+e^{(\alpha y_n+\beta)}, & \text{其它.}\end{cases}\qquad n=1,2,\cdots,N, \tag{0-19}$$

其中 $\alpha,\beta,\gamma,\delta$ 为变换参数.

显然,(0-19)式包括了大部分常用的数据变换公式,如恒等变换、线性变换、指数变换、对数变换、幂变换等.

第一章　实验数据的预处理和几类基本算法

§1.1　样本均值、样本方差和样本协方差的数值计算递推公式

设 $x_1,x_2,\cdots,x_n$ 是实际问题中提供的一组实验数据，多数场合下需要计算这组数据(样本)的均值和方差.

$$\bar{x}=\frac{1}{n}\sum_{k=1}^{n}x_k, \tag{1.1-1}$$

$$S^2=\frac{1}{n-1}\sum_{k=1}^{n}(x_k-\bar{x})^2\triangleq S_1^2, \tag{1.1-2}$$

或

$$S^2=\frac{1}{n-1}\Big(\sum_{k=1}^{n}x_k^2-n\bar{x}^2\Big)\triangleq S_2^2 \tag{1.1-3}$$

分别称为样本 $x_1,x_2,\cdots,x_n$ 的样本均值和样本方差，统称为样本矩. 随机变量的数字特征中，均值和方差唯一确定了二项分布(或泊松分布、正态分布、均匀分布)随机变量的基本规律. 另外，大量的实际问题也常常只需知道随机变量的均值和方差. 因此(1.1-1)、(1.1-2)、(1.1-3)式的实际计算是经常遇到的. 在理论上(1.1-1)、(1.1-2)、(1.1-3)式都是正确的公式，然而在实际计算中并不一定能得到满意的结果，特别是用(1.1-3)式计算样本方差时，计算的结果往往不令人满意. 下面用一个简单的例子加以说明.

取 $x_k=44444,k=1,2,\cdots,n$，无疑，对任何 n 值，样本均值 $\bar{x}$ 应该为常数 44444，对于(1.1-2)式及(1.1-3)式，样本方差都应该为 0. 我们分别取 $n=10,100,10000,20000$ 的不同样本量，取同一数据 $x_k=44444$ 在 IBM PC/XT 微机上试算，结果如下：

n	10	100	10000	20000
$\bar{x}$	44444	44444	44447.4	44435.03
S_1^2	0	0	11.52448	80.596
S_2^2	0	1158.465	−111999.1	1016750

计算结果表明，当 n 较大时，出现了怪结果，比如当 $n=10000$ 时，S_2^2 是相当大的负数值. 一般来说，人们并没有注意到这一点，即使是很熟悉统计理论的专家，有时也可能不能有效地解决实际问题，其原因不是概率统计理论上的错误，而是计算方法上的问题，是不知不觉地直接使用(1.1-1)、(1.1-2)和(1.1-3)式的结果. 那么，对于不同的原始数据(如 $x_i\neq x_j,i\neq j,i,j=1,2,\cdots,n$)在上述情况下，特别是 S_2^2，使得计算结果失去了意义. 由此，提出了这样一个问题，当样本容量 n 很大，且样本值 x_k 的有效数位又较多时，如何较准确地算出如均值和方差之类的统

计参数?下面我们介绍的统计计算方法可以达到这一目的,并且可以得到一系列的中间结果进行实时处理.值得注意的是,目前国内外不少的计算程序是直接按照(1.1-1)、(1.1-2)和(1.1-3)式进行计算的.当样本容量 n 不大且有效数位不多时,这类程序可以直接开发使用,但当样本容量 n 较大且有效数位较多时,这类程序则需改造使用.

一、一组样本值的情形

设 $x_1,x_2,\cdots,x_n$ 为一组样本观察值,我们用以下递推公式计算样本均值和样本方差:

$$\bar{x}^{(k)}=\bar{x}^{(k-1)}+\frac{1}{k}(x_k-\bar{x}^{(k-1)}), \tag{1.1-4}$$

$$S^{(k)}=S^{(k-1)}+(x_k-\bar{x}^{(k-1)})(x_k-\bar{x}^{(k)}), \tag{1.1-5}$$

其中定义

$$\bar{x}^{(k)}=\frac{1}{k}\sum_{t=1}^{k}x_t,\qquad k=1,2,\cdots,n, \tag{1.1-6}$$

$$S^{(k)}=\sum_{t=1}^{k}(x_t-\bar{x}^{(k)})^2,\quad k=1,2,\cdots,n, \tag{1.1-7}$$

初始值

$$\bar{x}^{(0)}=0,S^{(0)}=0.$$

最后结果:

样本均值 $\bar{x}=\bar{x}^{(n)}$, (1.1-8)

样本方差 $S^2=\frac{1}{n-1}S^{(n)}$. (1.1-9)

证明 1)

$$\begin{aligned}\bar{x}^{(k)}&=\frac{1}{k}\sum_{t=1}^{k}x_t=\frac{1}{k}\Big(\sum_{t=1}^{k-1}x_t+x_k\Big)=\frac{1}{k}\sum_{t=1}^{k-1}x_t+\frac{1}{k}x_k\\&=\frac{1}{k}\cdot(k-1)\cdot\frac{1}{k-1}\sum_{t=1}^{k-1}x_t+\frac{1}{k}x_k=\frac{k-1}{k}\bar{x}^{(k-1)}+\frac{1}{k}x_k\\&=\bar{x}^{(k-1)}-\frac{1}{k}\bar{x}^{(k-1)}+\frac{1}{k}x_k=\bar{x}^{(k-1)}+\frac{1}{k}(x_k-\bar{x}^{(k-1)}).\end{aligned}$$

2) 证明 $S^2=\frac{1}{n-1}S^{(n)}$,即证明差方和递推公式

$$\sum_{t=1}^{k}(x_t-\bar{x}^{(k)})^2=S^{(k-1)}+(x_k-\bar{x}^{(k-1)})(x_k-\bar{x}^{(k)})$$

成立.因为

$$\sum_{t=1}^{k}(x_k-\bar{x}^{(k-1)})(x_t-\bar{x}^{(k)})=(x_k-\bar{x}^{(k-1)})\sum_{t=1}^{k}(x_t-\bar{x}^{(k)})=0, \tag{1.1-10}$$

所以

$$\begin{aligned}\sum_{t=1}^{k}(x_t-\bar{x}^{(k)})^2&=\sum_{t=1}^{k}(x_t-\bar{x}^{(k)})(x_t-\bar{x}^{(k)})\\&\xlongequal{由(1.1-4)}\sum_{t=1}^{k}\Big[x_t-\bar{x}^{(k-1)}-\frac{1}{k}(x_k-\bar{x}^{(k-1)})\Big](x_t-\bar{x}^{(k)})\\&\xlongequal{由(1.1-10)}\sum_{t=1}^{k}(x_t-\bar{x}^{(k-1)})(x_t-\bar{x}^{(k)})\\&\xlongequal{由(1.1-10)}\sum_{t=1}^{k-1}(x_t-\bar{x}^{(k-1)})(x_t-\bar{x}^{(k)})+(x_k-\bar{x}^{(k-1)})(x_k-\bar{x}^{(k)})\\&\xlongequal{由(1.1-4)}\sum_{t=1}^{k-1}(x_t-\bar{x}^{(k-1)})\Big[(x_t-\bar{x}^{(k-1)})-\frac{1}{k}(x_k-\bar{x}^{(k-1)})\Big]\\&\quad+(x_k-\bar{x}^{(k-1)})(x_k-\bar{x}^{(k)})\end{aligned}$$

$$\xrightarrow{由(1.1\text{-}10)}\sum_{t=1}^{k-1}(x_t-\bar{x}^{(k-1)})^2+(x_k-\bar{x}^{(k-1)})(x_k-\bar{x}^{(k)})$$

$$\xrightarrow{由(1.1\text{-}10)}S^{(k-1)}+(x_k-\bar{x}^{(k-1)})(x_k-\bar{x}^{(k)}).$$

利用(1.1-5)和(1.1-9)式计算样本方差，避免了利用(1.1-3)式中$\sum_{k=1}^{n}x_k^2$在计算机中可能发生的溢出现象，提高了计算结果的效率和精度.这里，再一次提请注意，在实际计算过程中应尽量避免使用(1.1-3)式.如果希望计算总体方差的极大似然估计值，也可用递推公式

$$S_1^{(k)}=\frac{k-1}{k}(S_1^{(k-1)}+\frac{1}{k}(x_k-\bar{x}^{(k-1)})^2),\tag{1.1-11}$$

其中定义

$$\bar{x}^{(k)}=\frac{1}{k}\sum_{t=1}^{k}x_t,\qquad k=1,2,\cdots,n,\tag{1.1-12}$$

$$S_1^{(k)}=\frac{1}{k}\sum_{t=1}^{k}(x_t-\bar{x}^{(k)})^2,\quad k=1,2,\cdots,n,\tag{1.1-13}$$

初始值

$$\bar{x}^{(0)}=0,\quad S_1^{(0)}=0,$$

最后结果

$$\bar{x}=\bar{X}^{(n)},\tag{1.1-14}$$

$$S_1^2=S_1^{(n)}.\tag{1.1-15}$$

同样取$x_k=44444,k=1,2,\cdots,n,n$分别取10、100、10000、20000，在IBM PC/XT上试算的结果都是$\bar{x}=44444,S^2=0$.这样，对于相当大的n(如上亿个数据)且有效位很多(如8位)的不同数据，用上述递推式计算出来的S^2真正反应了原始数据的离散程序.

二、二元样本协方差的递推公式

实际问题中，除了需要计算某一变量的样本均值和方差以外，还需要计算某两个变量之间的协方差，进而计算相关系数以了解这两个变量之间(或二组数据之间)的相关程度.

设$x_1,x_2,\cdots,x_n$和$y_1,y_2,\cdots,y_n$为两组不同的样本，样本容量都为n，那么满足无偏性要求的样本均值和样本协方差为

$$\bar{x}=\frac{1}{n}\sum_{t=1}^{n}x_t,\tag{1.1-16}$$

$$\bar{y}=\frac{1}{n}\sum_{t=1}^{n}y_t,\tag{1.1-17}$$

$$\hat{v}_{xy}=\frac{1}{n-1}\sum_{t=1}^{n}(x_t-\bar{x})(y_t-\bar{y}).\tag{1.1-18}$$

定义

$$\bar{x}^{(k)}=\frac{1}{k}\sum_{t=1}^{k}x_t,\tag{1.1-19}$$

$$\bar{y}^{(k)}=\frac{1}{k}\sum_{t=1}^{k}y_t,\tag{1.1-20}$$

$$S_{xy}^{(k)}=\sum_{t=1}^{k}(x_t-\bar{x}^{(k)})(y_t-\bar{y}^{(k)}),\tag{1.1-21}$$

其中$k=1,2,\cdots,n$，有递推公式

$$S_{xy}^{(k)}=S_{xy}^{(k-1)}+(x_k-\bar{x}^{(k-1)})(y_k-\bar{y}^{(k)}),\tag{1.1-22}$$

其中 $k=1,2,\cdots,n$ 和 $\bar{x}^{(0)}=\bar{y}^{(0)}=S_{xy}^{(0)}=0$.

最后得样本协方差的无偏估计

$$\hat{v}_{xy}=\frac{1}{n-1}S_{xy}^{(n)}. \tag{1.1-23}$$

显然，当 $x_1,x_2,\cdots,x_n$ 和 $y_1,y_2,\cdots,y_n$ 是同一样本时，(1.1-23)式就是样本方差.(1.1-22)式的证明方法与(1.1-5)式的证明方法完全相同，这里不再重复.

三、p 元($p>2$)样本均值(向量)和协方差(矩阵)的递推公式

设

$$\begin{matrix} x_{11},x_{12},\cdots,x_{1p} \\ x_{21},x_{22},\cdots,x_{2p} \\ \cdots\cdots \\ x_{n1},x_{n2},\cdots,x_{np} \end{matrix} \tag{1.1-24}$$

为 p 元样本的 n 组观察资料，用向量表示为

$$\boldsymbol{x}_k=(x_{k1},x_{k2},\cdots,x_{kp})',\quad k=1,2,\cdots,n. \tag{1.1-25}$$

满足无偏性要求的样本均值(向量)和样本协方差(矩阵)为

$$\bar{\boldsymbol{x}}=\frac{1}{n}\sum_{k=1}^{n}\boldsymbol{x}_k, \tag{1.1-26}$$

$$\hat{V}=\frac{1}{n-1}\sum_{k=1}^{n}(\boldsymbol{x}_k-\bar{\boldsymbol{x}})(\boldsymbol{x}_k-\bar{\boldsymbol{x}})'. \tag{1.1-27}$$

定义

$$\begin{cases} \bar{\boldsymbol{x}}^{(k)}=\dfrac{1}{k}\displaystyle\sum_{t=1}^{k}\boldsymbol{x}_t, \\ S^{(k)}=\sum(\boldsymbol{x}_t-\bar{\boldsymbol{x}}^{(k)})(\boldsymbol{x}_t-\bar{\boldsymbol{x}}^{(k)})',k=1,2,\cdots,n; \end{cases}$$

$$\bar{\boldsymbol{x}}^{(0)}=\boldsymbol{0},\quad S^{(0)}=O,$$

易证有如下递推公式：

$$\begin{cases} \bar{\boldsymbol{x}}^{(k)}=\bar{\boldsymbol{x}}^{(k-1)}+\dfrac{1}{k}(\boldsymbol{x}_k-\bar{\boldsymbol{x}}^{(k-1)}), & (1.1\text{-}28) \\ S^{(k)}=S^{(k-1)}+(\boldsymbol{x}_k-\bar{\boldsymbol{x}}^{(k-1)})(\boldsymbol{x}_k-\bar{\boldsymbol{x}}^{(k)})', & (1.1\text{-}29) \end{cases}$$

$$k=1,2,\cdots,n.$$

最终有

$$\bar{\boldsymbol{x}}=\bar{\boldsymbol{x}}^{(n)}, \tag{1.1-30}$$

$$\hat{V}=\frac{1}{n-1}S^{(n)}. \tag{1.1-31}$$

数据的预处理包括计算均值、方差和协方差.采用递推算法，通常能提高数据预处理的效率和精度，如回归分析计算.特别是逐步回归分析计算中，开始的准备工作就是计算离差阵或相关阵，若这方面的计算精度不高，必然会影响后续的分析计算，甚至会使计算的结果毫无意义.

值得注意的是，利用递推公式(1.1-28)、(1.1-29)，顺便可以得到由于某种需要追加或删去一个样本时的最为方便的计算公式，不需从头算起.设从 n 个样本中删去 $\boldsymbol{x}_i$ 之后，相应的样本均值、协差阵、离差阵分别记为 $\bar{\boldsymbol{x}}(-i)$，$\hat{V}(-i)$，$S(-i)$，由(1.1-28)、(1.1-29)式易得

$$\bar{\boldsymbol{x}}(-i)=\bar{\boldsymbol{x}}+\frac{1}{n-1}(\bar{\boldsymbol{x}}-\boldsymbol{x}_i)\overset{\text{或}}{=\!=}\frac{n}{n-1}\left(\bar{\boldsymbol{x}}-\frac{\boldsymbol{x}_i}{n}\right), \tag{1.1-32}$$