TONGJI JISUAN FANGFA

统计计算方法

程兴新 曹 敏 编著

北京大学出版社

一九八九年十一月

内 容 提 要

本书较系统地介绍了统计计算的基本方法，并讲解了各种计算方法的构思及其内在联系。内容包括：概率分布函数及其分位点的计算方法；随机数的产生；矩阵计算及各种回归分析方法；试验设计中的统计分析方法；最优化方法，以及主成份分析、典型相关分析、因子分析和Fisher判别分析等多元统计分析方法。本书各章都配有例题、习题、上机作业及习题解答，并选入了较多的方法和算法以便于实际工作者查阅。

本书可作为理工院校数学系、应用数学系、概率统计系、计算机系的教材，也可以作为教师、研究生以及从事计算机、统计工作的实际工作者的参考书。

统计计算方法

程兴新 曹 敏 编著

责任编辑 徐信之

北京大学出版社出版

（北京大学校内）

北京大学印刷厂印刷

新华书店北京发行所发行 各地新华书店经售

850×1168毫米 32开本 9.625印张 240千字

1989年10月第一版 1989年10月第一次印刷

印数：0001—3,500册

ISBN 7-301-00794-9/O·142

定价：2.35元

序　言

统计计算是应用统计的一个分支，它是研究用统计方法进行数据处理时所需的计算方法，其目的是将统计方法与计算方法结合起来，以解决应用统计中出现的各种实际问题。

由于计算机的应用日益广泛，人们在越来越多的领域中使用统计分析的方法解决实际问题。对实际工作者来说，除了需要具备基本的概率统计知识以外，大量遇到的是如何运用统计方法的问题，以及如何解决在应用中出现的计算问题。目前介绍这方面内容的书籍比较缺乏，而且有的偏重于具体的计算程序，有的偏重于其理论依据，致使读者不能根据各自的需要从中灵活掌握解决实际问题的方法。编写本书旨在弥补这些欠缺，并使高等院校能有一本较系统地介绍统计计算的教科书。

本书较全面、系统地介绍了统计计算中的基本计算方法，使读者不仅能掌握这些方法，并且还能结合工作的实际需要，灵活运用这些方法编写出具体的程序，以解决各种实际问题。考虑到本书既是高等院校的教科书或教学参考书，又可作为实际工作者查阅的手册，所以在编写过程中我们尽量做到内容深入浅出，文字浅显易懂，并汇集了较多的方法和算法。同时我们力求从不同角度阐明各种方法间的内在联系，使各部分内容能有机地联系起来，尽可能地体现其特色。

全书共分七章。第一章结合数值积分、函数逼近以及方程求根等计算方法，介绍概率分布函数及其分位点的计算，并对一些常用分布给出其算法；第二章详细地讨论产生随机数的方法，给出常用分布的随机数的产生算法；第三章系统地介绍统计计算中所需的矩阵计算方法；第四章使用多种方法处理线性回归分析中的各种问题；第五章介绍试验设计中的统计分析方法及算法；第

六章讨论最优化方法及其在非线性回归分析中的应用，并介绍在统计应用中的一个重要的优化方法——EM 方法；第七章介绍主成份分析、典型相关分析、因子分析和判别分析等多元统计分析方法及算法。

本书第一、二、六章的内容具有相对的独立性，读者可根据需要选读其中某一章；因第三章是第四、五、七章的基础，读者可以先选读第三章的内容，再阅读后面的章节。我们假定读者已具备一般的工科高等数学知识。

本书在正文中还经常涉及到一些参考文献，其目的是引导读者对有关内容作进一步地研究。考虑到教学的需要，本书在各章末还选编了部分习题，以帮助读者加深对各章内容的理解。

在本书的编写过程中，北京大学数学系胡德焜副教授、周芝英副教授和雷功炎副教授仔细地审阅了初稿，并为本书的内容编排及细节的修改提出了许多建设性意见，在此一并致谢。

由于编写时间紧迫，我们水平有限，书中难免有不足之处，我们衷心地欢迎读者随时指教，以便再版时改正。

目　　录

第一章　分布函数及其分位点的计算

在讨论分布函数及其分位点的计算之前，我们先简单介绍一下与此有关的基本概念。

在实际工作中，我们经常要对一些试验对象进行观测，以便从中获得原始数据。这些观测对象可以是温度、压力，也可以是小球的个数或信号灯的颜色。每一个具体的或抽象的单个对象在数学上都可以看成是一个变量，对它的实际观测所得到的数据则看成是这个变量的取值。这些变量的取值往往受到某些随机因素的影响，使这些变量的取値带有不确定性或随机性，我们称这种取值带有随机性的变量为随机变量。若随机变量可以连续取值，则称该随机变量为连续型随机变量，如温度、压力及炮弹的弹着点位置等；若随机变量只能取离散值，则称该随机变量为离散型随机变量，如小球的个数及信号灯的颜色等。在数学上，我们用概率来描述随机变量取值具有怎样的随机性。设ξ是一个随机变量，ξ的取值小于或等于实数x的概率$P_r(\xi\leqslant x)$是一个随x变化的实函数，我们称$F(x)=P_r(\xi\leqslant x)$为ξ的分布函数。$P_r(\xi\leqslant x)$可以直观地认为是“ξ的取值小于等于x”这一事件发生的“可能性”的数量描述。对连续型随机变量，$F(x)$通常可写成

$$F(x)=\int_{-\infty}^{x} f(t)\mathrm{d}t, \tag{1.1}$$

其中$f(t)$称为连续型随机变量ξ的分布密度函数。当ξ是离散型随机变量时，它只能取离散型数值$x_1<x_2<\cdots<x_n<\cdots$。这时$\xi$的分布函数可以写成

$$F(x_k)=P_r(\xi\leqslant x_k)=\sum_{i=1}^{k} P_r(\xi=x_i), \tag{1.2}$$

其中随 i 变化的离散函数 $P_r(\xi=x_i)$ 称为离散型随机变量 ξ 的概率分布函数，简称分布函数。

分布函数不仅能反映随机变量可能取什么值，而且描述了随机变量取这些值的概率。因此，分布函数完全刻画了随机变量的“随机性”。基于这种观点分析客观事物的方法称为概率统计方法。刻画分布函数的两个数量指标是期望和方差。若 ξ 是连续型随机变量，其分布密度为 $f(t)$，分布函数为 $F(x)$，则 ξ 的期望和方差为

$$E(\xi)=\int_{-\infty}^{+\infty}x\mathrm{d}F(x), \qquad D(\xi)=\int_{-\infty}^{+\infty}(x-E\xi)^2\mathrm{d}F(x).$$

由分布函数的定义可以知道 $F(x)$ 的取值范围是 $0\leqslant F(x)\leqslant 1$。任取 $p\in[0,1]$，若存在 x_p，使

$$F(x_p)=p, \tag{1.3}$$

则称 x_p 为分布函数 $F(x)$ 的 p-分位点。

在对数据进行统计分析时，经常要计算分布函数 $F(x)$ 及其分位点。计算分布函数的方法之一是用分布函数的数表。即先把数表存入计算机内，需要时调出来使用。但这样做需要占据很多计算机内存，而且使用起来也不方便。另一种计算分布函数的方法是利用积分的严格公式，但它的程序较难，计算时所需时间也较长。比较好的方法是利用分布函数的近似公式进行计算，可先将计算分布函数的程序存入机内，需要时随时调出使用。下面我们就来讨论有关的计算方法，并且在最后给出一些常用的分布函数及分位点的计算公式。

§1 常用的分布函数

为了对分布函数有比较具体的印象，我们先给出一些常用分布函数的表达式。

1） 正态分布

正态分布的分布密度函数及分布函数分别为

$$f(x)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}\quad(-\infty<x<+\infty),$$

$$F(x)=\int_{-\infty}^{x}f(t)\mathrm{d}t,$$

其中μ和σ是参数。正态分布通常记为$N(\mu,\sigma^2)$。特别地，当$\mu=0$，$\sigma^2=1$时，称该正态分布为标准正态分布，通常记为$N(0,1)$，它的分布密度和分布函数分别记为$\phi(x)$，$\Phi(x)$。

2）t分布

t分布的分布函数为

$$T(x|n)=\int_{-\infty}^{x}f(t|n)\mathrm{d}t$$

$$=\frac{1}{\sqrt{n}B\left(\frac{1}{2},\frac{n}{2}\right)}\int_{-\infty}^{x}\left[1+\frac{t^2}{n}\right]^{-\frac{n+1}{2}}\mathrm{d}t$$

$$(-\infty<x<+\infty),$$

其中n称为自由度，$B(a,b)$是Beta函数：

$$B(a,b)=\int_0^1 t^{a-1}(1-t)^{b-1}dt\qquad(a>0,b>0).$$

3）Beta分布

Beta分布的分布函数$I_x(a,b)$是由Beta函数构成的。

$$I_x(a,b)=\frac{1}{B(a,b)}\int_0^x t^{a-1}(1-t)^{b-1}dt$$

$$(x\in[0,1],\quad a>0,\quad b>0),$$

其中a和b是参数。

4）Gamma分布

Gamma分布的分布函数$G(x|a,b)$是由Gamma函数构成的。

$$G(x|a,b)=\frac{1}{b^a\Gamma(a)}\int_0^x t^{a-1}\exp\left(-\frac{t}{b}\right)\mathrm{d}t$$

$$(a>0,\quad b>0,\quad x\geqslant 0),$$

其中$\Gamma(a)$是Gamma函数：

$$\Gamma(a)=\int_0^{+\infty}t^{a-1}\exp(-t)\mathrm{d}t.$$

Beta函数与Gamma函数之间有下面的关系

$$B(a,b)=\frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}.$$

5）χ^2分布

χ^2分布的分布函数为

$$H(x|n)=\left[2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)\right]^{-1}\int_0^x t^{\frac{n-2}{2}}\exp\left(-\frac{t}{2}\right)\mathrm{d}t$$

$$(x\geqslant 0,\quad n\text{ 为整数}).$$

通常称 n 为χ^2分布的自由度. 不难看出，χ^2分布实际上是 Gamma 分布的特例，即 $H(x|n)=G\left(x\middle|\frac{n}{2},2\right)$.

6） F分布

F 分布的分布函数$F(x|n_1,n_2)$为

$$F(x|n_1,n_2)=\frac{n_1^{\frac{n_1}{2}}n_2^{\frac{n_2}{2}}}{B\left(\frac{n_1}{2},\frac{n_2}{2}\right)}\int_0^x t^{\frac{n_1}{2}-1}(n_2+n_1t)^{-\frac{n_1+n_2}{2}}\mathrm{d}t$$

$$(x>0,\quad n_1>0,\quad n_2>0),$$

其中参数n_1和n_2称为F分布的自由度.

7）二项分布

二项分布是离散型随机变量的概率分布，它的分布密度为

$$f(x|n,p)=\binom{n}{x}p^x(1-p)^{n-x}\quad(x=0,1,\cdots,n),$$

其中$p\in(0,1)$和 n 是参数，并且

$$\binom{n}{x}=\frac{n!}{n!(n-x)!}.$$

8）Poisson分布

Poisson分布也是离散型随机变量的概率分布，它的分布密度

是

$$f(x|\lambda)=\frac{\lambda^x}{x!}e^{-\lambda} \qquad (x=0,1,\cdots),$$

其中$\lambda>0$是参数。

从上面给出的常用分布我们可以看出，它们的分布密度函数的计算比较简单，但要计算它们的分布函数却比较复杂，其中涉及到积分与求和运算。在实际工作中我们还会遇到很多比上述常用分布更为复杂的分布函数，因此有必要掌握一些求分布函数方面的基本的计算方法和计算技巧，这就是本章要介绍的主要内容。

§2 计算分布函数的一般方法

我们可以用各种数值逼近的方法计算分布函数。数值逼近方法的主要思想是，根据要计算的函数的复杂程度选取适当的结构简单的函数，用这些较简单的函数近似代替较复杂的函数进行计算，从而达到简化计算过程的目的。这里我们将介绍几种主要的数值逼近方法。由于分布函数实际上是一个积分，因此先讨论数值求积方法。

2.1 Newton-Cotes 型求积公式

我们的目的是计算积分$\int_a^b f(x)\mathrm{d}x$。显然，如果能用较简单的函数近似地代替$f(x)$，那么计算过程就会简化。由于多项式的积分很容易计算，所以我们选用$f(x)$在$[a,b]$上的插值多项式来近似代替$f(x)$。设已知$f(x)$在$[a,b]$上的若干点$x_0,x_1,\cdots,x_n$处（为方便起见，可选n个等分点）的值为$f(x_0),f(x_1),\cdots,f(x_n)$，定义$f(x)$在$[a,b]$上的插值多项式为

$$L_n(x)=\sum_{i=0}^{n} l_i(x)f(x_i), \tag{1.4}$$

其中

$$l_i(x)=\frac{\omega_{n+1}(x)}{(x-x_i)\omega'_{n+1}(x_i)},$$

$$\omega_{n+1}(x)=(x-x_0)(x-x_1)\cdots(x-x_n),$$

$L_n(x)$称为拉格朗日插值多项式。易知$L_n(x)$与$f(x)$在点$x_0,x_1,\cdots,x_n$上具有相同的函数值。将$f(x)$写成

$$f(x)=L_n(x)+R_n(x),\tag{1.5}$$

其中$R_n(x)$是用$L_n(x)$近似$f(x)$时的误差函数。可以证明，当$f(x)$有$n+1$阶有界导数时，$R_n(x)$可表成

$$R_n(x)=\frac{\omega_{n+1}(x)}{(n+1)!}f^{(n+1)}(\xi)\qquad(\xi\in[a,b])。\tag{1.6}$$

特别地，当 $f^{(n+1)}(x)\equiv 0$时，$R_n(x)\equiv 0$。也就是说，当$f(x)$是任意不高于n阶的多项式函数时，恒有

$$f(x)=L_n(x)。\tag{1.7}$$

这一事实非常重要。

对(1.5)式两边积分，得到

$$\int_a^b f(t)\mathrm{d}t$$

$$=\sum_{i=0}^{n}f(x_i)\int_a^b l_i(t)\mathrm{d}t+\int_a^b R_n(t)\mathrm{d}t.$$

据此得到积分计算公式

$$\int_a^b f(t)\mathrm{d}t\doteq\sum_{i=0}^{n}f(x_i)A_i,\quad A_i=\int_a^b l_i(t)\mathrm{d}t。\tag{1.8}$$

此计算公式的误差为

$$E=\int_a^b R_n(t)\mathrm{d}t.$$

(1.8)式称为Newton–Cotes 型积分公式，A_i 称为 Cotes 系数。由(1.7)式可以看出，当$f(x)$为任意不高于n阶的多项式函数时，公式(1.8)无误差地精确成立。一些常用的求积公式是 Newton–Cotes求积公式的特例。如，若取$n=1,x_0=a,x_1=b$,则(1.8)式变成梯型求积公式

$$\int_a^b f(t)\mathrm{d}t\doteq\frac{b-a}{2}(f(a)+f(b))。$$

又若取 $n=2$，$x_0=a$，$x_1=(a+b)/2$，$x_2=b$，则(1.8)式变为 Simpson 求积公式

$$\int_a^b f(t)\,\mathrm{d}t \doteq \frac{1}{6}(b-a)\left[f(a)+4f\left(\frac{a+b}{2}\right)+f(b)\right].$$

有关这方面的详细内容，可参见[8]。

公式(1.8)的计算精度主要取决于插值函数 $L_n(x)$ 与 $f(x)$ 的近似程度。当区间 $[a,b]$ 很宽而 n 很小时，$L_n(x)$ 与 $f(x)$ 会相差较大，从而导致计算精度降低；而当 n 足够大时，虽然 $L_n(x)$ 与 $f(x)$ 误差较小，但计算量会大大增加，并且由于 n 较大，故按(1.8)式求和时累计误差也越大，因此最终的计算精度及计算速度也不能令人满意。为了克服这些缺点，使计算的结果既能保证计算精度又能加快计算速度，我们可以使用复合求积方法：先将积分区间分割成若干个小区间，即 $[a,b]=\bigcup_{i=1}^{N} I_i$；然后在每一小区间 $I_i(i=1,\cdots,N)$ 上构造低阶的 Newton-Cotes 求积公式

$$\int_{I_i} f(t)\,\mathrm{d}t \doteq \sum_{j=0}^{M} f(x_j)A_j, \quad A_j=\int_{I_i} l_i(t)\,\mathrm{d}t;$$

最后通过求和

$$\int_a^b f(t)\,\mathrm{d}t \doteq \sum_{i=1}^{N}\int_{I_i} f(t)\,\mathrm{d}t \tag{1.9}$$

得到积分值。(1.9)式称为复合求积公式。读者不难由此公式，导出常用的复合梯形求积公式和复合 Simpson 求积公式。

复合求积公式在实际应用中的一个缺点是，必须事先由误差公式确定所分割的小区间的个数 N，而这样做实际上是很困难的。克服这一缺点的办法是：依次对不同的 $N(N=1,2,3,\cdots)$，由复合求积公式计算积分值，得到积分值序列 $T_0,T_1,T_2,\cdots$，当相邻的两个积分值 T_{i-1} 与 T_i 的误差足够小时停止计算，取 T_i 为最终的计算结果。在构造此类算法时应尽量减小重复计算量，例如在计算 T_i 时，应尽量利用计算 $T_0,T_1,\cdots,T_{i-1}$ 时得到的信息使算法简单。为了加快 $\{T_i\}$ 序列的收敛速度还可以使用Romberg外

推积分法。这一方法的具体步骤请读者参考[4]。

2.2　Gauss型求积公式

Gauss型求积公式是在 Newton-Cotes 型求积公式上发展起来的。我们知道，在Newton-Cotes型求积公式

$$\int_a^b f(t)\mathrm{d}t = \sum_{i=0}^{n} f(x_i)A_i \qquad (1.10)$$

中，插值点 x_i 是可以任意选取的。那么是否可以通过适当选取 $n+1$ 个插值点而使积分公式(1.10)的“精度”进一步提高呢？答案是肯定的。Gauss 型求积公式就是通过给出插值点的确定方法而得到的精度更高的Newton-Cotes型求积公式。

为了给出确定插值点的方法，先引进代数精度的概念。

定义　如果一个求积公式对任意阶数不高于 m 的多项式精确成立，那么称此求积公式的代数精度为 m。

从(1.7)式可以看出，当 $f(x)$ 为低于 n 阶的多项式时，$f(x)$ 恒等于其 $n+1$ 点插值多项式，故由此插值多项式构造的 Newton-Cotes 型积分公式无误差地精确成立，所以它的代数精度不小于 n。事实上可以证明，Newton-Cotes型求积公式的代数精度最大为 $2n+1$。在求插值多项式时，适当地选取插值点可提高求积公式的代数精度。事实上，若适当地选取 $n+1$ 个插值点，则可使Newton-Cotes型求积公式的代数精度提高，达到其最大精度 $2n+1$。我们将代数精度达到 $2n+1$ 的Newton-Cotes型 $n+1$ 点求积公式称为Gauss型 $n+1$ 点求积公式，并将相应的插值点称为Gauss型基点。为了寻找 Gauss 型求积公式，我们引进直交函数系的概念。

定义　如果两个函数 $Q(x)$ 和 $R(x)$ 满足

$$\int_a^b Q(x)R(x)\mathrm{d}x = 0,$$

则称 $Q(x)$ 与 $R(x)$ 在区间 $[a,b]$ 上正交。若函数列 $\{f_n(x)\}$ 中的任意两个函数都在区间 $[a,b]$ 上正交，则称 $\{f_n(x)\}$ 为区间 $[a,b]$ 上的正交函数系。

可以证明，若$\{\omega_n(x)\}$是区间$[a,b]$上的一个正交多项式系，则多项式$\omega_n(x)$的n个零点就可作为Gauss型n点求积公式中的Gauss型基点。这一事实的证明作为习题留给读者。据此事实，我们可以按下列步骤得到Gauss型求积公式：首先寻找$[a,b]$上的正交多项式系$\{\omega_n(x)\}$，然后取$\omega_n(x)$的n个零点作为Newton-Cotes n点求积公式中的n个插值点，这样得到的求积公式就是Gauss型n点求积公式。类似地可以导出计算形如$\int_a^b p(x)f(x)\mathrm{d}x$的Gauss型求积公式：

$$\int_a^b p(x)f(x)\mathrm{d}x=\sum_{j=1}^{n}A_jf(x_j)+E\doteq\sum_{i=1}^{n}A_if(x_i),\tag{1.11}$$

其中多项式系$\{\omega_k(x)\,|\,k=0,1,\cdots\}$在$[a,b]$上按下面的意义正交：

$$\int_a^b p(x)\omega_i(x)\omega_j(x)\mathrm{d}x=0\qquad(i\neq j).$$

Gauss型基点$x_0,x_1,\cdots,x_n$是$\omega_{n+1}(x)$的$n+1$个零点。又设$\omega_k(x)$中的最高次项x^k的系数为c_k，并且

$$\delta_k=\int_a^b p(x)\omega_k^2(x)\mathrm{d}x,$$

则积分公式(1.11)中的A_i和E可表为

$$A_i=\frac{-c_{n+1}\delta_n}{c_n\omega_{n+1}(x_i)\omega'_n(x_i)}\qquad(i=0,1,\cdots,n),$$

$$E=\frac{\delta_n}{c_n^2(2n)!}f^{(2n)}(\xi)\qquad(\xi\in(a,b)).$$

下面给出几个常用的Gauss型求积公式，这些公式仅供读者在实际应用时查阅。

1）Legendre-Gauss求积公式

区间为$[-1,1]$，权函数$P(x)\equiv1$，$\omega_i(x)$选为i阶Legendre多项式。据此得到的求积公式中的A_i及E为

$$A_i=\frac{-2}{(n+1)\omega_{n+1}(x_i)\omega'_n(x_i)},$$