概率统计计算

中国科学院计算中心概率统计组 编著

科学出版社

概率统计计算

中国科学院计算中心概率统计组 编著

科学出版社
1979

内 容 简 介

本书包括基础统计计算、多元统计分析、随机过程处理和概率统计模拟等四部分，共十章.

本书从概率统计模型建立入手，介绍必要的基本概念，着重讲计算方法并指出计算过程中常见问题的处理方法，最后给出算法语言 BCY 程序或框图以及数字例子.

概率统计计算

中国科学院计算中心概率统计组 编著

*

科学出版社出版

北京朝阳门内大街137号

朝阳六六七厂印刷

新华书店北京发行所发行　各地新华书店经售

*

1979年4月第一版　开本：787×1092 1/32

1979年4月第一次印刷　印张：15

印数：0001—47,300　字数：344,000

统一书号：13031·942

本社书号：1332·13—1

定价：1.55元

序　言

随着我国阶级斗争、生产斗争和科学实验三大革命运动的深入发展，在许多领域，如自动控制、工程设计、天文、气象、水文、地质、地震、农林、医药以及化工等等，越来越多地应用概率统计的方法解决各自的问题．与此相应，在电子计算机上进行概率统计计算的要求也与日俱增．本书正是根据这种需要编写的．

全书分为四部分，共十章．内容包括基础统计计算、多元统计分析、随机过程处理和概率统计模拟等数字计算机上一些常用的概率统计方法和算法．为阅读方便，各章之间将保持相对的独立性．本书着重讨论概率统计中的一些计算方法，此外也介绍一些有关的基本概念．各章大都附有计算程序或框图，并附有数字例子．书中的全部计算程序均采用 BCY 语言编制．对该语言不熟悉的读者，可参考书末的附录二．为方便读者，概率统计计算中常用的几个矩阵代数方面的计算程序，作为附录一给出．

本书可供各领域中使用概率统计方法的同志和大专院校有关专业的师生参考．

本书是我组多年工作的小结，是集体工作的成果．各章的执笔人是：张建中（第一、七、九、十章）、魏公毅（第二章）、肖文金（第三章）、杨自强（第四、五、六章和附录二）、王振华（第八章）、李风林（附录一）．此外，我组徐锺济、邹兴德、李维相同志也参加了有关工作．

在本书的编写过程中，曾得到所内外很多同志的支持和

鼓励，提供了很多宝贵的意见，谨此致谢.

由于我们的水平所限，书中一定会有谬误和欠妥之处，衷心希望读者批评和指正.

中国科学院计算中心概率统计组

1977 年 8 月于北京

目　　录

第一部分　基础统计计算

第二部分　多元统计分析

第四部分 概率统计模拟

第一部分　基础统计计算

第一章　实验数据的整理和检验

§1. 概　　论

在工农业生产、社会生活和科学研究过程中，经常会遇到各种类型的不同的实验数据．这些实验数据为我们认识事物的内在规律、研究事物之间的关系、预测事物的今后可能发展等一系列问题，提供了丰富的材料和科学的依据，是十分宝贵的．但是，要想从这样一些庞大的数据堆中找到有用的东西，得到可靠的结论，就必须很好地下一翻工夫，对实验数据进行认真的整理和必要的检验．概率统计计算中的一些计算方法和处理技巧，可以帮助我们对实验数据进行去粗取精、去伪存真的分析整理和统计检验，从而便于揭露问题中存在的矛盾，找到解决问题的线索或可能途径．

在这一章中，首先分析常用实验数据的一些基本特点，介绍今后计算处理中常用的一些概率统计概念；然后利用一些简单的概率统计算法，计算实验数据的基本统计特征，即通过对实验数据数字特征和分布特征的计算和统计检验，推断实验数据的理论统计参数的数值和统计分布的规律，为进一步深入进行实验数据的统计分析提供较为合理的数学模型；最后，利用估计得到的一些参数，指出实验数据中可能存在的异常点，为实验数据校正提供一定的统计依据．

§2. 实 验 数 据

要对实验数据进行有效的整理和检验，首先需要对实验

表 1.1 实验数据一例

n	1	2	3	4	5	6	7	8	9	10
0	9.94	10.05	10.22	10.00	9.89	9.87	9.95	10.05	9.94	10.01
10	10.13	10.05	10.16	9.96	10.07	9.89	10.09	9.98	9.97	9.83
20	10.06	10.07	9.91	10.07	9.91	9.98	9.97	10.03	9.98	9.99
30	10.02	10.02	10.07	10.05	10.04	10.06	9.91	10.00	9.98	10.02
40	10.20	10.10	10.20	10.03	10.14	9.99	10.05	10.03	9.94	10.09
50	9.97	10.14	9.97	10.02	10.10	9.89	10.13	10.02	10.18	9.90
60	10.10	10.09	10.00	10.07	10.03	9.95	9.91	10.01	10.00	9.92
70	10.01	10.11	9.97	9.88	9.91	9.77	9.98	9.95	10.11	9.98
80	9.99	10.07	10.00	10.10	9.93	10.20	10.06	9.72	10.04	10.03
90	9.88	9.90	10.05	10.01	9.93	9.70	9.99	10.00	9.88	9.98
100	10.02	10.15	9.97	10.17	10.11	10.01	9.91	10.03	10.24	9.90
110	10.00	9.98	10.00	9.78	10.00	10.17	9.89	9.86	10.04	10.00
120	9.90	10.15	10.15	9.89	9.89	10.08	9.87	9.97	10.04	9.86
130	10.00	10.02	10.21	9.97	9.96	9.95	9.86	9.92	9.85	9.84
140	9.95	9.79	10.18	9.79	9.92	10.07	9.93	10.18	10.10	9.83
150	10.09	9.96	10.14	9.94	10.04	10.02	9.86	10.07	9.97	10.11
160	9.98	9.96	10.02	9.89	10.12	10.08	10.07	9.89	9.98	10.07
170	10.07	10.17	10.07	9.86	9.98	10.06	9.99	9.89	10.03	10.19
180	9.91	9.87	9.96	9.88	10.03	10.05	10.00	10.14	10.12	10.11
190	10.17	9.99	9.96	10.02	10.20	9.83	10.11	9.95	9.78	9.96

续表 1.1

n	1	2	3	4	5	6	7	8	9	10
200	10.02	10.01	10.04	9.96	10.16	10.06	9.97	9.94	10.06	9.98
210	9.96	9.95	9.97	9.84	9.90	10.08	10.11	10.11	10.06	10.09
220	9.92	10.16	9.93	9.85	9.92	9.94	9.87	9.80	9.92	10.01
230	9.94	9.89	10.03	10.00	10.10	9.87	9.76	10.05	10.09	10.17
240	10.09	10.10	10.07	10.11	9.82	9.88	9.96	10.03	10.03	9.90
250	10.20	10.03	9.80	9.97	10.00	10.04	9.91	10.11	9.90	9.84
260	9.90	10.01	9.99	10.04	10.06	10.07	10.09	10.03	10.01	9.97
270	10.00	10.20	10.02	9.81	9.90	10.02	10.02	9.94	9.99	9.87
280	10.05	10.05	9.94	10.11	9.82	9.99	10.03	10.11	9.87	9.93
290	9.93	10.01	9.91	9.91	10.09	10.16	10.05	10.18	9.98	9.91
300	10.08	9.88	9.84	10.19	10.11	9.93	10.01	10.04	10.02	9.97
310	10.04	10.06	10.10	10.02	9.85	9.88	10.03	9.97	10.07	10.11
320	10.17	10.16	10.01	9.95	9.79	9.96	9.98	9.90	9.95	10.09
330	10.10	10.05	10.01	9.92	10.06	9.99	10.10	10.14	10.05	10.10
340	9.97	10.03	9.96	9.84	9.85	9.85	10.07	10.18	9.97	9.96
350	10.06	9.99	9.94	9.93	10.19	10.22	10.10	9.91	9.92	10.11
360	9.93	10.13	9.91	9.91	10.15	10.15	9.91	9.95	9.75	9.98
370	9.96	10.06	9.92	10.05	10.10	9.96	9.94	9.99	9.96	10.07
380	10.09	10.11	9.96	10.00	9.95	9.81	9.82	10.06	9.97	9.90
390	10.05	10.04	9.92	9.87	10.06	10.02	10.05	9.91	10.15	10.11

数据的基本特点有所了解和分析. 下面，用一个十分简单的例子，说明常见实验数据的一些基本特点.

假若对系统中的一个常量 l（如高度、重量等）进行测量. 每测量一次，得到 l 的一个测量值 x. 重复进行 400 次测量，得到常量 l 的 400 个实验数据，依次列于表 1.1 中.

对表 1.1 中的实验数据进行一些简单的分析以后，可以看出：

1. 在形式上，实验数据

$$x_1,\ x_2,\cdots,x_n,\cdots,x_N \tag{2.1}$$

以有限次数 N（在表 1.1 中，$N=400$）由离散化的实验结果给出.

2. 实验数据 x_n（$n=1,2,\cdots,N$）尽管是对系统中的同一个常量 l 进行多次测量得到的，实验结果的数值也不完全相同，也就是说，在实验数据中总存在实验误差. 这时，实验误差

$$\delta_n = x_n - l \qquad n=1,2,\cdots,N$$

按其性质大致分为三类[2,6].

（1）随机误差. 在实验数据中，随机误差是由一系列偶然因素引起的一类不易控制的测量误差. 在多次反复实验过程中，随机误差取值可大可小，可正可负，具有统计规律性，服从一定的概率分布. 随机误差在实验过程中是难免的，随着实验观测次数的增加，随机误差的算术平均值将愈来愈小，逐渐接近于零.

（2）系统误差. 把实验观测过程中服从确定性规律的误差称为系统误差. 在多数情况下，系统误差是一个常量. 在实验观测过程中或实验数据分析整理过程中，可以通过一定的方法，识别、消除这类实验误差. 实验数据中的系统误差和随机误差不同，不可能通过实验次数增加的算术平均进行消除.

(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点. 过失误差一般是由于实验观测系统测错、传错或记错等不正常的原因造成的. 在实验数据整理过程中，必须消除这类过失误差，舍弃实验数据中的异常点，否则会严重影响计算结果的准确度，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.

3. 在对实验数据进行了一定的分析整理和统计检验之后，可以看到，实验数据大都具有一定的统计规律性. 找出实验数据的统计规律，估计实验数据的基本统计参数，是概率统计计算要解决的基本问题之一. 以表1.1中的实验数据为例，它们围绕在10的上下对称地随机分布着，而且距10愈远点数愈少.

为今后讨论方便起见，下面引进概率统计计算中常用的几个基本概念[1,3,5,6].

1. 总体和子样. 我们把概率统计计算研究的全部元素组成的集合称为总体(总体也常称为母体)，组成总体的最小研究单位称为个体. 在上面的例子里，常量 l 的全部可能测量值组成我们研究的总体，而每一次的实验观测数据 x_n 就是一个个体.

通过实验观测，可以得到总体的部分结果，通常称为子样. 一个子样中包含个体的总数称为子样的容量，简称为样本量. 表1.1中的实验数据 x_n 组成一个样本量 $N=400$ 的子样.

2. 概率分布和经验分布. 在实际问题中，经常遇到的事件可以分为两种，即确定性事件和随机性事件. 在一组给定条件下，一定发生或一定不发生的事件分别称为必然事件和

不可能事件，是确定性的；在一组给定条件下，可能发生，也可能不发生的事件称为随机事件．为了研究随机事件的数量规律性，引入特征量 η，它在随机事件发生和不发生时取不同的数值，称为随机变量．

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

$$P\{\eta < x\} = F(x)$$

称为随机事件 $\{\eta < x\}$ 发生的概率．显然，$F(x)$ 是实数 x 的一个函数，称为随机变量 η 的概率分布函数，简称为分布函数．

设实验数据 (2.1) 相互独立地来自总体 η．把实验数据 (2.1) 按取值大小由小到大顺序排列，得到随机变量 η 的值序统计量如下：

$$x_1^{(N)} < x_2^{(N)} < \cdots < x_n^{(N)} < \cdots < x_N^{(N)} \qquad (2.2)$$

由值序统计量 $x_n^{(N)}$ 可以给出分布函数 $F(x)$ 的渐近统计估计

$$S_N(x) = \begin{cases} 0 & \text{当 } x \leqslant x_1^{(N)} \text{ 时} \\ \dfrac{n}{N} & \text{当 } x_n^{(N)} < x \leqslant x_{n+1}^{(N)} \text{ 时} \\ 1 & \text{当 } x > x_N^{(N)} \text{ 时} \end{cases} \qquad (2.3)$$

称为随机变量 η 的经验分布函数．经验分布函数 $S_N(x)$ 和分布函数 $F(x)$ 有着类似的性质．

3. 数字特征和统计量．在实际问题中，常常需要用几个有代表性的数字描述随机变量 η 的基本统计特征，一般把它称为随机变量 η 的数字特征；用以估计总体数字特征的函数 $g(x_1, x_2, \cdots, x_N)$ 由子样 (2.1) 给定，称为统计量．在 §3 中给出的算术均值 $\bar{x}$、方差 s^2 等都是统计量，它们分别给出总体数学期望 $E(\eta) = \mu$、方差 $E[\eta - E(\eta)]^2 = \sigma^2$ 等数字特征的估计值．

§3. 数字特征计算及其统计检验

实验数据

$$x_1, x_2, \cdots, x_n, \cdots, x_N \tag{3.1}$$

的数字特征计算，就是从(3.1)中计算一些有代表性的特征量，用以浓缩、简化实验数据(3.1)中的信息，使问题变得更加清晰、简单，易于理解和处理．这里给出的数字特征参数分别用来描述实验数据(3.1)取值的大致位置、离散程度、分布特征和相关特征等[3,4,5,9]．

1. 位置特征参数计算．实验数据(3.1)的位置特征参数是描述实验数据的平均位置和特定位置的，其中常用的有均值、分位数、极小值和极大值等．

均值是描述实验数据取值平均位置的，其中又有算术均值、几何均值和调和均值之分．在实际问题中，算术均值是最常用的．

算术均值，一般用 $\bar{x}$ 表示，是实验数据的代数和除以样本量 N，即

$$\bar{x} = \frac{1}{N}(x_1 + x_2 + \cdots + x_N) = \frac{1}{N}\sum_{n=1}^{N} x_n = \frac{1}{N}\sum_{n} x_n \tag{3.2}$$

这个值用来表示总体的平均水平．为简单计，在不致发生混淆时，今后记

$$\sum_{n=1}^{N} f(x_n) = \sum_{n} f(x_n)$$

其中 $f(x_n)$ 表示实验数据 x_n 的函数．

在数字计算机上，计算算术均值 $\bar{x}$ 的常用算法有[7]：

(1) 直接算法．直接利用算术均值定义的算法(3.2)，通过实验数据(3.1)求和平均得到 $\bar{x}$，简记为

$$\bar{x}_{(1)} = \frac{1}{N}\sum_n x_n \tag{3.3}$$

(2) 递推算法. 令 $\bar{x}_0 = 0$, 对 $n = 1, 2, \cdots, N$, 计算中间均值

$$\bar{x}_n = \frac{n-1}{n}\bar{x}_{n-1} + \frac{1}{n}x_n = \bar{x}_{n-1} + \frac{1}{n}(x_n - \bar{x}_{n-1}) \tag{3.4}$$

最后得到算术均值

$$\bar{x}_{(2)} = \bar{x}_N$$

(3) 二次均值算法

$$\bar{x}_{(3)} = \bar{x}_{(1)} + \frac{1}{N}\sum_n (x_n - \bar{x}_{(1)}) \tag{3.5}$$

比较上面三种不同的算法,不难看出,直接算法(3.3)的运算量最省;递推算法(3.4)可以进行实时处理,得到一系列的中间均值 $\bar{x}_n$; 二次均值算法考虑到数字计算机处理大量实验数据舍入误差的影响,由(3.5)计算中的第二项 $\sum_n (x_n - \bar{x}_{(1)})$ 集中了直接均值计算过程中的主要舍入误差, 从而提高了均值结果的计算精度.

显然, 二次均值算法(3.5)只在处理大量实验数据(3.1)时才有意义,但这种算法要两次处理原始数据,多要动用慢速存储,不太方便. 实际应用时,可以像简化均值计算的古典算法那样, 用部分实验数据(如只用超快速存储中的实验数据)计算一次均值 $\bar{x}_{(1)}$ 的渐近值或选用一个适当的接近于 $\bar{x}_{(1)}$ 的常量 c 作为参数,用

$$\bar{x}^*_{(3)} = c + \frac{1}{N}\sum_n (x_n - c) \tag{3.6}$$

代替(3.5), 计算均值 $\bar{x}$.

描述实验数据位置特征的参数,除常用的算术均值 $\bar{x}$ 外,还有极小值、极大值、众数、中位数和分位数等. 由于在数字