数据分析中的统计和计算方法

[西德]西格蒙德·布兰特 著
莫梧生 译　王宏禹 校

国防工业出版社

数据分析中的统计和计算方法

〔西德〕 西格蒙德·布兰特 著
莫梧生 译
王宏禹 校

国防工业出版社

内 容 简 介

概率统计方法已广泛应用于自动控制、天文学、气象学、水文学、地质学、地震学、物理学、电子学、经济学，以及农林、医药、化工等各个领域的数据分析中。并越来越多地应用电子计算机进行统计计算，解决各自的问题。

本书简短地而数学上又足够严格地介绍了概率统计学中的各种基本概念，并着重讲述计算方法和应用。结合计算方法和应用，给出了不少计算程序和数值例子。

本书可作为大学生和研究生的教科书，也可供数学家、物理学家、经济学家以及各工程领域的专业人员参考。

STATISTICAL AND COMPUTATIONAL
MMETHODS IN DATA ANaLYSIS
SIEGMUND BRANDT
NORTH-HOLLAND PUBLISHING COMPANY
2nd, revised edition 1976

*

数据分析中的统计和计算方法

〔西德〕 西格蒙德·布兰特 著

莫 梧 生 译

王 宏 禹 校

*

国防工业出版社 出版

新华书店北京发行所发行　各地新华书店经售

国防工业出版社印刷厂印装

*

787×1092 1/16 印张18 412千字

1983年10月第一版 1983年10月第一次印刷 印数：00,001—12,500册

统一书号：15034·2549 定价：2.25元

第一版序言

本书是根据作者 1967 至1968 年在海得堡（Heidelberg）大学给物理系学生及核子物理学者所作的讲演而编写成的。它阐述了那些与数据分析最有关系的数理统计的部份。本书可供科学、医学、工程和经济学中面临实验数据估算问题的学生和研究工作者参考。

本书是依照数理统计学使用者的观点编写的。并不过份强调数学上的严格性。另一方面，它不仅对不同实际应用的许多方法，而且还对统计方法的概念和原理作了说明。所介绍的部分内容是受物理学者所写或为他们而写的讲课笔记及评论文章的影响〔这些物理学者有伯克（Böck，1960）；奥雷尔（Orear，1958）；佐尔米茨（Solmitz，1964)〕。

这里假定读者已掌握了良好的微积分基础知识。其他必需的数学工具，特别是概率论，本书作了简短的复习。在介绍中，一个重要而基本的因素是使用矩阵表示法。它对许多问题，诸如最小二乘法，提供了很紧凑的说明。在附录中编写了矩阵计算的导论。

由于目前多数复杂的数据分析问题都靠计算机来处理，所以对这样一些场合，介绍了FORTRAN 语言所写的程序。在附录中阐述了 FORTRAN 程序设计语言的基本特征。它还包括了处理矩阵所需的子程序的一个短小程序库，这些子程序是以由日内瓦欧洲核研究中心的伯克（R. Böck）最初所编写的一组类似的程序为基础的。

作者希望本书不仅成为统计数据分析的导论，而且也能在日常的工作中得到使用。所以本书包括一些统计表以及较重要的公式的汇编，便于参考。

作者向海得堡大学的几位同事对本书所进行的研讨表示感谢，特别是谢（T. P. Shah）博士，他校阅了本书的手稿，并对本书的改进提出了许多有用的建议。作者也向（阿姆斯特丹市）坦纳（A. G. C. Tenner）博士表示感谢，他和作者对本书的组编进行了富有成果的讨论。还要感谢（海得堡市）伊米希（H. Immich）博士，他给作者提供了实例 8-2，11-1 及 11-2。感谢(魏茨拉市)弗伦克（H. Frenk）博士替作者获得了封里所印的木刻画。

S. 布兰特

（海得堡，1970年 1 月）

再版序言

在这次编写中对本书的一般概念——简短而充分的严格数学处理，并且主要强调应用——保持不变。但在直接应用的统计方法方面，特别是计算机程序，增加了很多新的内容。主要的新章节为

——蒙特卡罗法的基础（第五章第 5.3 节）；

——采样数据的近似数值分析和图解分析（第六章第 8 节）；

——线性回归用的 FORTRAN 程序（第十二章第 5 节）；

——时间序列分析（第十三章）。

它们都含有 FORTRAN 程序（直接应用于统计问题的程序的数目增加了三倍）。

此外，几种分布的褶积问题，在实际应用中是颇感麻烦的。本书作了更详细的讨论，并给出了正态分布的褶积的几个例子。采样一章现在增编了关于很小样本的一节。

练习题现在编在每章的结尾，并且专门有一节列出了它们的解。

作者向西根大学的同事对于所作有益的研讨表示感谢，特别是海因里希(W. Heinrich)博士，他审阅了各新章节的手稿。在此作者还要向沃伊西克（W. Wojcik）博士和约希基（H. Yoshiki）教授表示感谢，他们卓越地将本书译成波兰文和日文版。

S. 布兰特

（西根，1976 年 6 月）

常用符号表

$x, y, \xi, \eta, \cdots$	（普通）变量	S	统计量，估计量
$\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\xi}, \boldsymbol{\eta}, \cdots$	变量的向量	S_c	临界区
$\mathrm{x}, \mathrm{y}, \xi, \eta, \cdots$	随机变量	t	“学生”（Student's）分布的变量
$\mathbf{x}, \mathbf{y}, \boldsymbol{\xi}, \boldsymbol{\eta}, \cdots$	随机变量的向量	x_m	众数
$A, B, C, \cdots$	矩阵	$x_{0.5}$	中位数
B	偏置	x_q	分位数
$\mathrm{cov}(\mathrm{x}, \mathrm{y})$	协方差	$\bar{\mathrm{x}}$	样本均值
F	方差商	$\tilde{x}$	最大似然或最小二乘估计量
$f(x)$	概率密度	λ	一分布的参量
$F(x)$	分布函数	$\varphi(t)$	特征函数
$E(\mathrm{x})=\hat{x}$	均值，数学期望值	$\phi(x), \psi(x)$	正态分布的概率密度和分布函数
H	假设	$\phi_0(x), \psi_0(x)$	正态化高斯分布的概率密度和分布函数
H_0	原假设	$\sigma(\mathrm{x})=\Delta(\mathrm{x})$	标准偏差
L, l	似然函数	$\sigma^2(\mathrm{x})$	方差
$L(S_c, \lambda)$	运算特征	$\Omega(P)$	正态分布的逆分布
$M(S_c, \lambda)$	功效函数		
M	要极小化的函数		
$P(A)$	事件 A 的概率		
$\mathrm{s}^2, \mathrm{s}_{\mathrm{x}}^2$	样本方差		

目 录

例子目录表

FORTRAN 程序目录表

第一章　引　论

实验科学的每个分支，在通过早期的定性说明之后，本身就涉及对感兴趣的现象作定量的研究，即对观测量的研究。在实验的设计和性能测试后，下一步的一个重要任务就是精确计算和全面探讨所得到的数据。让我们列举几个典型的问题加以说明。

1．研究试验动物在各种药物影响下体重的增加。当对25只动物应用药物A以后，观测到体重平均增加5％。应用药物B于10只动物后，增3％。药物A是否更有效？平均值5％和3％实际上不能回答这个问题，这是因为较低的数值是由一只动物由于某种理由体重减轻所致。所以必须研究各个体重的分布以及它们以平均值为中心的散布。并且必须决定，是否所用试验动物的数目将能使人以一定的精度去区别两种药物的效果。

2．在晶体生长的实验中，确切维持不同分量的比例是必不可少的。从全部500个晶体，选择20个进行分析。对于其余480个晶体的成分能下什么样的结论？这个采样问题在生产控制、自动测量器件的可靠性检验以及民意测验中都会出现。

3．一定的实验结果得到之后，这就必须决定它是否与某些预测的理论值或以前的实验相矛盾。实验被用来做假设检验。

4．描述所测变量相互依赖关系的一般定律已知，但这定律的参量须从实验获得。例如，在放射性衰变中，每秒衰变的原子数目N随着时间按指数下降：$N(t)=$常量$\times\exp(-\lambda t)$。衰变常数λ及其测量误差都要利用许多观测量$N_1(t_1)$，$N_2(t_2)$，…来确定。这种参量估计的问题多半对许多实验工作者是很感兴趣的。

从这些例子看，数据分析的某些特征已变得很明显。特别是我们可以看到，一个实验的结果并不是唯一地由实验程序所确定，而是也受机会的影响：它是一个随机变量。这种随机的趋势或者根源于实验的性质（试验动物必须不同，放射性是一种随机现象），或者这是实验设备不可避免的不定性引起的结果，即观测误差。所以下一章要专门复习概率论的最重要的概念。

在第三章和第四章中，介绍随机变量。阐述随机变量的分布，并求诸如均值和方差等参量以表征这些分布。特别注意到几个随机变量的独立无关的性质。在第五章中，研究许多分布，这些分布都是应用中很感兴趣的。特别是正态分布或高斯分布的性质，本书详细地进行了讨论。

在实际中，一种分布必须由有限个观测量即一个样本来确定。在第六章中，考虑了采样的不同场合。并介绍FORTRAN程序，对经验数据作初步的数值处理和图形显示。样本函数，即含有各个观测量的函数，可用来估计分布的特性参量。推导了一个良好的估计所应满足的要求。在这个阶段，介绍数量χ^2。它是观测值与期望值之间偏差的平方和，所以是观测质量的一个合适的指标。

在第七章中所说的最大似然法，是现代统计分析的核心。它使人们去构造具有最佳性质的估计量。这种方法可用于探讨单参量和多参量场合，并列举了许多例子。

第八章专讲假设检验。它含有最常用的F检验、t检验以及χ^2检验，并概述了一般理论。

最小二乘法，它也许是最广泛使用的统计分析方法，是第九章的主题。在这一章中，详细讨论一般场合之前，对日常应用中所遇到的直接观测、间接观测和约束观测的特殊场合，进行了详细的阐述。提供了处理一般最小二乘问题的一个 FORTRAN 程序，并列举不同的例子说明其使用方法。每一个最小二乘问题都可表达成确定几个变量的一个函数的极小值的求值问题。这对最大似然概念为基础的所有参量估计也是真确的。在第十章中，介绍几种计算方法以获得这种极小值。

第十一章的方差分析可被看作是F检验的一个扩充。它被广泛地使用于生物和医学的研究中，用来研究实验条件上一个观测到的变量表示成为其它变量的相依关系，而不是检验其独立性。当有几个变量时，会引起颇为复杂的情况。本书利用一个FORTRAN程序计算了某些简单的数值计算例。

第十二章的线性回归是最小二乘法的一个特殊情况，故在第九章中已有所涉及。在计算机获得进展之前，通常只能处理线性最小二乘问题。对这种情况，发展了一个特殊的术语，到现在仍在使用。故看来有理由要用专门的一章来阐述这个题目。同时，也扩充了第九章的处理方法。例如，研究了一个解的置信区间的确定，以及回归与方差分析之间的关系。本章给出了一个线性回归通用的FORTRAN程序，并举例说明其使用方法。

在最后一章中，介绍了时间序列分析要素。如果数据是作为受控变量（通常是时间）的一个函数给出的，且对数据作为受控变量函数的特性缺乏理论上的预测时，就使用这个方法。它用来设法减少数据的统计波动而不损坏与受控变量真正的相依关系。由于时间序列分析中的计算工作颇为困难和繁冗，故也给出了一个FORTRAN程序。

第二章 概　　率

2-1. 实验，事件，样本空间

在本书中，由于我们所讨论的是来自实验的数据的分析，所以我们必须首先说明一个实验及其结果是指什么意思。作为实验室的工作，我们定义一个实验就是严格遵照一个预定程序进行试验，作为试验的结果，将获得一个数量或一组数量。这些数量在性质上有的是连续型的（如温度，长度，电流）有的是离散型的（如粒子数，一个人的生日，三种可能的颜色中的一种）。不论保持所规定的所有条件如何精确，一般，重复一个实验所得结果总是不同的。引起不同结果的原因，或是所研究的现象具有内在的统计性质，或是测量的精度有限。因此，对每一数量的所有可能结果总是跨越一有限区域。构成一个实验结果所有数量的这些区域的总体就建立了该实验的样本空间。由于在一个特定的实验中，对于所测的数量，难于和常常不可能确切确定可达的区域，所以实际上所用的样本空间比较大，且含有适于作为子空间的样本空间。我们将使用样本空间的这个不严格的概念。

例2-1：由连续型变量所覆盖的样本空间

在电阻器的生产中，重要的是将R值（电阻值，以Ω计量）以及N值（最大热损耗值，以W计量）维持在已知的数值上。对R和N的样本空间可以是由R轴及N轴所覆盖的一个平面。由于两个数量总是正的，这个平面的第一象限再一次成为一个样本空间。

例2-2：离散型变量所覆盖的样本空间

实际上，只要R和N维持在以额定值为中心的某个一定的区间内（例如$99\text{k}\Omega < R < 101\text{k}\Omega$，$0.49\text{W} < N < 0.60\text{W}$），而$R$和$N$的确切数值并不重要。如果是这种情况，则我们说，电阻器具有性质R_n，N_n。如果数值低于下限或高于上限，那么我们用－（或＋）代替下标n。因此，电阻和热功耗的可能数值为R_-，R_n，R_+及N_-，N_n，N_+。样本空间现由9点组成

$$R_-N_-,\ R_-N_n,\ R_-N_+,$$
$$R_nN_-,\ R_nN_n,\ R_nN_+,$$
$$R_+N_-,\ R_+N_n,\ R_+N_+。$$

常常，样本空间的一个或更多特定的子空间特别令人感兴趣。例如在例2-2中，点R_nN_n代表了电阻器满足产品规格的场合。我们给这样一些子空间取名，例如A，B，…，且我们说，如果一个实验的结果落入一个这样的子空间，则事件A（或B，C，…）已经出现。如果A未出现，则我们说它是互补事件或对立事件$\overline{A}$（不是A）。整个样本空间对应于每个实验中所出现的一个事件，我们称它为E。在本章的以下各节中，我们将要定义一个事件出现的概率是什么意思。

2-2. 概率的概念

让我们考虑最简单的实验，即掷硬币的实验。像掷骰子或玩纸牌的某些问题那样，它

没有实际意义，但对教学法来说，它却是有价值的。一个“普通”的硬币掷一次时，呈现“头像”的概率是什么？我们可直觉地认为此概率等于$\frac{1}{2}$。这是以下列假定为基础的，即样本空间中所有的点（这里只有两点：“头”和“尾”）都有相等的可能性，且按习惯，我们给事件E（这里是“头”或“尾”）以概率1。这种确定概率的方法只能应用于完全对称的实验，故很少有实际的用途。（但在统计物理学及量子统计学中，它是十分重要的，在这些学科中，所有容许的状态有相等的概率是很成功的理论的一个重要的先决条件。）

如果不存在这种完全的对称性——甚至是正常的“实际”硬币情况——则下列步骤似乎是合理的。在大量N的实验中，事件A被观测到出现了n次，我们定义

$$P(A)=\lim_{N\to\infty}\frac{n}{N}。\tag{2.1}$$

为事件A出现的概率。这个稍为不严格的概率的频率定义对实际应用来说是足够的了，虽然在数学上它还不够令人满意。利用这个定义的困难之一是需要作无穷次的实验，这当然不可能去做，甚至难于想像。虽然在本书中我们将实际上使用频率定义，但我们却要指出科尔莫戈罗夫（Kolmogorov〔1933〕）所提出的概率的公理的基本概念。一般所使用的公理的极小集如下。

（a）对于每个与一非负数相对应的事件A，其概率

$$P(A)\geqslant 0。\tag{2.2}$$

（b）事件E具有单位概率

$$P(E)=1。\tag{2.3}$$

（c）如果A和B为互相排斥的事件，则A或B（$A+B$）的概率为

$$P(A+B)=P(A)+P(B)。\tag{2.4}$$

从这些公理*，立即可得下列有用的结果。

从(b)和(c)

$$P(A+\overline{A})=P(A)+P(\overline{A})=1,\tag{2.5}$$

进一步与（a）结合一起

$$0\leqslant P(A)\leqslant 1。\tag{2.6}$$

从(c)，容易得到对于互斥事件A，B，C，…更一般的理论

$$P(A+B+C+\cdots)=P(A)+P(B)+P(C)+\cdots\tag{2.7}$$

这应注意，到此为止所提到的事件的唯一组合途径是“异或”。任何其他的组合必须由它来构成。在掷一个骰子当中，A可表示偶数的点，B表示奇数的点，C表示少于4点，D表示4点或多于4点。令人感兴趣的是对事件A或C的概率，它们显然不是互斥的。A和C记为AC，同样还有AD，BC及BD，它们是互相排斥的，并对A或C(有时写成$A\dotplus C$)求表达式$AC+AD+BC$。

值得注意的是公理并不为获得一个特殊的概率$P(A)$规定一种方法。

最后应该指出，概率一词常常用于普通语言中，其意义与我们所考虑的不同，甚至相反。我们的意思是主观概率。这里，一事件的概率是由我们相信其出现的程度给出的。这种概

* 有时定义(3.1)作为第四公理来介绍。

率的一个例子是："在1984年前在英吉利海峡下建筑一个隧道的概率是1／3"。作为另一个例子可考虑在核乳胶中由一个质子或介子（pion）所留下的某种轨迹的场合。人们常说"由一个介子引起的轨迹的概率为$\frac{1}{2}$"。但因事件已经发生，只有两种粒子中的一种会引起那种特殊的轨迹，所以所述概率虽然未知但无非是1或是0。

2-3．概率计算法则；条件概率

一个实验的结果具有性质A。我们现在要问：它也具有性质B的概率是什么？即在条件A之下B的概率是什么？我们定义这条件概率为

$$P(B|A)=\frac{P(AB)}{P(A)}。\tag{3.1}$$

因此

$$P(AB)=P(A)P(B|A)。\tag{3.2}$$

从图 2-1 所示的例子，可以看出，这个定义是合理的。这里如果某一点分别处于区域A或B内，则事件A或B出现。如果该点出现在重叠区域内，则事件AB（A和B）出现。设不同区域的面积与相应的事件的概率成正比，则在条件A下B的概率是AB和A的面积之商。即如果A被包含在B内，则在这个特殊情况下，商等于1；如果重叠面积等于零，则商等于零。

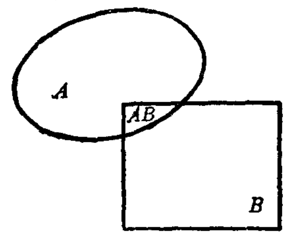

图2-1　条件概率的具体说明

利用条件概率，我们现在可以列出全概率法则的公式。一个实验能导致n个可能的互相排斥事件

$$E=A_1+A_2+\cdots+A_n\tag{3.3}$$

中的一个事件。具有性质B的任何事件的出现概率为

$$P(B)=\sum_{i=1}^{n}P(A_i)P(B|A_i),\tag{3.4}$$

这可从式（3.2）及（2.7）容易看出。

现在我们也能定义事件的独立性。如果两事件A及B，A出现并不改变B出现的概率，反之亦然；亦即如果

$$P(B|A)=P(B),\tag{3.5}$$

或，与式(3.2）一起，

$$P(AB)=P(A)P(B)。\tag{3.6}$$

则事件A及B可说是独立的。

一般，式（3.3）那种型式的几个分解式

$$\begin{aligned}E&=A_1+A_2+\cdots+A_n,\\E&=B_1+B_2+\cdots+B_m,\\&\vdots\\E&=Z_1+Z_2+\cdots+Z_l,\end{aligned}\tag{3.7}$$

如果对所有可能的组合α，β，…，ω，下列条件

$$P(A_\alpha B_\beta \cdots Z_\omega)=P(A_\alpha)P(B_\beta)\cdots P(Z_\omega) \tag{3.8}$$

得到满足，则所有事件是互相独立的。

练 习 题

2-1：从对称性确定概率

在一个年级中有n个学生。至少有两个学生的生日相同的概率是什么？回答下列提问以解上述问题。

（a）在一年中（365天），分布n个生日的可能性的数目N是什么？

（b）所有n个生日都不相同的可能性的数目N'是什么？

（c）所有生日都不相同的概率P_{diff}是什么？

（d）最后，至少有两个生日不同的概率是什么？

2-2：非互斥事件的概率

对非互斥事件A及B，已知概率$P(A)$，$P(B)$以及$P(AB)\neq 0$。观测A或B的概率$P(A+B)$是什么？作为例子，计算任意从一副52张卡片的纸牌抽一张卡片是一个A或一个方块的概率。

2-3：相依事件或独立事件

从一副纸牌抽出的一张卡片是一个A或一个方块的事件A及B是否是独立的。

（a）如果使用通常的52张卡片的纸牌？

（b）如果纸牌中加一张“百搭”？

〔利用方程式(3.6)〕

2-4：互补事件（或对立事件）

证明：如果A及B是独立的，则$\bar{A}$及$\bar{B}$是独立的。利用练习题2-2的结果，计算$P(\bar{A}\bar{B})$，并以$P(A)$，$P(B)$及$P(AB)$表示。

2-5：从大总体和小总体抽取的概率

有一只匣子，含有很大数目（>1000）的硬币。它们被分成三种A，B及C，并成为总数的20%，30%及50%。

（a）如果某个人随机地取一个硬币，则取A种，B种及C种的一个硬币的概率$P(A)$，$P(B)$，$P(C)$是什么？取两个硬币时，概率$P(AB)$，$P(AC)$，$P(BC)$，$P(AA)$，$P(BB)$，$P(CC)$，P（2个相等的硬币），P（2个不同的硬币）是什么？

（b）如果在匣子中有10个硬币（A种2个，B种3个，C种5个），则概率是什么？