高等院校信息科学系列教材

数据分析

范金城 梅长林 编著

科学出版社

高等院校信息科学系列教材

数据分析

范金城　梅长林　编著

科学出版社

2002

内 容 简 介

本书介绍了数据分析的基本内容与方法，其特点是既重视数据分析的基本理论与方法的介绍，又强调应用计算机软件 SAS 进行实际分析和计算能力的培养. 主要内容有：数据描述性分析、非参数方法、回归分析、主成分分析、判别分析、聚类分析、时间序列分析、Bayes 统计分析以及常用数据分析方法的 SAS 过程简介. 本书每章末附有大量实用、丰富的习题，并要求学生独立上机完成.

本书可作为高等院校信息科学及数理统计专业的本科生教材，也可供有关专业的研究生及工程技术人员参考.

图书在版编目(CIP)数据

数据分析/范金城，梅长林编著. —北京：科学出版社，2002
(高等院校信息科学系列教材)
ISBN 7-03-010458-7

Ⅰ. 数… Ⅱ. ①范… ②梅… Ⅲ. 统计分析(数学) Ⅳ. 0212.1

中国版本图书馆 CIP 数据核字(2002)第 035339 号

科学出版社 出版
北京东黄城根北街16号
邮政编码：100717
http://www.sciencep.com
新蕾印刷厂 印刷
科学出版社发行 各地新华书店经销
*
2002 年 7 月第 一 版 开本：720×1000
2002 年 7 月第一次印刷 印张：26 1/2
印数：1—5 000 字数：520 000

定价：32.00 元

(如有印装质量问题，我社负责调换〈路通〉)

序　言

1998年教育部进行高校专业调整时，设立了“信息与计算科学”专业.该专业的设立，受到很多高等院校的热烈响应.据不完全统计，几年来已有约280所院校招收了该专业的本科生，其中大部分院校计划开设信息科学方面的系列课程.

为了配合高等院校在学科专业设置上的改革与深化，来自几十所高等院校有关专业的部分领导和教师，于1999年、2000年召开了第一、二届“信息科学专业发展与学术研讨会”，与会者热烈讨论并探讨了许多关于信息学科的学科发展和建设的基本问题.会议一致认为教材建设是目前最为紧迫的任务，因此成立了教材编审协调组来组织该系列教材的编写.

2001年教材编写协调组召集了有多位经验丰富的教师和出版社参加的教材建设会议.会议明确了教材建设是一项长期的工作，并决定首先编写和出版本套教材来满足近期急需.为了保证教材的质量，会议对每本教材的要求、内容和大纲进行了具体研讨，并请具有多年教学经验的重点院校教授担任各教材的负责人.

为了贴近教学的实际，每本教材都配有习题或思考题，同时对内容也作了结构化安排，以便教师能根据实际情况部分选讲.本套教学用书不仅适用于教学，也可供相关读者参考.

在教材编写和出版过程中，作者对内容的取舍、章节的安排、结构的设计以及表达方式等方面多方听取意见，并进行了反复修改.在感谢作者们辛勤劳作的同时，编委会还特别感谢科学出版社的鞠丽娜编辑，她不辞辛劳，在统筹印刷出版、督促进度、征求意见、组织审校等方面做了大量工作.这套教材能在保证质量的前提下及时与读者见面，是和她的努力分不开的.

从长远的教学角度考虑，为了适应不同类型院校、不同要求的课程需要，教材编审协调组将不断组织教材的修订、编写(译)，从而使信息科学教学用书做到逐步充实、完善、提高和多样化.在此衷心希望采用该系列用书的教师、学生和读者对书中存在的问题及时提出修改意见和建议.

高等院校信息科学系列教材编委会

2002年3月

前　言

本书是高等院校信息科学专业本科生教材，也适用于数理统计专业的本科生. 本书包括数据分析的主要内容：数据描述性分析、非参数方法、回归分析、主成分分析、判别分析、聚类分析、时间序列分析、Bayes 统计分析和常用数据分析方法的 SAS 过程简介. 本书的特点是既重视数据分析的基本理论与方法的介绍，又重视应用 SAS 软件进行实际的分析计算.

数据分析是信息科学专业本科生重要的必修课，因此本书重视数据分析的基本理论与方法的介绍，详细叙述基本内容及算法. 本书的内容经过精选，对选入的内容详细予以介绍，并力求反映新颖内容. 其中数据描述性分析力求体现“让数据自身说话”；非参数方法模型具有一般性. 多元数据分析是数据分析极为重要的方面，书中主要介绍回归分析、主成分分析、判别分析、聚类分析. 因时间序列分析在自然、技术、经济等领域有极广泛的应用，因此介绍其基本内容. Bayes 统计分析作为数据分析的重要方法，也给以简单介绍.

本教材计划学时是 72 学时. 内容选择是模块式的，各校可以根据具体情况予以选择. 其中，数据描述性分析是基本的，必选的. 非参数方法可全学或选学一部分. 关于多元数据分析的几章中，回归分析是基本的；主成分分析、判别分析、聚类分析可全学或选学一部分；时间序列分析、Bayes 统计分析是两个单独的模块，可全学或选学其中的一个模块.

本书与 SAS 软件系统紧密结合. 书中大多数例题都由 SAS 软件分析计算. SAS 系统是大型集成应用软件系统，在数据分析与统计分析领域被誉为国际标准软件系统，并被广泛应用于各个领域. 通过对典型例题采用各种不同方法进行分析计算，以培养学生分析、解决实际问题的能力. 练习中有相当大的部分要通过 SAS 软件或其软件计算完成. 本书的第 9 章“常用数据分析方法的 SAS 过程简介”较系统地介绍了与本书内容有关的 SAS 过程，并结合本书部分例题进行编程. 若用 SAS 软件进行计算，应讲授该章，并让学生掌握计算技能. 对于采用其他软件进行计算的情况，当然要根据具体情况介绍其他软件. 因为本书主要介绍数据分析方法，即使使用其他软件，仍可使用本书. 由于 SAS 系统各种数据分析方法大都输出相应的 p 值，故本书精简大部分统计用表，仅列出几个常用统计数值表.

书中第1,5～8章由范金城编写,第2～4,9章由梅长林编写.

由于作者的水平所限,书中尚存在一些不妥之处,欢迎读者批评指正.

作者

于西安交通大学

2002年1月

目　录

第 1 章　数据描述性分析

1.1　数据的数字特征

数据分析研究的对象是数据，它们是 n 个观测值：

$$x_1, x_2, \cdots, x_n.$$

它们可以是从所要研究的对象的全体——总体中取出的，这 n 个观测值就构成一个样本. 在某些简单的实际问题中，这 n 个观测值就是所要研究对象的全体. 数据分析的任务就是要对这全部 n 个数进行分析，提取数据中包含的有用的信息. 如果数据是从总体抽出的样本，就要分析推断样本中包含的总体的信息.

数据作为信息的载体，当然要分析数据中包含的主要信息，即要分析数据的主要特征. 也就是说，要研究数据的数字特征. 对于数据的数字特征，要分析数据的集中位置、分散程度、数据的分布是正态还是偏态等等. 对于多元数据，还要分析多元数据的各个分量之间的相关性等等.

1.1.1　均值、方差等数字特征

一元数据的数字特征主要有下列几种. 设 n 个观测值为

$$x_1, x_2, \cdots, x_n,$$

其中 n 称为样本容量.

1. 均值

均值即是 $x_1, x_2, \cdots, x_n$ 的平均数：

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i. \tag{1.1}$$

均值表示数据的集中位置.

2. 方差、标准差与变异系数

方差是描述数据取值分散性的一个度量，它是数据相对于均值的偏差平方的平均：

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2. \tag{1.2}$$

方差的开方称为**标准差**. 方差的量纲与数据的量纲不一致，它是数据量纲的平方，

而标准差的量纲与数据量纲一致. 标准差为

$$s = \sqrt{s^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}. \tag{1.3}$$

刻画数据相对分散性的度量是**变异系数**:

$$CV = 100 \times \frac{s}{\bar{x}}(\%). \tag{1.4}$$

它是一个无量纲的量,用百分数表示.

与均值、方差有关的还有下列数字特征:

校正平方和

$$CSS = \sum_{i=1}^{n}(x_i - \bar{x})^2.$$

未校平方和

$$USS = \sum_{i=1}^{n}x_i^2.$$

3. *偏度与峰度*

偏度与峰度是刻画数据的偏态、尾重程度的度量. 它们与数据的矩有关. 数据的矩分为原点矩与中心矩.

k 阶原点矩

$$v_k = \frac{1}{n}\sum_{i=1}^{n}x_i^k. \tag{1.5}$$

k 阶中心矩

$$u_k = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^k. \tag{1.6}$$

显然,一阶原点矩 v_1 即均值. 二阶中心矩

$$u_2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

也称为方差. 而前述的方差 s^2 是 USS 除以 $n-1$,是为了保证估计总体方差时的无偏性.

偏度的计算公式为

$$g_1 = \frac{n}{(n-1)(n-2)s^3}\sum_{i=1}^{n}(x_i - \bar{x})^3$$

$$= \frac{n^2 u_3}{(n-1)(n-2)s^3}, \tag{1.7}$$

其中 s 是标准差. 偏度是刻画数据对称性的指标. 关于均值对称的数据其偏度为 0,右侧更分散的数据偏度为正,左侧更分散的数据偏度为负(图 1.1).

峰度的计算公式是

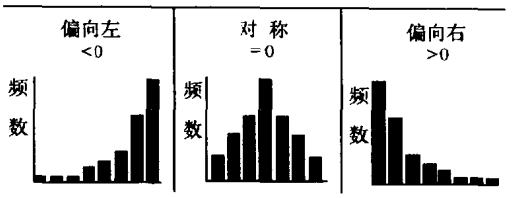

图 1.1

$$g_2=\frac{n(n+1)}{(n-1)(n-2)(n-3)s^4}\sum_{i=1}^{n}(x_i-\bar{x})^4-3\frac{(n-1)^2}{(n-2)(n-3)}$$

$$=\frac{n^2(n+1)u_4}{(n-1)(n-2)(n-3)s^4}-3\frac{(n-1)^2}{(n-2)(n-3)}. \tag{1.8}$$

当数据的总体分布为正态分布时，峰度近似为 0；当分布较正态分布的尾部更分散时，峰度为正，否则峰度为负. 当峰度为正时，两侧极端数据较多；当峰度为负时，两侧极端数据较少.

设观测数据是由总体 X 中取出的样本，总体的分布函数是 $F(x)$. 当 X 为离散分布时，总体的分布可由概率分布列刻画：

$$p_i=P\{X=x_i\},\qquad i=1,2,\cdots.$$

总体为连续分布时，总体的分布可由概率密度 $f(x)$ 刻画. 连续分布中最重要的是正态分布，它的概率密度 $\varphi(x)$ 及分布函数 $\Phi(x)$ 分别为

$$\varphi(x)=\frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right], \tag{1.9}$$

$$\Phi(x)=\int_{-\infty}^{x}\varphi(t)dt. \tag{1.10}$$

具有正态分布的总体称为正态总体.

上述数据的数字特征在数据为从某总体随机取出的样本时，即是样本的数字特征. 与样本数字特征对应的是总体的数字特征，它们分别是：

总体均值　　$\mu=\mathrm{E}(X)$，　　(1.11)

总体方差　　$\sigma^2=\mathrm{Var}(X)$，　　(1.12)

总体标准差　　$\sigma=\sqrt{\mathrm{Var}(X)}$，　　(1.13)

总体变异系数　　$\gamma=\frac{\sigma}{\mu}$，　　(1.14)

总体原点矩　　$\gamma_k=\mathrm{E}(X^k)$　　(k 阶)，　　(1.15)

总体中心矩　　$\mu_k=\mathrm{E}(X-\mu)^k$　　(k 阶)，　　(1.16)

总体偏度　　$G_1=\frac{\mu_3}{\sigma^3}$，　　(1.17)

总体峰度　　$G_2=\frac{\mu_4}{\sigma^4}-3$，　　(1.18)

这里，我们对偏度与峰度作进一步的说明.

总体偏度是度量总体分布是否偏向某一侧的指标. 对于对称的分布，偏度为0. 例如，对于正态分布，因 $\mu_3=0$，故 $G_1=0$. 若总体分布在右侧更为扩展，偏度为正；若总体分布在左侧更为扩展，偏度为负. 图 1.2 表示了偏度为正和偏度为负的概率密度的图像特点. 我们看到，总体偏度的这一特性与样本偏度的相应特性是相似的.

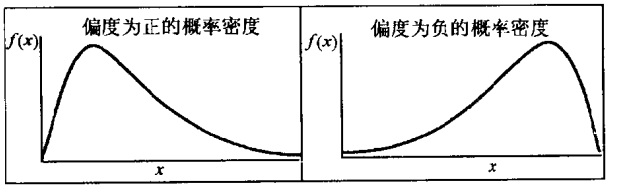

图 1.2

总体峰度是以同方差的正态分布为标准，比较总体分布尾部分散性的指标. 当总体分布是正态分布时，因 $\mu_4=3\sigma^4$，故总体峰度 $G_2=0$. 当 $G_2>0$ 时，总体分布中极端数值分布范围较广，此种分布称为粗尾的. 当 $G_2<0$ 时，两侧极端数据较少，此种分布称为细尾的(图 1.3).

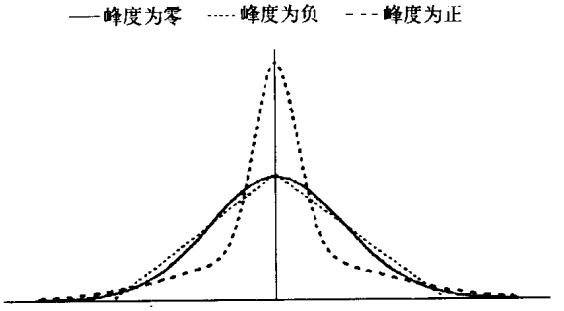

图 1.3

根据统计学的结果，样本数字特征是相应的总体数字特征的矩估计. 当总体数字特征存在时，相应的样本数字特征是总体数字特征的相合估计，从而当 n 较大时，有

$$\mu\approx\bar{x},$$

$$\sigma^2\approx s^2,$$

$$\sigma\approx s,$$

$$\gamma\approx CV,$$

$$\gamma_k\approx v_k,$$

$$\mu_k\approx u_k,$$

$$G_1\approx g_1,$$

$$G_2\approx g_2.$$

这里，特别要强调下列情况：当观测数据 $x_1,x_2,\cdots,x_n$ 是所要研究对象的全体时，

数据的分布即总体分布. 我们认为取得每一个观测数据 x_i 是等可能性的，即为 $\frac{1}{n}$；总体分布是离散均匀分布：

$$P\{X=x_i\}=\frac{1}{n}, \qquad i=1,2,\cdots,n.$$

对这种情况，数据数字特征即总体数字特征. 许多实际数据属于这种情况，它更能体现数据分析的特点——让数据本身说话. 实际上，我们也可以把这种情况看作取自确定性模型的数据，而上述数字特征仍有相应的统计意义.

计算数据的上述数字特征可以通过 SAS 系统 proc means 过程或 proc univariate 过程来实现.

例 1.1 从 19 个杆塔上的普通盘形绝缘子测得该层电导率(μs)的数据如下：

9.89	8.00	6.40	6.17	5.39	7.27	9.08
10.40	11.20	8.75	6.45	11.90	10.30	9.58
9.24	7.75	6.20	8.95	8.33		

计算均值、方差、标准差、变异系数、偏度、峰度.

解 通过计算，得

$$\bar{x}=8.487, \qquad s^2=3.406,$$
$$s=1.845, \qquad CV=21.745,$$
$$g_1=0.035, \qquad g_2=-0.852.$$

这里需要注意：

$$CV=100\times\frac{s}{\bar{x}}$$

取的是百分数. $g_1=0.035$，向右微偏. g_1,g_2 的绝对值较小，可以认为是来自正态总体的数据. 关于正态性的检验，以后还将介绍.

例 1.2 某单位对 100 名女学生测定血清总蛋白含量(g/L)，数据如下：

74.3	78.8	68.8	78.0	70.4	80.5	80.5	69.7	71.2	73.5
79.5	75.6	75.0	78.8	72.0	72.0	72.0	74.3	71.2	72.0
75.0	73.5	78.8	74.3	75.8	65.0	74.3	71.2	69.7	68.0
73.5	75.0	72.0	64.3	75.8	80.3	69.7	74.3	73.5	73.5
75.8	75.8	68.8	76.5	70.4	71.2	81.2	75.0	70.4	68.0
70.4	72.0	76.5	74.3	76.5	77.6	67.3	72.0	75.0	74.3
73.5	79.5	73.5	74.7	65.0	76.5	81.6	75.4	72.7	72.7
67.2	76.5	72.7	70.4	77.2	68.8	67.3	67.3	67.3	72.7
75.8	73.5	75.0	73.5	73.5	73.5	72.7	81.6	70.3	74.3
73.5	79.5	70.4	76.5	72.7	77.2	84.3	75.0	76.5	70.4

计算均值、方差、标准差、变异系数、偏度、峰度.

解 通过计算,得

$$\bar{x}=73.660, \qquad s^2=15.524,$$
$$s=3.940, \qquad CV=5.349,$$
$$g_1=0.061, \qquad g_2=0.034.$$

从计算结果看,偏度、峰度的绝对值皆较小,可以认为数据是取自正态总体的样本,即数据的总体分布是正态分布.

对于容量 n 较大的数据,有时可以采用分组的办法进行统计.这也可以通过SAS程序进行计算,可参看下例.

例 1.3 某电瓷厂的某种悬式绝缘子机电破坏负荷试验数据(单位:吨)分组表示如表 1.1.计算这批分组数据的均值、方差、标准差、变异系数、偏度、峰度.

表 1.1 绝缘子机电负荷破坏试验数据

组 段	组中值	组频数
5.5～6.0	5.75	4
6.0～6.5	6.25	3
6.5～7.0	6.75	15
7.0～7.5	7.25	42
7.5～8.0	7.75	49
8.0～8.5	8.25	78
8.5～9.0	8.75	50
9.0～9.5	9.25	31
9.5～10.0	9.75	5

解 通过计算,得

$$\bar{x}=8.100, \qquad s^2=0.629,$$
$$s=0.793, \qquad CV=9.789,$$
$$g_1=-0.382, \qquad g_2=0.057.$$

这批数据的容量 $n=277$.需要注意,对于这批分组数据,实际上是以组中值当成各组段中实际数据的代表(它不一定是实际数据),因此,以上算得的各数字特征是原始数据的数字特征的近似.

下面一个例子是确定型数据的类型,我们来进行分析.

例 1.4 1952～1997 年我国人均国内生产总值数据如表 1.2(单位:元).计算这批数据的数字特征.