实用数据分析方法

吴国富　安万福　刘景海　编

中国统计出版社

实用数据分析方法

吴国富　安万福　刘景海　编

中国统计出版社

（京）新登字041号

内 容 提 要

本书作为《微型机对话型数据统计分析方法程序集》的姊妹篇，全书共十章，内容有基础统计分析方法，各种随机数的产生，多元数据图表示的各种方法，非参数检验的各种方法，多元分析中的各种方法及数量化理论。该书注重介绍各种方法的实施步骤和应用范围，并扼要介绍其主要数学理论依据，但不作繁琐的推导证明。

本书可供从事应用数据分析方法的工程技术人员和管理人员学习及解决实际问题使用，也可供大专院校学生学习和实习参考。

实用数据分析方法

SHIYONG SHUJU FENXI FANGFA

吴国富　安万福　刘景海　编

*

中国统计出版社出版

（北京三里河月坛南街38号　100826）

新华书店北京发行所发行

北京市通县永乐印刷厂印刷

*

850×1168毫米　32开本　11.125印张　27万字

1992年10月第1版　1992年10月北京第1次印刷

印数：1—4000

ISBN 7-5037-0882-4/C·568

定价：11.00元

前　言

众所周知，无论是实验得到的数据，还是通过抽样或调查得到的数据，只有针对不同的数据，选用某些统计分析方法进行处理和分析，才能够揭示这些数据所反应的自然规律，进而获得解决问题的方法。因此，数据的各种统计分析方法，在各个学科中得到了越来越广泛的应用，特别是在地质、医学、气象、管理及各种社会调查的数据分析、处理等诸多方面的应用，成绩更为显著。

本书的内容与《微型机对话型数据统计分析方法程序集》一书相对应，那里侧重于程序的编写及调用，而这里则侧重于方法的介绍和应用，推导和证明较少。

本书是在1987年完成第一稿的基础上，经过中国科学院应用数学研究所多年来的教学与实践，几次修改后定稿。另外，我们还编制了与本书涉及的全部统计方法对应的应用软件(共2张盘，现已是第三版)，以供社会各部门、各单位需要。

本书主要包括三部分内容：

第一部分（1～5章）为基础方法部分，主要介绍各种检验，包括非参数检验、各种随机数的产生、各种图分析法等。

第二部分（6～9章）为多元统计分析方法部分，主要介绍多元回归分析、主成份分析、因子分析、判别分析、聚类分析的各种方法。

第三部分(10章)为数量化理论部分，包括数量化方法Ⅰ～Ⅳ，这些方法在管理及社会科学研究方面有着广泛的应用。

本书的目的在于给广大工程技术人员和管理人员在利用《微型机对话型数据统计分析方法程序集》解决实际问题时，能带来

一些方便。

限于我们的水平，错误和不当之处在所难免，欢迎广大读者批评指正。

编　者

1992.5

目　录

第一章　描述统计

第二章　估计、检验

第三章 非参数法

第四章 随机数

第五章 多元数据的图分析法

第六章 多元回归分析

第七章 判别分析

第八章 主成份分析与因子分析

第九章 聚类分析

第十章 数量化理论

第一章　描述统计

统计的方法，通常有描述性统计和推测性统计。描述统计就是把数据本身包含的信息，加以总结、概括、浓缩、简化，使问题变得更加清晰、简单，易于理解，便于处理。例如，表、图、数字特征等的计算就属于描述统计。

§1.1　特征值、方差、斜度、峰度

实验数据有如下三种形式：

（a）n个数据　x_1、x_2，…，x_n

（b）k个不同的数x_1、x_2，…，x_k以及每个数的频数f_1，f_2，…，f_k，列成表如下：

数据x_i	x_1	x_2	…	x_k	合　计
频数f_i	f_1	f_2	…	f_k	n

（c）给出分组和频数

分　组	组中值	频　数
$(a_0, a_1]$	m_1	f_1
$(a_1, a_2]$	m_2	f_2
⋮	⋮	⋮

续表

分　　组	组 中 值	频　数
$(a_{i-1}, a_i]$	m_i	f_i
⋮	⋮	⋮
(a_{k-1}, a_k)	m_k	f_k
合　计		n

其中$a_0<a_1<a_2<\cdots<a_k$，$m_i=(a_{i-1}+a_i)/2$，f_i表示数据落入第i组(a_{i-1}, a_i)内的频数。

1.1.1 总和与均值

对如上三种形式的数据，其总和、均值的表达式分别为：

总和：

$$(a)\quad T=x_1+x_2+\cdots+x_n=\sum_{i=1}^{n}x_i$$

$$(b)\quad T=f_1x_1+f_2x_2+\cdots+f_kx_k=\sum_{i=1}^{k}f_ix_i$$

$$(c)\quad T=f_1m_1+f_2m_2+\cdots+f_km_k=\sum_{i=1}^{k}f_im_i$$

均值：

$$(a)\quad \bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$$

$$(b)\quad \bar{x}=\frac{1}{n}\sum_{i=1}^{k}f_ix_i$$

$$(c)\quad \bar{x}=\frac{1}{n}\sum_{i=1}^{k}f_i m_i$$

1.1.2 最大值、最小值、中位数、众数

设n个数据按从小到大的顺序排列为

$$x_{(1)}\leqslant x_{(2)}\leqslant\cdots\leqslant x_{(n)}$$

则最大值即为$x_{(n)}$，最小值为$x_{(1)}$，而中位数

$$M=\begin{cases}x_{(\frac{n+1}{2})} & \text{当}n\text{为奇数时}\\ \frac{1}{2}\left[x_{(\frac{n}{2})}+x_{(\frac{n}{2}+1)}\right] & \text{当}n\text{为偶数时}\end{cases}$$

众数：在(b)、(c)情况下，使频数最大的x_i或m_i。

1.1.3 方差、标准差、均差

方差：

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

$$(b)\quad S^2=\frac{1}{n}\sum_{i=1}^{k}f_i(x_i-\bar{x})^2=\frac{1}{n}\sum_{i=1}^{k}f_i x_i^2-\bar{x}^2$$

$$(c)\quad S^2=\frac{1}{n}\sum_{i=1}^{k}f_i(m_i-\bar{x})^2=\frac{1}{n}\sum_{i=1}^{k}f_i m_i^2-\bar{x}^2$$

标准差：

$$S=\sqrt{S^2}$$

均差：

$$(a)\quad D=\frac{1}{n}\sum_{i=1}^{n}|x_i-\bar{x}|$$

$$(b)\quad D=\frac{1}{n}\sum_{i=1}^{k}f_i|x_i-\bar{x}|$$

$$(c)\quad D=\frac{1}{n}\sum_{i=1}^{k}f_i|m_i-\bar{x}|$$

1.1.4 极差、四分位极差

极差：

$$R=\text{最大值}-\text{最小值}=x_{(n)}-x_{(1)}$$

四分位极差：

$$QR=Q_3-Q_1$$

其中：Q_3——第三个四分位数（75%的点）。

Q_1——第一个四分位数（25%的点）。

四分位偏差：

$$QD=(Q_3-Q_1)/2$$

四分位变差：

$$QV=100(Q_3-Q_1)/(Q_3+Q_1)$$

四分位斜度：

$$QMS=|(Q_3-M)-(M-Q_1)|/(Q_3-Q_1)$$

其中M为中位数。

变异系数：

$$CV=S/\bar{x}\quad\text{（标准差/均值）}$$

1.1.5 斜度、峰度

斜度：

$$\gamma_1=-\mu_3/S^3$$

峰度：

$$\gamma_2=(\mu_4/S^4)-3$$

其中S为标准差，而μ_i对三种数据的定义如下：

$$(a)\quad \mu_i=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^i$$

$$(b)\quad \mu_j=\frac{1}{n}\sum_{i=1}^{k}f_i(x_i-\bar{x})^j \qquad (j=3,\ 4)$$

$$(c)\quad \mu_j=\frac{1}{n}\sum_{i=1}^{k}f_i(m_i-\bar{x})^j$$

由于正态分布的斜度、峰度皆为零，所以常把γ_1、γ_2做为偏离正态分布的尺度。

1.1.6 箱尾图

Tukey 提出的箱尾图由箱子和其上引出的两个尾组成（见图1.1），故有箱尾图之称。这种图用来表示在一定时间内一个班成绩的变化、物体位置的变化、原料的变化、产品标准的变化等。

例1.1 如下表1.1表示高一年级10门课程5次考试成绩，对此数据，取横坐标为测验次数，画出箱尾图。

我们对每次考试成绩，分别求出均值、标准差、最大值、最小值，便可画出箱尾图如下：

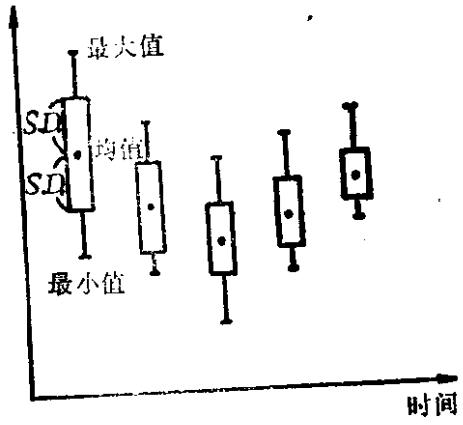
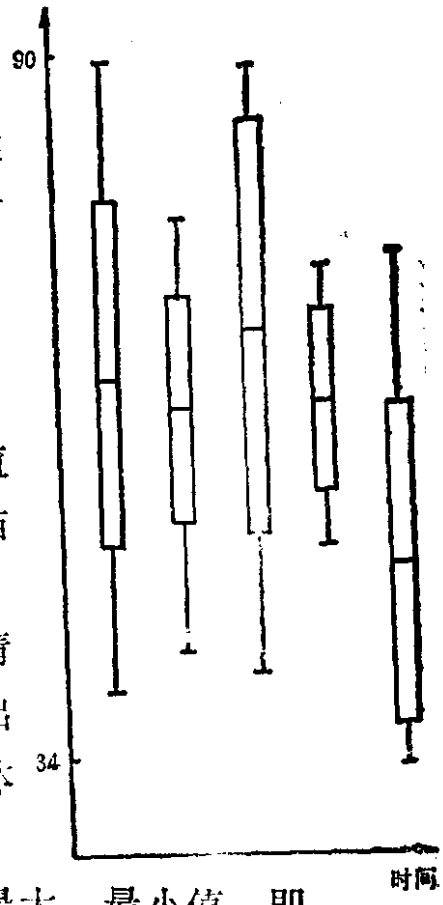


图1.1 箱尾图

表1.1 考试成绩

次数 \ 课目	A	B	C	D	E	F	G	H	I	J
1	77	56	61	54	60	68	90	39	68	75
2	63	69	77	66	65	52	58	43	66	64
3	78	58	82	65	86	90	60	41	45	80
4	66	58	60	73	66	63	64	51	74	63
5	51	42	75	68	46	40	44	38	34	58

§1.2 频数分布



在这一节里我们将介绍两种描述频数分布的方法：直方图法和树干叶法。

1.2.1 频数分布与直方图

实验数据x_1，x_2，…，x_n的均值$\bar{x}$、方差S^2等数字特征，概括地描述了实验数据的一些基本统计特征，但要进一步研究数据取值的分布情况，则需计算它们的经验分布，做出直方图就是一种最常用的方法，具体做法如下：

先用a，b分别表示该数据中的最大、最小值，即

$$a=\max_{1\leqslant i\leqslant n}\{x_i\},\qquad b=\min_{1\leqslant i\leqslant n}\{x_i\}$$

再把区间$[a,\ b]$分成k个等间隔的小区间：

$$[a_{i-1},\ a_i],\quad i=1,\ 2,\ \cdots,\ k$$

这里$a=a_0<a_1<a_2<\cdots<a_k=b$。这样的$k$个小区间把$n$个数据分成了$k$个组，且称每组的中间值$m_i=(a_{i-1}+a_i)/2$，$i=1,\ 2,\ \cdots,\ k$为该组的组中值，往往就以它来代表该组的数据。

然后统计落入各组内的数据个数f_i，它就是第i组$[a_{i-1},a_i]$的频数，显然应该有

$$f_1+f_2+\cdots+f_k=\sum_{i=1}^{k}f_i=n$$

若用n去除各组的频数，又可得各组的频率

$$f_i/n\qquad(i=1,\ 2,\ \cdots,\ k)$$