图书在版编目（CIP）数据

环境数据统计分析基础/程子峰，徐富春编著. —北京：化学工业出版社，2005.12
ISBN 7-5025-8137-5

Ⅰ.环… Ⅱ.①程…②徐… Ⅲ.数理统计-应用-环境管理-统计数据-统计分析（数学） Ⅳ.X32

中国版本图书馆 CIP 数据核字（2005）第 157955 号

环境数据统计分析基础

程子峰 徐富春 编著

责任编辑：管德存 邹 宁

责任校对：吴 静

封面设计：胡艳玮

*

化学工业出版社
环境·能源出版中心 出版发行

（北京市朝阳区惠新里 3 号 邮政编码 100029）

购书咨询：(010)64982530
(010)64918013

购书传真：(010)64982630

http://www.cip.com.cn

*

新华书店北京发行所经销

北京永鑫印刷有限责任公司印刷

三河市东柳万龙印装有限公司装订

开本 720mm×1000mm 1/16 印张 13¼ 字数 245 千字

2006 年 3 月第 1 版 2006 年 3 月北京第 1 次印刷

ISBN 7-5025-8137-5

定 价：28.00 元

前言

在环境保护工作中，人们得到了大量有关环境质量、污染物排放、生态指标、水文气象以及经济和社会发展的基础数据。这些数据的特点是量大、多元和具有不确定性。如何整理分析这些基础数据，并从这些基础数据中概括出一些描述环境的基本特征和概念；如何从这些数据中探讨和分析各种环境因素之间的关系，预测各种环境因素和环境系统的变化发展趋势，是从事环境数据分析、环境管理和环境科学研究工作的人员常常面临的问题。

作者在多年的环境保护工作实践中，利用概率论和数理统计的基本理论和计算方法来整理、分析环境数据，研究环境中各种因素之间的相互关系，揭示环境系统的内在规律，预测环境系统的变化趋势。实践表明，概率论和数理统计的基本理论和计算方法对于环境数据的分析是十分有用的。本书结合实际应用案例，主要介绍概率论和数理统计的基本理论和计算方法在环境数据分析中的应用。

本书可作为从事环境科学研究、环境数据分析和环境管理人员的参考书，也可作为高等院校师生教育、学习的参考教材。文中所述内容难免存在疏漏和不足，请广大读者批评指正。

作者

2005 年 8 月

目录

第一章　数据统计分析基础

在环境监测和科研工作中，我们从所布设的采样地点采集样品，进行分析测定，得到了大量的数据。分析、比较这些数据，从中获得有用的环境信息是环境监测和科研工作中的一个重要环节。环境数据统计分析的基础是概率论和数理统计方法。随着概率论和数理统计基础理论和计算方法在环境监测和科研系统中的普及和推广，越来越多的从事环境监测和科研的技术人员认识到数据统计分析的重要性。本章我们将扼要地介绍概率论和数理统计的基础概念及数据整理的一些方法，为以后各章节介绍数据统计分析方法提供基础知识。

第一节　概率论基本概念

一、随机事件

在生产实践和科学实验中，人们常常发现许多事件是不肯定的，在一定的条件下它可能发生，也可能不发生，即在一定条件下它的结果是不完全确定的。这类事件称为随机事件，简称事件，记作 A、B、C、…。例如抛掷一枚硬币，每次的结果都是不肯定的，A 表示“正面朝上”，B 表示“背面朝上”，则 A、B 就是两个随机事件。

概率论中将不可再分的事件称为基本事件。实际上，不可再分事件是针对试验目的而言的。如我们研究河流汞的污染水平，那么任何一种汞的浓度值都为一基本事件，总共有无穷多个基本事件；若研究汞的污染水平是为了了解河流的汞污染水平是否超标，那么就只有超标和未超标两个基本事件了。若干个基本事件组合而成的事件称为复合事件。

必然事件是随机事件的一个极端情况，它是指在一定条件下必然发生的事情，记作 Ω。例如，一年中四季的变化；海水每日两次的涨潮、落潮；压力为一个标准大气压和温度为 0℃时水会结冰等，都是必然事件。

不可能事件是随机事件的另一个极端情况，它是指在一定条件下必然不会发生的事件，记作 Φ。例如，我们上抛一个硬币，硬币必然不会飞离地球；当气压为一个标准大气压，温度大于 0℃时，水必然不会结冰等，都是不可能事件。

由全体基本事件组成的集合称为样本空间。随机事件在一次观察中是否发生，事前无法确定。但任何一次随机试验的结果必然出现在该样本空间中，因此样本空间本身作为一个事件是必然事件，因而样本空间亦可以以 $\boldsymbol{\Omega}$ 表示。类似的，不包含任何事件的集合也就是不可能事件，称为空集。空集亦可用 $\boldsymbol{\Phi}$ 表示。

【例 1-1】 检测空气样品中 SO_2、NO_x、CO 三项指标超标与否，可有 8 个基本事件，分别为：

A（SO_2 不超，NO_x 不超，CO 不超）　　E（SO_2 超，NO_x 超，CO 不超）

B（SO_2 不超，NO_x 不超，CO 超）　　F（SO_2 超，NO_x 不超，CO 超）

C（SO_2 不超，NO_x 超，CO 不超）　　G（SO_2 超，NO_x 超，CO 超）

D（SO_2 超，NO_x 不超，CO 不超）　　H（SO_2 不超，NO_x 超，CO 超）

二、事件的运算关系

随机事件是概率论和数理统计中最基本的概念。事件的概念是与一次随机试验出现的结果以及它所有可能出现的结果联系在一起的。事件和事件之间存在一定的联系。下面介绍事件之间最重要的一些关系和运算，这对我们理解数理统计的一些基本概念和计算方法是有帮助的。

1. 事件的基本关系

（1）包含　若事件 B 的发生必然导致事件 A 的发生，则称事件 A 包含事件 B 或事件 B 含于事件 A，记作 $A \supset B$ 或 $B \subset A$。

$B \subset A$ 相当于事件 B 中的每一个基本事件都包含在事件 A 之中，见图 1-1(a)。

显然空集可以作为任何集合的子集，而样本空间是所有基本事件的集合。因而对任何一个随机事件 A，总有：$\Omega \supset A \supset \Phi$。

（2）等价　如果两事件 A 和 B 同时发生或同时不发生，即 $A \supset B$ 且 $B \supset A$，则称事件 A 与事件 B 等价，记作 $A=B$。

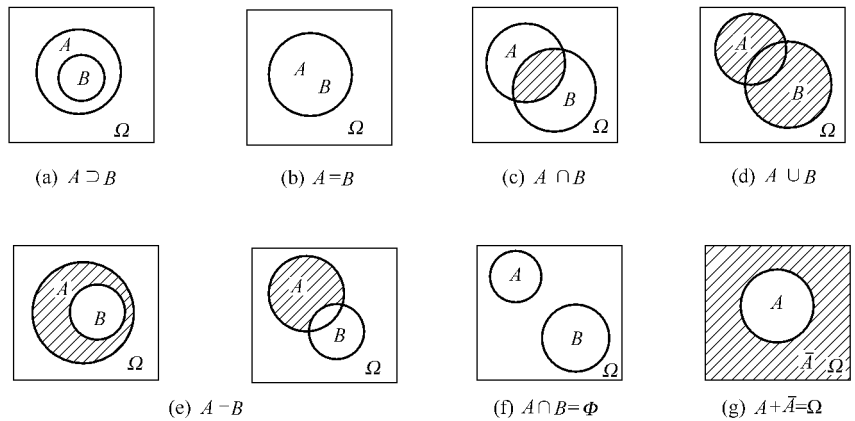
事件 A 与事件 B 等价意味着两事件是由完全相同的基本事件构成，A、B 不过是同一随机事件的两种不同的表达方式，见图 1-1（b）。

（3）积　表示两事件 A、B 同时发生的事件，称为事件 A 与事件 B 的积(或称为交)，记作 $A \cap B$（或 AB）。

$A \cap B$ 本身就是一个事件，它由既包含于事件 A 中又包含于事件 B 中的基本事件所构成，见图 1-1（c）的阴影部分。

（4）和　表示事件 A 与事件 B 至少有一个发生的事件，称为事件 A 与事件 B 的和，记作 $A \cup B$ 或 $A+B$。

$A \cup B$ 是由所有包含在事件 A 中的和包含在事件 B 中的基本事件构成的，见图 1-1（d）中的阴影部分。



(a) $A \supset B$　(b) $A=B$　(c) $A \cap B$　(d) $A \cup B$

(e) $A-B$　(f) $A \cap B=\Phi$　(g) $A+\bar{A}=\Omega$

图 1-1　事件的基本关系

如果事件 A_1，A_2，…，A_n 中至少有一个出现，则称为 A_1，A_2，…，A_n 的和，记作 $A_1+A_2+\cdots+A_n$ 或 $A_1 \cup A_2 \cup \cdots \cup A_n$ 或 $\sum_{i=1}^{n} A_i$ 或 $\bigcup_{i=1}^{n} A_i$ 。

（5）差　表示事件 A 发生而事件 B 不发生的事件，称为事件 A 与事件 B 的差，记作 $A-B$。

$A-B$ 是由所有包含于 A 而不包含于 B 中的基本事件组成，见图 1-1（e）的阴影部分。

对于任一事件 A，显然有：

$$A-A=\Phi; \qquad A-\Phi=A; \qquad A-\Omega=\Phi$$

（6）互斥　如果两个事件 A 和 B 不可能同时发生，则称事件 A 与事件 B 为互斥关系（或互不相容关系），必有 $A \cap B=\Phi$。

A、B 互斥表示这两个事件中不包含相同的基本事件，如图 1-1（f）所示。

（7）逆　必然事件 Ω 与事件 A 之差，称为事件 A 的逆事件或对立事件，记作 $\bar{A}$，即 $\bar{A}=\Omega-A$。它表示事件 A 不发生的事件。

A 与 $\bar{A}$ 事件互不相容，而且充满了整个样本空间，如图 1-1（g）所示。A 与 $\bar{A}$ 事件是互逆的。也就是说事件 A 也是事件 $\bar{A}$ 的逆事件，即 $A=\Omega-\bar{A}$。

（8）完备　如果事件 A_1，A_2，…，A_n，在每次观察中至少发生一个，即 $A_1 \cup A_2 \cup \cdots \cup A_n=\Omega$，则称事件 A_1，A_2，…，A_n 构成了一个事件完备组。

2. 事件的运算关系

事件运算满足如下关系。

(1) 交换律 $A\cup B=B\cup A$

$A\cap B=B\cap A$

(2) 结合律 $A\cup(B\cup C)=(A\cup B)\cup C$

$A\cap(B\cap C)=(A\cap B)\cap C$

(3) 分配律 $A\cap(B\cup C)=(A\cap B)\cup(A\cap C)$

$A\cup(B\cap C)=(A\cup B)\cap(A\cup C)$

分配律可以推广到多个或无限个事件的情况：

$$A\cap(\bigcup_{i=1}^{n}A_i)=\bigcup_{i=1}^{n}(A\cap A_i)$$

$$A\cup(\bigcap_{i=1}^{n}A_i)=\bigcap_{i=1}^{n}(A\cup A_i)$$

(4) 对偶性 对有限个或无穷多个 A_i，恒有：

$$\overline{\bigcup_i A_i}=\bigcap_i \overline{A}_i$$

$$\overline{\bigcap_i A_i}=\bigcup_i \overline{A}_i$$

(5) $A-B=A\cap\overline{B}$ 对任意两个随机事件都成立。

有兴趣的读者可以验证以上的 5 个运算关系。

三、概率的基本概念

如前所述，随机事件在一次试验中的结果是无法事先知道的。我们观察一个随机试验的结果时，一般来说，总可以发现有些事件出现的可能性大一些，有些事件出现的可能性小一些，有些事件发生的可能性基本相同。当我们进行多次重复试验和观察时，就可能对各种事件发生的可能性进行判断，了解这些事件出现的内在的统计规律。

随机事件发生的可能性大小用一个数值来表示，称为事件的概率，通常用字母 P 表示。如事件 A 的概率记作 $P(A)$。在概率论的发展史上，有各种定义概率和计算概率的方法，下面介绍两种最主要的方法。

1. 古典概率

由于概率是某一事件发生可能性大小的数量指标，我们很自然地联想，当完备事件组中的每一个基本事件是等可能出现时，事件 A 的概率可以用事件 A 所包含的基本事件数 k 与基本事件总数 n 的比值来计算，即：

$$P(A)=\frac{k}{n} \tag{1-1}$$

这种计算方法得到的概率称为古典概率。

古典概率除了要求所有基本事件是等可能外，还要求基本事件总数是有限的，即 n 不等于无穷大。

【例 1-2】 连续两次掷骰子，求两次出现的点数和为 10 的概率。

解 每掷一次骰子有 6 种可能结果，即 6 个基本事件，两次掷骰子有$6\times6=36$ 种可能结果，每一基本事件都是等可能的，可知完备组中的基本事件总数$n=36$。两次出现的点数和为 10 的可能结果有 3 个，即（4，6）、（5，5）、（6，4），因而 $k=3$。

由式（1-1）计算两次掷骰子点数和为 10 的概率为：

$$P=\frac{3}{36}=\frac{1}{12}$$

2. 统计概率

在古典概率的计算和定义中，完备事件组中每个基本事件是等可能出现的。这个条件在实践中常常不能满足，这时计算事件 A 的概率可以通过多次重复试验，观察事件 A 出现的频率来计算事件 A 出现的概率 $P(A)$。若重复试验的次数为 n，事件 A 出现的次数为 m，则事件 A 发生的频率为 m/n。用事件 A 出现的频率来表示事件 A 的概率，则有

$$P(A)=\frac{m}{n} \tag{1-2}$$

这种计算方法得到的概率称为统计概率。当试验次数 n 无限增大时，统计概率值呈现出稳定在某一数值的特征，称为概率的稳定性。它体现了事件 A 发生的可能性大小是事件本身固有特性的反映。

【例 1-3】 抽查某汽车制造厂生产的汽车整车噪声超标情况，抽查结果见表 1-1。

表 1-1　某汽车制造厂的汽车整车噪声抽查表

抽查汽车台数(n)	5	10	50	100	200	500	1000	2000
超标台数(m)	2	3	11	23	48	128	254	503
超标率(m/n)	0.4	0.33	0.22	0.23	0.24	0.256	0.254	0.252

从表 1-1 可见，车辆噪声超标率在 0.25 附近摆动，随着抽样台数增多，摆动幅度减小，超标概率稳定在 0.25。

3. 条件概率

按照概率的定义，事件 A 发生的概率为 $P(A)$，事件 B 发生的概率为$P(B)$，若事件 A、B 是有联系的，那么考虑在事件 A 已经发生的条件下，事件 B 发生的概率，称为事件 B 关于 A 的条件概率，记作$P(B/A)$。条件概率有如下的关

系式：

$$P\left(\frac{B}{A}\right)=\frac{P(AB)}{P(A)} \tag{1-3}$$

式中，$P(AB)$ 是事件 A、B 相交的事件出现的概率。

【例 1-4】 一盒子中有 100 个球，其中 60 个为白色球，40 个为红色球。并且这 100 个球中有 30 个标有号码，其中白球有 20 个标有号码。求从盒中抽得的白色球中，带有号码的球的概率。

解 这是一个条件概率问题。若事件 A 表示抽到白色球，事件 B 表示抽到带有号码的球，则从盒中抽球不外乎下列四种互斥的情况。

(1) 抽到白色带有号码的球为 AB。设属于这种情况的球的数目为 n_1。

(2) 抽到白色不带有号码的球为 $A\bar{B}$。设属于这种情况的球的数目为 n_2。

(3) 抽到红色带有号码的球为 $\bar{A}B$。设属于这种情况的球的数目为 n_3。

(4) 抽到红色不带有号码的球为 $\bar{A}\bar{B}$。设属于这种情况的球的数目为 n_4。

则上述抽到的白色球中，带有号码的球的概率 $P(B/A)$ 应为白色带有号码的球数 n_1 与全部白色球数 n_1+n_2 之比。即：

$$P\left(\frac{B}{A}\right)=\frac{n_1}{n_1+n_2}=\frac{\dfrac{n_1}{n_1+n_2+n_3+n_4}}{\dfrac{n_1+n_2}{n_1+n_2+n_3+n_4}}$$

由于 $P(AB)=\dfrac{n_1}{n_1+n_2+n_3+n_4}$；$P(A)=\dfrac{n_1+n_2}{n_1+n_2+n_3+n_4}$

则，$P\left(\dfrac{B}{A}\right)=\dfrac{P(AB)}{P(A)}$

在本问题中 $n_1=20$，$n_2=40$，因而抽到的白球中带有号码的球的概率为

$$P\left(\frac{B}{A}\right)=\frac{20}{20+40}=\frac{1}{3}$$

如果两事件 A、B 是相互完全独立的，即：

$$P(AB)=P(A)P(B)$$

此时，条件概率有如下关系式：$P\left(\dfrac{B}{A}\right)=\dfrac{P(AB)}{P(A)}=\dfrac{P(A)P(B)}{P(A)}=P(B)$

即事件 B 出现的概率与事件 A 是否已出现完全无关。

4. 概率的一些基本性质

概率论是进行数据统计分析的基础，了解概率的基本性质和运算规律对于正确理解和应用数据统计分析方法是有重要意义的。由上述定义的概率概念，我们可以引出有关概率的如下基本性质。

(1) 概率是非负数，即

$$P(A)\geqslant 0$$

若 A 是不可能事件，则

$$P(A)=P(\Phi)=0$$

(2) 概率不大于 1，即

$$P(A)\leqslant 1$$

若 A 是必然事件，则

$$P(A)=P(\Omega)=1$$

(3) 对任何两事件 A 和 B 有

$$P(A\cup B)=P(A)+P(B)-P(A\cap B)$$

若 A 和 B 互斥，则

$$P(A\cup B)=P(A)+P(B)$$

(4) 对任一事件 A 有

$$P(\bar{A})=1-P(A)$$

(5) 对任何两事件 A 和 B，若 $A\subset B$，则恒有

$$P(A)\leqslant P(B)$$

四、随机变量和分布函数

1. 随机变量的概念

在一定条件下进行重复试验，试验的结果随着随机因素的变化而变化，但又遵从一定的概率分布规律。这种随机试验的可能结果可以用一个变量 X 的数值来表示，称为随机变量。

随机变量是随机事件的数量化表征。通常将随机变量分为两类，即离散型随机变量和连续型随机变量。

• 离散型随机变量　如果随机变量 X 只能以一定的概率取分列的数值 x_1，x_2，…，x_n，则称这种变量为离散型随机变量。例如一年中实验室失控样品的数目是离散型随机变量，[例 1-3] 中整车噪声超标的汽车台数也是离散型随机变量。

• 连续型随机变量　如果随机变量 X 以一定概率的取值充满某一数值区间，即在某一数值区间中可任意取值，取值数量有任意多个，则称这种变量为连续型随机变量。例如检测河流中某断面的 pH 值，检测结果可以是一定 pH 范围内的任何一个值，是连续型随机变量。

2. 分布函数的概念和性质

一个随机变量取值的规律，称为该随机变量的分布，分布函数就是表征随机

变量分布的函数。给定随机变量 X，考虑 X 的值小于 x 的概率为 $P(X<x)$，显然它是 x 的函数，我们称其为随机变量 X 的分布函数。若记分布函数为$F(x)$，则有：

$$F(x)=P(X<x) \tag{1-4}$$

随机变量 X 落在某一数值区间 $[x_1, x_2)$ 内的概率为：

$$P(x_1\leqslant X<x_2)=P(X<x_2)-P(X<x_1)=F(x_2)-F(x_1) \tag{1-5}$$

式（1-5）表明，随机变量 X 的概率分布可由其分布函数确定。

分布函数有如下基本性质：

（1）在（$-\infty<x<+\infty$）的整个区间中，任一随机变量必满足：

$$0\leqslant F(x)\leqslant 1$$

（2）由于随机变量不取任何值的概率为零，因而有：

$$F(-\infty)=\lim_{x\to-\infty}F(x)=0$$

（3）随机变量能够取任何值为必然事件，因而有：

$$F(+\infty)=\lim_{x\to+\infty}F(x)=1$$

（4）当 $x_2>x_1$ 时，显然有概率 $P(X<x_2)\geqslant P(X<x_1)$，因而有：

$$F(x_2)\geqslant F(x_1) \qquad (\text{当 } x_2>x_1 \text{ 时})$$

这表明分布函数具有单调递增的性质。

五、分布密度函数

连续型随机变量的概率分布除了可以用分布函数 $F(x)$ 表示外，还可用分布密度函数表示。分布密度函数的定义是：连续型随机变量 X 的值落在单位区间内的概率，记作 $f(x)$。根据定义，可以得到分布密度和分布函数之间的关系：

$$F(x)=\int_{-\infty}^{x}f(x)\mathrm{d}x \tag{1-6}$$

或

$$f(x)=\lim_{\Delta x\to 0}\frac{F(x+\Delta x)-F(x)}{\Delta x} \tag{1-7}$$

可以证明随机变量的概率密度函数 $f(x)$ 有如下性质：

$$f(x)\geqslant 0 \qquad (-\infty<x<+\infty)$$

$$\int_{-\infty}^{+\infty}f(x)\mathrm{d}x=1$$

$$P(x_1\leqslant X\leqslant x_2)=F(x_2)-F(x_1)=\int_{x_1}^{x_2}f(x)\mathrm{d}x$$

【例 1-5】 已知一随机变量的概率密度函数为 $f(x)=\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{x^2}{2}\right)$，求随机变量数值落在区间 $[-1, 1]$ 的概率。

解　随机变量落在数值区间［－1，1］的概率为：

$$P(-1 \leqslant X \leqslant 1) = F(1) - F(-1) = \int_{-1}^{1} f(x)\mathrm{d}x = \int_{-1}^{1} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) \mathrm{d}x$$

上面的积分没有解析解。这是一个正态分布的情况，查正态分布表（见附表 1）可得：

$$P(-1 \leqslant X \leqslant 1) = 0.6826$$

六、随机变量的数字特征

随机变量的数字特征是反映随机变量的某方面特征的数值，如随机变量的取值中心；随机变量的取值的分散程度等。随机变量的数字特征中，最重要的是数学期望和方差。下面将分别予以介绍。

1. 数学期望（均值）

随机变量的数学期望（或称均值）μ 是反映随机变量取值的平均水平的特征数字，通常记作 $E(X)$ 或 $M(X)$。

对离散性随机变量 X，设其可能的取值为 x_k，其概率密度函数为：

$$P(X = x_k) = p_k \quad (k = 1, 2, 3, \cdots, n)$$

则随机变量的数学期望为：

$$\mu = E(X) = \frac{\sum_{k=1}^{n} x_k p_k}{\sum_{k=1}^{n} p_k} \tag{1-8}$$

由于 $\sum_{k=1}^{n} p_k = 1$，因而式（1-8）可简化为：

$$\mu = E(X) = \sum_{k=1}^{n} x_k p_k \tag{1-9}$$

连续型随机变量 X，若其分布密度函数为 $f(x)$，则其数学期望为：

$$\mu = E(X) = \frac{\int_{-\infty}^{+\infty} x f(x)\mathrm{d}x}{\int_{-\infty}^{+\infty} f(x)\mathrm{d}x} = \int_{-\infty}^{+\infty} x f(x)\mathrm{d}x \tag{1-10}$$

由上述离散型随机变量和连续型随机变量的定义可以看到，随机变量的数学期望实际上是该随机变量所有可能的取值以其相应的概率为权重的加权平均值。

数学期望有以下几个简单性质：

（1）$E(c) = c$

（2）$E(kX) = kE(X)$

(3) $E(X+c)=E(X)+c$

(4) $E(kX+c)=kE(X)+c$

式中的 k、c 均为常数。

【例 1-6】 试求概率密度函数如上例的 $f(x)=\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{x^2}{2}\right)$ 的随机变量的数学期望。

解 这是一连续型随机变量，由数学期望的定义式（1-10）可计算得：

$$E(X)=\int_{-\infty}^{+\infty}\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{x^2}{2}\right)x\mathrm{d}x=0$$

2. 方差和标准差

随机变量的方差和标准差是反映随机变量的取值相对于其数学期望的偏差程度的特征数字。

随机变量方差 σ^2 的定义为

$$\sigma^2=E[X-E(X)]^2$$

对于离散型随机变量，x_k 是可能取值，$P(X=x_k)=p_k$ 为概率分布（$k=1,2,3,\cdots,n$），则方差为：

$$\sigma^2=\sum_{k=1}^{n}[x_k-E(X)]^2\cdot p_k \tag{1-11}$$

若连续型随机变量的概率密度为 $f(x)$，则方差为：

$$\sigma^2=\int_{-\infty}^{+\infty}[x-E(X)]^2 f(x)\mathrm{d}x \tag{1-12}$$

随机变量的标准差的定义是方差开方，记作 σ。即有：

$$\sigma=\sqrt{\sigma^2}=\sqrt{E[X-E(X)]^2} \tag{1-13}$$

不难看出，随机变量的取值愈分散，标准差和方差值就愈大。

【例 1-7】 试求［例 1-6］中概率密度函数为 $f(x)=\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{x^2}{2}\right)$ 的随机变量的方差和标准差。

解 已知该随机变量的数学期望 $E(X)=0$，因而按方差定义式（1-12）有：

$$\begin{aligned}\sigma^2&=\int_{-\infty}^{+\infty}[x-E(X)]^2\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{x^2}{2}\right)\mathrm{d}x\\&=\int_{-\infty}^{+\infty}x^2\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{x^2}{2}\right)\mathrm{d}x\\&=\frac{1}{\sqrt{2\pi}}\left[-x\exp\left(-\frac{x^2}{2}\right)\Big|_{-\infty}^{+\infty}+\int_{-\infty}^{+\infty}\exp\left(-\frac{x^2}{2}\right)\mathrm{d}x\right]\\&=1\end{aligned}$$

显然标准差为：

$$\sigma=1$$

七、随机变量的几种重要的理论分布

前面我们已经介绍了分布函数的概念，这里我们要介绍在理论和实践应用中有重要意义的随机变量的几种理论分布。

1. 二项分布

二项分布是一种常见的离散型随机变量的理论分布。设一次试验中，试验结果对事件 A 只有两种可能，出现或不出现，二者必居其一，且每次试验结果彼此独立，即一次试验的结果完全不影响另一次试验的结果。若令事件 A 出现的概率为 p，事件 $\bar{A}$ 出现的概率（A 不出现）为 q，则显然有关系式：

$$q=1-p$$

进行 n 次试验事件 A 出现 k 次的概率为：

$$P=C_n^k p^k q^{n-k}=C_n^k p^k (1-p)^{n-k} \tag{1-14}$$

式（1-14）中，C_n^k 是 n 个元素中取 k 个元素的组合数。

$$C_n^k=\frac{n!}{k!\,(n-k)!} \tag{1-15}$$

由于式（1-14）中的概率 p 的表达式是二项式 $(p+q)^n$ 的展开式中的对应项，因而这一概率分布称为二项分布。

二项分布的分布函数 $F(x)$ 为

$$F(x)=P(x\leqslant k)=\sum_{i=0}^{k} C_n^i p^i (1-p)^{n-i} \tag{1-16}$$

二项分布的数学期望为：

$$E(X)=\sum_{k=0}^{n} kC_n^k p^k (1-p)^{n-k}=np \tag{1-17}$$

二项分布的方差为：

$$\sigma^2=\sum_{k=0}^{n} (k-np)^2 C_n^k p^k (1-p)^{n-k}=npq \tag{1-18}$$

【例 1-8】 已知某市的机动车辆整车噪声超标率为 40%。若任意抽查 10 辆汽车，问抽查到噪声超标车辆的概率分布。

解 这显然是一个二项分布问题，依题已知 $n=10$，$p=0.4$，$q=0.6$，即可按二项分布概率计算式（1-14），计算超标车辆数 0～10 台的概率见表 1-2。

显然 $k=4$ 的概率 P 有最大值，即抽到 4 辆汽车噪声超标的可能性最大。

表 1-2 抽查到的超标车辆数目与其相应的概率

超标车辆(k)	0	1	2	3	4	5	6	7	8	9	10
概率(P)	0.006	0.04	0.12	0.22	0.25	0.20	0.11	0.04	0.01	0.0016	0.0001

2. 正态分布

正态分布是一种具有重要理论和实践意义的连续型理论分布。在环境数据统计分析中，正态分布同样具有重要意义。例如，在环境分析测试中，人们经常要对分析误差进行分析，分析误差一般服从正态分布。一般而言，当随机变量受到很多随机因素的影响，而每一随机因素的影响很小，不起决定性作用时，具有这种特性的随机变量，一般服从正态分布。

正态分布的随机变量 X 具有如下的概率密度函数：

$$f(x)=\frac{1}{\sigma\sqrt{2\pi}}\exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] \tag{1-19}$$

式（1-19）中 μ，σ 是两个常数。可以证明，μ 是随机变量 X 的数学期望，σ 为标准差。均值为 μ，标准差为 σ 的正态分布通常以 $N(\mu,\sigma)$ 表示。

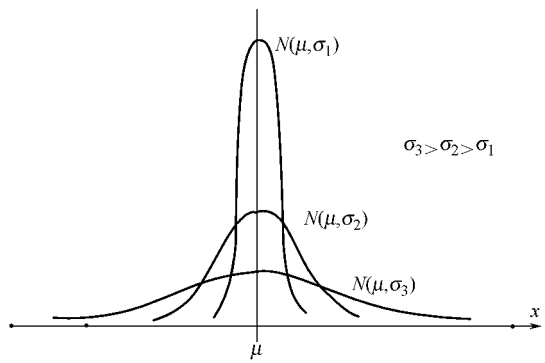


图 1-2 正态分布曲线

正态分布有如图 1-2 所示的分布曲线。

由图 1-2 可见，两个常数 μ，σ 是概率密度曲线的决定因素。均值 μ 决定了曲线的位置，随着 μ 的取值不同，整个曲线在 x 轴上平移。σ 决定了曲线的形状，σ 值越大曲线越低平，σ 值越小曲线越尖锐。

当 $\mu=0$，$\sigma=1$ 时的正态分布称作标准正态分布，通常表示为 $N(0,1)$。标准正态分布的密度函数为：

$$f(x)=\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{x^2}{2}\right) \tag{1-20}$$

显然这就是我们在［例 1-5］、［例 1-6］、［例 1-7］遇到的概率密度函数。

对于服从一般正态分布 $N(\mu,\sigma)$ 的随机变量 X，若用适当的数学变换$\left(z=\frac{x-\mu}{\sigma}\right)$，即可使 z 成为服从标准正态分布 $N(0,1)$ 的随机变量。同样通过反变

换 $x=\sigma z+\mu$ 亦可将标准正态分布还原为一般正态分布。因而，只要我们掌握了标准正态分布的性质，就可认为掌握了所有正态分布的性质。

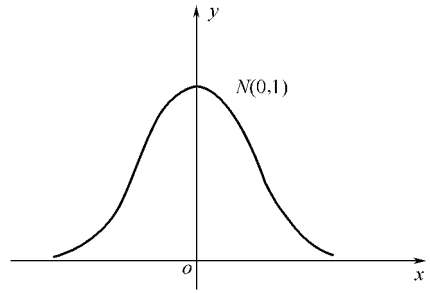


图 1-3　标准正态分布曲线

标准正态分布 $N(0，1)$ 具有如下主要性质。

(1) 由标准正态分布的定义可知，标准正态分布的数学期望 $\mu=0$，标准差 $\sigma=1$。

(2) 在均值处，即 $x=0$ 处，标准正态分布概率密度函数 $f(x)$ 具有最大值。

(3) 标准正态分布的概率密度曲线（见图 1-3）是以 $x=0$ 为对称轴的曲线，曲线以下的总面积为 1，即全概率为 1。

由于正态分布在理论和实践上的重要性，数理统计学家已将标准正态分布随机变量落在不同区间内的概率计算结果列成表格（见附表 1），以供读者查阅。其中在统计学中最重要的几个区间内的概率为：

$$P(-1<X<1)=0.683$$
$$P(-1.96<X<1.96)=0.950$$
$$P(-2.58<X<2.58)=0.990$$
$$P(-3<X<3)=0.997$$

3. 对数正态分布

环境统计数据分析中经常遇到的另一类分布是对数正态分布。顾名思义，对数正态分布是指若随机变量的对数服从正态分布，则称该随机变量服从对数正态分布。对数正态分布的概率密度函数为：

$$f(x)=\frac{1}{x\sigma\sqrt{2\pi}}\exp\left[-\left(\frac{\ln x-\mu}{\sigma}\right)^2\right] \tag{1-21}$$

式中，μ 和 σ 分别为随机变量 $\ln x$ 的均值和标准差。实际上，只要将服从对数正态分布的随机变量 x 取对数 $y=\ln x$，则变量 y 为服从正态分布的随机变量。

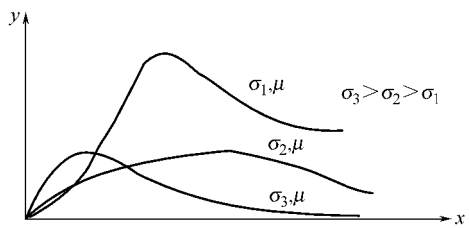


图 1-4　对数正态分布曲线

显然，服从对数正态分布的随机变量 X 的取值范围为 $[0，+\infty)$。全概率为 1，即：

$$\int_0^{+\infty}\frac{1}{x\sigma\sqrt{2\pi}}\exp\left[-\left(\frac{\ln x-\mu}{\sigma}\right)^2\right]\mathrm{d}x=1$$

概率密度函数的曲线见图 1-4。

服从对数正态分布的例子在环境数

据资料的整理中经常可以发现。如某些城市空气中 SO_2 的浓度、一些地区土壤中某些金属元素的含量等，都服从对数正态分布。

第二节 统计学基础

一、总体和个体

在统计学中将研究对象的所有可能的观测结果称为总体，总体中的一个单元称为个体。显然总体是所有个体的集合。例如要研究某一地区一年中空气中 SO_2 日均污染水平，则该年中每日的 SO_2 平均值是一个个体，而一年中所有 SO_2 日均值组成总体。总体和个体的内涵随研究问题改变而改变。在上述例子中，如要研究该地区一日内 SO_2 小时均值污染水平，则这一天中，每小时的 SO_2 均值为一个个体，一日中所有 SO_2 小时均值组成该研究问题的总体。总体可以是有限的。上述两个例子，显然总体包含的个体是有限的，这种总体称为有限总体。总体也可以是无限的。我们研究某河流 BOD_5 浓度的沿程分布。由于该河流沿程的点为无限多个，因而该研究总体有无限多个个体，这种总体称为无限总体。

二、样本

从总体中抽取一部分个体称为总体的一个样本。样本中所含个体的数目称为样本的大小（或称样本容量）。

总体的性质是由各个个体的性质决定的。当我们了解了总体中每一个个体的性质，我们就掌握了总体的性质。但是要做到这一点常常是很困难的，有时甚至是不可能的。因为通常总体包含的个体数目非常多，有时甚至是无限的，不可能对每个个体的性质加以测定。

例如研究某一河流的水质状况，我们不可能对整条河流的每一滴水（总体）化验，而是在一些断面采集一些水样（样本）进行分析，从而了解整个河流的水质情况。有时总体所包含的个体数目虽然是有限的，但是当我们为确定个体的性质所做的测定或试验是破坏性的，我们也不可能对总体所含的每个个体进行测定。由于上述原因，人们总是从总体中抽取样本，通过分析样本的性质来了解总体性质。研究样本，通过样本来了解总体成为统计学中重要的研究内容。

三、样本的频数分布

样本的频数是指将样本数据在取值范围内分成若干区间，统计数据落入每个区间内的次数。频数与样本数之比称为相对频数，亦称频率。对数据进行分组，得到的每组的频数或相对频数称为数据的频数分布。频数分布能较为完整地反映

实验数据的统计性质。因此在进行数据整理时，频数分布是通常采用的方法。

对数据进行频数分布时，一般有以下几个基本步骤。

(1) 指定进行频数分布区间的上下限　找出观察数据的最大值和最小值。以最大值作为频数分布区间的上限，最小值作为下限。

(2) 确定频数分布的组数　频数分布的组数根据观察数据的数目确定。不宜太少，也不宜太多。一般以 5～15 组为宜。

(3) 确定每组的界限　根据所确定的组数进一步可以确定每组的界限。一般采用等区间分组，即每组组距相等。特殊情况可做一些调整。需要注意的是，每组界限的取值一般比原始数据的精度高一位，这样可以避免实验数据落在界限上。

统计实验数据落入每组内的次数，即得到了频数分布。

【例 1-9】 某城市用网格法进行城市环境噪声普查，共设置 249 个测点。将 249 个测点数据按 5dB(A) 一档分档列于表 1-3。

表 1-3　某城市环境噪声普查结果

dB(A)	40.00～44.99	45.00～49.99	50.00～54.99	55.00～59.99	60.00～64.99	65.00～69.99	70.00～74.99	75.00～79.99
测点数(频数)	11	22	40	82	43	31	15	5
相对频数	4.4%	8.8%	16.1%	32.9%	17.3%	12.5%	6.0%	2.0%

频数分布亦可用相对频数作频数分布图表示（见图 1-5）。

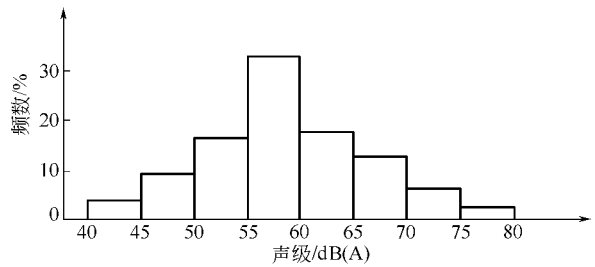


图 1-5　某城市环境噪声频数分布图

频数分布图表示的优点是直观，由图 1-5，该城市的环境噪声分布状况可一目了然。

相对频数分布也常用饼形图来描绘。每一扇形部分与相应的每一组相对频数相对应。例如，某城市一年空气质量优、良、轻微污染、中度污染和重污染天数的相对频数分布为：

优	良	轻微污染	中度污染	重污染
0.25	0.32	0.28	0.10	0.05

其相对频数亦可用饼形图来表示（见图 1-6）。

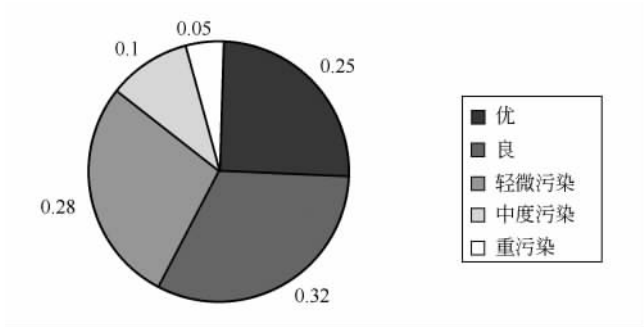


图 1-6 某城市空气质量相对频数分布图

四、样本的特征数

在上一节中，我们介绍了随机变量的数字特征概念，并具体介绍了随机变量最重要的两个特征数字——数学期望（均值）和方差。随机变量的特征数字反映的是总体的数据统计性质，在实际问题中，它往往是未知的。为了了解总体的性质，人们通常从总体中抽取样本，通过分析样本数据的统计性质来推断总体的性质。因而样本数据的统计性质，如数据分布的中心趋势、数据分布的分散程度、数据分布的形状等在数据分析的实践中具有重要的意义。反映样本数据统计性质的一些数值称为样本的特征数。由于从总体中抽取样本可以以不同方式进行，样本的特征数随样本抽取方式的不同而不同，它们本身是变量。而对某一指定的总体而言，它的数字特征为常数，这是样本特征数和总体特征数的不同之处。下面将介绍样本主要的一些特征数。

1. 算术平均值

设样本的 n 个测定值为 x_1，x_2，…，x_n，其算术平均值 $\bar{x}$ 的定义为所有测定值的总和除以测定次数，其计算式为

$$\bar{x}=\frac{x_1+x_2+\cdots+x_n}{n}=\frac{1}{n}\sum_{i=1}^{n}x_i \tag{1-22}$$

算术平均值是最常用的反映样本数据中心趋势的特征数。但它易受样本数据中特大或特小值的影响。对于服从正态分布的数据，算术平均值代表了数据中心趋势的典型水平。对于不服从正态分布的数据，算术平均值往往不反映数据中心趋势的典型水平，这是我们在应用算术平均值对样本进行统计分析时需要注意的问题。

【例 1-10】 对某一植物样品的有机砷含量进行了 7 次独立的检测，检测结果见表 1-4。求其算术平均值。

表 1-4　某一植物样品有机砷含量检测

次　序	1	2	3	4	5	6	7
浓度/(mg/L)	17.21	17.20	17.23	17.19	17.18	17.20	17.19

显然，该样品的有机砷含量的算术平均值为：

$$\bar{x}=\frac{1}{7}(17.21+17.20+17.23+17.19+17.18+17.20+17.19)=17.20$$

在本例中，样本服从正态分布，因而算术平均值代表了样本的中心趋势。

2. 几何平均值

几何平均值的定义是 n 个测定值乘积的 n 次方根。设 n 个样本的测定值为 x_1，x_2，…，x_n，其几何平均值 $\overline{x_G}$ 的计算式为：

$$\overline{x_G}=\sqrt[n]{x_1x_2\cdots x_n} \tag{1-23}$$

利用对数形式表示几何均值在计算上更为方便。将式（1-23）两侧取对数可得几何均值的对数计算形式

$$\lg \overline{x_G}=\frac{1}{n}\sum_{i=1}^{n}\lg x_i \tag{1-24}$$

取 $\lg \overline{x_G}$ 的反对数即可得到几何均值 $\overline{x_G}$。

对于服从对数正态分布的样本，几何均值比算术均值更能反映样本的中心趋势。需要注意的一点是，对数据中有零或负值的样本，不能计算样本的几何均值。

有兴趣的读者可以证明，对任一样本，若样本的几何平均值存在，该值总是小于或等于该样本的算术平均值。如［例 1-10］中植物样品中有机砷含量的几何平均值为：

$$\overline{x_G}=\sqrt[7]{17.21\times17.20\times17.23\times17.19\times17.18\times17.20\times17.19}=17.20$$

一般来说，样本越分散，$\overline{x_G}$ 比 $\bar{x}$ 就小得越多。

3. 加权算术平均值

一组样本数为 n 的测定值的加权算术平均值是该 n 个测定值的加权总和除以 n 个权重总和。设样本的 n 个测定值为 x_1，x_2，…，x_n；其相应的权重为 ω_1，ω_2，…，ω_n；则该样本的加权算术平均值 $\overline{x_w}$ 为：

$$\overline{x_w}=\frac{\sum_{i=1}^{n}\omega_i x_i}{\sum_{i=1}^{n}\omega_i} \tag{1-25}$$

【例 1-11】 某河流四个断面上测得污染物酚的浓度为 0.036mg/L，0.024mg/L，0.023mg/L，0.019mg/L。每个断面所代表的河段长度分别为 3.4km，5.6km，2.7km，4.3km，计算该河流酚污染平均水平。

以断面所代表的河段长度为权重因子，以加权算术平均值来反映该河流的酚污染平均水平，由式（1-25）可有：

$$\overline{x_w}=\frac{3.4\times0.036+5.6\times0.024+2.7\times0.023+4.3\times0.019}{3.4+5.6+2.7+4.3}=0.025$$

又例，城市中 n 条交通干线的噪声污染水平分别测得为 L_1，L_2，…，L_n，每条交通干线的长度分别为 l_1，l_2，…，$l_n$km，则城市交通噪声平均污染水平可由以下加权算术平均值求得：

$$\overline{L}=\frac{\sum_{i=1}^{n}L_i l_i}{\sum_{i=1}^{n}l_i}$$

这是评价城市道路交通噪声污染的重要公式。

4. 中位数

样本的中位数是指当样本的 n 个测定值从小到大排列时，居于中间位置的那个值。将样本的 n 个测定值由小到大顺序排列，并由 1 到 n 编号。若 n 为奇数，则样本的中位数是第$\frac{n+1}{2}$个值，即为居于中间位置的测定值。若 n 为偶数，中位数介于第$\frac{n}{2}$个与第$\frac{n}{2}+1$个测定值之间，如无特殊规定，中位数即取这两个测定值的算术平均值。

中位数也是表征样本中心趋势的特征数。它与算术平均值相比，算术平均值易受特大值或特小值的影响，中位数则不受测定值中特大值或特小值的影响。正因为这一点，在偏态分布中，中位数比算术平均值更好地代表样本的中心趋势。

环境噪声的一个很常用的评价量——统计声级中的 L_{50} 基本上就是一个中位数。L_{50} 代表有 n 次测定值的一个样本中，有 50％的测定值大于比值。

5. 众数

众数是指样本中出现频数最高的测定值。一般来说，众数也是反映样本中心趋势的特征数，但它与算术平均值和中位数不完全相同。对正态分布的样本，众数与算术平均值和中位数重合，而对对数正态分布的样本，众数与几何平均值相同。

众数除了定量表征样本的中心趋势外，在很多场合仅用于定性反映样本的特

征，而在对样本进行进一步统计分析时其作用受到某些限制。

某些样本的分布可以具有几个局部众数，在这种情况下，分布称为多峰分布。

6. 极差

极差的定义是样本测定值中最大值与最小值之差，一般用 R 表示，即

$$R=x_{最大}-x_{最小} \tag{1-26}$$

极差是反映样本测定值离散程度的特征数。由于极差由样本中两个极端值所确定，因而它易受到极大值或极小值波动的影响。在一些问题中，用极差来表征样本测定值的离散程度是较为粗糙的方法。但是由于极差计算简便，一目了然，在样本数据的统计分析中仍有许多应用。在后面要介绍的样本测定值离群值的检验中，就有极差的应用。

7. 平均偏差

设样本的 n 个测定值为 x_1，x_2，…，x_n，算术平均值为 $\bar{x}$。每个测定值与算术平均值的差为 $x_1-\bar{x}$，$x_2-\bar{x}$，…，$x_n-\bar{x}$，称为偏差。显然，由于偏差有正有负，这些偏差的总和为零，即：

$$\sum_{i=1}^{n}(x_i-\bar{x})=0$$

样本平均偏差的定义是样本偏差绝对值的算术平均值，平均偏差用字母 D 表示，即有：

$$D=\frac{\sum_{i=1}^{n}|x_i-\bar{x}|}{n} \tag{1-27}$$

平均偏差可以直观反映测定值离散程度的特征数，但由于包含了绝对值运算，平均偏差在统计中的应用受到很大的限制。

有兴趣的读者可以发现，如果取中位数为原点，即计算测定值与中位数的差，平均偏差有极小值。但是一般都选择算术平均值为原点计算。

［例 1-10］中植物样品中有机砷含量的样本平均偏差为：

$$\begin{aligned}D=&\frac{1}{7}(|17.21-17.20|+|17.20-17.20|+|17.23-17.20|+|17.19-\\&17.20|+|17.18-17.20|+|17.20-17.20|+|17.19-17.20|)\\=&0.011\end{aligned}$$

8. 方差和标准差

设样本的 n 个测定值为 x_1，x_2，…，x_n，其平均值为 $\bar{x}$。方差的定义是每

个测定值与算术平均值差的平方和除以 $n-1$，通常用 S^2 表示。根据方差的定义，S^2 的计算式为：

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \tag{1-28}$$

标准差为方差的平方根，通常用 S 表示。标准差的计算式为：

$$S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{1-29}$$

方差和标准差避免了平均偏差中绝对值运算不便的困难，是表征样本离散程度最常用的特征数。在数据统计分析中有广泛的应用。为了计算方便，方差计算式通常可以化简为式（1-30）：

$$S^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n}x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n}x_i\right)^2\right] \tag{1-30}$$

读者可证明，式（1-30）与式（1-28）完全等价。

应当注意，方差和标准差亦有以下的定义：

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

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

即用偏差的平方和除以 n 而不是 $n-1$。这两种定义都是可行的，重要的是要明确采用了哪一种定义。

［例 1-10］中植物样品的有机砷含量的样本方差为：

$$\begin{aligned}S^2 =& \frac{1}{7-1}[(17.21-17.20)^2+(17.20-17.20)^2+\\&(17.23-17.20)^2+(17.19-17.20)^2+\\&(17.18-17.20)^2+(17.20-17.20)^2+(17.19-17.20)^2]\\=& 2.7\times10^{-4}\end{aligned}$$

标准差为：

$$S=\sqrt{2.7\times10^{-4}}=0.016$$

9. 变异系数

标准差虽然反映了样本的离散程度，但对两个算术平均值不同的样本，仅用标准差就不能比较它们的离散程度。这样就需引入变异系数，通常用 C. V. 表示。它是样本的标准差相对于样本平均值的百分比。若样本的平均值和标准偏差分别为 $\bar{x}$ 和 S，则样本的变异系数的计算式为：

$$C.V.=\frac{S}{\bar{x}}\times 100\% \tag{1-31}$$

变异系数是一个无量纲的数值，它表征的数据相对离散程度与数据的绝对单位无关，因而在比较单位不同的样本之间离散程度差别时，变异系数有广泛的应用。

【例 1-12】 用方法一检测一植物样品的有机砷含量为 17.20mg/L，标准差为 0.016mg/L。用方法二检测另一种植物样品的有机砷含量为 3.21mg/L，标准差为 0.007mg/L。显然，方法一的标准差比方法二的标准差大，但这不能说明方法一的误差大。比较变异系数，有：

$$C.V.(\text{方法一})=\frac{0.016}{17.20}\times 100\%=0.093\%$$

$$C.V.(\text{方法二})=\frac{0.007}{3.21}\times 100\%=0.22\%$$

显然，方法二的变异系数较大，亦即方法二的检测精密度要较方法一的差。

10. 偏倚系数和峰凸系数

在环境数据统计分析的实践中，我们会发现两样本测定值有相同的均值和标准偏差，但其分布曲线的形状可以有相当大的差别。为进一步描述样本的分布特征，引入偏倚系数和峰凸系数两个特征数。

(1) 偏倚系数　在计算了样本的均值和标准偏差后，我们有了样本中心趋势和分散程度的数量概念，但还不能确定样本分布形状是对称的还是偏倚的。偏倚系数的定义是每个测定值与平均值差的三次方的均值除以样本标准差的三次方，通常记作 S.C.。偏倚系数反映了样本分布形状的偏倚方向和程度。设样本的 n 个测定值为 x_1，x_2，…，x_n，均值和标准差分别为 $\bar{x}$ 和 S，则由定义，偏倚系数的计算式为：

$$S.C.=\frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^3}{S^3} \tag{1-32}$$

若计算得到的 S.C. 为正值，样本分布曲线偏向均值的右边，称为正偏分布；反之分布曲线偏向均值的左边，称为负偏分布。

(2) 峰凸系数　峰凸系数表征数据分布曲线的陡峭程度。峰凸系数的定义是每个测定值与均值差的四次方的均值除以样本标准差的四次方，通常记作 K.C.。设样本的 n 个测定值为 x_1，x_2，…，x_n，均值和标准差分别为 $\bar{x}$ 和 S，则由定义，峰凸系数的计算式为：

$$K.C.=\frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^4}{S^4} \tag{1-33}$$

分布曲线的陡峭程度通常是与正态分布曲线的陡峭程度比较而言的。服从正态分布的样本峰凸系数为3。若某一样本测定值，其计算所得的峰凸系数值大于3，则其分布曲线陡峭程度比正态分布曲线大；反之，则其分布曲线陡峭程度比正态分布曲线的小。

五、抽样方法

前面我们介绍了从总体中抽取一部分个体称为总体的一个样本。样本是总体的一个子集。由于我们研究一个现象的总体性质是通过分析样本的数字特征得到的，因此样本获取方法无疑是十分重要的。

根据使用的抽样方法，抽样可分为概率抽样和非概率抽样。概率抽样是可以计算取得每个可能样本的概率的抽样方法。非概率抽样是不知道取得的每个可能样本的概率的抽样方法。如果我们要对通过样本做出估计的精度做出说明，必须用概率抽样方法。非概率抽样的优点是成本低、容易完成，其缺点是不能对估计精度做出正确的说明。此处主要介绍常用的概率抽样方法：简单随机抽样、分层简单随机抽样、整群抽样和系统抽样。

1. 简单随机抽样

简单随机抽样的定义为从一个容量为 N 的有限总体中抽取得到一个容量为 n 的简单随机样本，使每一个容量为 n 的可能样本，都有相同的概率被抽中。

用简单随机抽样方法进行抽样，首先建立一个总体中所有个体的名册，然后根据随机数表进行抽样。使用随机数表，可以保证抽样总体中的每个个体都有相同的概率被抽中。

当从一个容量为 N 的有限总体中，抽取一个容量为 n 的简单随机样本时，其均值及其标准差的估计值为：

$$\overline{x}=\frac{\sum_{i=1}^{n}x_i}{n} \tag{1-34}$$

$$S_{\overline{x}}=\sqrt{\frac{N-n}{N}}\left(\frac{S}{\sqrt{n}}\right) \tag{1-35}$$

若用 X 表示总体总量估计值，则可用式（1-36）表示：

$$X=N\overline{x} \tag{1-36}$$

在抽样调查中，样本容量的选择是一个重要问题。样本容量的选择需要对经费和精度进行权衡。较大的样本可以提供较高的精度，但费用较多。在经费允许的条件下，样本容量应该是足够大，以满足所要求的精度水平。通常，选择样本容量的方法是首先规定所需要的精度，然后确定满足精度的最小样本容量。在我

们估计总体均值时，如果要求允许误差 B 为均值标准差的 2 倍，由标准差公式（1-35）则有：

$$B=2\sqrt{\frac{N-n}{N}}\left(\frac{S}{\sqrt{n}}\right) \tag{1-37}$$

解式（1-37），可得到样本 n 的估计值

$$n=\frac{NS^2}{N\left(\frac{B^2}{4}\right)+S^2} \tag{1-38}$$

由式（1-38）可知，一旦给出了所需要的精度水平，便可以得到满足所需要精度水平的样本值 n。但是，对一个实际研究问题而言，除了规定所需要的允许误差 B 外，还必须知道样本的标准差 S 或方差 S^2。而样本方差 S^2 只有在得到实际样本时才可以算出。为了解决此问题，可以用两步抽样的方法来估计方差 S^2。第一步，抽取部分样本 n_1，按式（1-36）计算，可得到方差 S^2 的估计值，再将此值代入式（1-38），计算出所要求的样本容量 n。若 $n_1>n$，则可认为样本抽取已满足要求；若 $n_1<n$，则可再补抽样本，以满足允许误差要求。

2. 分层简单随机抽样

在分层简单随机抽样中，首先将总体划分为 H 个层，然后从第 h 层中抽取一个容量为 n_h 的简单随机样本。由这 H 个简单随机样本，可得出总体、均值、总体总量、均值的标准差等各种总体参数的估计值。一般说来，各层内的差异比层间的差异小，则分层简单随机抽样可得到更大的精度。层的划分，可根据所研究对象内在的性质差异、类别以及事先对研究对象的初步研究或以往的经验进行。

在分层抽样中，总体均值的估计值是各层样本平均值的加权平均值，所用权重为总体在各层的比重，其计算式如下：

$$\overline{X}=\sum_{h=1}^{H}\left(\frac{N_h}{N}\right)\overline{x_h} \tag{1-39}$$

式中，$\overline{x_h}$ 为第 h 层中样本的均值。

对分层简单随机样本，均值的标准差的计算式为：

$$S=\sqrt{\frac{1}{N^2}\sum_{h=1}^{H}N_h(N_h-n_h)\frac{S_h^2}{n_h}} \tag{1-40}$$

总体总量估计值的计算式为：

$$X=N\overline{x} \tag{1-41}$$

对分层简单随机抽样，我们可以用两个阶段过程来选择样本容量。第一步确定总样本容量 n，第二步确定各层应分配的样本容量 n_h。也可以第一步确定每层

样本容量 n_h，第二步通过各层样本容量相加得到总样本容量 n。确定总样本容量 n 及其分配，可对所要研究的总体参数提供必要的精度。然而，有时对某些层，样本容量没有达到满足层内估计量所需要的精度的数量，则这些层内的样本容量需要向上调整。一般而言，层内样本容量和方差较大的层应分配较多的样本数。而对于费用给定的前提下，为了获得更多的信息，抽样成本较大的层应分配较少的样本数。我们在进行各层样本数分配时一般要考虑三个重要因素：各层的样本容量、各层内的样本方差、各层抽取样本的费用。

在许多抽样调查中，抽样成本在各层近似相等。这时我们可以忽略抽样成本。对满足给定精度并使抽样成本达到最低要求，可采用著名的 Neyman 分配法，其将样本总容量 n 分配到各层的计算式如下：

$$n_h = n\left(\frac{N_h S_h}{\sum_{h=1}^{H} N_h S_h}\right) \tag{1-42}$$

式（1-42）表明，分配到各层的样本数受各层容量和标准差的影响，而且在进行分配前，必须先确定样本总容量 n。对于给定的允许误差 B，我们可使用式（1-43）确定样本总容量：

$$n = \frac{\sum_{h=1}^{H} N_h S_h^2}{N^2\left(\frac{B^2}{4}\right) + \sum_{h=1}^{H} N_h S_h^2} \tag{1-43}$$

3. 整群抽样

整群抽样需要将总体、各个体分为 N 组（也称作群），使总体中每个个体只属于某一群。

整群抽样和分层抽样都将总体划分为组，因此这两种抽样过程感觉上是相似的。但是，选择整群抽样与分层抽样的原因是不同的。当群内个体存在差异时，整群抽样可提供较好的结果。理想的情形是每一群是整个总体的一个缩影，在这种情形下，抽取很少的群就可以提供关于整个总体特征的信息。

为介绍整群抽样中总平均值、标准差和总体总量的计算公式，我们使用如下符号定义：

N—总体的群数

n—样本中选出的群数

M_i—第 i 个群的样本数

M—总体样本数，即 $M=M_1+M_2+\cdots+M_N$

$\overline{M}$—每一群的平均样本数，即 $\overline{M}=\frac{M}{N}$

x_i—第 i 群所有特征值（或称观察值）总量

$\overline{x_c}$—总体均值估计的计算值

则有，总体均值估计的计算公式：

$$\overline{x_c}=\frac{\sum_{i=1}^{n}x_i}{\sum_{i=1}^{n}M_i} \tag{1-44}$$

总体标准差估计的计算公式：

$$S_{\overline{x_c}}=\sqrt{\left(\frac{N-n}{Nn\overline{M}^2}\right)\frac{\sum_{i=1}^{n}(x_i-\overline{x_c}M_i)^2}{n-1}} \tag{1-45}$$

及总体总量的计算公式：

$$X=M\overline{x_c} \tag{1-46}$$

4. 系统抽样

对某些抽样情况，特别是总体容量很大的研究对象，可以用系统抽样来代替简单的随机抽样。例如需要从容量 10000 的总体中抽取一个容量为 50 的样本。我们可以从总体中每 200（10000/50）个个体中抽取一个个体。这种情况的系统样本，是从第一组 200 个个体中随机抽取一个个体。根据选中的第一个个体位置，隔 200 个位置，在第二组 200 个个体中抽取第二个个体。以此类推，我们可得到从总体容量为 10000 的研究对象中，抽出容量为 50 的系统抽样样本。

第二章　统计检验

第一节　统计检验的基本概念

在分析整理环境数据时，我们经常会遇到以下的几类问题。

① 环境科研和监测工作中所获得的一组数据，个别值与平均值偏差较大，这些值是否合理，是否应该舍弃。

② 一组环境数据有何规律，服从什么样的概率分布。

③ 在两种不同的条件下，例如某地区供热锅炉改造前和改造后，获得两组监测数据。需要做出判断，这两组数据的差异是由于随机因素引起的偶然误差还是存在明显的差异，即锅炉改造对环境质量是否产生了明显的影响。

④ 根据某地区的环境监测数据，回答该地区环境质量是否满足环境质量标准。

要解决这些问题，就需要用统计检验的方法，对科研和监测所得到的数据进行统计分析，在保证犯错误的可能性小于某个概率（例如5%或1%）的条件下，对需要做出判断的问题做出明确的回答。

统计检验又称作假设检验，或统计假设检验。统计检验的基本思想是对需要做出判断的问题先做出假设的结论，然后利用实际数据，按一定的统计计算方法计算，根据计算结果检验所做的假设是否合理，最后决定接受或否定假设的结论。

统计检验的基本步骤如下。

(1) 建立统计假设　针对需要做出判断的问题，假设其具有或不具有某种性质。通常用 H_0 表示待检验的基本假设，称作原假设；用 H_1 表示对立假设或备选假设，即当 H_0 被否定时将接受的假设。

基本假设和备选假设的设立有两种类型。一类是我们只检验待判断问题的某种性质是否与某一特定性质相同，如总体的均值 μ 是否等于某一特定值 μ_0，而不关心二者的大小，则假设 H_0 为 $\mu=\mu_0$，H_1 为 $\mu\neq\mu_0$。它意味着无论 μ 大于或小于 μ_0 都将否定假设 H_0，因而称作双侧检验。另一类是检验 μ 值大于或者小于某一特定数值 μ_0，则假设 H_0 为 $\mu\geqslant\mu_0$，H_1 为 $\mu<\mu_0$（或 H_0 为 $\mu\leqslant\mu_0$，H_1 为

$\mu>\mu_0$）。它意味着只有当 μ 小于（或大于）μ_0 时将否定假设 H_0，因而称作单侧检验。

（2）选择判别假设成立与否的显著性水平　由于统计检验对所判别问题的判断是在一定概率保证下做出的，因而必须给出概率保证的水平，即显著性水平，通常用 α 表示。进行统计检验时，显著性水平常选用 $\alpha=0.01$ 或 $\alpha=0.05$ 两种。

（3）选择合适的统计计算方法进行计算　进行统计检验时，由于所需判断的问题的性质各异，需要选择合适的统计计算公式进行计算。

（4）对统计假设做出判断　通常统计学家已计算出各种判别统计结果的临界值表，根据统计计算结果与相应显著性水平的临界值比较即可判断统计假设成立与否。

需要指出的是，我们做出的判断不是绝对的，而是在一定概率保证条件下的判断。当我们选择显著性水平 $\alpha=0.01$ 时，意味着我们对统计假设做出的判断犯错误的可能性小于 0.01，而 0.01 可能犯错误的概率在统计学中认为是极小的概率，因而可以认为我们做出的判断是极显著成立的。当选择显著性水平为 $\alpha=0.05$ 时，由于统计学中认为 0.05 可能犯错误的可能性是小概率，因此可以认为所做出的判断是显著成立的。根据不同的具体问题可以选择不同的显著性水平 α，但必须保证 α 的取值是小概率，否则不能做出判断。

由上面介绍的统计判断步骤可知，在做出判断时是有可能犯两类性质的错误的。一类是事情本身成立，即假设 H_0 是真实的，而我们却判断其不成立，否定了假设 H_0，称作“去真”错误或第一类错误。另一类是事情本身不成立，即假设 H_0 是错误的，而我们判断其成立，接受了假设 H_0，称作“存伪”错误或第二类错误。为了防止犯这两类错误，可以将显著性水平 α 选得很小，例如 $\alpha=0.001$，但即便如此，也仍然存在犯上述两类错误的可能，只是犯错误的可能性更小一些。

第二节　离群值的检验

在分析一组环境数据时，常常会发现某些离群值。顾名思义，离群值与其他数值在数量上有较显著的差异。离群值又称异常值、可疑值或极端值。离群值的产生可能是由于各种随机因素的影响使其偏离群体，也有可能是由于过失误差，如分析测试过程中的失误或数据传递过程中的失误等原因造成的。因此需要进行统计判断，确定其离群的性质，然后决定该值的舍取。如果保留了由于过失误差而造成的离群值，会得到偏离实际情况的错误结果；如果剔除了由于随机因素引起的离群值，同样会得到虚假“较高精度”的错误结果。因此，离群值的检验是

环境数据统计分析中一个重要的判别问题。

离群值的统计检验包括单个离群值的检验、多个离群值的检验以及多组数据平均值或方差的检验等。统计检验的基本思想在于，给定一个显著性水平 α（如 $\alpha=0.01$），并确定一个临界值，凡是超过这个临界值的误差就认为是过失误差，而不是随机误差，予以剔除。下面将分别介绍离群值的几种主要的检验方法。

一、单个离群值的检验

单个离群值的检验是指在一组测试结果中检验决定一个离群值的取舍。通常采用格拉布斯法（Grubbs）进行检验。

将 n 个观察值 x_1，x_2，…，x_n，按数值由小到大顺序排列，$x(1)\leqslant x(2)\leqslant\cdots\leqslant x(n)$。用于判断最大离群值的格拉布斯检验统计量 G_n 计算公式为：

$$G_n=\frac{x(n)-\bar{x}}{S} \tag{2-1}$$

用于判断最小离群值的格拉布斯统计量 G_1 的计算公式为：

$$G_1=\frac{\bar{x}-x(1)}{S} \tag{2-2}$$

式（2-1）和式（2-2）中，$\bar{x}$ 为 n 个观察值的平均值，S 为 n 个观察值的标准差。计算所得到的 G_n 或 G_1 与表 2-1 中所列的格拉布斯检验临界值比较。当 G_n 或 G_1 大于表 2-1 中所列的临界值，则认为该离群值是异常的（显著性水平 $\alpha=0.05$）或者高度异常的（显著性水平 $\alpha=0.01$），予以剔除，否则该离群值属正常范围之内，应予保留。

格拉布斯检验用于单个离群值的检验。对于一组测量结果而言，一般只进行一次格拉布斯检验，而不进行连续剔除检验。

格拉布斯检验临界值表是这样使用的：根据已知的样本数 n 以及所选择的显著性水平 α（0.01 或 0.05），在表中查找相应的值。

例如：

$$G(0.01,10)=2.410$$

$$G(0.05,30)=2.745$$

$$G(0.01,100)=3.600$$

值得注意的是：在计算标准差 S 时，可疑值也要计算在内。标准差取定义

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

表 2-1 格拉布斯检验临界值 $G(\alpha, n)$ 表

n	显著性水平 α				n	显著性水平 α			
	0.05	0.025	0.01	0.005		0.05	0.025	0.01	0.005
3	1.153	1.155	1.155	1.155	31	2.759	2.924	3.119	3.253
4	1.463	1.481	1.492	1.496	32	2.773	2.938	3.135	3.270
5	1.672	1.715	1.749	1.764	33	2.786	2.952	3.150	3.286
6	1.822	1.887	1.944	1.973	34	2.799	2.965	3.164	3.301
7	1.938	2.020	2.097	2.139	35	2.811	2.979	3.178	3.316
8	2.032	2.126	2.221	2.274	36	2.823	2.991	3.191	3.330
9	2.110	2.215	2.323	2.387	37	2.835	3.003	3.204	3.343
10	2.176	2.290	2.410	2.482	38	2.846	3.014	3.216	3.356
11	2.234	2.355	2.485	2.564	39	2.857	3.025	3.228	3.369
12	2.285	2.412	2.550	2.636	40	2.866	3.036	3.240	3.381
13	2.331	2.462	2.607	2.699	41	2.877	3.046	3.251	3.393
14	2.371	2.507	2.659	2.755	42	2.887	3.057	3.261	3.404
15	2.409	2.549	2.705	2.806	43	2.896	3.067	3.271	3.415
16	2.443	2.585	2.747	2.852	44	2.905	3.075	3.282	3.425
17	2.476	2.620	2.785	2.894	45	2.914	3.085	3.292	3.435
18	2.504	2.651	2.821	2.932	46	2.923	3.094	3.302	3.445
19	2.532	2.681	2.854	2.968	47	2.931	3.103	3.310	3.455
20	2.557	2.709	2.884	3.001	48	2.940	3.111	3.319	3.464
21	2.580	2.733	2.912	3.031	49	2.948	3.120	3.329	3.474
22	2.603	2.758	2.939	3.060	50	2.956	3.128	3.336	3.483
23	2.624	2.781	2.963	3.087	60	3.025	3.199	3.411	3.560
24	2.644	2.802	2.987	3.112	70	3.082	3.257	3.471	3.622
25	2.663	2.822	3.009	3.135	80	3.130	3.305	3.521	3.673
26	2.681	2.841	3.029	3.157	90	3.171	3.347	3.563	3.716
27	2.698	2.859	3.049	3.178	100	3.207	3.383	3.600	3.754
28	2.714	2.876	3.068	3.199					
29	2.730	2.893	3.085	3.218					
30	2.745	2.908	3.103	3.236					

【例 2-1】 在一固定的运转状态下测量一鼓风机的声功率，总共进行了 9 次独立的测量，测量结果见表 2-2，问第 4 次测量值是否应当剔除？

表 2-2 ［例 2-1］中鼓风机声功率的测量结果

测量次数	1	2	3	4	5	6	7	8	9
声功率级	92.5	93.1	91.9	96.8	92.8	92.3	93.0	93.9	92.6

解 第 4 次测量值也是最大值，为 96.8。测定结果的均值为：

$$\bar{x}=\frac{1}{9}(92.5+93.1+91.9+96.8+92.8+92.3+93.0+93.9+92.6)$$

$$=93.2$$

标准差为：

$$S=1.46$$

计算统计量：$G_n=(96.8-93.2)/1.46=2.466$

查格拉布斯临界值表 2-1 得：$G(0.01, 9)=2.323$

显然有 $G_n>G(0.01, 9)$，因而可以判断第 4 次测量值高度异常，属于过失误差，应予以剔除。

二、多个离群值的检验

多个离群值的检验是指在一组分析测试所得结果中检验决定几个离群值的取舍。多个离群值的检验通常采用狄克逊法（Dixon）进行检验。

狄克逊法是应用了极差比的方法，得到简化而严密的结果。为使判断的准确率高，不同的样本数应采用不同的极差比计算。n 次测量的观察值 x_1，x_2，…，x_n 按大小顺序排列为：

$$x(1)\leqslant x(2)\leqslant\cdots\leqslant x(n-1)\leqslant x(n)$$

用于判别最小值 $x(1)$ 或最大值 $x(n)$ 是否为离群值的统计计算公式列于表 2-3 中。

表 2-3　狄克逊检验统计量 D 计算公式

样本范围	可疑值为最小值	可疑值为最大值	样本范围	可疑值为最小值	可疑值为最大值
3～7	$\frac{x(2)-x(1)}{x(n)-x(1)}$	$\frac{x(n)-x(n-1)}{x(n)-x(1)}$	11～13	$\frac{x(3)-x(1)}{x(n-1)-x(1)}$	$\frac{x(n)-x(n-2)}{x(n)-x(2)}$
8～10	$\frac{x(2)-x(1)}{x(n-1)-x(1)}$	$\frac{x(n)-x(n-1)}{x(n)-x(2)}$	14～30	$\frac{x(3)-x(1)}{x(n-2)-x(1)}$	$\frac{x(n)-x(n-2)}{x(n)-x(3)}$

狄克逊检验统计量计算结果与表 2-4 中所列的狄克逊检验临界值比较，当计算值大于表中临界值时，即 $D_{\min}$（或 $D_{\max}$）$>D(\alpha, n)$ 时，则认为该离群值是显著异常的（显著性水平 $\alpha=0.05$）或者是高度显著异常的（显著性水平 $\alpha=0.01$），可考虑予以剔除。

表 2-4　狄克逊检验临界值 $D(\alpha, n)$ 表

n \ α	0.01	0.05	n \ α	0.01	0.05	n \ α	0.01	0.05
			11	0.709	0.619	21	0.555	0.478
			12	0.660	0.583	22	0.544	0.468
3	0.994	0.970	13	0.638	0.557	23	0.535	0.459
4	0.926	0.829	14	0.670	0.586	24	0.526	0.451
5	0.821	0.710	15	0.647	0.565	25	0.517	0.443
6	0.740	0.628	16	0.627	0.546	26	0.510	0.436
7	0.680	0.569	17	0.610	0.529	27	0.502	0.429
8	0.717	0.608	18	0.594	0.514	28	0.495	0.423
9	0.672	0.564	19	0.580	0.501	29	0.489	0.417
10	0.635	0.530	20	0.567	0.489	30	0.483	0.412

狄克逊法用于多个离群值的检验意味着对一组环境数据可以连续使用该法逐个判别和剔除离群值，直至不能检验出离群值时为止。

【例 2-2】 试用狄克逊法检验［例 2-1］中声功率测量的离群值。

解 声功率的 9 次测量结果按大小顺序排列为：

$x(1)=91.9$，$x(2)=92.3$，$x(3)=92.5$，$x(4)=92.6$，$x(5)=92.8$，

$x(6)=93.0$，$x(7)=93.1$，$x(8)=93.9$，$x(9)=96.8$

此处，可疑值为最大值，样本数为 9，依照狄克逊法检验公式，最大值离群检验为：

$$D_{\max}=\frac{x(n)-x(n-1)}{x(n)-x(2)}$$

因而有

$$D_{\max}=(96.8-93.9)/(96.8-92.3)$$
$$=0.644$$

查狄克逊检验临界值 $D(\alpha, n)$ 表得

$$D(0.01,9)=0.672;\ D(0.05,9)=0.564$$

显然有 $D_{\max}>D(0.05, 9)$，因而可以判断可疑值为显著异常，应考虑剔除。

读者可用同样的方法继续检验 $x(1)$ 和 $x(8)$ 是否异常，可以发现 $x(1)$ 和 $x(8)$ 并非离群值，不能剔除。

三、多组测定数据平均值离群的检验

n 组测定数据得到 n 个平均值 $\bar{x}_1$，$\bar{x}_2$，…，$\bar{x}_n$，将其按大小顺序排列 $\bar{x}(1)\leqslant\bar{x}(2)\leqslant\cdots\leqslant\bar{x}(n)$。多组测定数据平均值可用格拉布斯检验法判断这组平均值中的一个离群值，其统计量计算公式为：

(1) 用于判别最大平均值是否离群：

$$G_n=\frac{\bar{x}(n)-\bar{\bar{x}}}{S} \tag{2-3}$$

(2) 用于判断最小平均值是否离群：

$$G_1=\frac{\bar{\bar{x}}-\bar{x}(1)}{S} \tag{2-4}$$

式 (2-3) 和式 (2-4) 中 $\bar{\bar{x}}$ 为 n 个平均值的平均值，S 为 n 个平均值的标准差。

将以上两式的计算结果与表 2-1 中的格拉布斯检验的临界值 $G(\alpha, n)$ 进行比较。若计算值大于表中临界值时，则认为该平均值离群是显著的（显著性水平 0.05）或高度显著的（显著性水平 0.01），可以剔除，否则就不能认为是离群值，应予保留。

一组平均值中多个离群值的判别则采用狄克逊法进行检验。狄克逊法统计量计算公式如表 2-3 所示，只需把公式中的观察值用一个平均值替代，即用 $\bar{x}(i)$ 替代 $x(i)$ 即可对一组平均值中多个离群值进行检验。判别的临界值如表 2-4 所示。同样将计算结果与表 2-4 的临界值进行比较，从而判断该值离群显著与否。

【例 2-3】 实验室质量控制的方法之一是让多个实验室测量分析同一种样品，12 个实验室对一土样含铅量的分析结果见表 2-5。

表 2-5　［例 2-3］土样含铅量分析结果

实验室编号	1	2	3	4	5	6	7	8	9	10	11	12
铅含量/(mg/L)	68.5	66.4	69.3	62.9	64.5	66.8	50.2	64.9	66.2	79.5	65.0	63.2

若该土样的含铅量并非已知，试从以上分析结果中判断哪些实验室的分析结果为异常。

解　已知每一实验室的分析结果都是多次分析的平均值。

将 12 个实验室的分析结果由小到大排列为：

$$\bar{x}(1)=50.2,\ \bar{x}(2)=62.9,\ \bar{x}(3)=63.2,$$
$$\bar{x}(4)=64.5,\ \bar{x}(5)=64.9,\ \bar{x}(6)=65.0,$$
$$\bar{x}(7)=66.2,\ \bar{x}(8)=66.4,\ \bar{x}(9)=66.8,$$
$$\bar{x}(10)=68.5,\ \bar{x}(11)=69.3,\ \bar{x}(12)=79.5$$

本问题中，$n=12$，有两个可疑值，$\bar{x}(1)$ 与 $\bar{x}(12)$。因而应采用狄克逊检验法。

(1) 首先检验最大值 $\bar{x}(n)$：

根据狄克逊计算公式表 2-3：

$$D_{\max}=\frac{x(12)-x(10)}{x(12)-x(2)}=(79.5-68.5)/(79.5-62.9)$$
$$=0.663$$

查狄克逊临界值表 2-4 得：$D(0.01,12)=0.660$，因而有 $D_{\max}>D(0.01,12)$，可以判断该值为高度显著异常，应予剔除。

(2) 再检验最小值 $\bar{x}(1)$，此时 $n=11$。

根据狄克逊公式表 2-3。

$$D_{\min}=\frac{x(3)-x(1)}{x(10)-x(1)}=(63.2-50.2)/(68.5-50.2)=0.710$$

查狄克逊临界值表 2-4 得：

$$D(0.01,11)=0.709$$

因而 $D_{\min}>D(0.01,11)$，可知 $\bar{x}(1)$ 为高度显著异常值，应予剔除。

有兴趣的读者还可继续检验剩下来的 10 个样本。结论是这 10 个平均值均不为离群值，不能剔除。

由检验可知，第 7 号实验室与第 10 号实验室的测量分析结果异常，应对分析方法、操作和记录等步骤作检查分析，找出产生系统误差的原因。

四、多组测定数据标准差离群的检验

多组测定数据标准差离群的检验通常用科克兰（Cochran）检验法进行。科

表 2-6 科克兰检验法的临界值表 $C(l, n)$

给定水平的组数 l	每组的测定结果数									
	$n=2$		$n=3$		$n=4$		$n=5$		$n=6$	
	0.01	0.05	0.01	0.05	0.01	0.05	0.01	0.05	0.01	0.05
2	—	—	0.995	0.975	0.979	0.939	0.959	0.906	0.937	0.877
3	0.993	0.967	0.942	0.871	0.883	0.798	0.834	0.746	0.793	0.707
4	0.968	0.906	0.864	0.768	0.781	0.684	0.721	0.629	0.676	0.590
5	0.928	0.841	0.788	0.684	0.696	0.598	0.633	0.544	0.588	0.506
6	0.883	0.781	0.722	0.616	0.626	0.532	0.564	0.480	0.520	0.445
7	0.838	0.727	0.664	0.561	0.568	0.480	0.508	0.431	0.466	0.397
8	0.794	0.680	0.615	0.516	0.521	0.438	0.463	0.391	0.423	0.360
9	0.754	0.638	0.573	0.478	0.481	0.403	0.425	0.358	0.387	0.329
10	0.718	0.602	0.536	0.445	0.447	0.373	0.393	0.331	0.357	0.303
11	0.684	0.571	0.504	0.417	0.418	0.348	0.366	0.308	0.332	0.281
12	0.653	0.541	0.475	0.392	0.392	0.326	0.343	0.288	0.310	0.262
13	0.624	0.515	0.450	0.371	0.369	0.307	0.322	0.271	0.291	0.243
14	0.599	0.492	0.427	0.352	0.349	0.291	0.304	0.255	0.274	0.232
15	0.575	0.471	0.407	0.335	0.332	0.276	0.288	0.242	0.259	0.220
16	0.553	0.452	0.388	0.319	0.316	0.262	0.274	0.230	0.246	0.208
17	0.532	0.434	0.372	0.305	0.301	0.250	0.261	0.219	0.234	0.198
18	0.514	0.418	0.356	0.293	0.288	0.240	0.249	0.209	0.223	0.187
19	0.496	0.403	0.343	0.281	0.276	0.230	0.238	0.200	0.214	0.181
20	0.480	0.389	0.330	0.270	0.265	0.220	0.229	0.192	0.205	0.174
21	0.465	0.377	0.318	0.261	0.255	0.212	0.220	0.185	0.197	0.167
22	0.450	0.365	0.307	0.252	0.246	0.204	0.212	0.178	0.189	0.160
23	0.437	0.354	0.297	0.243	0.238	0.197	0.204	0.172	0.182	0.155
24	0.425	0.343	0.287	0.235	0.230	0.191	0.197	0.166	0.176	0.149
25	0.413	0.334	0.278	0.228	0.222	0.185	0.189	0.160	0.170	0.144
26	0.402	0.325	0.270	0.221	0.215	0.179	0.184	0.155	0.164	0.140
27	0.391	0.316	0.262	0.215	0.209	0.173	0.179	0.150	0.159	0.135
28	0.382	0.308	0.255	0.209	0.202	0.168	0.173	0.146	0.154	0.131
29	0.372	0.300	0.248	0.203	0.196	0.164	0.168	0.142	0.150	0.127
30	0.363	0.293	0.241	0.198	0.191	0.159	0.164	0.138	0.145	0.124
31	0.355	0.286	0.235	0.193	0.186	0.155	0.159	0.134	0.141	0.120
32	0.347	0.280	0.229	0.188	0.181	0.151	0.155	0.131	0.138	0.117
33	0.339	0.273	0.224	0.184	0.177	0.147	0.151	0.127	0.134	0.114
34	0.332	0.267	0.218	0.179	0.172	0.144	0.147	0.124	0.131	0.111
35	0.325	0.262	0.213	0.175	0.168	0.140	0.144	0.121	0.127	0.108
36	0.318	0.256	0.208	0.172	0.165	0.137	0.140	0.118	0.124	0.106
37	0.312	0.251	0.204	0.168	0.161	0.134	0.137	0.116	0.121	0.103
38	0.306	0.246	0.200	0.164	0.157	0.131	0.134	0.113	0.119	0.101
39	0.300	0.242	0.196	0.161	0.154	0.129	0.131	0.111	0.116	0.099
40	0.294	0.237	0.192	0.158	0.151	0.126	0.128	0.108	0.114	0.097

克兰检验是一种方差均匀性检验方法，它用 l 组测定结果中的最大方差与 l 组测定结果的方差和之比值与临界值比较，判别该组测定结果的方差离群与否。

设有 l 组测定值，每组 n 次测定结果的标准为 S_1，S_2，…，S_l，将其按大小顺序排列 $S(1)\leqslant S(2)\leqslant\cdots\leqslant S(l)$，并记最大标准差 $S(l)$ 为 S_{max}。科克兰检验统计量 C 计算公式为：

$$C=\frac{S_{max}^2}{\sum_{i=1}^{l}S_i^2} \tag{2-5}$$

表 2-6 中为科克兰检验临界值表。将由式（2-5）计算所得结果与表 2-4 中的临界值进行比较，当计算值大于表中的临界值时，则认为该组测定结果标准差离群是显著的(显著性水平 $\alpha=0.05$) 或高度显著的(显著性水平 $\alpha=0.01$)。

用科克兰检验法对多组测定数据标准差离群与否的检验可连续进行。例如 l 组测定数据的标准差经过检验剔除一个最大的标准差，可继续对所剩的 $l-1$ 组测定数据的最大标准差检验。如此进行下去，直至不能检验出离群值为止。

科克兰检验要求 l 组测定值的每组测量次数 n 相同。在实际应用中，由于数据的多余、缺漏或剔除而使得每组的测量次数不尽相同，这时候的 n 应取绝大多数实验组的测量次数。

【例 2-4】 6 个实验室分析测试同一水样的酚的浓度，要求每一实验室测试 5 次后求平均，分析结果见表 2-7。

表 2-7 ［例 2-4］中水样酚浓度的测试分析结果/(mg/L)

次数 / 实验室号	1	2	3	4	5	均值	标准差
1	3.15	3.18	3.21	3.18	3.20	3.18	0.023
2	3.04	3.17	3.26	3.02	3.11	3.12	0.098
3	3.13	3.17	3.12	3.15	3.14	3.14	0.019
4	3.09	3.11	3.08	3.10	3.09	3.09	0.011
5	3.20	3.21	3.20	3.23	3.19	3.21	0.015
6	3.16	3.15	3.18	3.15	3.11	3.15	0.025

试用科克兰法检验离群值。

解　已知 $S_{max}=0.098$

由式（2-5），科克兰检验的统计量为：

$C=0.098^2/(0.023^2+0.098^2+0.019^2+0.011^2+0.015^2+0.025^2)=0.838$

查科克兰检验临界值表得

$$C(6,5,0.01)=0.564$$

显然有 $C>C(6,5,0.01)$

因而 2 号实验室测量结果的标准差高度显著异常，应予以剔除。

读者可以证明，剩下的 5 个实验室测量结果的标准差不异常，检验可以结束。

无论采用哪一种检验方法，离群值应是个别的或极少量的，否则应从数据产生的每一个环节中找原因。

五、离群值的剔除

经检验后确定了某些数据为离群值，这为剔除这些值提供了统计学上的依据。但是在实际剔除这些离群值前，还需要做认真分析，查找离群值产生的原因。

离群值产生的原因是十分复杂的，有些经检验后确定的离群值可能是反映环境质量真实情况的数值，因此不能单纯地依赖检验结果来剔除数据，而需要从技术上检查离群值出现的原因。例如是否由于测试中的疏忽、测试步骤不正确、试剂使用不当、样品运输贮存不当、实验数据记录错误、数据处理不正确等。当能够找到离群值产生的原因时，剔除离群值就比较有把握。因此，根据离群值检验的结果并结合实际技术原因分析来决定离群值的取舍，是较为科学和客观的。

第三节 μ 检验法

在某地区采样点采集了一系列样品，经分析测试获得环境分析测试数据，对数据进行分析时，人们常需要回答这样一个问题：根据环境分析测试数据，该采样点所代表区域的环境质量是否超过环境标准？这就是说要根据样本的测定值推断总体的平均值是小于、等于还是大于某一确定的值。同样，在环境数据分析时，也经常遇到根据甲、乙两个采样点得到的环境数据，判别甲、乙两地环境污染水平是相当或存在显著差异的问题，这是根据两样本的测定值判别两总体水平相同或显著不同的问题。

解决总体均值的判别问题常用的统计检验方法有 μ 检验法和 t 检验法。μ 检验法用于总体标准差已知的情况，t 检验法用于总体标准差未知的情况。一般而言，总体标准差已知的情况是不多的，因而 μ 检验法作为总体平均值的检验不如 t 检验法应用广泛。然而，对大样本（$n>30$）而言，以样本的

标准差 S 代替总体标准差 σ 误差不大，因而 μ 检验法仍是一种可以利用的方法。对于小样本($n<30$)，用 μ 检验法对总体均值进行统计检验时误差较大，一般不采用。遇到这种情况，可以利用 t 检验法进行统计检验。t 检验法将在下一节中介绍。

需要指出的是，用 μ 检验法或 t 检验法对总体的平均值进行统计检验时，总体必须是服从正态分布的。对于大样本，不管总体服从什么分布，根据概率论中的中心极限定理，可以认为样本均值 $\bar{x}$ 渐近服从正态分布。对于小样本，在可能的条件下判别其服从的分布类型。若为正态分布，可用 t 检验法对总体的平均值进行统计检验。

一、由一个样本检验总体的平均值

在某一采样点获得一组监测数据，通过统计检验确定该测点所代表的区域环境污染水平是否超过环境标准（或某一确定值），这就是由一个样本检验总体平均值的问题。区域的环境污染水平是我们要研究的总体，总体的平均值通过样本测定值进行推断。

用 μ 表示总体平均值，μ_0 表示某一确定的值（例如环境质量标准），$\bar{x}$ 表示样本的平均值，σ 表示总体的标准差（在大样本时，可用样本的标准差 S 代替），n 表示样本数。一个样本检验总体平均值的 μ 检验法的具体步骤可按表 2-8 进行。

表 2-8　μ 检验法统计检验表

统计检验步骤	双侧检验	单侧检验
建立统计假设	$H_0:\mu=\mu_0$ $H_1:\mu\neq\mu_0$	$H_0:\mu\geqslant\mu_0$　$H_1:\mu<\mu_0$ 或　$H_0:\mu\leqslant\mu_0$　$H_1:\mu>\mu_0$
选择显著性水平 α	0.05 或 0.01	0.05 或 0.01
计算统计量 μ	$\mu=\dfrac{\bar{x}-\mu_0}{\dfrac{\sigma}{\sqrt{n}}}$	$\mu=\dfrac{\bar{x}-\mu_0}{\dfrac{\sigma}{\sqrt{n}}}$
确定临界值 μ_α	查附表 1 得临界值 $\mu_{\frac{\alpha}{2}}$	查附表 1 得临界值 μ_α
统计判别	若 $\lvert\mu\rvert\leqslant\mu_{\frac{\alpha}{2}}$，接受 H_0	若 $\mu\geqslant-\mu_\alpha$ 接受 H_0 或若 $\mu\leqslant\mu_\alpha$ 接受 H_0
	若 $\lvert\mu\rvert>\mu_{\frac{\alpha}{2}}$，否定 H_0， 接受备选假设 H_1	若 $\mu<-\mu_\alpha$ 否定 H_0 接受 H_1； 或若 $\mu>\mu_\alpha$ 否定 H_0 接受 H_1

表 2-8 中 μ_α 和 $\mu_{\frac{\alpha}{2}}$ 的定义是这样的：

双侧检验：$P(\lvert\mu\rvert\leqslant\mu_{\frac{\alpha}{2}})=1-\alpha$

单侧检验：$P(\mu \leqslant \mu_\alpha)=1-\alpha$

【例 2-5】 检验一污水处理厂的出水的氯化物浓度，共采了 20 个水样。设水样的氯化物浓度遵从正态分布。问：(1) 若标准差为 30mg/L，水样的氯化物平均浓度为 253mg/L，出水的氯化物浓度是否达到设计的 250mg/L 标准。(2) 若标准差为 20mg/L，平均浓度为 261mg/L，出水的氯化物浓度是否超过 250mg/L 的标准。

解 (1) 这是一个双侧检验的问题。

已知：$\sigma=30\text{mg/L}$，$\bar{x}=253\text{mg/L}$，$\mu_0=250\text{mg/L}$，$n=20$，H_0：$\mu=\mu_0$

$$\mu=\frac{253-250}{\frac{30}{\sqrt{20}}}=0.447$$

选取显著性水平 $\alpha=0.05$，查附表 1 可得

$$\mu_{\frac{\alpha}{2}}=1.960$$

显然有：$|\mu|=0.447<\mu_{\frac{\alpha}{2}}=1.960$

即假设 H_0 不能推翻，出水的氯化物浓度达到设计标准。

(2) 这是单侧检验问题。

已知：$\sigma=20\text{mg/L}$，$\bar{x}=261\text{mg/L}$，$\mu_0=250\text{mg/L}$，$n=20$，假设 H_0：$\mu \leqslant \mu_0$

$$\mu=\frac{261-250}{\frac{20}{\sqrt{20}}}=2.460$$

选取显著性水平 $\alpha=0.05$，查附表 1 可得

$$\mu_\alpha=1.645$$

显然有：$\mu=2.460>\mu_\alpha=1.645$

即否定了原假设 H_0，因而有 $\mu>\mu_0$，即出水的氯化物浓度超过了 250mg/L 的标准。

二、由两个样本检验两总体平均值的一致性

在两个采样点得到两组监测数据，通过统计检验确定两采样点所代表的两个区域环境污染水平是否相等，这是由两个样本检验两总体平均值一致性的问题。用 μ_1、μ_2 表示两总体的平均值，$\bar{x}_1$、$\bar{x}_2$ 表示两样本的平均值，σ_1、σ_2 表示两总体的标准差（在 $n>30$ 的大样本时可用样本的标准差 S_1、S_2 代替），n_1、n_2 表示两测点的样本数。

用 μ 检验法对两总体平均值的一致性的检验可按表 2-9 μ 检验法统计表的步骤进行。表中的 μ 是一个遵从标准正态分布 $N(0,1)$ 的随机变量（若总体不是正态的，则当 n_1、n_2 相当大时，可满足条件）。μ_α 的定义与前同。

表 2-9　μ 检验法统计检验表

统计检验步骤	双侧检验	单侧检验
建立统计假设	$H_0:\mu_1=\mu_2$ $H_1:\mu_1\neq\mu_2$	$H_0:\mu_1\geqslant\mu_2$　$H_1:\mu_1<\mu_2$ 或　$H_0:\mu_1\leqslant\mu_2$　$H_1:\mu_1>\mu_2$
选择显著性水平 α	0.05 或 0.01	0.05 或 0.01
计算统计量 μ	$\mu=\dfrac{\overline{x}_1-\overline{x}_2}{\sqrt{\dfrac{\sigma_1^2}{n_1}+\dfrac{\sigma_2^2}{n_2}}}$	$\mu=\dfrac{\overline{x}_1-\overline{x}_2}{\sqrt{\dfrac{\sigma_1^2}{n_1}+\dfrac{\sigma_2^2}{n_2}}}$
确定临界值 μ_α	查附表 1 得 $\mu_{\frac{\alpha}{2}}$	查附表 1 得 μ_α
统计判别	若 $\lvert\mu\rvert\leqslant\mu_{\frac{\alpha}{2}}$，接受 H_0	若 $\lvert\mu\rvert\leqslant\mu_\alpha$，接受 H_0
	若 $\lvert\mu\rvert>\mu_{\frac{\alpha}{2}}$，否定 H_0，接受 H_1	若 $\lvert\mu\rvert>\mu_\alpha$，否定 H_0，接受 H_1

【例 2-6】 检测两片不同降水环境下树林的生长情况，已知林高符合正态分布。随机从两片树林中抽取容量为 15 的两个样本如下(单位：m)。

样本 1：9.6，9.1，10.2，10.5，8.8，10.8，9.5，8.1，10.8，8.4，10.1，8.5，8.9，11.2，8.3

样本 2：11.3，10.8，9.3，8.8，12.4，12.5，11.6，10.8，9.5，9.2，10.1，12.0，11.7，11.2，12.4

若两总体的标准差分别为 $\sigma_1=0.98$m，$\sigma_2=1.14$m，试以 0.05 的显著性水平检验不同的降水环境是否造成了这两片林区的生长差异。

解　已知 $\sigma_1=0.98$，$\sigma_2=1.14$

$$\overline{x}_1=9.52,\quad \overline{x}_2=10.91$$

采用双侧检验，假设 H_0：$\mu_1=\mu_2$

统计量为：

$$\mu=\frac{9.52-10.91}{\sqrt{\dfrac{0.98^2}{15}+\dfrac{1.14^2}{15}}}=\frac{-1.39}{0.388}=-3.58$$

$\alpha=0.05$，查附表 1 得：

$$\mu_{0.025}=1.96$$

显然有：$\lvert\mu\rvert=3.58>\mu_{0.025}=1.96$

即否定假设 H_0。表明由于降水环境不同使得两片林区的生长存在差异。

第四节　t 检验法

如上节所述，在进行总体平均值一致性检验时，如果样本数量较小($n<30$)，采用 μ 检验法用样本标准差代替总体标准差时误差较大，这时总体平均值

统计检验应采用 t 检验法。由于 t 检验法不要求总体的标准差已知，所以在总体平均值一致性的检验中，特别是在样本数量较少时应用很广。

一、由一个样本检验总体平均值

与 μ 检验法相同，由一个样本检验总体的平均值是指在一个采样点得到的一组监测数据，通过统计检验确定该测点代表的区域环境污染水平是否超过某一确定的值。与 μ 检验法不同的是 t 检验法无须已知总体的标准差。用 μ 表示总体的平均值，μ_0 表示某一确定的值（例如环境质量标准），$\bar{x}$ 表示样本的平均值，S 表示样本的标准差，n 为样本数。t 检验法对总体平均值的检验可按 t 检验法统计检验表进行，见表 2-10。

表 2-10 t 检验法统计检验表

统计检验步骤	双侧检验	单侧检验
建立统计假设	$H_0:\mu=\mu_0$ $H_1:\mu\neq\mu_0$	$H_0:\mu\geqslant\mu_0$　$H_1:\mu<\mu_0$ 或　$H_0:\mu\leqslant\mu_0$　$H_1:\mu>\mu_0$
选择显著性水平 α	0.05 或 0.01	0.05 或 0.01
计算统计量 t	$t=\dfrac{\bar{x}-\mu_0}{\dfrac{S}{\sqrt{n}}}$	$t=\dfrac{\bar{x}-\mu_0}{\dfrac{S}{\sqrt{n}}}$
计算自由度 df	$df=n-1$	$df=n-1$
确定临界值 t_α	查附表 3 得 t_α	查附表 3 得 $t_{2\alpha}$
统计判别	若 $\lvert t\rvert\leqslant t_\alpha$，接受 H_0	若 $\lvert t\rvert\leqslant t_{2\alpha}$，接受 H_0
	若 $\lvert t\rvert>t_\alpha$，否定 H_0，接受 H_1	若 $\lvert t\rvert>t_{2\alpha}$，否定 H_0，接受 H_1

附表 3 是 t 分布的双侧分位数（t_α）表。对于单侧检验，给定的显著性水平 α 应查临界值为 $t_{2\alpha}$，查该表时还应考虑相应的自由度 df（$df=n-1$）。

【例 2-7】 从河流的某一断面采得 20 个水样，测得酚的浓度（mg/L）为：0.027，0.025，0.031，0.030，0.023，0.028，0.034，0.026，0.024，0.028，0.024，0.032，0.024，0.033，0.029，0.031，0.023，0.032，0.025，0.031。

试用 0.05 的显著性水平检验该断面酚的平均污染水平是否达到了 0.030mg/L。

解 假设该断面的酚污染水平达到了 0.030mg/L。

即假设 H_0：$\mu=\mu_0$

由样本数据计算得：

$$n=20，\ \bar{x}=0.028\text{mg/L}$$

标准差：$S=0.0036$

自由度：$df=n-1=19$

统计量为：

$$t=\frac{0.028-0.030}{\frac{0.0036}{\sqrt{20}}}=-2.486$$

查附表 3 可得（$\alpha=0.05$ 时）

$$t_{\alpha}=2.093$$

显然有：$|t|=2.486>t_{\alpha}=2.093$，则否定 H_0。

即该断面的酚污染水平不等于 0.030mg/L。

同样可用单侧检验，设 H_0：$\mu\geqslant\mu_0$

统计量：$t=-2.486$

临界值：$t_{2\alpha}=1.792$（$\alpha=0.05$）

显然有：$t=-2.486<-t_{2\alpha}=-1.792$

则否定 H_0，即该断面的酚污染水平低于 0.030mg/L。

二、由两个样本检验两总体平均值的一致性

当两总体标准差未知，但可以认为其相等，即 $\sigma_1=\sigma_2$ 时，由两个样本检验两总体平均值的一致性，通常用 t 检验法。μ_1 和 μ_2 分别表示两总体的平均值，$\bar{x}_1$ 和 $\bar{x}_2$ 表示两样本的平均值，S_1 和 S_2 表示两样本的标准差，n_1 和 n_2 表示两样本的样本数。t 检验法对两总体平均值的一致性检验可按以下的 t 检验法统计检验表 2-11 中的程序进行。

表 2-11　t 检验法统计检验表

统计检验步骤	双侧检验	单侧检验
建立统计假设	$H_0:\mu_1=\mu_2$　$H_1:\mu_1\neq\mu_2$	$H_0:\mu_1\geqslant\mu_2$　$H_1:\mu_1<\mu_2$ 或 $H_0:\mu_1\leqslant\mu_2$　$H_1:\mu_1>\mu_2$
选择显著性水平 α	0.05 或 0.01	0.05 或 0.01
计算统计量 t	$t=\dfrac{\bar{x}_1-\bar{x}_2}{S_0\sqrt{\dfrac{1}{n_1}+\dfrac{1}{n_2}}}$ 其中：$S_0=\sqrt{\dfrac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}}$	
计算自由度 df	$df=n_1+n_2-2$	
确定临界值 t_{α}	查附表 3 得 t_{α}	查附表 3 得 $t_{2\alpha}$
统计判别	若 $\|t\|\leqslant t_{\alpha}$，接受 H_0	若 $\|t\|\leqslant t_{2\alpha}$，接受 H_0
	若 $\|t\|>t_{\alpha}$，否定 H_0 接受 H_1	若 $\|t\|>t_{2\alpha}$，否定 H_0 接受 H_1

【例 2-8】 在进行环境噪声对居民睡眠影响的研究中，对生活在 50dB（A）和 55dB（A）噪声环境的居民分别抽查 10 人次，睡眠调查结果见表 2-12。

表 2-12 ［例 2-8］中居民睡眠时间调查结果/h

噪声级 \ 小时 \ 居民	1	2	3	4	5	6	7	8	9	10
50dB(A)	6.9	5.8	6.3	5.0	7.3	8.0	7.1	7.2	6.8	6.6
55dB(A)	6.4	7.2	6.1	6.7	5.6	8.2	7.5	6.9	7.0	6.6

试用 0.05 的显著性水平检验 50dB（A）和 55dB（A）的环境噪声对居民睡眠影响是否有差异。

解 假设两种环境噪声对居民睡眠影响无差异，即假设 H_0：$\mu_1=\mu_2$

由调查数据可得：

$$\bar{x}_1=6.7,\ \bar{x}_2=6.82$$

$$S_1=0.842,\ S_2=0.730$$

$$n_1=n_2=10,\ \mathrm{d}f=2n-2=18$$

统计量为：

$$t=\frac{\bar{x}_1-\bar{x}_2}{\sqrt{\frac{S_1^2+S_2^2}{n}}}=\frac{6.7-6.82}{\sqrt{\frac{0.842^2+0.730^2}{10}}}=-0.341$$

查附表 3 得（$\alpha=0.05$ 时）

$$t_\alpha=2.101$$

显然有：

$$|t|=0.341<t_\alpha=2.101$$

接受假设 H_0，即表明 50dB（A）和 55dB（A）的环境噪声对居民睡眠的影响无显著性差异。

第五节 χ^2 检验

在前面两节中，我们用 μ 检验法和 t 检验法对总体的一个特征数——平均值进行了统计检验。在一些情况下，我们还需要对总体的另一个重要的特征数——方差进行检验。例如在建立一种新的环境分析测试方法时，需要判断该方法的精密度是否能够满足预定的要求，这就是一个判断总体方差与一已知值之间是否存在差异的问题。

χ^2 检验是一种检验总体方差与某一确定值差异的重要方法，主要包括下述两方面内容。

一、已知总体的平均值检验总体的方差

当总体的平均值为已知时，可用 χ^2 检验法来检验总体的方差 σ^2 是否等于、小于或大于某一已知常数，用 μ_0 表示总体的均值，σ^2 表示总体的方差，σ_0^2 表示

已知常数，x_1，x_2，…，x_n 表示样本的 n 个测定值，自由度为 $df=n$。χ^2 检验对总体方差的检验可按表 2-13 的统计检验程序进行。

表 2-13　χ^2 检验法统计检验表

统计检验步骤	双侧检验	单侧检验	
建立统计假设	$H_0:\sigma^2=\sigma_0^2$ $H_1:\sigma^2\neq\sigma_0^2$	$H_0:\sigma^2\geqslant\sigma_0^2$ $H_1:\sigma^2<\sigma_0^2$	$H_0:\sigma^2\leqslant\sigma_0^2$ $H_1:\sigma^2>\sigma_0^2$
选择显著性水平 α	0.05 或 0.01	0.05 或 0.01	
计算统计量 χ^2	$\chi^2=\frac{1}{\sigma_0^2}\sum_{i=1}^{n}(x_i-\mu_0)^2$	$\chi^2=\frac{1}{\sigma_0^2}\sum_{i=1}^{n}(x_i-\mu_0)^2$	
计算自由度 df	$df=n$	$df=n$	
确定临界值 χ_α^2	查附表 4 得临界值 $\chi_{\frac{\alpha}{2}}^2$ 和 $\chi_{1-\frac{\alpha}{2}}^2$	查附表 4 得临界值 $\chi_{1-\alpha}^2$	查附表 4 得临界值 χ_α^2
统计判别	若 $\chi_{\frac{\alpha}{2}}^2\geqslant\chi^2\geqslant\chi_{1-\frac{\alpha}{2}}^2$，接受假设 H_0	若 $\chi^2\geqslant\chi_{1-\alpha}^2$，接受假设 H_0	若 $\chi^2\leqslant\chi_\alpha^2$，接受假设 H_0
	若 $\chi^2\geqslant\chi_{\frac{\alpha}{2}}^2$ 或 $\chi^2\leqslant\chi_{1-\frac{\alpha}{2}}^2$，否定假设 H_0，接受备选假设 H_1	若 $\chi^2<\chi_{1-\alpha}^2$，否定假设 H_0，接受备选假设 H_1	若 $\chi^2>\chi_\alpha^2$，否定假设 H_0，接受备选假设 H_1

在实际应用中，总体平均值已知的情况是不多的，而使用较多的是总体平均值未知的情况。总体平均值未知时的 χ^2 检验在下面介绍。

二、总体平均值未知检验总体的方差

当总体的平均值未知时，用样本的平均值代替总体的均值，同样可用 χ^2 检验来检验总体的方差是否等于、小于或大于某一已知常数。用 σ^2 表示总体的方差，σ_0^2 表示已知常数，x_1，x_2，…，x_n 表示样本的 n 个测定值，$\bar{x}$ 表示样本的均值。χ^2 检验对总体方差的检验可按 χ^2 检验法统计检验表进行（见表 2-14）。

表 2-14　χ^2 检验法统计检验表

统计检验步骤	双侧检验	单侧检验	
建立统计假设	$H_0:\sigma^2=\sigma_0^2$ $H_1:\sigma^2\neq\sigma_0^2$	$H_0:\sigma^2\geqslant\sigma_0^2$ $H_1:\sigma^2<\sigma_0^2$	$H_0:\sigma^2\leqslant\sigma_0^2$ $H_1:\sigma^2>\sigma_0^2$
选择显著性水平 α	0.05 或 0.01	0.05 或 0.01	
计算统计量 χ^2	$\chi^2=\frac{1}{\sigma_0^2}\sum_{i=1}^{n}(x_i-\bar{x})^2$	$\chi^2=\frac{1}{\sigma_0^2}\sum_{i=1}^{n}(x_i-\bar{x})^2$	
计算自由度 df	$df=n-1$	$df=n-1$	
确定临界值 χ_α^2	查附表 4 得临界值 $\chi_{\frac{\alpha}{2}}^2$ 和 $\chi_{1-\frac{\alpha}{2}}^2$	查附表 4 得临界值 $\chi_{1-\alpha}^2$	查附表 4 得临界值 χ_α^2
统计判别	若 $\chi_{\frac{\alpha}{2}}^2\geqslant\chi^2\geqslant\chi_{1-\frac{\alpha}{2}}^2$，接受假设 H_0	若 $\chi^2\geqslant\chi_{1-\alpha}^2$，接受假设 H_0	若 $\chi^2\leqslant\chi_\alpha^2$，接受假设 H_0
	若 $\chi^2>\chi_{\frac{\alpha}{2}}^2$ 或 $\chi^2<\chi_{1-\frac{\alpha}{2}}^2$，否定假设 H_0，接受备选假设 H_1	若 $\chi^2<\chi_{1-\alpha}^2$，否定假设 H_0，接受备选假设 H_1	若 $\chi^2>\chi_\alpha^2$，否定假设 H_0，接受备选假设 H_1

留心的读者可以发现，表 2-14 与表 2-13 极其相似，仅有的两个差别是表 2-14中用样本均值代替了总体平均值，自由度取 $n-1$。

【例 2-9】 评价交通噪声污染的一项重要指标是噪声的涨落，通常可用标准差表示。某一路段行车高峰期的声级涨落标准差为 3.50dB（A），一季度后在该路段的同一行车高峰期测得一组交通噪声数据如下：

70.2	72.1	71.1	69.1	68.0	67.1	72.2	74.1	75.3
73.1	70.6	70.8	70.2	69.2	68.0	66.8	67.3	68.4

试按 0.05 的显著性水平检验声级涨落的标准差有无变化。

解 上述问题为 χ^2 检验问题。

假设 H_0：$\sigma=3.50$，H_1：$\sigma\neq3.50$

由测量数据计算得：

$$\bar{x}=70.2，\ n=18，\ df=17$$

统计量为：$\chi^2=\dfrac{1}{3.5^2}\sum_{i=1}^{18}(x_i-70.2)^2=8.431$

选定显著性水平 $\alpha=0.05$，在 χ^2 分布临界值表（见附表 4）中查得：

$$\chi^2_{17}(0.05)=27.587；\ \chi^2_{17}(0.95)=8.672$$

显然有：$\chi^2=8.431<\chi^2_{17}(0.95)=8.672$

否定假设 H_0，即交通噪声标准差发生了变化，交通噪声较前平稳。

第六节 *F* 检验

在用 t 检验法对两总体的平均值的一致性进行检验时，并不要求两总体的方

表 2-15 *F* 检验法统计检验表

统计检验步骤	双侧检验	单侧检验	
建立统计假设	$H_0:\sigma_1^2=\sigma_2^2$ $H_1:\sigma_1^2\neq\sigma_2^2$	$H_0:\sigma_1^2\leqslant\sigma_2^2$ $H_1:\sigma_1^2>\sigma_2^2$	$H_0:\sigma_1^2\geqslant\sigma_2^2$ $H_1:\sigma_1^2<\sigma_2^2$
选择显著性水平 α	0.05 或 0.01	0.05 或 0.01	
计算统计量 F	$F=\dfrac{S_1^2}{S_2^2}$	$F=\dfrac{S_1^2}{S_2^2}$	$F=\dfrac{S_2^2}{S_1^2}$
计算自由度 df_1,df_2	$df_1=n_1-1$ $df_2=n_2-1$	$df_1=n_1-1$ $df_2=n_2-1$	$df_1=n_2-1$ $df_2=n_1-1$
确定临界值 F_α	查附表 5 得 $F_{\frac{\alpha}{2}}$，$F_{1-\frac{\alpha}{2}}$	查附表 5 得 F_α	
统计判别	若 $F_{1-\frac{\alpha}{2}}\leqslant F\leqslant F_{\frac{\alpha}{2}}$ 接受 H_0	若 $F\leqslant F_\alpha$，接受 H_0	
	若 $F>F_{\frac{\alpha}{2}}$ 或 $F<F_{1-\frac{\alpha}{2}}$，否定 H_0 接受 H_1	若 $F>F_\alpha$，否定 H_0 接受 H_1	

差已知，但却有一个先决的假定，两总体的方差应一致。这个假定是需要进行验证的。F 检验是在两总体的平均值和方差未知的条件下，对两总体的方差进行一致性检验的重要方法。

假设两总体分别服从正态分布，σ_1^2 和 σ_2^2 表示两总体的方差，S_1^2 和 S_2^2 表示两样本的方差，n_1 和 n_2 表示两样本的样本数。F 检验对两总体方差一致性的检验可按表 2-15 的统计检验步骤进行。

表 2-15 中的临界值 $F_\alpha(df_1,df_2)$ 可在附表 5 中查得，需要指出的是临界值 $F_\alpha(df_1,df_2)$ 和 $F_{1-\alpha}(df_2,df_1)$ 有如下关系：

$$F_\alpha(df_1,df_2)=\frac{1}{F_{1-\alpha}(df_2,df_1)}$$

在实际计算中，不要将两样本自由度的顺序搞错了。

【例 2-10】 两个路段的交通噪声测量数据如下［单位 dB(A)］：

路段 1	70.2	72.1	71.1	69.1	68.0	67.1	72.2	74.1	75.3
	73.1	70.6	70.8	70.2	69.2	68.0	66.8	67.3	68.4
路段 2	65.5	67.3	68.4	70.5	72.1	69.0	64.1	66.3	
	62.0	61.1	64.4	68.6	67.2	68.1	69.5	65.0	
	63.1	64.9	61.4	60.5	63.9	66.8	69.1		

试用 0.05 的显著性水平检验两路段的交通噪声涨落是否存在差异。

解　道路交通噪声服从正态分布。

假设　H_0：$\sigma_1^2=\sigma_2^2$　　H_1：$\sigma_1^2\neq\sigma_2^2$

实际计算结果如下：

$n_1=18$，$\bar{x}_1=70.2$，$S_1^2=6.076$，$n_2=23$，$\bar{x}_2=66.0$，$S_2^2=10.087$

统计量为：

$$F=\frac{S_2^2}{S_1^2}=1.660$$

给定显著性水平 $\alpha=0.05$，$df_1=22$，$df_2=17$

查附表 5 得：$F_{0.05}(22,17)=2.21$

$$F_{0.95}(22,17)=\frac{1}{F_{0.05}(17,22)}=\frac{1}{2.11}=0.474$$

显然有：$F_{0.95}(22,17)<F<F_{0.05}(22,17)$

接受假设 H_0，即认为两路段的交通噪声声级涨落情况无显著差异。

第七节　总体分布类型的统计检验

在讨论总体平均值或方差一致性的各种检验方法时，我们都假设总体有某种已知形式的理论分布，如正态分布。严格地说，这种假设是需要予以证明的。在

实际环境数据的统计分析中，有时也需要对总体的分布类型进行检验。例如，为了选择反映总体中心趋势的特征数，需要知道总体服从正态分布还是对数正态分布，或是其他类型的偏态分布等。

由样本推断总体的分布类型称为分布的假设检验。检验总体服从某种分布的方法很多，本节介绍几种常用的检验方法。

一、概率纸法

概率纸法是应用概率纸判断统计数据是否服从正态分布或对数正态分布的常用的方法。概率纸是一种按某种分布函数而设计的坐标纸，检验统计数据是否服从正态分布时使用正态概率纸（见图 2-1），检验统计数据是否服从对数正态分布时使用对数正态概率纸（见图 2-2）。

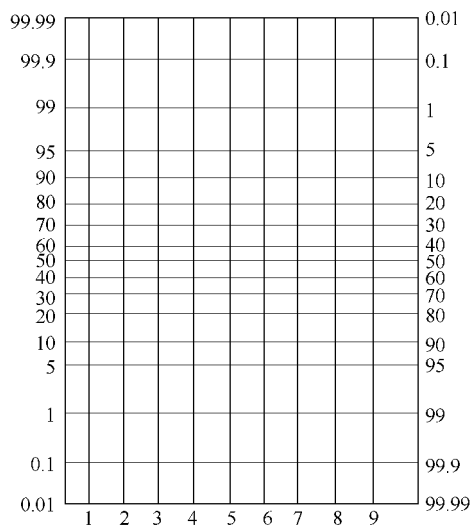


图 2-1　正态概率纸

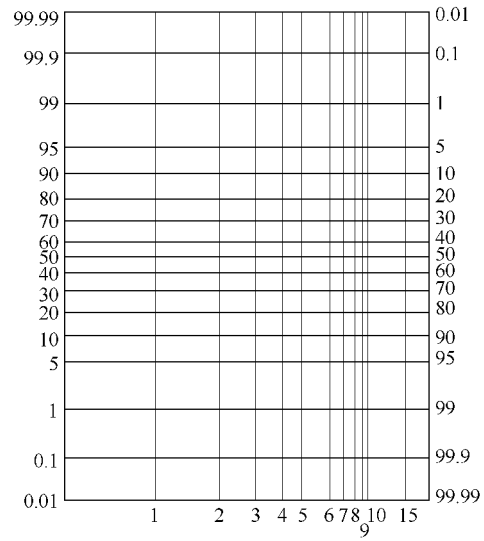


图 2-2　对数正态概率纸

正态概率纸是这样设计的：标准正态分布的分布函数为：

$$F(x)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x}\exp\left(-\frac{\mu^2}{2}\right)\mathrm{d}\mu$$

以 x 为横坐标，$F(x)$ 为纵坐标，使 $[x_1,F(x_1)]$，$[x_2,F(x_2)]$，…，$[x_n,F(x_n)]$ 这几个点连成一条直线。因而要建立这样的坐标系：横坐标轴按 x_1，x_2，…，x_n 等间隔作刻度，纵坐标轴以 $F(x_1)$，$F(x_2)$，…，$F(x_n)$ 不等间隔作刻度，用这样的特殊坐标系制成坐标纸，称之为正态概率纸。

对数正态概率纸的设计与正态概率纸相仿，横坐标轴只需按 $\ln x_1$，$\ln x_2$，…，$\ln x_n$ 等间隔作刻度即可。

由图 2-1 和图 2-2 我们可以看到，概率纸的纵坐标为累积频率（%），横坐标由测定值的单位而定的普通等分刻度或自然对数刻度。将统计数据和相应数值的累积频率在正态概率纸或对数正态概率纸上标点，若这些点分布在一条直线上，则统计数据服从正态分布或对数正态分布。现以正态概率纸为例，介绍概率纸检验统计数据正态分布的具体步骤：

① 将样本的 n 个值 x_1，x_2…，x_n 按由小到大的次序排列，得 $x(1)\leqslant x(2)\leqslant\cdots\leqslant x(i)\leqslant\cdots\leqslant x(n)$，$i$ 为顺序排列后各数值的秩次数。

② 按累积频率计算式$\frac{i}{n}\times 100\%$计算每个数值相应的累积频率。

③ 根据样本数值及其对应的累积频率值，在正态概率纸上标点。

④ 根据点的分布状况，判别数据是否属正态分布：若点的分布呈直线趋势，则认为数据服从正态分布；若点的分布呈曲线分布或其他无规律的势态，则认为数据不服从正态分布。

应用概率纸判别数据的分布类型是一种简便易行的方法，但精确程度较差。在实践中我们经常会发现，在概率纸上点的分布形状不十分明显，使我们对数据是否服从正态分布或对数正态分布的判断把握不大，为了能够得到较为确切的结论，需要用更严谨一些的方法对总体的分布进行检验。

【例 2-11】 在［例 2-10］中假设道路交通噪声的声级是正态分布，试检验路段 2 的测定数据，以证明交通噪声确属正态分布。

解　可将路段 2 的测定数据由小到大排列如下：

60.5　61.1　61.4　62.0　63.1　63.9　64.1　64.4　64.9　65.0
65.5　66.3　66.8　67.2　67.3　68.1　68.4　68.6　69.0　69.1
69.5　70.5　72.1

按 2dB（A）一档将频数和累积频率列表见表 2-16。

表 2-16　［例 2-11］中交通噪声频数和累积频率

声级	61(60.62)	63(62.64)	65(64.66)	67(66.68)	69(68.70)	71(70.72)	73(72.74)
频数	4	2	5	4	6	1	1
累积频数	4	6	11	15	21	22	23
累积频率/%	17.4	26.1	47.8	65.2	91.3	95.6	100.00

以各档声级的中值点为横坐标，以相对应的累积频率为纵坐标，在正态概率纸上标点，共可标出 7 个点，如图 2-3 所示。显然，这几个点基本上集中在一条直线上，也就证明了交通噪声是符合正态分布的。

从正态概率纸上可以粗略给出均值 μ 及标准差 σ 的估计值。根据正态分布函数的性质，50%概率对应的横坐标即为 μ 值。而 15.87%概率对应的横坐标应为 $\mu-\sigma$。［例 2-11］中，由右图可估算出：

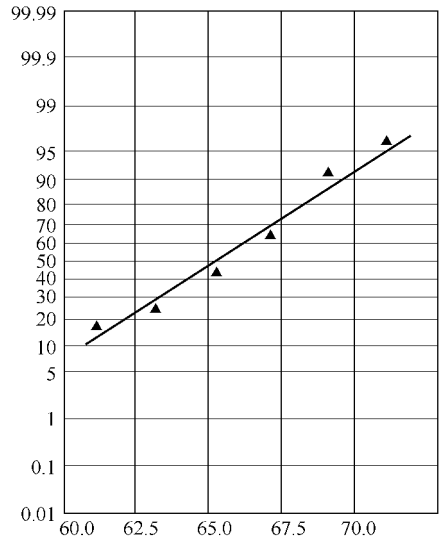


图 2-3 ［例 2-11］图

$\mu\approx65.2$，$\sigma\approx65.2-61.3=3.9$

二、W 检验

当样本数不大时，可以应用 W 检验法（又称 Shapiro-Wilk）来检验总体是否服从正态分布。W 检验法的检验步骤如下：

（1）将样本的 n 个数据按大小顺序排列

$$x_1\leqslant x_2\leqslant\cdots\leqslant x_n$$

（2）建立统计假设

H_0：总体服从正态分布

H_1：总体不服从正态分布

（3）选择显著性水平 α

一般选 $\alpha=0.05$ 或 $\alpha=0.01$

（4）计算统计量 W

$$W=\frac{b^2}{(n-1)S^2} \tag{2-6}$$

式中，n 为样本数，S 为样本的标准差，b 值按式（2-7）计算：

$$b=a_1(x_n-x_1)+a_2(x_{n-1}-x_2)+\cdots+a_k(x_{n+1-k}-x_k) \tag{2-7}$$

式中当 n 为偶数时，$k=\frac{n}{2}$；当 n 为奇数时，$k=\frac{n-1}{2}$，并且 x_{k+1} 不列入式中计算。系数 a_k，a_{k-1}，…，a_2，a_1 在附表 6 中给出。

（5）确定临界值 w_α

根据选定的显著性水平 α 和样本数 n，可在附表 7 中查得 W 检验法的临界值 w_α。

（6）判断假设成立与否

若计算所得的统计量 W 值大于临界值 w_α（即 $W>w_\alpha$），则在显著性水平 α 条件下，接受假设 H_0，即认为总体服从正态分布；

若计算所得的统计量 W 值小于临界值 w_α（即 $W<w_\alpha$），则在显著性水平 α 条件下，否定假设 H_0，接受备选假设 H_1，即总体不服从正态分布。

检验总体是否服从对数正态分布的方法和步骤与上述检验总体是否服从正态分布的方法基本相同。惟一不同之处是，检验总体是否服从对数正态分布时，需要将每个数据进行对数变换（自然对数），然后按上述检验正态分布的步骤进行。

【例 2-12】 试用 W 检验法检验［例 2-11］的交通噪声符合正态分布，显著

性水平选择 $\alpha=0.05$。

解　上例的交通噪声样本的两个参数为：

$$\bar{x}=\frac{\sum_{i=1}^{23}x_i}{23}=66.0\text{dB(A)}$$

$$S^2=10.087;\ n=23$$

由 W 检验法定义，由式（2-7）计算 b 值：

$$\begin{aligned}b&=\sum_{i=1}^{k}a_i(x_{n+1-i}-x_i)\\&=0.4542\times(72.1-60.5)+0.3126(70.5-61.1)+\cdots+\\&\quad 0.0228(66.8-65.5)\\&=14.714\end{aligned}$$

b 值计算中系数 a_k 按 $n=23$ 从附表 6 中查得。

统计量：

$$W=\frac{b^2}{(n-1)S^2}=\frac{14.714^2}{22\times10.087}=0.9756$$

由 $n=23$，α 取值 0.05 查附表 7 得临界值 $w_\alpha=0.914$。

显然有 $W=0.9756>w_\alpha=0.914$，接受假设 H_0，即在显著性水平 0.05 条件下接受该路段交通噪声服从正态分布的假设。

三、皮尔逊 χ^2 检验

为判断一个样本所属的母体是否服从某一理论分布，可用皮尔逊 χ^2 检验法进行检验。皮尔逊 χ^2 检验不只限于检验总体是否服从正态分布，而适用于各种类型的分布进行检验判断。假设所研究的样本的母体分布函数为 $F_0(x)$，皮尔逊 χ^2 检验统计量的计算公式为：

$$\chi^2=\sum_{i=1}^{k}\frac{(f_i-np_i)^2}{np_i}\tag{2-8}$$

式中，f_i 为实际观察频数；np_i 为理论频数；k 为划分的区间数。

根据皮尔逊定理，不论母体分布 $F_0(x)$ 是什么分布，由式（2-8）计算所得的统计量以自由度 $k-r-1$ 的 χ^2 分布为极限分布，其中 r 为 $F_0(x)$ 分布中的参数数目。

因此，当样本充分大时，可以认为统计量 $\chi^2=\sum_{i=1}^{k}\frac{(f_i-np_i)^2}{np_i}$ 服从 χ^2 分布。在实际应用皮尔逊 χ^2 检验法对总体的分布进行检验时，样本数 $n>50$，即可认为满足样本数充分大的条件。在对样本分组时，以每组内所含的样本数不少于 5

为宜。如果出现组内样本频数小于5的情况，需重新调整分组的区间或将相邻的组予以合并，以保证每组内样本的频数不少于5。下面以检验总体是否服从正态分布为例，说明皮尔逊 χ^2 检验的方法和步骤。

(1) 建立统计假设：H_0 表示总体分布服从正态分布。H_1 表示总体分布不服从正态分布。

(2) 将样本数据按大小划分为 k 组，确定每组内观察值的频数 f_i，要求每组内的频数 f_i 不少于5，若某组内频数 f_i 小于5，则需要将相邻的组予以合并以保证组内 f_i 不少于5。

(3) 计算样本的平均值 $\bar{x}$ 和标准差 S。

(4) 以每组区间的下限值 x_i 按式 (2-9) 变换成相应的理论标准正态分布的变量。

$$\mu_i=\frac{x_i-\bar{x}}{S} \tag{2-9}$$

(5) 由附表1查得 μ_i 的正态分布概率 $F(\mu_i)$。

(6) 计算每组的理论频数：

$$np_i=n[F(\mu_{i+1})-F(\mu_i)] \tag{2-10}$$

(7) 选择显著性水平 α，一般选 $\alpha=0.05$ 或 $\alpha=0.01$。

(8) 计算皮尔逊 χ^2 检验的统计量 χ^2。

$$\chi^2=\sum_{i=1}^{k}\frac{(f_i-np_i)^2}{np_i}$$

(9) 确定临界值 χ^2_α。

对于正态分布，皮尔逊 χ^2 检验的自由度为 $df=k-r-1=k-2-1$。由附表4可查得自由度为 $k-3$、显著性水平为 α 的临界值 χ^2_α。

(10) 判断假设成立与否。

若计算所得的统计量 χ^2 值小于临界值 χ^2_α，则在显著性水平 α 条件下，接受假设 H_0，即总体服从正态分布；若计算所得的统计量 χ^2 值大于临界值 χ^2_α，则在显著性水平 α 条件下，否定假设 H_0，接受备选假设 H_1，即总体不服从正态分布。

【例 2-13】 某城市的网格法环境噪声普查共获得200个测定数据，按5dB一档分组可有表2-17所示的数据。

表 2-17 [例 2-13] 中环境噪声普查结果

声级范围	40～45	45～50	50～55	55～60	60～65	65～70	70～75
频数	16	28	46	69	24	10	7

试检验该城市的环境噪声级是否服从正态分布。

解 假设 H_0：环境噪声服从正态分布，按所给数据，算得：

$$\bar{x}=55.4；\ S=7.025$$

由附表 1 查得 $\mu_i=\dfrac{x_i-\bar{x}}{S}$的正态分布概率 $F(\mu_i)$ 见表 2-18。

表 2-18 ［例 2-13］中正态分布概率值

x_i	40	45	50	55	60	65	70	75
$F(\mu_i)$	0.0143	0.0694	0.2206	0.4801	0.7454	0.9147	0.9812	0.9974

即可计算出理论频数 $np_i=n[F(\mu_{i+1})-F(\mu_i)]$

统计量为：

$$\begin{aligned}\chi^2 &= \sum_{i=1}^{7}\frac{(f_i-np_i)^2}{np_i}\\ &=\frac{(16-11.02)^2}{11.02}+\cdots+\frac{(7-3.24)^2}{3.24}\\ &=15.93\end{aligned}$$

选择显著性水平 $\alpha=0.05$，自由度 df 为：

$$df=7-2-1=4$$

由附表 4 查得临界值为：

$$\chi^2_{(4)0.05}=9.488$$

显然有：$\chi^2=15.93>\chi^2_\alpha=9.488$，否定原假设，即该城市的环境噪声不服从正态分布。

四、柯尔莫哥洛夫正态检验

上面介绍的皮尔逊 χ^2 检验是以小区间内的频数分布与理论频数分布的差异来检验总体是否服从某种分布。而柯尔莫哥洛夫检验则是对样本的每一个值考虑它与理论分布的偏差，从而判别其是否服从某种分布。

容量为 n 的样本的频数分布为 $F_n(x)$，假设其总体的理论分布为 $F_0(x)$，是一完全确定的连续型分布函数。考虑样本的每一个值与理论分布的偏差，柯氏提出的检验统计量为：

$$D_n=\underset{-\infty<x<+\infty}{SUP}|F_n(x)-F_0(x)| \tag{2-11}$$

式中，D_n 表示样本频数分布与所假设的总体理论分布之差的所有点中的最大绝对值。由柯尔莫哥洛夫定理可知，对任意的 $y>0$ 有：

$$\lim_{n\to+\infty}P\{D_n\sqrt{n}<y\}=k(y)=\sum_{-\infty}^{+\infty}(-1)^k\exp(-2k^2y^2) \tag{2-12}$$

这一定理表明，当 n 足够大时（例如 $n>30$），可以认为 $D_n\sqrt{n}$近似服从正态分布 $k(y)$，而正态分布 $k(y)$ 的分布是已知的，可以在附表 1 查到。这样可以根

据显著性水平 α 查附表 1，找到满足 $k(y_\alpha)=1-\alpha$ 的临界值 y_α，比较统计量 $D_n\sqrt{n}$ 和 y_α 的大小可判断总体是否服从某种理论分布。

仍以检验总体是否服从正态分布为例，说明柯尔莫哥洛夫检验的具体步骤。

(1) 建立统计假设　设 H_0 表示总体服从正态分布，H_1 表示总体不服从正态分布。

(2) 选择显著性水平 α　一般取 $\alpha=0.05$ 或 $\alpha=0.01$。

(3) 将样本的 n 个数据按大小顺序排列：

$$x_1\leqslant x_2\leqslant\cdots\leqslant x_n$$

(4) 计算样本的频数分布

$$F_n(x_i)=\frac{i-1}{n}$$

式中，i 为该样本值相应的顺序号数。

(5) 计算样本的平均值 $\bar{x}$ 和标准差 S。

(6) 将样本的每一个值 x_i 按下式进行变换：

$$\mu_i=\frac{x_i-\bar{x}}{S}$$

然后从正态分布表（附表 1）查得对应于 μ_i 的理论分布值 $F_0(\mu_i)$。

(7) 计算统计量　求出样本每一频数分布值 $F_n(x_i)$ 与相应的理论分布值 $F_0(x_i)$ 的差值 $d(i)=|F_i(x_i)-F_0(x_i)|$，并从中找出最大值：

$$D_n=\underset{x}{SUP}\ |F_i(x)-F_0(x),F_n(x)-F_0(x)|$$

(8) 确定临界值 y_α　由显著性水平 α 查附表 1，查得相应于 $k(y_\alpha)$ 的临界值 y_α。

(9) 判断假设成立与否。

若 $D_n\sqrt{n}<y_\alpha$，接受假设 H_0，即接受总体服从正态分布的假设。

若 $D_n\sqrt{n}>y_\alpha$，则否定假设 H_0，接受备选假设 H_1，即认为总体不服从正态分布。

【例 2-14】 考察我国城市大气污染情况，抽查了 34 个城市的大气监测资料，SO_2 的浓度（mg/m^3）分布如下。

0.019	0.040	0.054	0.070	0.078	0.094	0.121	0.149	0.274
0.025	0.046	0.058	0.071	0.079	0.102	0.129	0.161	0.391
0.029	0.049	0.061	0.072	0.088	0.111	0.134	0.182	
0.032	0.052	0.066	0.074	0.091	0.118	0.138	0.215	

试验证我国城市大气的 SO_2 浓度符合对数正态分布。

解　假设 H_0：SO_2 浓度符合对数正态分布，需对上表的 SO_2 浓度做变换：

$$y_i=\lg x_i$$

可计算得：　　$\bar{y}=-1.088$　　$S=0.283$

参量变换：　　$\mu_i=\dfrac{y_i-\bar{y}}{S}$并查附表 1 可得 $F_0(\mu_i)$

$$F_n(y_i)=\frac{i}{n}，\text{此处 } n=36，$$

因而有表 2-19。

表 2-19　[例 2-14] 中正态分布概率值

$F_0(\mu_i)$	0.013	0.034	0.056	0.075	0.136	0.189	0.215	0.245	0.264
$F_n(y_i)$	0	0.028	0.056	0.083	0.111	0.139	0.167	0.194	0.222
$d(i)$	0.013	0.006	0.000	0.008	0.025	0.050	0.048	0.049	0.042
$F_0(\mu_i)$	0.298	0.326	0.371	0.397	0.405	0.413	0.425	0.440	0.472
$F_n(y_i)$	0.250	0.278	0.306	0.333	0.361	0.389	0.417	0.444	0.472
$d(i)$	0.048	0.048	0.065	0.064	0.044	0.024	0.008	0.004	0.000
$F_0(\mu_i)$	0.481	0.488	0.548	0.568	0.587	0.633	0.681	0.716	0.726
$F_n(y_i)$	0.500	0.528	0.556	0.583	0.611	0.639	0.666	0.694	0.722
$d(i)$	0.019	0.040	0.008	0.015	0.024	0.006	0.015	0.022	0.004
$F_0(\mu_i)$	0.751	0.776	0.788	0.821	0.851	0.891	0.932	0.969	0.991
$F_n(y_i)$	0.750	0.778	0.806	0.833	0.861	0.889	0.917	0.944	0.972
$d(i)$	0.001	0.002	0.018	0.012	0.010	0.002	0.015	0.025	0.019

显然有：$D_n=0.065$

取显著性水平 $\alpha=0.10$，查附表 1 有：

$$\frac{y_\alpha}{\sqrt{n}}=\frac{0.805}{\sqrt{36}}=0.134$$

因而：$D_n=0.065<\dfrac{y_\alpha}{\sqrt{n}}=0.134$

即接受假设 H_0，可以认为我国城市的大气 SO_2 浓度分布符合对数正态分布。

在柯尔莫哥洛夫正态检验中，当样本容量 n 大于 35 时，临界值是用 $1.36/\sqrt{n}$计算得到。当样本容量小于 35 时，y_n 可查表 2-20 得到。

表 2-20　y_n 值表

n	y_n	n	y_n	n	y_n
11	0.391	16	0.328	25	0.270
12	0.375	17	0.318	30	0.240
13	0.361	18	0.309	35	0.230
14	0.349	19	0.301		
15	0.338	20	0.294		

第八节 符号检验法和秩和检验法

在本章前面几节介绍的 μ 检验法、t 检验法以及 F 检验法等，都要求对总体的分布做出假设。在假定总体具有某种形式分布的条件下，用样本数据对总体的参数（如平均值、标准差）进行统计检验。由于这些统计检验都涉及总体的参数，因而称为参数统计检验。在实际统计分析工作中，有时要求对一些没有特定形式的分布或分布形式不知道的总体进行比较，这时需要采用不必对总体分布做出假定的统计检验方法。由于这种统计检验是在总体的分布之间而不涉及总体的参数，因而称为非参数统计检验法。前面介绍的皮尔逊 χ^2 检验总体的理论分布就是一种非参数统计检验。非参数统计检验不仅能在上述情况下应用，而且能在某些资料不能简单用数量表示的场合下应用。如环境监测中经常碰到的“检出”、“未检出”，“超标”、“未超标”以及某些按人为法则规定的“秩序”、“大小”等，也可应用非参数统计检验法进行检验。本节主要讨论非参数统计检验中符号检验法和秩和检验法两种方法。

一、符号检验法

符号检验法是检验两样本所代表的总体分布是否一致的简单、直观的方法。符号检验法通过对两样本的波动趋势程度是否相同来判别两样本所代表的总体分布是否一致，因而符号检验法对两样本的数据或资料有一定的要求，它要求两样本的数据或资料必须是两两对应的。顾名思义，符号检验法是将两样本对应数据比较结果用符号来表示。例如用符号“+”、“−”和“0”分别表示两样本对应数据中较大值、较小值和两值相等。

符号检验法的检验步骤如下：

(1) 建立统计假设。

H_0：两样本所代表的总体具有相同的分布。

H_1：两样本所代表的总体具有的分布不相同。

(2) 选择显著性水平 α，一般选 $\alpha=0.05$ 或 $\alpha=0.01$。

(3) 用符号表示两样本对应数据的比较情况。

“+”表示大于或优于，“−”表示小于或劣于，“0”表示相等。当两样本对应数值相等时，即取符号“0”时，在检验过程中舍弃这对数据。

(4) 计算统计量 γ：

$$\gamma=\min\{n_+,n_-\} \tag{2-13}$$

式中，n_+、n_- 分别表示两样本比较出现的“+”和“−”的次数。$\min\{\ \}$ 表示取 $\{\ \}$ 中的最小值。

（5）确定临界值 γ_α　由给定的显著性水平 α，在附表 8 中查出临界值 γ_α。

（6）判断假设成立与否。

若 $\gamma>\gamma_\alpha$，则在显著性水平 α 的条件下，接受 H_0，即认为两样本的总体分布相同。

若 $\gamma\leqslant\gamma_\alpha$，则在显著性水平 α 的条件下，否定假设 H_0，接受备选假设 H_1，即认为两样本总体分布不相同。

【例 2-15】 为了比较两个污灌区的土壤含镉量是否存在明显的差异，从两个灌区中随机各抽取 20 个土壤样品进行测试。测试结果列表见表 2-21。表中的正号和负号表示灌区甲的检测值与相应的灌区乙的检测值的大小比较。

表 2-21　［例 2-15］土壤中 Cd 含量/(mg/kg)

灌区甲	4.24	4.02	4.38	3.85	5.01	4.83	3.24	4.75	4.64	5.23
灌区乙	3.82	4.16	4.38	3.64	4.97	4.06	5.17	4.83	5.26	4.58
符号	+	−	0	+	+	+	−	−	−	+
灌区甲	3.67	4.11	4.25	4.91	5.41	5.26	5.09	3.82	4.28	4.17
灌区乙	4.29	3.66	3.90	5.41	3.87	5.06	4.44	4.68	4.87	4.69
符号	−	+	+	−	+	+	+	−	−	−

试在显著水平 $\alpha=0.05$ 下检验这两个灌区中土壤镉含量是否服从同一分布。

解　首先做假设。

假设 H_0：两灌区中土壤的含镉量服从同一分布。

由表可见：$n=19$（舍弃一相等项）

$$n_+=10,\ n_-=9$$

$$\gamma=\min\{n_+,n_-\}=9$$

由附表 8 查得：

$$\gamma_{0.05}(19)=4$$

显然有：$\gamma=9>\gamma_{0.05}(19)=4$

接受原假设 H_0，即认为两灌区土壤的含镉量分布相同。

二、秩和检验法

秩和检验法也是一种两样本总体分布是否一致的非参数检验法，它比符号检验法的精度要高一些。

秩和检验法的基本思想是，对样本量分别为 n_1 和 n_2 的两样本，按数值大小次序混合排列，并顺序编上号码，所编的号码称为该数值的秩。最小的秩为 1，最大的秩为（n_1+n_2）。如果两样本的总体服从同一理论分布，则两样本的观察

值在多数情况下是相间排列的。若定义样本观察值秩之和为该样本的秩和，那么两样本的秩和差别不会太大，一样本的秩和比另一样本的秩和大得多或小得多的情况应是极少发生的情况，即小概率事件。因此，可以确定出秩和的上下界限 T_α''和 T_α'，与统计量 T 进行比较，从而判断两样本总体的分布是否一致。在秩和检验法中，一般采用样本容量较小的样本进行统计判别。

秩和检验法的检验步骤如下：

(1) 建立统计假设。

H_0：两样本代表的总体具有相同的分布。

H_1：两样本代表的总体具有不同的分布。

(2) 选择显著性水平 α，一般选 $\alpha=0.05$ 或 $\alpha=0.01$。

(3) 计算统计量 T，将两样本按数值大小次序混合排列，并顺序编上号码——即秩。计算样本容量较小的样本的秩和 T。

(4) 确定临界值 $T_\alpha(n_1, n_2)$。由附表 9 查得秩和的下限值 $T_\alpha'(n_1, n_2)$ 和秩和的上限值 $T_\alpha''(n_1, n_2)$。

(5) 判别假设成立与否。

若 $T_\alpha'<T<T_\alpha''$，则在显著性水平 α 条件下，接受假设 H_0，即认为两样本总体的分布相同。

若 $T\geqslant T_\alpha''$或 $T\leqslant T_\alpha'$，则在显著性水平 α 条件下，否定假设 H_0，接受备选假设 H_1，即认为两样本的总体分布不相同。

【例 2-16】 甲、乙两人对一气体样的一氧化碳浓度同时进行测定，得到数据如表 2-22 所列。

表 2-22 ［例 2-16］CO 浓度测定结果

甲	1.52	1.48	1.50	1.47	1.55	1.46	1.58	1.40	
乙	1.51	1.46	1.43	1.54	1.39	1.58	1.41	1.59	1.45

试在显著性水平 $\alpha=0.05$ 下检验两人的分析结果有无差异。

解 首先把所给的数据按大小顺序排列成表 2-23。

表 2-23 ［例 2-16］数据排序

秩	1	2	3	4	5	6,7	8	9
甲		1.40				1.46	1.47	1.48
乙	1.39		1.41	1.43	1.45	1.46		
秩	10	11	12	13	14	15,16	17	
甲	1.50		1.52		1.55	1.58		
乙		1.51		1.54		1.58	1.59	

假设 H_0：两人分析结果无差异。

统计量为：

$$T=2+6.5+8+9+10+12+14+15.5=77$$

这里，1.46 和 1.58 两个值甲、乙均有，分别并列在第 6、7 位和第 15、16 位上，计算秩的可取平均数 6.5 和 15.5。

查附表 9 可得上下临界值($n_1=8$，$n_2=9$)

$$T_\alpha''(n_1,n_2)=T''_{0.05}(8,9)=90$$

$$T_\alpha'(n_1,n_2)=T'_{0.05}(8,9)=54$$

显然有：$54<77<90$

接受原假设 H_0，即认为两人分析结果无显著性差异。

第三章　方差分析

在环境科研和监测工作中，常常需要对测试结果进行分析，以判断各种测试因素是否对试验结果产生显著影响。例如测量某一污染物浓度时，不同的实验室、不同的测量仪器、不同的分析方法、不同的操作人员等种种因素都会对测试结果产生影响。方差分析就是判断这些影响是否显著的重要方法。

方差分析中将不同的实验室、不同的分析方法、不同的操作人员等称为方差分析中的因素，每个因素中可选取不同的浓度水平进行试验，这些浓度水平在方差分析中称为因素内的水平。在因素及水平确定的条件下，可进行一次或多次试验，试验次数在方差分析中称为重复数。通过因素、水平和重复数的不同组合，可以有单因素不同水平无重复、单因素不同水平有重复、多因素不同水平无重复和多因素不同水平有重复等不同组合形式。

方差分析的基本思想是：将测定数据的总变异（方差）分解为因素间的变异和因素内不同水平间的变异。通过比较因素在不同水平间的变异，分析不同水平的选取是否对测定结果产生影响。或者通过因素间的变异的比较分析各因素对分析结果产生的影响及因素间的交互作用。方差分析的目的是要确定是否存在影响测试结果的系统误差，即确定不同因素间或同一因素中不同水平间是否存在实质性的差异。需要指出的是，在方差分析中，假定所研究的对象都是服从正态分布的。这种假设显然未经证明，但考虑环境监测的测试结果受多种随机因素的影响，可以认为其总体的分布服从正态分布。此外，方差分析中假定各水平试验总体的方差都相等，尽管这些方差通常是未知的。

本章主要讨论方差分析中最简单、最基本同时也是应用最广的单因素方差分析和多因素方差分析。有了这方面的知识，也就有了对更多因素的方差分析的理解基础，需要时可进一步查阅有关方差分析的文献和专著。

第一节　单因素方差分析

只研究一个因素对试验结果有无显著影响时，是方差分析中最简单的情况，称为单因素方差分析。对所选择确定的试验因素，可选择不同的试验水平，每个水平内的重复数可以相同，也可以不相同。下面对单因素方差分析各水平内重复

数相同和重复数不相同的情况分别予以介绍。

一、各水平内重复数相等的单因素方差分析

设所选定的试验因素 A 有 k 个水平，每个水平进行了 n_i 次重复测定，n_i 即为各水平内的重复数。对各水平内重复数相等的单因素方差分析，显然应有$n_1=n_2=\cdots=n_k=n$。现在要检验 k 个水平测定结果的各个平均值间有无显著差异，即判别不同水平的选取是否会对测试结果产生实质性的影响。若测定结果以 x_{ij} 表示，下标 i 表示水平号，j 表示重复数号，则 x_{ij} 表示第 i 个水平的第 j 次测定结果。当我们选定水平数为 k，重复数为 n 时，则显然有 $i=1, 2, \cdots, k$；$j=1, 2, \cdots, n$。

单因素重复数相等的方差分析步骤如下。

(1) 建立统计假设。

假设 H_0：各水平的总体平均值无显著差异，即 $\mu_1=\mu_2=\cdots=\mu_k$；

假设 H_1：各水平的总体平均值有显著差异。

(2) 选择显著性水平 α，通常选择 $\alpha=0.01$ 或 $\alpha=0.05$。

(3) 计算组间均方和组内均方。

组间均方记作 S_1：

$$S_1=\frac{1}{k-1}\sum_{i=1}^{k}n(\bar{x}_i-\bar{\bar{x}})^2$$

组内均方记作 S_2：

$$S_2=\frac{1}{k(n-1)}\sum_{i=1}^{k}\sum_{j=1}^{n}(x_{ij}-\bar{x}_i)^2$$

上两式中：

$$\bar{x}_i=\frac{1}{n}\sum_{j=1}^{n}x_{ij};\quad \bar{\bar{x}}=\frac{1}{kn}\sum_{i=1}^{k}\sum_{j=1}^{n}x_{ij}$$

(4) 计算统计量

$$F=\frac{S_1}{S_2}$$

(5) 计算自由度

$$df_1=k-1$$

$$df_2=k(n-1)$$

(6) 确定临界值 F_α　由选定的显著性水平 α，查 F 检验的临界值表(附表 5)，得自由度为 $[k-1, k(n-1)]$ 的临界值 F_α。

(7) 统计判别。

若 $F\leqslant F_\alpha$，则在选定的显著性水平 α 条件下，接受假设 H_0，即认为各水平

总体的平均值无显著差异。

若 $F>F_{\alpha}$，则在选定的显著性水平 α 条件下，否定假设 H_0，接受备选假设 H_1，即认为各水平总体的平均值存在显著差异，在实际工作中有 $(1-\alpha)\%$ 的把握认为该因素对试验的结果有显著的影响。

在方差分析中，计算各种方差贡献需要进行各种累加平方和的计算，在实验数据量较大的时候，容易发生差错。为了使计算过程条理清楚，避免差错，方差分析通常通过列表方式进行。表 3-1 和表 3-2 是单因素不同水平重复数相等的实验数据计算和方差分析表。

表 3-1　单因素不同水平重复数相等的实验数据计算表

水平	实验数据	$\sum_{j=1}^{n} x_{ij}$	$\left(\sum_{j=1}^{n} x_{ij}\right)^2$	$\sum_{j=1}^{n} x_{ij}^2$
1	$x_{11},x_{12},\cdots,x_{1n}$	$\sum_{j=1}^{n} x_{1j}$	$\left(\sum_{j=1}^{n} x_{1j}\right)^2$	$\sum_{j=1}^{n} x_{1j}^2$
2	$x_{21},x_{22},\cdots,x_{2n}$	$\sum_{j=1}^{n} x_{2j}$	$\left(\sum_{j=1}^{n} x_{2j}\right)^2$	$\sum_{j=1}^{n} x_{2j}^2$
⋮	⋮	⋮	⋮	⋮
k	$x_{k1},x_{k2},\cdots,x_{kn}$	$\sum_{j=1}^{n} x_{kj}$	$\left(\sum_{j=1}^{n} x_{kj}\right)^2$	$\sum_{j=1}^{n} x_{kj}^2$
$\sum$		$\sum_{i=1}^{k}\sum_{j=1}^{n} x_{ij}$	$\sum_{i=1}^{k}\left(\sum_{j=1}^{n} x_{ij}\right)^2$	$\sum_{i=1}^{k}\sum_{j=1}^{n} x_{ij}^2$

记：

$$P=\frac{1}{kn}\left(\sum_{i=1}^{k}\sum_{j=1}^{n} x_{ij}\right)^2$$

$$Q=\frac{1}{n}\sum_{i=1}^{k}\left(\sum_{j=1}^{n} x_{ij}\right)^2$$

$$R=\sum_{i=1}^{k}\sum_{j=1}^{n} x_{ij}^2$$

$S_1=\frac{1}{k-1}(Q-P)$ 为组间方差来源，

$S_2=\frac{1}{k(n-1)}(R-Q)$ 为组内方差来源，

$S=\frac{1}{kn-1}(R-P)$ 为总方差。

表 3-2 单因素不同水平重复数相等的方差分析表

方差来源	自由度	均方	统计量	临界值	统计判别
组间	$k-1$	S_1	$F=\frac{S_1}{S_2}$	$F_\alpha[k-1,k(n-1)]$	若 $F<F_\alpha$，接受 H_0 若 $F\geqslant F_\alpha$，否定 H_0
组内	$k(n-1)$	S_2			

【例 3-1】 实验室质量控制工作中，令 4 个操作人员对同一环境水样的铜元素含量进行 10 次重复测定，测定结果见表 3-3。

表 3-3 ［例 3-1］中铜元素含量测定结果/(μg/L)

操作者＼重复数	1	2	3	4	5	6	7	8	9	10
甲	22.0	22.5	21.7	23.1	22.8	21.5	21.4	22.9	23.5	21.2
乙	21.8	20.9	22.7	21.2	20.2	20.7	21.1	22.0	21.5	20.6
丙	21.9	23.2	23.8	22.9	24.0	22.8	21.2	22.7	23.4	23.8
丁	22.1	22.8	21.6	21.7	22.4	23.0	23.2	22.0	21.8	23.0

操作人员使用同一套测量仪器和测量方法。试用 0.05 的置信水平判断操作人员是否对测定结果有显著影响？

解 这显然是单因素方差分析的问题。

水平数 $k=4$

重复数 $n=10$，重复数相同。

假设 H_0：各水平的总体平均值无显著差异。

显著性水平 $\alpha=0.05$

由测定结果（表 3-3）可计算得：

$$P=\frac{1}{kn}\left(\sum_{i=1}^{k}\sum_{j=1}^{n}x_{ij}\right)^2=19740.25$$

$$Q=\frac{1}{n}\sum_{i=1}^{k}\left(\sum_{j=1}^{n}x_{ij}\right)^2=19755.11$$

$$R=\sum_{i=1}^{k}\sum_{j=1}^{n}x_{ij}^2=19776.24$$

$$S_1=\frac{1}{k-1}(Q-P)=4.953$$

$$S_2=\frac{1}{k(n-1)}(R-Q)=0.587$$

统计量为：$F=\frac{S_1}{S_2}=8.44$

自由度：$df_1=k-1=3$

$$df_2=k(n-1)=36$$

由附表 5 可查得：

$$F_{0.05}(3,36)=2.87$$

显然有：$F=8.44>F_\alpha(df_1,df_2)=2.87$

即在显著性水平 $\alpha=0.05$ 下否定原假设 H_0。也就是说操作人员对实验结果有显著性影响。

二、各水平内重复数不相等的单因素方差分析

单因素方差分析中除了各水平内重复数相等的情况外，有时由于样品破坏或丢失，离群值的剔除等原因，使各水平内重复数不完全相同，这时需要采用各水平内重复数不相等的方差分析方法。各水平内重复数不相等的单因素方差分析的方法和步骤与重复数相同的单因素方差分析基本相同。惟一的区别是由于重复数 n_i 不全相等，使计算公式略有差异。重复数不相等的单因素方差分析的计算及分析步骤如表 3-4 和表 3-5 所示。

表 3-4　单因素不同水平重复数不相同的实验数据计算表

水平	实验数据	重复次数	$\sum_{j=1}^{n_i}x_{ij}$	$\frac{1}{n_i}\left(\sum_{j=1}^{n_i}x_{ij}\right)^2$	$\sum_{j=1}^{n_i}x_{ij}^2$
1	$x_{11},x_{12},\cdots,x_{1n_1}$	n_1	$\sum_{j=1}^{n_1}x_{1j}$	$\frac{1}{n_1}\left(\sum_{j=1}^{n_1}x_{1j}\right)^2$	$\sum_{j=1}^{n_1}x_{1j}^2$
2	$x_{21},x_{22},\cdots,x_{2n_2}$	n_2	$\sum_{j=1}^{n_2}x_{2j}$	$\frac{1}{n_2}\left(\sum_{j=1}^{n_2}x_{2j}\right)^2$	$\sum_{j=1}^{n_2}x_{2j}^2$
⋮	⋮	⋮	⋮	⋮	⋮
k	$x_{k1},x_{k2},\cdots,x_{kn_k}$	n_k	$\sum_{j=1}^{n_k}x_{kj}$	$\frac{1}{n_k}\left(\sum_{j=1}^{n_k}x_{kj}\right)^2$	$\sum_{j=1}^{n_k}x_{kj}^2$
$\sum$		$N=\sum_{i=1}^{k}n_i$	$\sum_{i=1}^{k}\sum_{j=1}^{n_i}x_{ij}$	$\sum_{i=1}^{k}\frac{1}{n_i}\left(\sum_{j=1}^{n_i}x_{ij}\right)^2$	$\sum_{i=1}^{k}\sum_{j=1}^{n_i}x_{ij}^2$

类似地，记：

$$P=\frac{1}{N}\left(\sum_{i=1}^{k}\sum_{j=1}^{n_i}x_{ij}\right)^2$$

$$Q=\sum_{i=1}^{k}\frac{1}{n_i}\left(\sum_{j=1}^{n_i}x_{ij}\right)^2$$

$$R=\sum_{i=1}^{k}\sum_{j=1}^{n_i}x_{ij}^2$$

表 3-5 单因素不同水平重复数不相等的方差分析表

方差来源	均方	自由度	统计量	临界值	统计判断
组间	$S_1=\frac{Q-P}{k-1}$	$k-1$			若 $F<F_\alpha$，接受 H_0
组内	$S_2=\frac{R-Q}{n-k}$	$n-k$	$F=\frac{S_1}{S_2}$	$F_\alpha(k-1,n-k)$	若 $F\geqslant F_\alpha$，否定 H_0
总和	$S=\frac{R-P}{n-1}$	$n-1$			

【例 3-2】 5 种不同的管道消声器，消声量的测定结果见表 3-6。

表 3-6 ［例 3-2］管道消声器消声量测定结果

消声器 \ 重复数	1	2	3	4	5	6	7
甲	17.1	18.3	17.5	17.7			
乙	17.4	18.4	16.8	17.9	18.8		
丙	19.2	18.1	18.4	17.3			
丁	16.5	16.8	17.8	18.5	17.4	16.1	
戊	17.3	18.0	17.5				

试判断这 5 种管道消声器的消声量有无显著性的差异？

解 本例中消声器的结构及消声原理的不同可看作是一个因素的不同水平情况。因而本例是单因素不同水平重复数不相等的方差分析问题。

假设 H_0：5 种消声器之消声量无显著差异，

取显著性水平 $\alpha=0.05$

水平数 $k=5$

总重复数 $n=22$

由测量结果（表 3-6）可计算得：

$$P=\frac{1}{n}\left(\sum_{i=1}^{k}\sum_{j=1}^{n_i}x_{ij}\right)^2=6871.16$$

$$R=\sum_{i=1}^{k}\sum_{j=1}^{n_i}x_{ij}^2=6883.44$$

$$Q=\sum_{i=1}^{k}\frac{1}{n_i}\left(\sum_{j=1}^{n_i}x_{ij}\right)^2=6874.12$$

$$S_1=\frac{Q-P}{k-1}=0.7399$$

$$S_2=\frac{R-Q}{n-k}=0.5482$$

统计量为：

$$F=\frac{S_1}{S_2}=1.35$$

自由度：

$$df_1=k-1=4$$

$$df_2=n-k=17$$

查附表 5 可得临界值

$$F_{0.05}(4,17)=2.96$$

显然有：$F=1.35<F_{0.05}(4,17)=2.96$

即在 95%的显著性水平之下接受原假设 H_0，也就是说这 5 种消声器的消声量大致相同。

第二节　双因素方差分析

在双因素方差分析中，我们研究的是在试验中存在两个变化因素的问题。需要考虑的问题是每种因素对试验结果产生的影响；两种因素中，哪种因素的影响更大一些；因素和因素之间是否存在交互作用等问题。

设 A 和 B 为独立可变化的因素，因素 A 分为 l 个水平，因素 B 分为 m 个水平，现考虑因素 A 和因素 B 对试验结果的影响。在双因素条件下，试验共有 $l\times m$ 种可能的组合。根据每种组合选择重复试验次数不同，可以有无重复双因素方差分析，重复数相等的双因素方差分析和重复数不相等的双因素方差分析等。下面分别予以介绍。

一、无重复双因素方差分析

无重复双因素方差分析是双因素方差分析中最简单的情况。设因素 A 分为 l 个水平，因素 B 分为 m 个水平，x_{ij} 表示 A 因素处于 i 水平，B 因素处于 j 水平的试验结果。全部试验结果的总方差可分解为三个部分：因素 A 各水平间方差，因素 B 各水平间方差以及随机方差，分别记作 S_A、S_B、S_E。通过统计分析，可以判别因素 A 或因素 B 是否对试验结果产生显著的影响。无重复双因素方差分析的具体步骤如下。

(1) 建立统计假设。

假设 H_0：因素 A（或因素 B）对试验结果没有显著的影响。

假设 H_1：因素 A（或因素 B）对试验结果有显著的影响。

(2) 选择显著性水平 α，一般选 $\alpha=0.05$ 或 $\alpha=0.01$。

(3) 计算三种方差　按表 3-7 及表 3-8 的方法计算 S_A、S_B 及 S_E。

(4) 计算统计量

$$F_A=\frac{S_A}{S_E} \qquad F_B=\frac{S_B}{S_E}$$

(5) 选择确定临界值并作统计判别　根据确定的显著性水平 α，并根据相应的自由度查附表 5 得临界值。

表 3-7　无重复双因素实验数据计算表

因素 A 的水平	因素 B 的水平	$\sum_{j=1}^{m}x_{ij}$　$\left(\sum_{j=1}^{m}x_{ij}\right)^2$　$\sum_{j=1}^{m}x_{ij}^2$
	$1,2,\cdots,j,\cdots,m$	
1	$x_{11},x_{12},\cdots,x_{1j},\cdots,x_{1m}$	$\sum_{j=1}^{m}x_{1j}$　$\left(\sum_{j=1}^{m}x_{1j}\right)^2$　$\sum_{j=1}^{m}x_{1j}^2$
2	$x_{21},x_{22},\cdots,x_{2j},\cdots,x_{2m}$	$\sum_{j=1}^{m}x_{2j}$　$\left(\sum_{j=1}^{m}x_{2j}\right)^2$　$\sum_{j=1}^{m}x_{2j}^2$
⋮	⋮	⋮
i	$x_{i1},x_{i2},\cdots,x_{ij},\cdots,x_{im}$	$\sum_{j=1}^{m}x_{ij}$　$\left(\sum_{j=1}^{m}x_{ij}\right)^2$　$\sum_{j=1}^{m}x_{ij}^2$
⋮	⋮	⋮
l	$x_{l1},x_{l2},\cdots,x_{lj},\cdots,x_{lm}$	$\sum_{j=1}^{m}x_{lj}$　$\left(\sum_{j=1}^{m}x_{lj}\right)^2$　$\sum_{j=1}^{m}x_{lj}^2$
		$\sum_{i=1}^{l}\sum_{j=1}^{m}x_{ij}$　$\sum_{i=1}^{l}\left(\sum_{j=1}^{m}x_{ij}\right)^2$　$\sum_{i=1}^{l}\sum_{j=1}^{m}x_{ij}^2$
$\left(\sum_{i=1}^{l}x_{i1}\right)^2\left(\sum_{i=1}^{l}x_{i2}\right)^2\cdots\left(\sum_{i=1}^{l}x_{ij}\right)^2\cdots\left(\sum_{i=1}^{l}x_{im}\right)^2$		$\sum_{j=1}^{m}\left(\sum_{i=1}^{l}x_{ij}\right)^2$

记：

$$P=\frac{1}{lm}\left(\sum_{i=1}^{l}\sum_{j=1}^{m}x_{ij}\right)^2$$

$$Q=\frac{1}{m}\sum_{i=1}^{l}\left(\sum_{j=1}^{m}x_{ij}\right)^2$$

$$R=\frac{1}{l}\sum_{j=1}^{m}\left(\sum_{i=1}^{l}x_{ij}\right)^2$$

$$T=\sum_{i=1}^{l}\sum_{j=1}^{m}x_{ij}^2$$

表 3-8　无重复双因素方差分析表

方差来源	自由度	均方	统计量	临界值	统计判别
因素 A	$m-1$	$S_A=\frac{Q-P}{l-1}$	$F_A=\frac{S_A}{S_E}$	$F_\alpha[l-1,(l-1)(m-1)]$	若 $F_A\leqslant F_\alpha$，接受 H_0
因素 B	$m-1$	$S_B=\frac{R-P}{m-1}$	$F_B=\frac{S_B}{S_E}$	$F_\alpha[m-1,(l-1)(m-1)]$	若 $F_B\leqslant F_\alpha$，接受 H_0
随机因素	$(l-1)\cdot(m-1)$	$S_E=\frac{(T-Q-R+P)}{(l-1)(m-1)}$			
总和	$lm-1$	$S_T=\frac{T-P}{lm-1}$			

【例 3-3】 在酸雨成因的研究中，为观察大气湿度和紫外光强对二氧化硫转化的影响，选取三种不同的紫外光照射强度和四种不同的大气湿度进行二氧化硫转化实验，实验结果见表 3-9。

表 3-9 ［例 3-3］中 SO_2 转化实验结果

光强＼湿度	1	2	3	4
1	4.8	5.1	5.7	6.5
2	5.3	5.6	6.4	7.8
3	6.0	6.3	6.9	9.3

试用方差分析，在显著性水平 0.05 下，检验紫外光强和大气湿度对二氧化硫的转化是否有显著影响？

解　首先建立统计假设。

假设 H_0：紫外光强和大气湿度两因素对二氧化硫的转化均无显著影响。

已知 $l=3$，$m=4$，方差计算如下：

$$P=\frac{1}{3\times4}\left(\sum_{i=1}^{3}\sum_{j=1}^{4}x_{ij}\right)^2=477.54$$

$$Q=\frac{1}{4}\sum_{i=1}^{3}\left(\sum_{j=1}^{4}x_{ij}\right)^2=482.67$$

$$R=\frac{1}{3}\sum_{j=1}^{4}\left(\sum_{i=1}^{3}x_{ij}\right)^2=488.72$$

$$T=\sum_{i=1}^{3}\sum_{j=1}^{4}x_{ij}^2=494.83$$

$$S_A=\frac{1}{2}(Q-P)=5.13$$

$$S_B=\frac{1}{3}(R-P)=11.18$$

$$S_E=\frac{1}{6}(T+P-Q-R)=0.163$$

统计量为：

$$F_A=\frac{S_A}{S_E}=31.47$$

$$F_B=\frac{S_B}{S_E}=68.59$$

自由度为：

$$df_A=l-1=2$$
$$df_B=m-1=3$$
$$df_E=(l-1)(m-1)=6$$

显著性水平为：$\alpha=0.05$

由附表 5 查得：

$$F_{0.05}(2,6)=5.14$$
$$F_{0.05}(3,6)=4.76$$

显然有：$F_A=31.47>F_{0.05}(2,6)=5.14$

$$F_B=68.59>F_{0.05}(3,6)=4.76$$

因而否定原假设 H_0，即紫外光照强度和大气湿度均对二氧化硫的转化有显著的影响。

二、重复数相等的双因素方差分析

实践表明，因素 A 和因素 B 对试验结果的影响，除了每个因素单独对试验结果起作用外，两因素各水平之间的搭配还会对试验结果起作用，通常称为交互作用。

设因素 A 分成 l 个水平，因素 B 分成 m 个水平。在因素 A 的任一水平 i 和因素 B 的任一水平 j 的配合条件下，做等重复 n 次试验。则全部试验共有 lmn 个数据，每个数据记作 x_{ijk}。全部试验结果的总方差 S_T 可分解成四个部分：因素 A 的水平间方差 S_A，因素 B 的水平间方差 S_B，因素 A 和因素 B 的交互作用方差 $S_{A\times B}$和随机方差 S_E。通过统计分析可判别因素 A 或因素 B 或因素 A×B 是否对试验结果产生显著的影响。

因素 A 或因素 B 对试验结果影响的含义较清楚，它是指因素 A 或因素 B 单独的效应。当因素 A 和因素 B 间不存在交互作用的时候，因素 A 对试验结果的影响与因素 B 取什么水平无关。同样因素 B 对试验结果的影响与因素 A 取什么水平无关。当因素 A 和因素 B 之间存在交互作用的时候，除了因素 A 和因素 B

单独的效应外，还存在交互作用效应，即因素 A 对试验结果的影响与因素 B 取什么水平有关，同样因素 B 对试验结果的影响与因素 A 取什么水平有关。因素之间的交互作用效应可正可负，即因素 A 与因素 B 组合时的效应可以比无交互作用时大，也可以比无交互作用时小。

重复数相等的双因素方差分析的步骤如下。

(1) 建立统计假设。

假设 H_0：因素 A（或因素 B，或因素 A×B）对试验结果无显著影响。

假设 H_1：因素 A（或因素 B，或因素 A×B）对试验结果有显著影响。

(2) 选择显著性水平 α，一般选 $\alpha=0.05$ 或 $\alpha=0.01$。

(3) 计算统计量　按数据计算表 3-10、表 3-11 和方差分析表 3-12 进行。

(4) 选择确定临界值并做统计判断，按表 3-12 进行。

表 3-10　双因素等重复实验数据计算表（Ⅰ）

A	B	试验结果(重复 n 次)	$x_{ij}=\sum_{k=1}^{n}x_{ijk}$	$x_{ij}^2=\left(\sum_{k=1}^{n}x_{ijk}\right)^2$	$\sum_{k=1}^{n}(x_{ijk})^2$
A_1	B_1	$x_{111},x_{112},\cdots,x_{11n}$	x_{11}	x_{11}^2	$\sum_{k=1}^{n}(x_{11k})^2$
	B_2	$x_{121},x_{122},\cdots,x_{12n}$	x_{12}	x_{12}^2	$\sum_{k=1}^{n}(x_{12k})^2$
	⋮	⋮	⋮	⋮	⋮
	B_m	$x_{1m1},x_{1m2},\cdots,x_{1mn}$	x_{1m}	x_{1m}^2	$\sum_{k=1}^{n}(x_{1mk})^2$
⋮	⋮	⋮	⋮	⋮	⋮
A_l	B_1	$x_{l11},x_{l12},\cdots,x_{l1n}$	x_{l1}	x_{l1}^2	$\sum_{k=1}^{n}(x_{l1k})^2$
	B_2	$x_{l21},x_{l22},\cdots,x_{l2n}$	x_{l2}	x_{l2}^2	$\sum_{k=1}^{n}(x_{l2k})^2$
	⋮	⋮	⋮	⋮	⋮
	B_m	$x_{lm1},x_{lm2},\cdots,x_{lmn}$	x_{lm}	x_{lm}^2	$\sum_{k=1}^{n}(x_{lmk})^2$
			$\sum_{i=1}^{l}\sum_{j=1}^{m}x_{ij}$	$\sum_{i=1}^{l}\sum_{j=1}^{m}x_{ij}^2$	$\sum_{i=1}^{l}\sum_{j=1}^{m}\sum_{k=1}^{n}(x_{ijk})^2$

表 3-11 双因素等重复实验数据计算表（Ⅱ）

A 的水平	$B_1, B_2, \cdots, B_m$		
A_1	$x_{11}, x_{12}, \cdots, x_{1m}$	$\sum_{j=1}^{m} x_{1j}$	$\left(\sum_{j=1}^{m} x_{1j}\right)^2$
A_2	$x_{21}, x_{22}, \cdots, x_{2m}$	$\sum_{j=1}^{m} x_{2j}$	$\left(\sum_{j=1}^{m} x_{2j}\right)^2$
⋮	⋮	⋮	⋮
A_l	$x_{l1}, x_{l2}, \cdots, x_{lm}$	$\sum_{j=1}^{m} x_{lj}$	$\left(\sum_{j=1}^{m} x_{lj}\right)^2$
	$\sum_{i=1}^{l} x_{i1}, \sum_{i=1}^{l} x_{i2}, \cdots, \sum_{i=1}^{l} x_{im}$		$\sum_{i=1}^{l}\left(\sum_{j=1}^{m} x_{ij}\right)^2$
	$\left(\sum_{i=1}^{l} x_{i1}\right)^2, \left(\sum_{i=1}^{l} x_{i2}\right)^2, \cdots, \left(\sum_{i=1}^{l} x_{im}\right)^2$	$\sum_{j=1}^{m}\left(\sum_{i=1}^{l} x_{ij}\right)^2$	

表 3-12 双因素等重复方差分析表

方差来源	平方和	自由度	均方	统计量	临界值	统计判别
因素 A	$SS_A = Q - P$	$l-1$	$S_A = \frac{SS_A}{l-1}$	$F_A = \frac{S_A}{S_E}$	$F_\alpha[l-1, lm(n-1)]$	$F_A > F_\alpha$，否定假设 H_0，接受假设 H_1；$F_A \leqslant F_\alpha$，接受假设 H_0
因素 B	$SS_B = R - P$	$m-1$	$S_B = \frac{SS_B}{m-1}$	$F_B = \frac{S_B}{S_E}$	$F_\alpha[m-1, lm(n-1)]$	$F_B > F_\alpha$，否定假设 H_0，接受假设 H_1；$F_B \leqslant F_\alpha$，接受假设 H_0
因素 A×B	$SS_{A\times B} = T - Q - R + P$	$(l-1)(m-1)$	$S_{A\times B} = \frac{SS_{A\times B}}{(l-1)(m-1)}$	$F_{A\times B} = \frac{S_{A\times B}}{S_E}$	$F_\alpha[(l-1)(m-1), lm(n-1)]$	$F_{A\times B} > F_\alpha$，否定假设 H_0，接受假设 H_1；$F_{A\times B} \leqslant F_\alpha$，接受假设 H_0
随机因素	$SS_E = W - T$	$lm(n-1)$	$S_E = \frac{SS_E}{lm(n-1)}$			
总和	$SS_T = W - P$	$lmn-1$				

记：

$$P=\frac{1}{lmn}\left(\sum_{i=1}^{l}\sum_{j=1}^{m}x_{ij}\right)^2$$

$$T=\frac{1}{n}\sum_{i=1}^{l}\sum_{j=1}^{m}x_{ij}^2$$

$$W=\sum_{i=1}^{l}\sum_{j=1}^{m}\sum_{k=1}^{n}x_{ijk}^2$$

$$Q=\frac{1}{mn}\sum_{i=1}^{l}\left(\sum_{j=1}^{m}x_{ij}\right)^2$$

$$R=\frac{1}{ln}\sum_{j=1}^{m}\left(\sum_{i=1}^{l}x_{ij}\right)^2$$

代入表 3-12，即可进行双因素等重复方差分析。

【例 3-4】　在［例 3-3］的酸雨成因研究中，若光强与湿度的每一交叉水平重复两次试验，结果见表 3-13。

表 3-13　［例 3-4］中 SO_2 转化实验结果

光强＼湿度	1	2	3	4
1	4.9　4.6	4.8　5.3	5.7　5.8	6.5　6.3
2	5.2　5.4	5.9　5.5	6.3　6.7	8.1　7.4
3	6.3　5.8	6.1　6.5	6.8　6.9	9.0　9.7

试用方差分析，在显著性水平 0.05 下，检验紫外光强和大气湿度两因素对二氧化硫转化是否有显著影响，以及两因素是否存在交互作用？

解　这显然是一个双因素等重复分析问题，重复数 $n=2$。

已知 $l=3$，$m=4$

假设 H_0：紫外光强和大气湿度两因素对二氧化硫转化无显著影响，两因素无交互作用。

按表 3-10、表 3-11、表 3-12 的方法，方差计算如下：

$$P=\frac{1}{3\times4\times2}\left(\sum_{i=1}^{3}\sum_{j=1}^{4}x_{ij}\right)^2=956.34$$

$$T=\frac{1}{2}\sum_{i=1}^{3}\sum_{j=1}^{4}x_{ij}^2=991.24$$

$$W=\sum_{i=1}^{3}\sum_{j=1}^{4}\sum_{k=1}^{2}x_{ijk}^2=992.37$$

$$Q=\frac{1}{4\times 2}\sum_{i=1}^{3}\left(\sum_{j=1}^{4}x_{ij}\right)^2=967.23$$

$$R=\frac{1}{3\times 2}\sum_{j=1}^{4}\left(\sum_{i=1}^{3}x_{ij}\right)^2=977.98$$

$$S_A=\frac{1}{2}(Q-P)=5.45$$

$$S_B=\frac{1}{3}(R-P)=7.21$$

$$S_{A\times B}=\frac{1}{6}(T-Q-R+P)=0.395$$

$$S_E=\frac{1}{12}(W-T)=0.094$$

统计量为：

$$F_A=\frac{S_A}{S_E}=57.98$$

$$F_B=\frac{S_B}{S_E}=76.70$$

$$F_{A\times B}=\frac{S_{A\times B}}{S_E}=4.20$$

选定显著性水平 $\alpha=0.05$，查附表 5 得：

$$F_{0.05}(2,12)=3.89$$

$$F_{0.05}(3,12)=3.49$$

$$F_{0.05}(6,12)=3.00$$

统计判断，显然有：

$$F_A=57.98>F_{0.05}(2,12)=3.89$$

$$F_B=76.70>F_{0.05}(3,12)=3.49$$

$$F_{A\times B}=4.20>F_{0.05}(6,12)=3.00$$

否定原假设 H_0，即紫外光强和大气湿度显著影响二氧化硫转化，两因素没有相互加强的交互作用。

第三节　系统分组方差分析

在进行环境科研和监测调查时，经常采用按系统分层采样的调查方法。例如，在进行全国粮食中有机农药污染调查时，从全国各县中首先选取一批县，然后在这些县中选取一批乡，每个选中的乡内又选几个自然村，然后确定采样点采集样品进行分析。这种方法称为系统分组法。由于系统分组是在不同层次上进行

分组，不同层次因素并不是平行的，次一层因素的效应受高一层次分组的影响，因此系统分组方差分析与各因素平行的多因素方差分析有所不同。

设采样调查先按因素 A 分组为 A_1，A_2，…，A_l 个水平，然后在每个水平 A_i 中按因素 B 分组为 B_{i1}，B_{i2}，…，B_{im} 个水平，在因素 A_i 和因素 B_{ij} 组合条件下，做 n 次试验，得试验数据 x_{ijk}（$i=1,2,\cdots,l$；$j=1,2,\cdots,m$；$k=1,2,\cdots,n$），系统分组方差分析步骤如下。

（1）建立统计假设。

假设 H_0：在因素 A 与因素 B 的组合 $A_i\times B_{ij}$ 条件下，因素 A（或因素 B）影响不显著。

假设 H_1：在因素 A 与因素 B 的组合 $A_i\times B_{ij}$ 条件下，因素 A（或因素 B）对试验结果有显著影响。

（2）选择显著性水平 α，一般选 $\alpha=0.05$ 或 $\alpha=0.01$。

（3）计算统计量，见数据计算表 3-14 及方差分析表 3-15。

（4）选择确定临界值并做统计判断，见方差分析表 3-15。

表 3-14　系统分组实验数据计算表

A	B	试验结果(重复 n 次)	$x_{ij}=\sum_{k=1}^{n}x_{ijk}$	$\sum_{j=1}^{m}x_{ij}$	$\left(\sum_{j=1}^{m}x_{ij}\right)^2$	$x_{ij}^2=\left(\sum_{k=1}^{n}x_{ijk}\right)^2$	$\sum_{k=1}^{n}x_{ijk}^2$
A_1	B_{11}	$x_{111},x_{112},\cdots,x_{11n}$	x_{11}	$\sum_{j=1}^{m}x_{1j}$	$\left(\sum_{j=1}^{m}x_{1j}\right)^2$	x_{11}^2	$\sum_{k=1}^{n}x_{11k}^2$
	B_{12}	$x_{121},x_{122},\cdots,x_{12n}$	x_{12}			x_{12}^2	$\sum_{k=1}^{n}x_{12k}^2$
	⋮	⋮	⋮			⋮	⋮
	B_{1m}	$x_{1m1},x_{1m2},\cdots,x_{1mn}$	x_{1m}			x_{1m}^2	$\sum_{k=1}^{n}x_{1mk}^2$
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
A_l	B_{l1}	$x_{l11},x_{l12},\cdots,x_{l1n}$	x_{l1}	$\sum_{j=1}^{m}x_{lj}$	$\left(\sum_{j=1}^{m}x_{lj}\right)^2$	x_{l1}^2	$\sum_{k=1}^{n}x_{l1k}^2$
	B_{l2}	$x_{l21},x_{l22},\cdots,x_{l2n}$	x_{l2}			x_{l2}^2	$\sum_{k=1}^{n}x_{l2k}^2$
	⋮	⋮	⋮			⋮	⋮
	B_{lm}	$x_{lm1},x_{lm2},\cdots,x_{lmn}$	x_{lm}			x_{lm}^2	$\sum_{k=1}^{n}x_{lmk}^2$
			$\sum_{i=1}^{l}\sum_{j=1}^{m}x_{ij}$		$\sum_{i=1}^{l}\left(\sum_{j=1}^{m}x_{ij}\right)^2$	$\sum_{i=1}^{l}\sum_{j=1}^{m}x_{ij}^2$	$\sum_{i=1}^{l}\sum_{j=1}^{m}\sum_{k=1}^{n}x_{ijk}^2$

记：

$$P=\frac{1}{lmn}\left(\sum_{i=1}^{l}\sum_{j=1}^{m}x_{ij}\right)^2$$

$$Q=\frac{1}{mn}\sum_{i=1}^{l}\left(\sum_{j=1}^{m}x_{ij}\right)^2$$

$$T=\frac{1}{n}\sum_{i=1}^{l}\sum_{j=1}^{m}x_{ij}^2$$

$$W=\sum_{i=1}^{l}\sum_{j=1}^{m}\sum_{k=1}^{n}x_{ijk}^2$$

表 3-15 系统分组方差分析表

方差来源	平方和	自由度	均方	统计量	置信限	统计判别
因素 A	$SS_A=Q-P$	$l-1$	$S_A=\frac{SS_A}{l-1}$	$F_{AB}=\frac{S_A}{S_B}$	$F_\alpha[l-1,l(m-1)]$	$F_{AB}>F_\alpha$，否定假设 H_0，接受假设 H_1 $F_{AB}\leqslant F_\alpha$，接受假定 H_0
因素 B	$SS_B=T-Q$	$l(m-1)$	$S_B=\frac{SS_B}{l(m-1)}$	$F_B=\frac{S_B}{S_E}$	$F_\alpha[l(m-1),lm(n-1)]$	$F_B>F_\alpha$，否定假设 H_0，接受假设 H_1 $F_B\leqslant F_\alpha$，接受假定 H_0
随机因素	$SS_E=W-T$	$lm(n-1)$	$S_E=\frac{SS_E}{lm(n-1)}$			
总和	$SS_T=W-P$	$lmn-1$				

【例 3-5】 为了检测某一水系的 pH 值，将该水系分做了 3 个河段，每一河段取 5 个断面进行水质检测，每一断面重复 6 次采样检测，pH 值的检测结果如表 3-16 所示。

在显著性水平 $\alpha=0.05$ 之下，检测河水的 pH 值是否明显随断面或河段而改变。

表 3-16 ［例 3-5］中河水 pH 值检测结果

河段	A_1					A_2					A_3				
断面	B_{11}	B_{12}	B_{13}	B_{14}	B_{15}	B_{21}	B_{22}	B_{23}	B_{24}	B_{25}	B_{31}	B_{32}	B_{33}	B_{34}	B_{35}
1	8.50	8.25	8.13	8.17	8.24	8.21	8.10	8.10	7.80	8.05	7.51	7.80	7.40	7.57	7.33
2	8.45	8.46	8.08	8.08	8.06	8.07	8.03	8.00	7.83	7.94	7.98	7.75	7.53	7.85	8.00
3	8.41	8.46	8.22	8.16	8.22	8.16	7.64	8.05	7.95	7.43	7.80	7.43	7.65	7.86	8.00
4	8.42	8.39	8.31	8.13	8.21	8.42	7.96	8.01	7.90	7.98	8.10	7.92	7.56	7.65	8.05
5	8.36	8.35	8.28	8.24	8.08	8.28	8.01	8.11	8.08	8.04	8.04	7.71	7.95	7.98	7.91
6	8.44	8.52	8.25	8.20	8.38	8.35	8.18	8.08	8.18	8.19	7.95	8.07	7.83	7.74	8.09

解 首先做统计假设：

假设 H_0：水质的 pH 值不随断面或河段而改变。

已知：$l=3$，$m=5$，$n=6$

计算得：

$x_{11}=50.58$；$x_{12}=50.43$；$x_{13}=49.27$；$x_{14}=48.98$；$x_{15}=49.19$

$x_{21}=49.49$；$x_{22}=47.92$；$x_{23}=48.35$；$x_{24}=47.74$；$x_{25}=47.63$

$x_{31}=47.38$；$x_{32}=46.68$；$x_{33}=45.92$；$x_{34}=46.65$；$x_{35}=47.38$

$$P=\frac{1}{3\times5\times6}\left(\sum_{i=1}^{3}\sum_{j=1}^{5}x_{ij}\right)^2=5817.58$$

$$Q=\frac{1}{5\times6}\sum_{i=1}^{3}\left(\sum_{j=1}^{5}x_{ij}\right)^2=5821.06$$

$$T=\frac{1}{6}\sum_{i=1}^{3}\sum_{j=1}^{5}x_{ij}^2=5822.07$$

$$W=\sum_{i=1}^{3}\sum_{j=1}^{5}\sum_{k=1}^{6}x_{ijk}^2=5824.16$$

方差为：

$$S_A=\frac{1}{2}(Q-P)=1.74$$

$$S_B=\frac{1}{12}(T-Q)=0.084$$

$$S_E=\frac{1}{75}(W-T)=0.028$$

统计量为：

$$F_{AB}=\frac{S_A}{S_E}=62.14$$

$$F_B=\frac{S_B}{S_E}=3.00$$

显著性水平 $\alpha=0.05$，查附表 5 得临界值：

$$F_\alpha[l-1,l(m-1)]=F_{0.05}(2,12)=19.41$$

$$F_\alpha[l(m-1),lm(n-1)]=F_{0.05}(12,75)=2.36$$

显然有：

$$F_{AB}=62.14>F_{0.05}(2,12)=19.41$$

$$F_B=3.00>F_{0.05}(12,75)=2.36$$

否定原假设 H_0，即可以认为不同河段，不同断面水质的 pH 值显著不同。

第四章　回归分析

在对环境数据进行统计分析的时候，我们常常要探讨各种量之间的相互关系，建立各类变量之间的各种联系。一般而言，各种量之间的相互关系可以分成两大类，一类是确定性的联系，一类是非确定性的联系。确定性的联系一般以确定的函数关系表示，这种联系在各门学科中大量存在。例如，理想气体的温度 T、体积 V、压力 P 之间的联系是以如下确定的函数关系相联系的：$PV=nRT$。非确定性的联系，通常由于变量受一些随机因素的影响，使诸变量之间的关系不是惟一确定的。但是在这些变量之间存在一定的统计关系，从大量的试验或统计中，我们能找到这些变量之间的某种规律性。这种规律性虽然不是某种确定的因果关系，但是对于我们认识事物的规律性却是很有帮助的。回归分析正是我们解决这类问题的有用方法。由一个或一组随机变量来估计或预测另一个或一组随机变量的值所建立的数学模型及所做的统计分析称之为回归分析。这就是说，回归分析的任务是寻找诸变量之间所服从的统计关系或数学模型，并且要确定做出这种统计关系时的准确度有多大。

第一节　一元线性回归

一、一元线性回归方程的建立

一元线性回归是回归分析中最简单也是最常用的回归问题。一元线性回归要解决的是两个变量之间，并且这两个变量之间的关系表现为具有线性关系的问题。

确定两个变量之间的相互关系，最简单和最直观的方法是在坐标纸上作图。我们将两个变量中的一个变量 X 作为自变量，另一个变量 Y 作为因变量，对应两个变量的每一组数据（x_i，y_i）在图中以一个点表示，这种图称为散点图。从散点图中可以看出两个变量之间的大致关系。

【例 4-1】 某生产过程中，排放污染物的温度与浓度的实际测定值见表 4-1，由这些测定值作散点图，并观察温度和浓度之间的关系。

由表 4-1 实际测定的数据，在坐标纸上做成散点图，如图 4-1 所示。

表 4-1 ［例 4-1］中温度与浓度的关系

温度 x_i/℃	45	50	55	60	65	70	75	80	85	90
污染物浓度 y_i/%	43	45	48	51	55	57	59	63	66	68

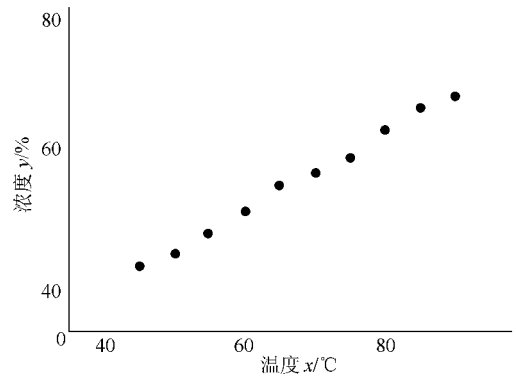


图 4-1 污染物排放浓度与温度关系

由图 4-1 可以看出，在所测得的实验数据范围内，污染物排放浓度与温度两个变量大致成直线关系。由散点图发现两变量具有某种直线关系，我们自然想到可以用直线方程来表示变量 X 和变量 Y 之间的关系：

$$\hat{Y}=a+bX \tag{4-1}$$

式（4-1）称为变量 X 和变量 Y 的回归方程，a、b 为回归方程中的参数，也称回归系数。由于对应于因变量 Y 的自变量 X 为一个，并且 X 和 Y 之间为线性关系，则式（4-1）称为一元线性回归方程。如果能够根据实际监测数据确定参数 a 和 b，那么直线方程也就完全确定了。但是我们知道平面上的直线有无穷多条，因此，可以做出很多条直线来表示变量 X 和变量 Y 之间的关系。因而我们必须建立一种方法能够从许多直线中找出一条最接近所有实验数据的直线，这条直线是我们所需要的回归直线。

从实验数据中得到最佳拟合直线，即回归直线的通常采用的方法是最小二乘法。用最小二乘法解得的回归直线来表示变量 X 和变量 Y 之间的关系时，所产生的误差比其他任何直线都要小。下面介绍如何用最小二乘法确定回归方程。

一般地说，自变量 X 与因变量 Y 的对应测定值可表示为：

X	X_1	X_2	…	X_n
Y	Y_1	Y_2	…	Y_n

表示变量 X 和变量 Y 间直线关系的回归方程可记为：

$$\hat{Y}_i=a+bX_i \tag{4-2}$$

则观察所得的测定值与根据方程（4-2）计算所得的计算值之间存在的误差 δ_i 可表示为：

$$\delta_i=Y_i-\hat{Y}_i=Y_i-a-bX_i \tag{4-3}$$

则对所有测定值与计算值之间的误差平方总和可表示为：

$$Q=\sum_{i=1}^{n}\delta_i^2=\sum_{i=1}^{n}(Y_i-a-bX_i)^2 \tag{4-4}$$

式（4-4）中，总误差之所以用平方和的形式表示是因为每个测定值与计算值之间的差值有正有负。若将误差单纯相加，由于正负误差抵消使所得到的总误差不能代表实际的总误差。而平方总和的形式能避免这个问题，并平方和的形式在数学运算中也较为简单，因而总误差用平方和的形式来表示。所谓最小二乘法，就是要求总误差在平方和最小意义下，得到回归方程（4-1）中的参数 a 和 b。

根据数学分析中求极值的原理，保证总误差平方和最小的回归方程中参数 a、b 应满足下列方程组：

$$\begin{cases}\dfrac{\partial Q}{\partial a}=0\\[2ex]\dfrac{\partial Q}{\partial b}=0\end{cases}\tag{4-5}$$

将方程（4-4）代入方程（4-5），可得：

$$\begin{cases}-2\sum\limits_{i=1}^{n}(Y_i-a-bX_i)=0\\[2ex]-2\sum\limits_{i=1}^{n}(Y_i-a-bX_i)X_i=0\end{cases}\tag{4-6}$$

解方程（4-6）得：

$$\begin{cases}a=\dfrac{\sum\limits_{i=1}^{n}Y_i}{n}-b\dfrac{\sum\limits_{i=1}^{n}X_i}{n}=\overline{Y}-b\overline{X}\\[2ex]b=\dfrac{\sum\limits_{i=1}^{n}X_iY_i-n\overline{X}\,\overline{Y}}{\sum\limits_{i=1}^{n}X_i^2-n\overline{X}^2}\end{cases}\tag{4-7}$$

式（4-7）中，$\overline{X}$ 和 $\overline{Y}$ 分别为 X 和 Y 实测值的平均值。

求解得到了回归系数 a 和 b，即确定了回归方程（4-1）。

在实际运算中，为了计算方便，可设

$$\begin{cases}l_{xx}=\sum\limits_{i=1}^{n}X_i^2-\dfrac{1}{n}\left(\sum\limits_{i=1}^{n}X_i\right)^2=\sum\limits_{i=1}^{n}X_i^2-n\overline{X}^2\\[2ex]l_{yy}=\sum\limits_{i=1}^{n}Y_i^2-\dfrac{1}{n}\left(\sum\limits_{i=1}^{n}Y_i\right)^2=\sum\limits_{i=1}^{n}Y_i^2-n\overline{Y}^2\\[2ex]l_{xy}=\sum\limits_{i=1}^{n}X_iY_i-\dfrac{1}{n}\left(\sum\limits_{i=1}^{n}X_i\right)\left(\sum\limits_{i=1}^{n}Y_i\right)=\sum\limits_{i=1}^{n}X_iY_i-n\overline{X}\,\overline{Y}\end{cases}\tag{4-8}$$

则有：

$$\begin{cases} b=\dfrac{l_{xy}}{l_{xx}} \\ a=\overline{Y}-b\overline{X} \end{cases} \tag{4-9}$$

方程（4-9）具体计算步骤通常是列表进行的。以［例 4-1］为例，具体介绍回归方程中回归系数的计算过程（见表 4-2）。

表 4-2 回归直线方程计算表

样品序号	温度/℃	浓度/%	X^2	Y^2	XY
1	45	43	2025	1849	1935
2	50	45	2500	2025	2250
3	55	48	3035	2304	2640
4	60	51	3600	2601	3060
5	65	55	4225	3025	3575
6	75	59	5625	3481	4425
7	70	57	4900	3249	3990
8	80	63	6400	3969	5040
9	85	66	7225	4356	5610
10	90	68	8100	4624	6120
Σ	675	555	47625	31483	38645
均值	67.5	55.5			

则由式（4-8）有：

$$l_{xy}=38645-10\times 67.5\times 55.5=1182.5$$

$$l_{xx}=47625-10\times 67.5^2=2062.5$$

$$l_{yy}=31483-10\times 55.5^2=680.5$$

将上述结果代入式（4-9），得：

$$\begin{cases} b=\dfrac{l_{xy}}{l_{xx}}=\dfrac{1182.5}{2062.5}=0.573 \\ a=\overline{Y}-b\overline{X}=55.5-0.573\times 67.5=16.82 \end{cases}$$

则最终得可回归方程：

$$Y=16.82-0.573X \tag{4-10}$$

二、一元线性回归方程的统计检验

前面我们通过计算建立了两个变量之间的一元线性回归方程。从计算过程中，细心的读者可以注意到，回归方程的建立对于变量 X 和 Y 之间的相互关系，并没有任何要求。这就是说，回归方程的建立过程，并没有解决变量 X 和 Y 之

间是否存在统计意义下的真实的线性相关关系问题。为了检验判断变量 X 和 Y 之间是否确实存在线性相关关系，需要对回归方程进行统计检验。

1. 相关系数及其显著性检验

相关系数是反映两变量之间线性相关程度的量，通常用字母 γ 表示。从统计意义上讲，当两变量之间的线性相关程度大于某一程度时，才能认为所得到的回归方程在统计上是有意义的，因此，相关系数可以作为一个指标判断回归方程在统计上是否有意义。

记变量 X 和 Y 的 n 对测定值为 X_i 和 Y_i（$i=1, 2, \cdots, n$），其平均值为 $\overline{X}$ 和 $\overline{Y}$，变量 X 和 Y 间相关系数的定义为：

$$\gamma = \frac{\sum_{i=1}^{n}(X_i-\overline{X})(Y_i-\overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\overline{X})^2 \cdot \sum_{i=1}^{n}(Y_i-\overline{Y})^2}} \tag{4-11}$$

为了计算方便，式（4-11）可以表示为：

$$\gamma=\frac{l_{xy}}{\sqrt{l_{xx}l_{yy}}} \tag{4-12}$$

相关系数的取值范围是 $0\leqslant|\gamma|\leqslant 1$，当 $|\gamma|=1$ 时，所有的测定值全部落在回归直线上，称为完全线性相关；当 $|\gamma|=0$ 时，所有测定值在散点图上毫无规则的分布，称全无线性相关。$|\gamma|$ 值越接近 1，变量 X 和 Y 的线性相关程度越大。

附表 2 给出了在不同显著水平下，判别两变量之间线性相关的临界值。当计算所得相关系数 $|\gamma|$ 大于表中相应的值时，所建立的回归方程才有意义。γ 的值可能为正值或负值，只要其绝对值大于表中相应的判别值时，可认为两变量 X 和 Y 是显著相关的。只是当 γ 为正值时，称 X 和 Y 正相关；γ 为负值时，称 X 和 Y 负相关。

对用［例 4-1］中的数据建立的回归方程（4-10），用相关系数检验温度和浓度两变量之间的相关性。

（1）建立统计假设　假设 H_0：变量 X（温度）和 Y（浓度）线性相关。

（2）选择确定显著性水平 α，选 $\alpha=0.01$。

（3）计算统计量

$$\gamma=\frac{l_{xy}}{\sqrt{l_{xx}l_{yy}}}=\frac{1182.5}{\sqrt{2062.5\times 680.5}}=0.998$$

（4）确定临界值 γ_α　在显著性水平 $\alpha=0.01$，自由度 $n-2=8$ 时，查附表 2，得临界值 $\gamma_\alpha=0.765$。

（5）统计判别　由于 $\gamma=0.998$ 大于临界值 $\gamma_\alpha=0.765$，则在显著性水平 $\alpha=0.01$ 的条件下，接受假设 H_0，即变量 X（温度）和 Y（浓度）是显著线性相关的。

2. 回归方程的统计检验

为了说明回归方程在统计上是否有意义，需要对回归方程进行统计检验。对回归方程进行统计检验的思路是比较所得到的回归直线与实际观察值的接近程度。为了定量地判别实际观察值与回归直线的接近程度，我们用因变量 Y 的观察值与平均值之差 $Y_i-\overline{Y}$，即离差来表示因变量的波动程度。总的波动程度可用所有观察值的离差平方和表示：

$$S_{总}=\sum_{i=1}^{n}(Y_i-\overline{Y})^2 \tag{4-13}$$

式中，$S_{总}$ 称为因变量 Y 的总离差平方和，总离差平方和 $S_{总}$ 是由两个方面原因引起的。一是自变量 X 的取值不同；另一原因是实验观察过程中，各种其他因素的影响。因此我们可以将总离差平方和 $S_{总}$ 分解成两个部分。

$$\begin{aligned}S_{总}&=\sum_{i=1}^{n}(Y_i-\overline{Y})^2\\&=\sum_{i=1}^{n}\left[\left(Y_i-\hat{Y}_i\right)+\left(\hat{Y}_i-\overline{Y}\right)\right]^2\\&=\sum_{i=1}^{n}\left(Y_i-\hat{Y}_i\right)^2+\sum_{i=1}^{n}\left(\hat{Y}_i-\overline{Y}\right)^2+2\sum_{i=1}^{n}\left(Y_i-\hat{Y}_i\right)\left(\hat{Y}_i-\overline{Y}\right)\end{aligned}$$

上式中右边第三项为零，则有

$$S_{总}=\sum_{i=1}^{n}\left(Y_i-\hat{Y}_i\right)^2+\sum_{i=1}^{n}\left(\hat{Y}_i-\overline{Y}\right)^2 \tag{4-14}$$

记：

$$U=\sum_{i=1}^{n}\left(\hat{Y}_i-\overline{Y}\right)^2$$

$$Q=\sum_{i=1}^{n}\left(Y_i-\hat{Y}_i\right)^2$$

则

$$S_{总}=Q+U \tag{4-15}$$

式（4-15）中，Q 称为残差平方和，它的大小反映了实验过程中各种随机因素引起的波动对总离差平方和的贡献。U 称为回归平方和，它是由自变量取值不同引起总离差平方和的变化。图 4-2 表示实验数据与回归直线的离差 $Y_i-\hat{Y}_i$、

回归直线与平均值差 $\hat{Y}_i-\overline{Y}$ 和总离差 $Y_i-\overline{Y}$ 之间的关系，从而可以了解总离差平方和、残差平方和和回归平方和的意义和相互关系。

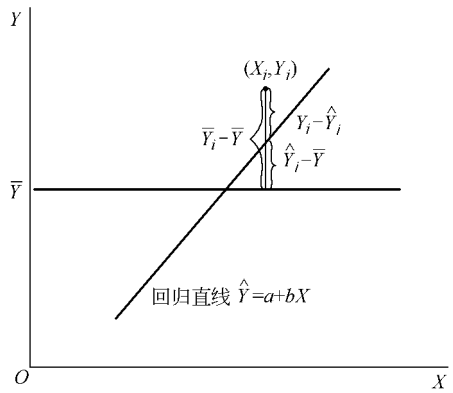


图 4-2　总离差的分解示意图

对一组观察数据而言，由于自变量 X 的取值已定，则回归平方和 U 亦为定值。残差平方和 Q 的大小则反映了除自变量取值以外所有因素对总离差平方和的影响。当残差平方和 Q 在总离差平方和中所占的比例越小时，或者说回归平方和 U 在总离差平方和中所占比例越大时，回归方程与观察结果的拟合程度越好。我们定义一个统计判别量 F 来反映 U 和 Q 的相互比例关系并用于回归方程的统计检验

$$F=\frac{\dfrac{U}{m}}{\dfrac{Q}{n-m-1}} \tag{4-16}$$

式（4-16）中，m 为回归变量的自由度，即自变量的数目；n 为观察值的组数；$n-m-1$ 为残差平方和的自由度。对一元回归而言，$m=1$，则式（4-16）可简化为

$$F=\frac{\dfrac{U}{1}}{\dfrac{Q}{n-2}} \tag{4-17}$$

以由［例 4-1］中的数据回归得到的回归方程（4-10）为例，介绍回归方程统计检验的步骤。

（1）建立统计假设。

假设 H_0：变量 Y 与变量 X 之间没有线性关系。

假设 H_1：变量 Y 与变量 X 之间存在线性关系。

（2）选择显著性水平 α，选 $\alpha=0.01$。

（3）计算统计量

$$F=\frac{\dfrac{U}{1}}{\dfrac{Q}{n-2}}=\frac{\dfrac{669.68}{1}}{\dfrac{2.58}{10-2}}=2076.6$$

（4）确定临界值 F_α　在显著性水平 $\alpha=0.01$ 条件下，在附表 5 中查得自由度 $df_1=1$，$df_2=10-2=8$ 的临界值 $F_\alpha=11.26$。

（5）统计判别 $F>F_\alpha$，否定假设 H_0，接受假设 H_1，即变量 Y 和变量 X 之间的线性关系是显著的，方程（4-10）从统计意义上讲是合理的。

第二节 可化成线性回归的曲线回归

在第一节中，我们讨论了解决一元线性回归问题的方法。实际工作中，面对的实验数据常常不表现为直接的线性关系，而是某种形式的曲线关系。由于线性回归参数的计算较为简单，因此我们在解决变量之间表现为曲线关系的问题时，常常首先考虑能否通过某种形式的变换，将变量之间的曲线关系转化为直线关系。实际上，变量之间的许多形式的曲线关系是可以通过适当的变换转化为线性关系的。

解决可化成线性回归的曲线回归问题的步骤如下。

① 确定变量之间函数关系的类型。变量之间函数关系类型可以根据变量之间关系的理论分析或者通过观察数据散点图的分布形状及特点来确定。

② 选择适当的变换，将变量之间的曲线关系转化为直线关系。

③ 确定变换后函数关系中的未知参数。由于变换后，变量之间的关系已成线性关系，因而可用第一节中介绍的线性回归方程中参数确定的方法加以解决。

④ 将所得到的线性关系中的参数值通过反变换变换成变量间曲线关系的形式，得到曲线方程。

⑤ 对所得到的曲线方程进行统计检验，判别方程在统计意义上是否合理。

下面介绍几种常用的可化成线性回归的曲线方程。

1. 幂函数 $y=ax^b$

幂函数 $y=ax^b$ 的曲线如图 4-3 所示。设 $Y=\lg y$，$X=\lg x$，则幂函数方程 $y=ax^b$ 化为

$$Y=\lg a+bX$$

令 $C=\lg a$，则有

$$Y=C+bX \tag{4-18}$$

式（4-18）即为标准的直线方程形式。

2. 指数函数 $y=ae^{bx}$

指数函数 $y=ae^{bx}$ 的曲线如图 4-4 所示。设 $X=x$，$Y=\ln y$，则指数函数方程 $y=ae^{bx}$ 化为 $Y=\ln a+bX$

令 $c=\ln a$，则有

$$Y=c+bX \tag{4-19}$$

式（4-19）为标准的直线方程形式。

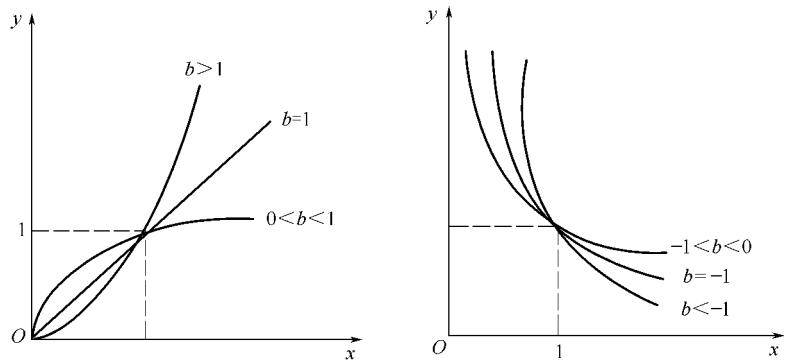


图 4-3　幂函数 $y=ax^b$ 曲线

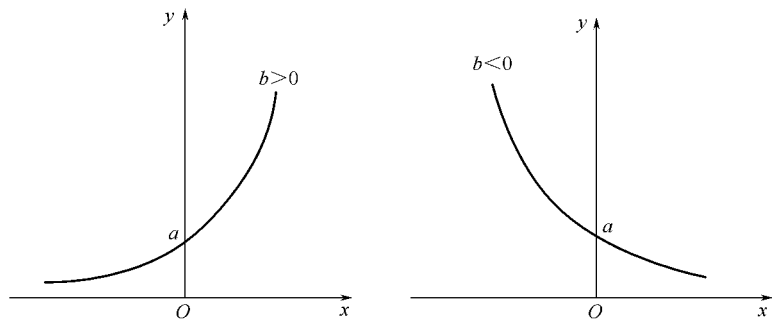


图 4-4　指数函数 $y=ae^{bx}$ 曲线

3. 对数函数 $y=a+b\lg x$

对数函数 $y=a+b\lg x$ 的曲线如图 4-5 所示。设 $Y=y$，$X=\lg x$，则对数函数方程 $y=a+b\lg x$ 化为

$$Y=a+bX \tag{4-20}$$

式（4-20）即为标准的直线方程形式。

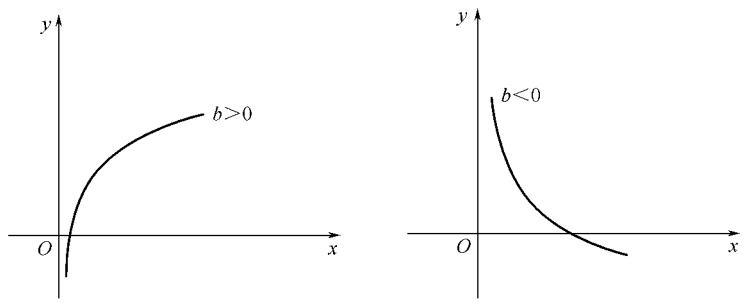


图 4-5　对数函数 $y=a+b\lg x$ 曲线

4. 双曲线函数$\frac{1}{y}=a+\frac{b}{x}$

双曲线函数$\frac{1}{y}=a+\frac{b}{x}$的曲线如图 4-6 所示。设$Y=\frac{1}{y}$，$X=\frac{1}{x}$，则双曲线函数方程$\frac{1}{y}=a+\frac{b}{x}$化为

$$Y=a+bX \tag{4-21}$$

式（4-21）即为标准的直线方程形式。

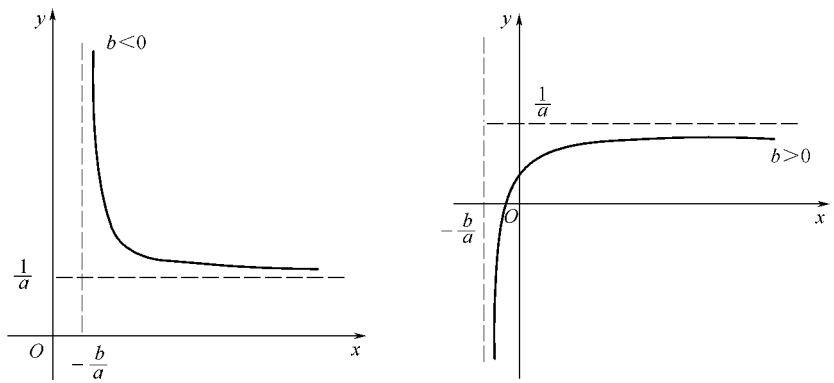


图 4-6 双曲线函数$\frac{1}{y}=a+\frac{b}{x}$

下面通过一实际例子来说明解决此类问题的方法。

【例 4-2】 某河流中，距排放源不同距离采样测定得到污染物酚浓度数据见表 4-3。

表 4-3 ［例 4-2］中河流中酚浓度测定结果

距离 x/km	0	0.2	0.9	1.4	1.8	2.2	3.0	4.2	5.2
酚浓度 y/(mg/L)	0.036	0.032	0.031	0.024	0.019	0.017	0.018	0.017	0.013

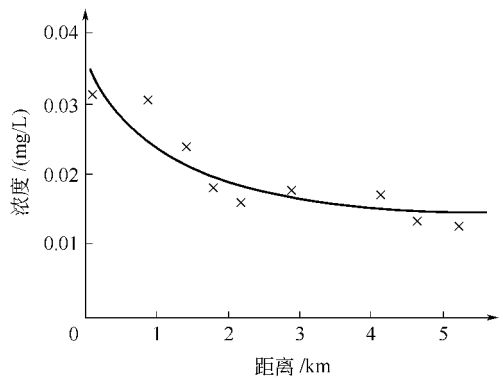


图 4-7 污染物酚浓度与距离关系

由上述实验数据建立距离和浓度间的回归方程。

根据本节介绍的解决变量间回归方程的步骤解此问题。

1. 确定变量之间函数关系的类型

由实际监测得到的数据作散点图（见图 4-7）。从图中的曲线形状看近似地与图 4-4 中的曲线形状相似。因而可初步确定酚浓度 y 与排

放源距离 x 之间服从指数函数关系，$y=ae^{bx}$。

2. 选择适当的变换，将变量之间的曲线关系转化为直线关系

设 $Y=\ln y$，$X=x$，$c=\ln a$，则有直线方程：$Y=c+bX$

3. 将实验数据按 2 中的设定进行变换，见表 4-4

表 4-4　数据变换

$X=x$	0	0.2	0.9	1.4	1.8	2.2	3.0	4.2	5.2
y	0.036	0.032	0.031	0.024	0.019	0.017	0.018	0.017	0.013
$Y=\ln y$	−3.32	−3.44	−3.47	−3.73	−3.96	−4.07	−4.02	−4.07	−4.34

4. 确定直线回归方程 $Y=c+bX$ 中的参数 c 和 b

按第一节中介绍的方法作回归直线方程计算表（见表 4-5）。

表 4-5　[例 4-2] 回归直线方程计算表

序号	X	Y	X^2	Y^2	XY
1	0	−3.32	0	11.02	0
2	0.2	−3.44	0.04	11.83	−0.688
3	0.9	−3.47	0.81	12.04	−3.12
4	1.4	−3.73	1.96	13.91	−5.22
5	1.8	−3.96	3.24	15.68	−7.13
6	2.2	−4.07	4.84	16.56	−8.95
7	3.0	−4.02	9.0	16.16	−12.06
8	4.2	−4.07	17.64	16.56	−17.09
9	5.2	−4.34	27.04	18.84	−22.57
总和	18.9	−34.42	64.57	132.6	−76.83
均值	2.1	−3.82			

由式（4-7）计算得回归直线方程中的参数

$$b=\frac{\sum_{i=1}^{n}X_iY_i-n\overline{X}\,\overline{Y}}{\sum_{i=1}^{n}X_i^2-n\overline{X}^2}=\frac{-76.83-9\times2.1\times(-3.82)}{64.57-9\times2.1^2}=-0.186$$

$$c=\overline{Y}-b\overline{X}=-3.82-(-0.186)\times2.1=-3.43$$

5. 通过反变换得曲线方程中参数

$$a=e^{c}=e^{-3.43}=0.032$$

则有污染物酚浓度和排放距离之间的指数函数函数关系

$$y=0.032e^{-0.186x} \tag{4-22}$$

6. 对回归所得的方程进行统计检验

(1) 建立统计假设

假设 H_0：回归方程式（4-22）在统计意义上不是显著成立的。

假设 H_1：回归方程式（4-22）在统计意义上是显著成立的。

(2) 选择显著性水平 $\alpha=0.01$。

(3) 计算统计量 F　回归平方和 U 和残差平方和 Q 的计算见表 4-6。

表 4-6　回归平方和 U 和残差平方和 Q 的计算表

序号	实验值 y_i	计算值 $\hat{y}_i$	$\hat{y}_i-\overline{y}$	$(\hat{y}_i-\overline{y})^2$	$y_i-\hat{y}_i$	$(y_i-\hat{y}_i)^2$
1	0.036	0.032	0.009	0.000081	0.004	0.000016
2	0.032	0.0308	0.0078	0.000061	0.0012	0.0000014
3	0.031	0.0270	0.004	0.000016	0.004	0.000016
4	0.024	0.0247	0.0017	0.0000029	−0.0007	0.00000049
5	0.019	0.0229	−0.0001	0.00000001	−0.0039	0.000015
6	0.017	0.0213	−0.0017	0.0000029	−0.0043	0.000018
7	0.018	0.0183	−0.0047	0.000022	−0.0003	0.00000009
8	0.017	0.0147	−0.0083	0.000069	0.0023	0.0000053
9	0.013	0.0122	−0.0108	0.00012	0.0008	0.00000064
	$\overline{y}=0.023$			$\sum=0.00037$		$\sum=0.000073$
	$U=\sum(\hat{y}_i-\overline{y})^2=0.00037$				$Q=\sum(y_i-\hat{y}_i)^2=0.000073$	

则由式（4-17）：

$$F=\frac{U}{\frac{Q}{n-2}}=\frac{0.00037}{\frac{0.000073}{9-2}}=35.48$$

(4) 确定临界值 F_α　在显著性水平 $\alpha=0.01$ 条件下，在附表 5 中查得自由度 $df_1=1$ 和 $df_2=9-2=7$ 的临界值 $F_\alpha=12.25$。

(5) 统计判别　由于 $F>F_\alpha$，则在显著性水平 0.01 条件下，否定假设 H_0，接受备选假设 H_1，即回归方程式（4-22）在统计意义上是显著成立的。

第三节　多元线性回归

在实际环境科研和监测工作中，经常遇到需要建立因变量与多个自变量关系的问题。即建立多个自变量之间的数学模型。这就要研究多元回归问题。多元回归问题与一元回归问题一样，也有线性回归和非线性回归两类问

题。本节主要讨论多元线性回归问题。多元非线性回归问题在本章第五节中讨论。

解决多元线性回归问题的原理和步骤与一元线性回归问题基本相同，也是用最小二乘法确定回归方程中的参数，只是在计算过程上复杂一些。下面介绍多元线性回归问题的解法。

一、多元线性回归方程的建立

在对变量 y 及 x_1，x_2，…，x_m 同时做了 n 次观察后，获得了 n 组观察数据：

y	$x_1, x_2, \cdots, x_m$
y_1	$x_{11}, x_{12}, \cdots, x_{1m}$
y_2	$x_{21}, x_{22}, \cdots, x_{2m}$
⋮	⋮
y_n	$x_{n1}, x_{n2}, \cdots, x_{nm}$

假设这 n 组实验数据所反映的 y 与 x_1，x_2，…，x_m 之间的关系可以用下述回归方程表示：

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_m x_m \tag{4-23}$$

式（4-23）中，b_0，b_1，…，b_m 为回归方程中的回归系数。回归系数用最小二乘法进行估算，即回归系数应满足使残差平方和 Q 值达到最小。

$$Q = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \Rightarrow \text{最小}$$

即：

$$Q = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_{1i} - b_2 x_{2i} - \cdots - b_m x_{mi})^2 \Rightarrow \text{最小}$$

由微分学中求极值原理，Q 对 b_0，b_1，…，b_m 的偏导数应为零，即：

$$\begin{cases} \dfrac{\partial Q}{\partial b_0} = 0 \\ \dfrac{\partial Q}{\partial b_1} = 0 \\ \vdots \\ \dfrac{\partial Q}{\partial b_m} = 0 \end{cases} \tag{4-24}$$

即：

$$\begin{cases}\dfrac{\partial}{\partial b_0}\left[\sum\limits_{i=1}^{n}(y_i-b_0-b_1x_{1i}-b_2x_{2i}-\cdots-b_mx_{mi})^2\right]=0\\ \dfrac{\partial}{\partial b_j}\left[\sum\limits_{i=1}^{n}(y_i-b_0-b_1x_{1i}-b_2x_{2i}-\cdots-b_mx_{mi})^2\right]=0\\ j=1,2,\cdots,m\end{cases}\tag{4-25}$$

对式（4-25）进行微分运算并化简可以得到：

$$\begin{cases}\sum\limits_{i=1}^{n}(y_i-b_0-b_1x_{1i}-b_2x_{2i}-\cdots-b_mx_{mi})=0\\ \sum\limits_{i=1}^{n}(y_i-b_0-b_1x_{1i}-b_2x_{2i}-\cdots-b_mx_{mi})x_{ji}=0\\ j=1,2,\cdots,m\end{cases}\tag{4-26}$$

令：$\bar{y}=\dfrac{1}{n}\sum\limits_{i=1}^{n}y_i$

$$\bar{x}_j=\frac{1}{n}\sum_{i=1}^{n}x_{ji}\qquad j=1,2,\cdots,m$$

代入式（4-26）则有：

$$\begin{cases}b_0=\bar{y}-b_1\bar{x}_1-b_2\bar{x}_2-\cdots-b_m\bar{x}_m\\ \sum[y_i-\bar{y}-b_1(x_{1i}-\bar{x}_1)-b_2(x_{2i}-\bar{x}_2)-\cdots-b_m(x_{mi}-\bar{x}_m)](x_{ji}-\bar{x}_i)=0\\ j=1,2,\cdots,m\end{cases}\tag{4-27}$$

令：$$l_{kj}=l_{jk}=\sum_{i=1}^{n}(x_{ki}-\bar{x}_k)(x_{ji}-\bar{x}_j)\quad k,j=1,2,\cdots,m$$

$$l_{jy}=\sum_{i=1}^{n}(y_i-\bar{y})(x_{ji}-\bar{x}_j)\qquad j=1,2,\cdots,m$$

代入式（4-27），并将式（4-27）中的第一式代入第二式则有：

$$\begin{cases}l_{11}b_1+l_{12}b_2+\cdots+l_{1m}b_m=l_{1y}\\ l_{21}b_1+l_{22}b_2+\cdots+l_{2m}b_m=l_{2y}\\ \qquad\vdots\\ l_{m1}b_1+l_{m2}b_2+\cdots+l_{mm}b_m=l_{my}\end{cases}\tag{4-28}$$

式（4-28）称为回归方程（4-23）的正规方程。解正规方程即可得到回归系数 b_1，b_2，…，b_m。

由式（4-28）可以看到，正规方程为一 m 元的线性方程组，为了得到 m 个回归系数，必须要解此 m 元线性方程组。线性方程组的解法很多，此处我们介

绍解线性方程组的两种常用的方法，行列式法和高斯-约当消去法。

正规方程（4-28）的行列式解一般以下述形式表达：

$$\begin{cases} b_1=\dfrac{\begin{vmatrix} l_{1y} & l_{12} & \cdots & l_{1m} \\ l_{2y} & l_{22} & \cdots & l_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ l_{my} & l_{m2} & \cdots & l_{mm} \end{vmatrix}}{\Delta} \\ b_2=\dfrac{\begin{vmatrix} l_{11} & l_{1y} & \cdots & l_{1m} \\ l_{21} & l_{2y} & \cdots & l_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ l_{m1} & l_{my} & \cdots & l_{mm} \end{vmatrix}}{\Delta} \\ \vdots \\ b_m=\dfrac{\begin{vmatrix} l_{11} & l_{21} & \cdots & l_{1y} \\ l_{21} & l_{22} & \cdots & l_{2y} \\ \vdots & \vdots & \vdots & \vdots \\ l_{m1} & l_{m2} & \cdots & l_{my} \end{vmatrix}}{\Delta} \end{cases} \tag{4-29}$$

式（4-29）中：

$$\Delta=\begin{vmatrix} l_{11} & l_{21} & \cdots & l_{m1} \\ l_{21} & l_{22} & \cdots & l_{m2} \\ \vdots & \vdots & \vdots & \vdots \\ l_{m1} & l_{m2} & \cdots & l_{mm} \end{vmatrix}$$

行列式解法的优点是解的表达式简明，易于为人们理解接受。其缺点是当行列式阶数比较高时，计算较繁。因而在实际解线性方程组的计算中，仅对低阶的线性方程组，采用行列式解法，以二元线性方程组为例，式（4-29）可表示为：

$$\begin{cases} b_1=\dfrac{\begin{vmatrix} l_{1y} & l_{12} \\ l_{2y} & l_{22} \end{vmatrix}}{\begin{vmatrix} l_{11} & l_{12} \\ l_{21} & l_{22} \end{vmatrix}}=\dfrac{l_{1y}l_{22}-l_{2y}l_{12}}{l_{11}l_{22}-l_{12}^2} \\ b_2=\dfrac{\begin{vmatrix} l_{11} & l_{1y} \\ l_{21} & l_{2y} \end{vmatrix}}{\begin{vmatrix} l_{11} & l_{12} \\ l_{21} & l_{22} \end{vmatrix}}=\dfrac{l_{2y}l_{11}-l_{1y}l_{21}}{l_{11}l_{22}-l_{12}^2} \end{cases} \tag{4-30}$$

高斯-约当消去法求解正规方程组的基本思想是对正规方程增广系数矩阵逐

步实行消去变换。正规方程的增广系数矩阵为：

$$\boldsymbol{L}^{(0)}=\begin{pmatrix} l_{11} & l_{12} & \cdots & l_{1m} & l_{1y} \\ l_{21} & l_{22} & \cdots & l_{2m} & l_{2y} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ l_{m1} & l_{m2} & \cdots & l_{mm} & l_{my} \end{pmatrix}$$

第 S 步对第 K 个未知数消去变换的公式为：

$$\begin{cases} \text{当 } i=k \text{ 时，} l_{ij}^{(s)}=\dfrac{l_{kj}^{(s-1)}}{l_{kj}^{(s-1)}} \\ \text{当 } i\neq k \text{ 时，} l_{ij}^{(s)}=l_{ij}^{(s-1)}-\dfrac{l_{ik}^{(s-1)}l_{kj}^{(s-1)}}{l_{kk}^{(s-1)}} \end{cases} \tag{4-31}$$

二、多元线性回归方程建立的应用举例

下面以一个二元线性回归问题作为实例，介绍多元线性回归问题的实际计算步骤。

【例 4-3】 两工厂排放有机污染物量为 x_1 和 x_2，在工厂排放口下游某段面监测得 BOD_5 浓度 y 的数据见表 4-7。

表 4-7 ［例 4-3］中 BOD_5 测定结果

样品序号	1	2	3	4	5	6	7	8	9	10	11	12
x_1/(kg/h)	24	26	41	55	60	67	25	79	70	55	45	33
x_2/(kg/h)	23	21	24	25	24	26	25	25	24	25	25	23
y/(mg/L)	210	206	260	244	271	285	270	265	234	241	258	230

建立污染物排放量 x_1，x_2 和 BOD_5 浓度 y 之间的线性回归方程。

设 BOD_5 浓度 y 和污染物排放量 x_1，x_2 之间存在如下线性关系。

$$\hat{y}=b_0+b_1x_1+b_2x_2 \tag{4-32}$$

对二元线性回归问题，式（4-32）可写成：

$$\begin{cases} b_0=\bar{y}-b_1\bar{x}_1-b_2\bar{x}_2 \\ b_1\sum\limits_{k=1}^{n}(x_{1k}-\bar{x}_1)^2+b_2\sum\limits_{k=1}^{n}(x_{1k}-\bar{x}_1)(x_{2k}-\bar{x}_2)=\sum\limits_{k=1}^{n}(x_{1k}-\bar{x}_1)(y_k-\bar{y}) \\ b_1\sum\limits_{k=1}^{n}(x_{2k}-\bar{x}_2)(x_{1k}-\bar{x}_1)+b_2\sum\limits_{k=1}^{n}(x_{2k}-\bar{x}_2)^2=\sum\limits_{k=1}^{n}(x_{2k}-\bar{x}_2)(y_k-\bar{y}) \end{cases} \tag{4-33}$$

正规方程的系数可表达为：

$$\begin{cases} l_{ij}=l_{ji}=\sum\limits_{k=1}^{n}(x_{jk}-\overline{x}_j)(x_{ik}-\overline{x}_i) \\ l_{iy}=\sum\limits_{k=1}^{n}(x_{ik}-\overline{x}_i)(y_k-\overline{y}) \end{cases} \tag{4-34}$$

为计算方便，通常将式（4-34）表达成为如下形式：

$$\begin{cases} l_{ij}=l_{ji}=\sum\limits_{k=1}^{n}x_{ik}x_{jk}-\dfrac{1}{n}\left(\sum\limits_{k=1}^{n}x_{ik}\right)\left(\sum\limits_{k=1}^{n}x_{jk}\right) \\ l_{iy}=\sum\limits_{k=1}^{n}x_{ik}y_k-\dfrac{1}{n}\left(\sum\limits_{k=1}^{n}x_{ik}\right)\left(\sum\limits_{k=1}^{n}y_k\right) \end{cases} \tag{4-35}$$

则式（4-33）可写成：

$$\begin{cases} b_0=\overline{y}-b_1\overline{x}_1-b_1\overline{x}_2 \\ l_{11}b_1+l_{12}b_2=l_{1y} \\ l_{21}b_1+l_{22}b_2=l_{2y} \end{cases} \tag{4-36}$$

式（4-36）可用行列式法解［见式（4-30）］，［例 4-3］的具体计算见表 4-8。

表 4-8　二元线性回归的计算

样品序号	x_1	x_2	y	x_1^2	x_1x_2	x_2^2	x_1y	x_2y	y^2
1	24	23	210	576	552	529	5040	4830	44100
2	26	21	206	676	546	441	5356	4326	42436
3	41	24	260	1681	984	576	10660	6240	67600
4	55	25	244	3025	1375	625	13420	6100	59536
5	60	24	271	3600	1440	576	16260	6504	73441
6	67	26	285	3721	1586	676	17385	7410	81225
7	75	25	270	5625	1875	625	20250	6750	72900
8	79	25	265	6241	1975	625	20935	6625	70225
9	70	24	234	4900	1680	576	16380	5616	54756
10	55	25	241	3025	1375	625	13255	6025	58081
11	45	25	258	2025	1125	625	11610	6450	66564
12	33	23	230	1089	759	529	7590	5290	52900
Σ	624	290	2974	36184	15272	7028	158141	72166	743764

$$n=12,\ \overline{x}_1=\frac{\sum\limits_{i=1}^{12}x_{1i}}{n}=52,\ \overline{x}_2=\frac{\sum\limits_{i=1}^{12}x_{2i}}{n}=24.17,\ y=\frac{\sum\limits_{i=1}^{12}y_i}{n}=247.83$$

$$l_{11}=\sum_{i=1}^{12}x_{1i}^2-\frac{1}{n}\left(\sum_{i=1}^{12}x_{1i}\right)^2=3736$$

$$l_{12}=\sum_{i=1}^{12}x_{1i}x_{2i}-\frac{1}{n}\left(\sum_{i=1}^{12}x_{1i}\right)\left(\sum_{i=1}^{12}x_{2i}\right)=192$$

$$l_{22}=\sum_{i=1}^{12}x_{2i}^2-\frac{1}{n}\left(\sum_{i=1}^{12}x_{2i}\right)^2=19.67$$

$$l_{1y}=\sum_{i=1}^{12}x_{1i}y_i-\frac{1}{n}\left(\sum_{i=1}^{12}x_{1i}\right)\left(\sum_{i=1}^{12}y_i\right)=3493$$

$$l_{2y}=\sum_{i=1}^{12}x_{2i}y_i-\frac{1}{n}\left(\sum_{i=1}^{12}x_{2i}\right)\left(\sum_{i=1}^{12}y_i\right)=294.33$$

$$l_{yy}=\sum_{i=1}^{12}y_i^2-\frac{1}{n}\left(\sum_{i=1}^{12}y_i\right)^2=6707.67$$

$$b_1=\frac{l_{1y}l_{22}-l_{2y}l_{12}}{l_{11}l_{22}-l_{12}^2}=0.333$$

$$b_2=\frac{l_{2y}l_{11}-l_{1y}l_{21}}{l_{11}l_{22}-l_{12}^2}=11.718$$

$$b_0=\bar{y}-b_1\bar{x}_1-b_2\bar{x}_2=-52.646$$

由表 4-8 的计算结果，得到［例 4-3］两工厂排放有机污染物量 x_1、x_2 与排放口下游监测段面 BOD_5 浓度 y 之间的线性回归方程：

$$y=-52.646+0.333x_1+11.718x_2 \tag{4-37}$$

三、多元线性回归方程的统计检验

为了判别回归方程在统计意义上成立与否，需要对所得的回归方程（4-37）进行统计检验。

1. *F* 检验

我们已知可用 F 检验来判别回归方程在统计上是否合理。F 检验统计量 F 的计算公式如式（4-16）所示。

首先计算回归平方和 U 和残差平方和 Q：

$$U=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2=4611.29$$

$$Q=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2=2096.48$$

计算统计量 F［见式（4-16）］：

$$F=\frac{\dfrac{U}{m}}{\dfrac{Q}{n-m-1}}=\frac{\dfrac{4611.29}{2}}{\dfrac{2096.48}{12-2-1}}=9.90$$

选择显著性水平 $\alpha=0.01$，在附表 5 中查得自由度 $df_1=2$、$df_2=n-m-1=9$ 时的临界值 $F_{0.01}=8.02$。由于 $F>F_{0.01}$，则回归方程（4-37）在统计意义上是显著成立的。

2. 复相关系数检验

我们已知除了用 F 检验来判别回归方程在统计上是否合理外，也可以用相关系数的办法来检验回归效果的优劣。与一元线性回归问题略为不同的是，对多元线性回归而言，采用复相关系数来判别回归方程在统计上是否合理。复相关系数 R 的定义为：

$$R=\sqrt{1-\frac{Q}{S_{总}}}=\sqrt{\frac{U}{S_{总}}} \tag{4-38}$$

对给定的观察值而言，总离差平方和 $S_{总}$ 是不变的，残差平方和 Q 值大则回归平方和 U 值小，反之 Q 值小则 U 值大。回归平方和含义是回归方程中全部自变量的“方差贡献”。因此，复相关系数 R 就是这种贡献在总误差平方和中所占的比例。复相关系数越接近 1 则回归效果越好。应该指出的是，复相关系数 R 与回归方程中自变量个数 m 及观察组数 n 有关。当 n 相对于 m 并不是很大时，常常有较大的 R，特别是当 $n=m+1$ 时，即使 m 个自变量与变量 y 毫无关系，亦有 $R=1$。在实际计算中，要注意 n 与 m 的适当比例，一般认为 n 至少是 m 的 5～10 倍。

回归方程（4-37）的复相系数为：

$$R=\sqrt{1-\frac{Q}{S_{总}}}=\sqrt{1-\frac{2096.48}{2096.48+4611.29}}=0.829$$

查附表 2，在显著性水平 $\alpha=0.01$ 时，相关系数的临界值 $R_{0.01}=0.708$。由于 $R>R_{0.01}$，因而回归方程（4-37）在统计意义上是显著成立的。

在多元线性组回归的实际运算中，F 检验或复相关系数的检验只需选一种进行即可。通常情况下两种检验结果的结论是一致的。在例 4-3 中，我们用两种统计检验的方法对回归方程进行检验的目的是为了介绍这两种统计检验的方法。如前所述，在实际多元线性回归问题的运算中，同时用两种方法进行统计检验是没有必要的。

3. 单个自变量回归效果的检验

上面介绍的多元回归方程的统计检验考虑的是全体自变量的回归效果的检验。然而，在多元回归方程中引进的多个自变量，有时互相并不完全独立，即某一个自变量与因变量的线性联系常常可以部分或全部通过另一些自变量与因变量的线性联系得到反映。它意味着该自变量对因变量的影响很小，并不很重要。在

我们进行回归运算时，总希望我们所引入的自变量对因变量有显著影响。因此，需要检验各个自变量对因变量的贡献。需要注意的是，某个自变量对因变量的影响可以从其余变量总体上对因变量的线性联系中得到反映，但它并不意味着该自变量与因变量线性不相关。该自变量有可能和因变量单独相关是显著的，因此不能通过对各个自变量与因变量的单相关检验的显著性来判别该自变量对因变量的影响。而是通过从所有自变量中去掉某个自变量后，回归平方和减少的程度来判别此自变量对因变量的影响大小。

若要检验 m 个自变量 x_1，x_2，…，x_m 中第 j 个自变量 x_j 的回归效果。我们可按照多元线性回归方程建立的步骤分别建立包含自变量 x_j 的线性回归方程：

$$\hat{y}=b_0+\sum_{i=1}^{m} b_i x_i \tag{4-39}$$

和不包含 x_j 的线性回归方程：

$$\hat{y}'=b_0+\sum_{\substack{i=1\\ i\neq j}}^{m} b_i x_i \tag{4-40}$$

计算回归方程（4-39）的回归平方和 U 和回归方程（4-40）的回归平方和 U'

$$U=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2$$

$$U'=\sum_{i=1}^{n}(\hat{y}_i'-\bar{y})^2$$

U'表示除 x_j 以外的 $m-1$ 个自变量的回归平方和贡献；U 表示所有 m 个自变量的回归平方和贡献。因此，可用 $U-U'$表示 m 个自变量中的 x_j 的回归平方和贡献。$U-U'$的数值越大，自变量 x_j 在回归方程（4-39）中占的地位越重要，计算统计量 F：

$$F=\frac{U-U'}{\dfrac{Q}{n-m-1}} \tag{4-41}$$

选择显著水平 α，若 $F\geqslant F_\alpha$（1，$n-m-1$），则有统计判别：x_j 在回归方程（4-39）中贡献是显著的，反之 x_j 在回归方程（4-39）中贡献不显著。

以［例 4-3］为例，检验回归方程（4-37）中，变量 x_2 贡献显著与否。

在［例 4-3］中，我们已建立了包含变量 x_2 的回归方程（4-37）。按线性回归方程建立步骤建立不包含变量 x_2 的线性回归方程得：

$$\hat{y}'=199.21+0.935x_1 \tag{4-42}$$

计算式（4-42）回归平方和：

$$U'=\sum_{i=1}^{n}(\hat{y}_i'-\bar{y})^2=3392$$

计算统计量 F'：

$$F'=\frac{(U-U')(n-m-1)}{Q}$$

$$=\frac{3392\times(12-2-1)}{2096.48}$$

$$=14.56$$

选择显著性水平 $\alpha=0.01$，在附表 5 中查得自由度 $df_1=1$、$df_2=9$ 时 $F_{0.01}=10.6$。由于 $F'>F_{0.01}$，则我们有统计判别：变量 x_2 在回归方程（4-37）中贡献是显著的。在引入变量 x_1 后，有必要继续引进 x_2，x_2 能显著的改善回归效果。

第四节　逐步回归

在进行多元回归分析时，我们将预先选定的自变量全部引入方程，然后经过统计检验，确定各个自变量的方差贡献是否显著，即确定各自变量对回归结果是否有显著的影响，从而将对回归结果影响不显著的自变量剔除，重新建立回归方程。但是，由于剔除了一个或几个自变量，重新建立回归方程时需重新计算回归系数，计算工作量相当大。特别是当自变量数目较多的时候，计算工作相当繁重。为了简化计算，我们希望对自变量进行挑选，将对回归结果影响大的首先选入回归方程，并检验选入回归方程的每个变量方差贡献的显著性，对贡献不显著的加以剔除，从而保证回归方程中只保留方差贡献显著的自变量。逐步回归就是这样一种方法。它根据各个自变量的重要性的大小，每步挑选一个重要变量进入回归方程。第一步是在所有可供挑选的变量中选出一个变量，并使它组成的一元回归方程将比其他量有更大的回归平方和（或等价的更小的残差平方和）。第二步是在未选入的变量中选一个这样的变量，它与已选入的那个变量组成的二元回归方程将比其他任何一个变量与已选入变量组成的二元方程有更大的回归平方和。一般地说，第 l 步是在未选中的变量中选这样一个变量，它与已选入变量组成的 l 元回归方程将比其余任何一个变量与已选入变量组成的 l 元回归方程有更大的回归平方和。为保证每一步选入回归方程的变量真正重要，还应该对即将选入的变量做显著性检验。当检验通过时才进行下一步的计算。如果检验不显著，挑选变量的工作便告结束。在实践中，选入变量的方式及检验方式不同，有各种逐步回归的方法。本节介绍的一种方法不仅考虑到按变量贡献的大小逐一选出重要的变量，而且还考虑到较早选入回归方程的某些变量，有可能随着其后另一些

变量的选入而失去原有的重要性，这就是说，逐步回归逐个引入变量后，对已选入的变量也逐个重新检验，将其中变为不显著的变量剔除，直到所有显著的变量都包括在回归方程中为止。

一、逐步回归的计算步骤

设有自变量 x_1，x_2，…，x_m 及因变量 y 的对应观察数据如表 4-9 所示。

表 4-9　x_i 与 y 的对应观察数据

实验序号	x_1	x_2	…	x_m	y
1	x_{11}	x_{21}	…	x_{m1}	y_1
2	x_{12}	x_{22}	…	x_{m2}	y_2
⋮	⋮	⋮	⋮	⋮	⋮
n	x_{1n}	x_{2n}	…	x_{mn}	y_n

（1）计算相关矩阵　在逐步回归的计算中，通常采用标准化的量，使计算简便，并可避免自变量由于量纲不同而取值不同对回归结果产生的影响。当我们采用标准化量进行回归运算时，正规方程（4-28）中的系数不以 l_{ij} 表示，而变换成相关系数 r_{ij}。

原始量纲的数值 x_{ij}、y_i 与标准化后的数值 X_{ij}、Y_i 之间的换算关系式如下：

$$X_{ik}=\frac{x_{ik}-\bar{x}_k}{S_k}\qquad k=1,2,\cdots,m \tag{4-43}$$

$$Y_i=\frac{y_i-\bar{y}}{S_y} \tag{4-44}$$

式（4-43）和式（4-44）中，

$$\bar{x}_k=\frac{1}{n}\sum_{i=1}^{n}x_{ik}$$

$$\bar{y}=\frac{1}{n}\sum_{i=1}^{n}y_i$$

$$S_k=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_{ik}-\bar{x}_k)^2}$$

$$S_y=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})^2}$$

由式（4-43）和式（4-44），我们可随时将标准化的量换算成原量纲的量。

按式（4-11）计算各个自变量和因变量之间、各自变量相互之间的线性相关系数 r_{iy}、r_{ij}，得相关系数矩阵 $\boldsymbol{R}^{(0)}$：

$$\boldsymbol{R}^{(0)}=\begin{pmatrix} r_{11} & r_{12} & \cdots & r_{1m} & r_{1y} \\ r_{21} & r_{22} & \cdots & r_{2m} & r_{2y} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ r_{m1} & r_{m2} & \cdots & r_{mm} & r_{my} \\ r_{y1} & r_{y2} & \cdots & r_{ym} & r_{yy} \end{pmatrix} \tag{4-45}$$

相关系数矩阵实际上是标准化后的正规方程组的增广系数矩阵。逐步回归不是解此正规方程组得到 m 元回归方程，而是要从 m 个变量中选择对因变量方差贡献最大者，逐步引入回归方程。

(2) 从所有自变量中选择对因变量方差贡献最大的一个引入回归方程。

要从 m 个自变量中引入一个对因变量方差贡献最大的变量，需要计算每个自变量的方差贡献：

$$u_i^{(0)}=\frac{r_{iy}^2}{r_{ii}} \tag{4-46}$$

并找出方差贡献最大的变量 x_{k1}：

$$u_{k1}^{(0)}=\max\{u_i^{(0)}\} \tag{4-47}$$

用 F 检验法检验所引入变量方差贡献的显著性。首先确定置信水平。F 检验的显著性水平需要根据问题的具体情况来确定。若希望引入较多的变量，显著性水平可取得大一些，例如可取 $\alpha=0.10$；若希望引入的变量少一些，显著性水平可取得小一些，例如可取 $\alpha=0.05$，甚至 $\alpha=0.01$。F 检验的临界值除了和显著性水平有关外，还与自由度有关。在逐步回归中，由于引入回归方程的变量不断变化，自由度也相应地在变化。但在逐步回归中为了计算方便，特别用计算机进行计算时，通常根据原始数据估计可能选入的变量数，确定自由度，得到相应的临界值 F^* 作为 F 检验判别值。

计算统计量 F：

$$F_{k1}=\frac{u_{k1}^{(0)}(n-2)}{r_{yy}^{(0)}-u_{k1}^{(0)}} \tag{4-48}$$

若 F_{k1} 大于临界值 F^*，则变量 x_{k1} 对因变量 y 方差贡献显著，x_{k1} 引入回归方程在统计意义上成立；若 F_{k1} 小于临界值 F^*，则变量 x_{k1} 对因变量 y 方差贡献不显著，x_{k1} 引入回归方程在统计意义上不成立。若方差贡献最大的变量都未能引入回归方程，则其他变量就更达不到显著性水平，计算可以停止。

假设 F_{k1} 大于 F^*，x_{k1} 引入回归方程成立，则继续进行下一计算步骤。

(3) 除 x_{k1} 外，在余下的变量中选择对因变量方差贡献最大者进入回归方程。

首先对相关系数矩阵 $\boldsymbol{R}^{(0)}$ 按下式进行变换：

$$r_{ij}^{(1)}=\begin{cases}\dfrac{1}{r_{k1k1}} & (i=j=k1)\\ \dfrac{r_{k1j}}{r_{k1k1}} & (i=k1,j\neq k1)\\ -\dfrac{r_{ik1}}{r_{k1k1}} & (i\neq k1,j=k1)\\ r_{ij}-\dfrac{r_{ik1}r_{k1j}}{r_{k1k1}} & (i,j\neq k1)\end{cases} \tag{4-49}$$

计算除变量 x_{k1} 以外的其余变量对因变量 y 的方差贡献：

$$u_i^{(1)}=\frac{(r_{iy}^{(1)})^2}{r_{ii}^{(1)}} \tag{4-50}$$

并从中找出方差贡献最大的变量 x_{k2}

$$u_{k2}^{(1)}=\max\{u_i^{(1)}\} \tag{4-51}$$

用 F 检验法检验所引入变量 x_{k2} 方差贡献的显著性，计算统计量：

$$F_{k2}=\frac{u_{k2}^{(1)}(n-3)}{r_{yy}^{(1)}-u_{k2}^{(1)}} \tag{4-52}$$

并检验其方差贡献的显著性。若 F_{k2} 大于临界值 F^*，则 x_{k2} 引入回归方程。

(4) 判断已引入回归方程的变量是否由于引入新的变量而成为方差贡献不显著的自变量。

计算由于新引入变量 x_{k2} 后原先已引入的变量 x_{k1} 的方差贡献和统计量 F：

$$u_{k1}^{(1)}=\frac{[r_{k1y}^{(1)}]^2}{r_{k1k1}^{(1)}} \tag{4-53}$$

$$F_{k1}^{(1)}=\frac{u_{k1}^{(1)}(n-3)}{r_{yy}^{(1)}} \tag{4-54}$$

将计算所得的 $F_{k1}^{(1)}$ 值与临界值 F^* 比较，检验变量 x_{k1} 方差贡献是否仍然显著，决定其保留与否。

(5) 重复步骤 (3) 和步骤 (4)，继续引进第三个变量 x_{k3}，并对已引入的变量重新进行检验。

重复步骤 (5)，直至所有方差贡献显著的变量都引入回归方程，未引入回归方程的变量方差贡献不显著，即既无变量可引入也无变量可剔除，逐步回归变量选择运算结束。

(6) 计算实际回归系数　上述逐步回归步骤都是按标准化量进行计算的。最后得到的回归方程需要化成原量纲时，回归系数应是实际回归系数。若引入回归方程的变量为 x_j ($j=1, 2, \cdots, l$)，实际回归系数由式 (4-55) 求得：

$$b_j=\frac{\sqrt{S_{yy}^2}}{\sqrt{S_{jj}^2}}\cdot\gamma_{jy}^{(L)} \tag{4-55}$$

式（4-55）中，L 表示相关系数矩阵 $\boldsymbol{R}$ 经 L 步变换。$r_{jy}^{(L)}$ 是 L 步变换后相关系数矩阵中的元素。

回归方程常数项为：

$$b_0=\overline{y}-\sum_{j=1}^{l}b_j\overline{x}_j \tag{4-56}$$

复相关系数 R 和残差平方和 Q 也可以由 L 步变换后的相关系数矩阵元素中求得：

$$R=\sqrt{1-r_{yy}^{(L)}} \tag{4-57}$$

$$Q=\frac{S_{yy}r_{yy}^{(L)}}{n-l-1} \tag{4-58}$$

二、逐步回归的计算举例

【例 4-4】 测得某城镇二氧化硫日均浓度 y（$\mu g/m^3$）和四个主要污染源二氧化硫排放量 x_1，x_2，x_3，x_4（kg/h）原始数据如表 4-10 所示。

表 4-10 ［例 4-4］中 SO_2 浓度测定结果

序号	y	x_1	x_2	x_3	x_4
1	115	13	7	26	19
2	198	15	11	40	34
3	137	21	8	29	17
4	216	19	12	15	23
5	223	27	11	13	27
6	191	32	10	21	15
7	117	17	8	18	16
8	194	26	10	35	23
9	106	14	6	14	18
10	255	28	13	21	34
11	187	19	9	13	29
12	193	12	10	19	38
13	156	23	8	25	17
14	247	28	11	33	32
15	153	21	9	18	19

试用逐步回归方法建立该镇二氧化硫日均浓度与污染源排放量之间的回归方程。

解

1. 计算各变量原始数据的平均值和标准偏差

项目	x_1	x_2	x_3	x_4	y
平均值	21	9.53	22.67	24.07	179.2
标准偏差	6.20	1.92	8.45	7.68	47.0

2. 计算变量 x_1，x_2，x_3，x_4 及 y 两两之间的相关系数矩阵 $\boldsymbol{R}^{(0)}$

$$\boldsymbol{R}^{(0)}=\begin{bmatrix}1 & 0.4915 & 0.0914 & -0.0811 & 0.5819\\ 0.4915 & 1 & 0.1261 & 0.6412 & 0.9410\\ 0.0914 & 0.1261 & 1 & 0.1722 & 0.1336\\ -0.0811 & 0.6412 & 0.1722 & 1 & 0.7009\\ 0.5819 & 0.9410 & 0.1336 & 0.7009 & 1\end{bmatrix}$$

3. 选择 F 统计检验的临界值 F^*

本例选择显著性水平 $\alpha=0.10$，估计可能选入回归方程的变量为三个，查 F 统计分布表（附表 5）在显著性水平 $\alpha=0.10$，自由度 $df_1=1$，$df_2=15-3-1=11$ 的条件下，F 检验临界值 $F^*=3.23$。

4. 计算各变量的方差贡献并选出最大者

$$u_1^{(0)}=\frac{[r_{1y}^{(0)}]^2}{r_{11}^{(0)}}=0.5819^2=0.3386$$

$$u_2^{(0)}=\frac{[r_{2y}^{(0)}]^2}{r_{22}^{(0)}}=0.9410^2=0.8855$$

$$u_3^{(0)}=\frac{[r_{3y}^{(0)}]^2}{r_{33}^{(0)}}=0.1336^2=0.0178$$

$$u_4^{(0)}=\frac{[r_{4y}^{(0)}]^2}{r_{44}^{(0)}}=0.7009^2=0.4913$$

方差贡献最大者为 $u_2^{(0)}=0.8855$。

5. 检验所选入变量方差贡献是否显著，按式（4-48）计算统计量

$$F_2^{(0)}=\frac{u_2^{(0)}(n-2)}{r_{yy}^{(0)}-u_2^{(0)}}=\frac{0.885\times(15-2)}{1-0.885}=100.04$$

由于 $F_2^{(0)}>F^*$，因此 x_2 方差贡献显著，应将 x_2 引入回归方程。

6. 继续挑选新的变量引入回归方程

按式（4-49）对相关系数矩阵 $\boldsymbol{R}^{(0)}$ 进行变换，得新的系数矩阵 $\boldsymbol{R}^{(1)}$：

$$\boldsymbol{R}^{(1)}=\begin{bmatrix}0.7584 & -0.4915 & 0.0294 & -0.3962 & 0.1194\\ -0.4915 & 1 & 0.1261 & 0.6412 & 0.9410\\ 0.0294 & 0.1261 & 0.9841 & 0.0913 & 0.0149\\ -0.3962 & 0.6412 & 0.0913 & 0.5889 & 0.0975\\ 0.1194 & 0.9410 & 0.0149 & 0.0915 & 0.1145\end{bmatrix}$$

计算除变量 x_2 以外各变量的方差贡献并选出最大者：

$$u_1^{(1)}=\frac{[r_{1y}^{(1)}]^2}{r_{11}^{(1)}}=\frac{0.1194^2}{0.7584}=0.0188$$

$$u_3^{(1)}=\frac{[r_{3y}^{(1)}]^2}{r_{33}^{(1)}}=\frac{0.0149^2}{0.9841}=0.0002$$

$$u_4^{(1)}=\frac{[r_{4y}^{(1)}]^2}{r_{44}^{(1)}}=\frac{0.0975^2}{0.5889}=0.0161$$

方差贡献最大者为 $u_1^{(1)}=0.0188$。

检验新选入的变量 x_1 和已选入变量 x_2 的方差贡献：

$$F_1^{(1)}=\frac{u_1^{(1)}(n-3)}{r_{yy}^{(1)}-u_1^{(1)}}=\frac{0.0188\times(15-3)}{0.1145-0.0188}=2.36$$

$$F_2^{(1)}=\frac{u_2^{(1)}(n-3)}{r_{yy}^{(1)}}=\frac{0.8855\times(15-3)}{0.1145}=92.80$$

由于 $F_1^{(1)}$ 小于临界值 F^*，故变量 x_1 方差贡献不显著，不能引入回归方程。由于 x_1 是余下变量中方差贡献最大的变量，因此再没有变量可引入回归方程。回归方程只能引入一个变量 x_2。

7. 计算实际回归系数

由式（4-55），可计算实际回归系数：

$$\hat{b}_2=\frac{\sqrt{S_{yy}^2}}{\sqrt{S_{22}^2}}r_{2y}^{(1)}=\frac{\sqrt{47.0^2}}{\sqrt{1.92^2}}\times 0.9410=23.03$$

常数项为：

$$\hat{b}_0=\bar{y}-\hat{b}_2\bar{x}_2=179.2-23.03\times 9.53=-40.28$$

其回归方程为：

$$y=-40.28+23.03x_2$$

由于逐步回归运算过程中，对引入的变量已进行了统计检验，并证明检验结果在统计意义上是显著成立的，因而对最终得到的回归方程无需再进行统计检验。

本例中，我们得到的逐步回归结果表明，该镇的二氧化硫日均浓度主要受变量 x_2，即污染源 2 的影响，而和变量 x_1、x_3、x_4 关系不甚密切。经进一步分析了解到，污染源 1、3、4 虽然排放二氧化硫的量与污染源 2 相当，甚至更多一些，但其分布在该镇下风向，且与该镇距离较远，因而污染源 1、3、4 排放二氧化硫对城镇内二氧化硫浓度影响不明显，而污染源 2 在城镇上风向，且排放高度较低，因而它对该镇二氧化硫浓度有显著影响。通过该例的运算也告诉我们回归统计分析仅反映了数据统计上的联系，数据真正内在联系还要靠对具体问题进行深入分析，才能发现统计分析所得到结论的内在涵义。

第五节 非线性回归分析

前面我们介绍了解决线性回归问题的方法。也介绍通过变量变换将一些曲线回归问题化为线性回归问题的方法。然而在我们遇到的实际问题中，还会遇到即使采用各种变量变换的方法，也无法将曲线回归问题化成线性回归问题。因此需要讨论解决非线性回归问题的方法。与解决线性回归问题一样，非线性回归问题的主要任务也是要确定回归方程中的参数，即建立数学模型，并判断所建立的非线性数学模型准确程度有多大。

一般非线性回归方程，其参数无法直接求解。通常采用的办法是逐次逼近的方法，即通过搜索逐步逼近参数的最佳值，最终使误差达到所希望的范围内。下面介绍两种常用的估计非线性回归方程中参数的逐步搜索逼近的方法。

一、高斯-牛顿法

记 m 个变量 x_1，x_2，…，x_m 和变量 y 的 n 组实验数据如下：

y	x_1	x_2	…	x_m
y_1	x_{11}	x_{21}	…	x_{m1}
y_2	x_{12}	x_{22}	…	x_{m2}
⋮	⋮	⋮	⋮	⋮
y_n	x_{1n}	x_{2n}	…	x_{mn}

变量 y 和 m 个变量 x_1，x_2，…，x_m 间的一般关系可表示为：

$$\hat{y}=f(x_1,x_2,\cdots,x_m;\ b_1,b_2,\cdots,b_k) \tag{4-59}$$

式（4-59）中，b_1，b_2，…，b_k 为 K 维待估参数，f 表示变量 x_i 和 y 间的非线性函数关系。

非线性回归问题同样是在最小平方和意义下确定 k 个待估参数，即使：

$$J=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2\Rightarrow\text{最小} \tag{4-60}$$

式（4-60）中，J 为目标函数。

搜索法的过程是先给待估参数 b_j（$j=1$，2，…，k）一个初值，记为 $b_j^{(0)}$，并记初值与真值之差为 Δb_j（真值为未知的）。这时有：

$$b_j=b_j^{(0)}+\Delta b_j \qquad j=1,2,\cdots,k \tag{4-61}$$

这样，我们将确定真值 b_j 的问题转化为确定修正值 Δb_j 的问题。为了确定 Δb_j，可将函数在 $b_j^{(0)}$ 附近做台劳级数展开，并略去 Δb_j 二次及二次以上项，则有：

$$f(x_1,x_2,\cdots,x_m;\ b_1,b_2,\cdots,b_k)\approx f_{i0}+\frac{\partial f_{i0}}{\partial b_1}\Delta b_1+\frac{\partial f_{i0}}{\partial b_2}\Delta b_2+\cdots+\frac{\partial f_{i0}}{\partial b_k}\Delta b_k \tag{4-62}$$

式（4-62）中：

$$f_{i0}=f[x_{1i},x_{2i},\cdots,x_{mi};\ b_1^{(0)},b_2^{(0)},\cdots,b_k^{(0)}]$$

$$\frac{\partial f_{i0}}{\partial b_j}=\frac{\partial f(x_1,x_2,\cdots,x_m;\ b_1,b_2,\cdots,b_k)}{\partial b_i}\Bigg|_{\substack{x_l=x_{li},(l=1,2,\cdots,m)\\ b_j=b_j^{(0)},(j=1,2,\cdots,k)}}$$

则式（4-60）可记为：

$$J=\sum_{i=1}^{n}[y_i-f_i(x_1,x_2,\cdots,x_m;\ b_1^{(0)}+\Delta b_1,b_2^{(0)}+\Delta b_2,\cdots,b_k^{(0)}+\Delta b_k)]^2$$

$$\approx\sum_{i=1}^{n}\left[\ y_i-f_{i0}-\sum_{j=1}^{k}\frac{\partial f_{i0}}{\partial b_j}\Delta b_j\ \right]^2\Rightarrow\text{最小}$$

令 $R_i=y_i-f_{i0}$

则有：$J=\sum_{i=1}^{n}\left[\ R_i-\sum_{j=1}^{k}\frac{\partial f_{i0}}{\partial b_j}\Delta b_j\ \right]^2\Rightarrow$最小

由微分学中求极值原理，目标函数 J 的极小值应满足下列方程组：

$$\begin{cases}\dfrac{\partial J}{\partial\Delta b_1}=0\\ \dfrac{\partial J}{\partial\Delta b_2}=0\\ \vdots\\ \dfrac{\partial J}{\partial\Delta b_j}=0\end{cases} \tag{4-63}$$

即：

$$\begin{cases}\dfrac{\partial J}{\partial\Delta b_1}=-2\sum\limits_{i=1}^{n}\left(\ R_i-\sum\limits_{j=1}^{k}\dfrac{\partial f_{i0}}{\partial b_j}\Delta b_j\ \right)\dfrac{\partial f_{i0}}{\partial b_1}=0\\ \dfrac{\partial J}{\partial\Delta b_2}=-2\sum\limits_{i=1}^{n}\left(\ R_i-\sum\limits_{j=1}^{k}\dfrac{\partial f_{i0}}{\partial b_j}\Delta b_j\ \right)\dfrac{\partial f_{i0}}{\partial b_2}=0\\ \qquad\qquad\vdots\\ \dfrac{\partial J}{\partial\Delta b_k}=-2\sum\limits_{i=1}^{n}\left(\ R_i-\sum\limits_{j=1}^{k}\dfrac{\partial f_{i0}}{\partial b_j}\Delta b_j\ \right)\dfrac{\partial f_{i0}}{\partial b_k}=0\end{cases} \tag{4-64}$$

整理式（4-64）得：

$$\begin{cases} \sum_{i=1}^{n}\sum_{j=1}^{k}\frac{\partial f_{i0}}{\partial b_j}\Delta b_j\left(\frac{\partial f_{i0}}{\partial b_1}\right)=\sum_{i=1}^{n}R_i\left(\frac{\partial f_{i0}}{\partial b_1}\right) \\ \sum_{i=1}^{n}\sum_{j=1}^{k}\frac{\partial f_{i0}}{\partial b_j}\Delta b_j\left(\frac{\partial f_{i0}}{\partial b_2}\right)=\sum_{i=1}^{n}R_i\left(\frac{\partial f_{i0}}{\partial b_2}\right) \\ \vdots \\ \sum_{i=1}^{n}\sum_{j=1}^{k}\frac{\partial f_{i0}}{\partial b_j}\Delta b_j\left(\frac{\partial f_{i0}}{\partial b_k}\right)=\sum_{i=1}^{n}R_i\left(\frac{\partial f_{i0}}{\partial b_k}\right) \end{cases} \tag{4-65}$$

即：

$$\begin{pmatrix} \sum_{i=1}^{n}\frac{\partial f_{i0}}{\partial b_1}\frac{\partial f_{i0}}{\partial b_1} & \sum_{i=1}^{n}\frac{\partial f_{i0}}{\partial b_2}\frac{\partial f_{i0}}{\partial b_1} & \cdots & \sum_{i=1}^{n}\frac{\partial f_{i0}}{\partial b_k}\frac{\partial f_{i0}}{\partial b_1} \\ \sum_{i=1}^{n}\frac{\partial f_{i0}}{\partial b_1}\frac{\partial f_{i0}}{\partial b_2} & \sum_{i=1}^{n}\frac{\partial f_{i0}}{\partial b_2}\frac{\partial f_{i0}}{\partial b_2} & \cdots & \sum_{i=1}^{n}\frac{\partial f_{i0}}{\partial b_k}\frac{\partial f_{i0}}{\partial b_2} \\ \vdots & \vdots & \vdots & \vdots \\ \sum_{i=1}^{n}\frac{\partial f_{i0}}{\partial b_1}\frac{\partial f_{i0}}{\partial b_k} & \sum_{i=1}^{n}\frac{\partial f_{i0}}{\partial b_2}\frac{\partial f_{i0}}{\partial b_k} & \cdots & \sum_{i=1}^{n}\frac{\partial f_{i0}}{\partial b_k}\frac{\partial f_{i0}}{\partial b_k} \end{pmatrix}\begin{pmatrix} \Delta b_1 \\ \Delta b_2 \\ \vdots \\ \Delta b_k \end{pmatrix}=\begin{pmatrix} \sum_{i=1}^{n}\frac{\partial f_{i0}}{\partial b_1}R_i \\ \sum_{i=1}^{n}\frac{\partial f_{i0}}{\partial b_2}R_i \\ \vdots \\ \sum_{i=1}^{n}\frac{\partial f_{i0}}{\partial b_k}R_i \end{pmatrix} \tag{4-66}$$

令：

$$\boldsymbol{A}_0=\begin{pmatrix} \frac{\partial f_1}{\partial b_1} & \frac{\partial f_1}{\partial b_2} & \cdots & \frac{\partial f_1}{\partial b_k} \\ \frac{\partial f_2}{\partial b_1} & \frac{\partial f_2}{\partial b_2} & \cdots & \frac{\partial f_2}{\partial b_k} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial f_n}{\partial b_1} & \frac{\partial f_n}{\partial b_2} & \cdots & \frac{\partial f_n}{\partial b_k} \end{pmatrix}$$

$$\boldsymbol{H}=\begin{pmatrix} R_1 \\ R_2 \\ \vdots \\ R_n \end{pmatrix} \qquad \Delta\boldsymbol{B}=\begin{pmatrix} \Delta b_1 \\ \Delta b_2 \\ \vdots \\ \Delta b_k \end{pmatrix}$$

则式（4-66）可简明地记为：

$$\boldsymbol{A}_0^{\mathrm{T}}\boldsymbol{A}_0\Delta\boldsymbol{B}=\boldsymbol{A}_0^{\mathrm{T}}\boldsymbol{H} \tag{4-67}$$

式（4-67）中，矩阵 $\boldsymbol{A}_0$ 右上角符号 T 表示对矩阵 $\boldsymbol{A}_0$ 作转置运算。

解式（4-67）得：

$$\Delta\boldsymbol{B}=(\boldsymbol{A}_0^{\mathrm{T}}\boldsymbol{A}_0)^{-1}\boldsymbol{A}_0^{\mathrm{T}}\boldsymbol{H} \tag{4-68}$$

式（4-68）中（$\boldsymbol{A}_0^{\mathrm{T}}\boldsymbol{A}_0$）右上角符号－1表示对矩阵$\boldsymbol{A}_0^{\mathrm{T}}\boldsymbol{A}_0$进行求逆运算。有关矩阵转置和求逆的定义及运算法则读者有兴趣可从有关的线性代数书本中查阅。

以解得的$\Delta\boldsymbol{B}=\begin{pmatrix}\Delta b_1\\ \Delta b_2\\ \vdots\\ \Delta b_k\end{pmatrix}$修正$b_j^{(0)}$，得经一次修正后的$b_j^{(1)}$值，将$b_j^{(1)}$作为新的初值，重复上述步骤进行运算，得新的修正值$b_j^{(2)}$。每一步修正即是进行了一次搜索，直至满足判别条件：

$\max|\Delta b_j|<\mathrm{eps}$（$j=1,2,\cdots,k$）时，搜索过程结束。判别条件中eps是人为规定的允许误差。搜索结束时所得到的b_j（$j=1,2,\cdots,k$）值即认为是非线性回归方程中的待估参数值。

上述通过逐步搜索确定回归方程中待估参数的方法称为高斯-牛顿法。

二、麦夸尔特法

麦夸尔特法是一种改进的高斯-牛顿法。我们已知高斯-牛顿法计算过程中，为了估计回归方程中的参数值，需要反复迭代搜索。在每一次迭代过程中，b_j值虽然还不是真值，但总是希望比修正前的值要接近于真值。实际计算可能发生两种情况：一种是逐次迭代的结果越来越接近真值，称为“迭代收敛”；也可能发生逐次迭代结果离真值越来越远或在真值附近振荡，称为“迭代发散”。迭代收敛与发散除了和数学模型本身结构有关外，初值$b_j^{(0)}$的选取，起着重要作用。对于不少问题，选取一个较好的初值不是十分容易的，麦夸尔特法改进了高斯-牛顿法的迭代过程，放宽了初值选取的限制程度，使迭代过程收敛性更好一些。

麦夸尔特法与高斯-牛顿法的差别仅在于确定修正值Δb_j的方程组不同。麦夸尔特法在高斯-牛顿法修正值方程组的系数矩阵的对角线上加一个“阻尼因子”λ，方程组（4-67）变为：

$$(\boldsymbol{A}_0^{\mathrm{T}}\boldsymbol{A}_0+\lambda\boldsymbol{P})\Delta\boldsymbol{B}=\boldsymbol{A}_0^{\mathrm{T}}\boldsymbol{H} \tag{4-69}$$

式（4-69）中，$\boldsymbol{P}$为单位矩阵。当$\lambda=0$时，麦夸尔特法退化为高斯-牛顿法。

麦夸尔特法搜索过程的基本步骤与高斯-牛顿法相同，不同的是在运算过程中有时需要调整λ的值。一般而言，当选取λ值较大时，所得到的修正值Δb_j较小，这意味着这一步搜索过程迈步很小。而我们在搜索过程中总是希望搜索过程快一些，即每一步搜索迈步大一些。因此在确认搜索过程收敛时，总是尽可能选取λ值小一些，使搜索过程加快。

三、非线性回归结果的统计检验

非线性回归结果的统计检验与线性回归结果的统计检验基本相同，可用 F 检验来判别所得到的回归方程在统计意义上成立与否。其统计量 F 的计算公式与式（4-16）相同，即：

$$F=\frac{\frac{U}{m}}{\frac{Q}{n-m-1}} \tag{4-70}$$

非线性回归结果的统计检验也可用相关比 R^* 进行检验，相关比 R^* 的计算在形式上与复相关系数 R 相同：

$$R^*=\sqrt{1-\frac{Q}{S_{总}}} \tag{4-71}$$

需要指出的是，即使对只包含一个自变量的非线性回归方程而言，相关比不能像相关系数那样区分正负号。对非线性回归问题而言，自变量的不同变化范围可能会有不同的相关形式，在某一区间可能表现为正相关，而在另一区间可能表现为负相关。因此，对自变量的整个变化范围来说，无法用简单的正相关或负相关来表示。但无论如何，R^* 的大小反映了所得到的回归方程在统计意义上显著与否，它在客观上反映了所得到的回归方程与实验数据的拟合程度。

第五章　聚类分析基础

在环境科研和监测工作中，除了分析自变量与因变量之间的统计关系，建立变量之间的回归方程外，还广泛地存在根据环境调查和监测所得的数据将调查对象进行分类的问题，这是统计学中数字分类的问题。数字分类问题与回归分析问题一样，都是研究多个变量之间的统计联系，只是要回答的问题与解决问题的数学方法不同，本章将介绍对数字进行统计分类的常用方法——聚类分析方法。聚类分析是多元分析的一个分支，它是根据分类对象的数字指标，定量地确定它们之间的亲疏关系，并根据它们之间的相似程度进行分类。进行聚类分析时，选取不同的分类单元可以有不同的分类形式，最基本的分类单元有两种：一种以样品为分类对象，它是将各参数（变量）比较相近的样品归为同一类，表征这些样品具有相似的特征和结构，这种分类形式称为 Q 型分类；另一种是以变量为分类对象，变量归为同一类是因为它们在各个样品中的分布状况相似，变量之间具有一定的相关性，这种分类形式称为 R 型分类。

聚类分析过程主要是确定分类对象之间的相似性量度及选用恰当的分类方法。因为可供选择的公式和方法很多，而分类的结果又往往因不同公式和方法的选用而有一定的差异，因此聚类分析只是帮助我们从大量且复杂的数据中提供更多信息的一种手段。为了得到正确的分类结果，要根据分类的目的、对象及实际问题的环境特征选取恰当的公式和分类方法。

第一节　相似性量度指标

样品之间的相似程度是通过各变量的数据来确定的，变量的数据可用不同的测量尺度来表示。常用的表示变量的测量尺度有下列三类。

（1）间隔尺度　变量用连续的量来表示，有确定的数值。

（2）有序尺度　变量的量度没有明确的数量表示，只有次序关系没有数量关系。如表示降雨量大小时采用大雨、中雨或小雨，即为有序尺度。

（3）名义尺度　变量的量度既没有数量表示也没有次序关系。如表示单个人的性别、民族等，即为名义尺度。

在本章讨论的聚类分析问题，主要是针对具有间隔尺度的数据分类问题。

一、距离和相似系数

设有 N 个样品，每个样品观察了 M 个参数（变量），用 x_{ij} 表示第 i 个样品第 j 个变量的数值。则原始数据可表示为原始数据表形式（见表 5-1）。

表 5-1 原始数据表

样品	x_1	x_2	…	x_m
1	x_{11}	x_{12}	…	x_{1m}
2	x_{21}	x_{22}	…	x_{2m}
⋮	⋮	⋮	⋮	⋮
N	x_{n1}	x_{n2}	…	x_{nm}

以每个样品作为分类单元（Q型分类），则每个样品中包含了 M 个变量的观察值。从向量空间的观点看，样品为 M 维空间的一个点。为了将样品分类，需要定义样品之间相似程度的数量指标。常用的度量样品之间相似程度的数量指标主要有距离系数（简称距离）和相似系数两类，此外，在实践中还发展了其他一些定义分类单元相似程度的数量指标，下面分别予以介绍。

1. 距离

将具有不同变量值的样品看作多维空间中的点，则距离即为多维空间中点和点之间的距离，一般用 d_{ij} 表示。点和点之间的距离可以以不同的形式来定义，我们可以有种种不同内涵的距离定义和表达式，但一般要求所定义的距离应满足以下四个条件：

（1）$d_{ij}=0$　　当 $i=j$ 时

（2）$d_{ij}\geqslant 0$　　对一切 i，j

（3）$d_{ij}=d_{ji}$　　对一切 i，j

（4）$d_{ij}\leqslant d_{ik}+d_{kj}$　　对一切 i，j

常用的距离定义是明考斯基距离，其定义表达式如下：

$$d_{ij}(q)=\left[\sum_{k=1}^{m}|x_{ik}-x_{jk}|^{q}\right]^{\frac{1}{q}} \tag{5-1}$$

当 $q=1$ 时，式（5-1）简化为：

$$d_{ij}(1)=\sum_{k=1}^{m}|x_{ik}-x_{jk}| \tag{5-2}$$

式（5-2）定义的距离称为绝对距离。

当 $q=2$ 时，式（5-1）简化为：

$$d_{ij}(2)=\sqrt{\sum_{k=1}^{m}(x_{ik}-x_{jk})^{2}} \tag{5-3}$$

式（5-3）定义的距离称为欧氏距离。

明考斯基距离的意义比较直观，特别是其简化形式绝对距离和欧氏距离，意义更为明显。但它有一个缺点，即当样品的各变量单位不相同时其对

距离的影响不同。例如，土壤样品中各种元素的含量选用不同单位时对距离的影响不同。为了克服各变量量纲不同对距离产生影响不同的问题，可将原始数据进行标准化变换。设 x_{ik} 为变量 k 的第 i 个观察值，其平均值 $\overline{x}_k$ 和标准差 S_k 为：

$$\overline{x}_k=\frac{1}{n}\sum_{i=1}^{n}x_{ik} \tag{5-4}$$

$$S_k=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_{ik}-\overline{x}_k)^2} \tag{5-5}$$

式（5-4）和式（5-5）中，n 为观察次数。原始数据 x_{ik} 标准化变换式为：

$$x'_{ik}=\frac{x_{ik}-\overline{x}_k}{S_k} \tag{5-6}$$

用标准化后的数据 x'_{ik} 计算距离，消除了变量量纲不同对距离产生的影响的差异。

需要指出的是，上述距离指的是样品之间的距离，在聚类分析中还有类与类之间的距离概念。类与类之间的距离概念在下一节系统聚类法中予以介绍。

2. 相似系数法

描述分类单元之间相似程度的另一种数量指标是相似系数。相似系数常用于以变量为分类单元（R 型分类）的聚类分析问题。用 C_{ij} 表示变量 x_i 和 x_j 的相似系数，一般 C_{ij} 应满足如下条件：

（1）$|C_{ij}|\leqslant 1$　　对一切 i，j

（2）$C_{ij}=C_{ji}$　　对一切 i，j

（3）$C_{ij}\neq\pm 1$　　相当于 $x_i\neq ax_j$，a 为非零常数。

变量 x_i 和 x_j 的相似系数的定义为：

$$C_{ij}=\frac{\sum_{k=1}^{n}(x_{ki}-\overline{x}_i)(x_{kj}-\overline{x}_j)}{\sqrt{\left[\sum_{k=1}^{n}(x_{ki}-\overline{x}_i)^2\right]\left[\sum_{k=1}^{n}(x_{kj}-\overline{x}_j)^2\right]}} \tag{5-7}$$

式（5-7）与式（4-11）比较可见，相似系数实际上是变量 x_i 和 x_j 的相关系数。$|C_{ij}|$ 越接近 1，则 x_i 和 x_j 的关系越密切；$|C_{ij}|$ 越接近 0，x_i 和 x_j 的关系越疏远。

二、其他公式

除了上面介绍的表征分类单元相似程度的距离系数和相似系数外，还有很多其他形式的定义公式。

（1）数量积

$$r_{ij}=\begin{cases}1(\text{当 } i=j \text{ 时})\\ \sum_{k=1}^{m}\frac{x_{ik}\cdot x_{jk}}{M}(\text{当 } i\neq j \text{ 时})\end{cases} \tag{5-8}$$

式（5-8）中，M 为一满足下述条件的任意选择的正数：$M\geqslant\max\left(\sum_{k=1}^{m}x_{ik}\cdot x_{jk}\right)$。其中 $\max\left(\sum_{k=1}^{m}x_{ik}\cdot x_{jk}\right)$表示取括号中所出现数值中的最大值。

（2）最大最小方法

$$r_{ij}=\frac{\sum_{k=1}^{m}\min(x_{ik},x_{jk})}{\sum_{k=1}^{m}\max(x_{ik},x_{jk})} \tag{5-9}$$

式（5-9）中，min(　) 表示取括号中所出现数值中的最小值。

（3）算术平均最小方法

$$r_{ij}=\frac{\sum_{k=1}^{m}\min(x_{ik},x_{jk})}{\frac{1}{2}\sum_{k=1}^{m}(x_{ik}+x_{jk})} \tag{5-10}$$

（4）几何平均最小方法

$$r_{ij}=\frac{\sum_{k=1}^{m}\min(x_{ik},x_{jk})}{\sum_{k=1}^{m}(x_{ik}\cdot x_{jk})^{\frac{1}{2}}} \tag{5-11}$$

（5）绝对值指数方法

$$r_{ij}=\mathrm{e}^{-\sum_{k=1}^{m}|x_{ik}-x_{jh}|} \tag{5-12}$$

（6）绝对值倒数方法

$$r_{ij}=\begin{cases}1(\text{当 } i=j \text{ 时})\\ \frac{M}{\sum_{k=1}^{m}|x_{ik}-x_{jk}|}(\text{当 } i\neq j \text{ 时})\end{cases} \tag{5-13}$$

式（5-13）中，M 为保证 $0\leqslant r_{ij}\leqslant 1$ 的适当选取的正数。

（7）绝对值减数方法

$$r_{ij}=\begin{cases}1\text{（当 }i=j\text{ 时）}\\ 1-c\sum_{k=1}^{m}|x_{ik}-x_{jk}|\text{（当 }i\neq j\text{ 时）}\end{cases}\tag{5-14}$$

式（5-14）中，c 为保证 $0\leqslant r_{ij}\leqslant 1$ 的适当选取的数值。

第二节　系统聚类法

在定量地确定了分类单元之间的相似程度后就需要对分类单元进行分类。常用的分类方法有系统聚类法，逐步聚类法和模糊聚类法。逐步聚类法和模糊聚类法在后两节中分别介绍，本节主要讨论系统聚类法。

系统聚类法是目前国内使用最多的一种方法，其基本思想是：先将 N 个样品各自成一类，然后规定样品之间的距离和类与类之间的距离，选择距离最小的一对并成一个新类，计算新类和其他类的距离，再将距离最近的两类合并。这样，每次缩减一类，直至所有的样品都成为一类为止，即完成了整个归并过程。类与类之间的距离可以有不同的方法定义，因而就有了不同的系统聚类方法。下面讨论常用的系统聚类方法：最短距离法、最长距离法、重心法和类平均法。

一、最短距离法

用 d_{ij} 表示样品 i 和 j 的距离，用 G_1，G_2，…表示类。最短距离法中定义的两类之间的距离用两类间最近样品的距离来表示，类 G_p 和 G_q 的距离用 D_{pq} 表示，则

$$D_{pq}=\min(d_{ij})\quad i\in G_p,j\in G_q\tag{5-15}$$

最短距离法聚类的步骤如下。

（1）规定样品之间的距离，计算样品两两距离的对称表 $\boldsymbol{D}(0)$，$\boldsymbol{D}(0)$ 亦称距离矩阵。

（2）选择 $\boldsymbol{D}(0)$ 中的最小元素，假定其为 D_{pq}，则将 G_p 和 G_q 合并成一新类 G_r，则 $G_r=\{G_p,G_q\}$。

（3）计算新类 G_r 与其他类的距离

$$D_{rk}=\min_{\substack{i\in G_r\\ j\in G_k}}(d_{ij})=\min\left[\min_{\substack{i\in G_p\\ j\in G_k}}(d_{ij}),\min_{\substack{i\in G_q\\ j\in G_k}}(d_{ij})\right]\tag{5-16}$$

将距离矩阵 $\boldsymbol{D}(0)$ 中的 p、q 行和 p、q 列删去，加上第 r 行、r 列，得到新

的距离矩阵 $\boldsymbol{D}(1)$。

（4）对 $\boldsymbol{D}(1)$ 重复（2）、（3）两步得到距离矩阵 $\boldsymbol{D}(2)$，直至所有元素聚成一类为止。

如果某一步 $\boldsymbol{D}(k)$ 中最小的元素不止一个，则对应这些最小元素的类可以同时合并。

表 5-2 $\boldsymbol{D}(0)$ 表

	G_1	G_2	G_3	G_4	G_5	G_6
G_2	1					
G_3	4	3				
G_4	6	5	2			
G_5	8	7	4	2		
G_6	9	8	5	3	1	

【例 5-1】 设有 6 个样品，每个样品测定一个变量，测定的结果是 1、2、5、7、9、10。用最短距离法进行分类。

1. 样品间距离采用绝对距离 $d_{ij}(1)$［见式（5-2)］进行计算，得到样品两两间距离矩阵 $\boldsymbol{D}(0)$，见表 5-2。

2. $\boldsymbol{D}(0)$ 中的最小元素是 1。对应的元素 $d_{12}=d_{56}=1$，则将 G_1 和 G_2 并成 $G_7=\{G_1, G_2\}$，G_5 和 G_6 并成 $G_8=\{G_5, G_6\}$。

3. 计算新类与其他类的距离，得距离矩阵 $\boldsymbol{D}(1)$，见表 5-3。

表 5-3 $\boldsymbol{D}(1)$ 表

	G_7	G_3	G_4
G_3	3		
G_4	5	2	
G_8	7	4	2

4. $\boldsymbol{D}(1)$ 中的最小元素为 2，对应元素为 $d_{34}=d_{48}=2$，于是组成新类 $G_9=\{G_3, G_4, G_8\}$。

5. 计算 G_9 与各类的距离，得到距离矩阵 $\boldsymbol{D}(2)$，见表 5-4。

表 5-4 $\boldsymbol{D}(2)$ 表

	G_7
G_9	3

将 G_9 和 G_7 合并成一类 $G_{10}=\{G_9, G_7\}$，这时所有的样品归为一类，聚类过程中止。

上述聚类过程可以用图表示（见图 5-1），表示聚类过程的图称为聚类图。

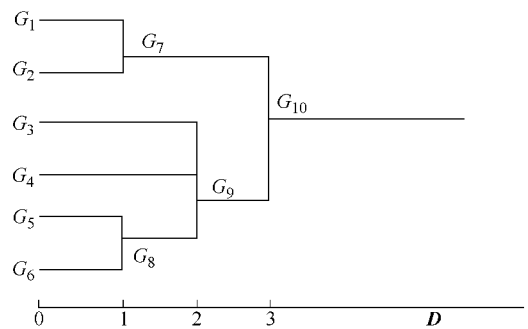


图 5-1 ［例 5-1］用最短距离法得到的聚类图

二、最长距离法

如果类与类之间的距离用两类间的最远距离来定义，则称为最长距离法。其类间距离定义式为：

$$D_{pq}=\max_{\substack{i\in G_p\\ j\in G_q}}(d_{ij}) \tag{5-17}$$

最长距离法和最短距离法的并类步骤完全一样，也是样品先自成一类，然后将最小距离的两类合并。设某一步将类 G_p 和类 G_q 合并成类 G_r，则 G_r 和 G_k 的距离为 $D_{rk}=\max\{D_{pk}, D_{qk}\}$ 再找最小距离并类，直至所有元素都成为一类为止。以［例 5-1］为例用最长距离法聚类的步骤如下：

（1）样品间的距离用绝对距离计算，得样品两两间距离矩阵 $\boldsymbol{D}(0)$，即为表 5-2。

（2）$\boldsymbol{D}(0)$ 中的最小元素为 1，对应元素为 $D_{12}=D_{56}=1$，则将 G_1 和 G_2 并成 $G_7=\{G_1, G_2\}$，G_5 和 G_6 并成 $G_8=\{G_5, G_6\}$。

表 5-5　$\boldsymbol{D}(1)$ 表

	G_7	G_3	G_4
G_3	4		
G_4	6	2	
G_8	9	5	3

（3）用最长距离法计算新类与其他类的距离，得距离矩阵 $\boldsymbol{D}(1)$，见表 5-5。

（4）$\boldsymbol{D}(1)$ 中的最小元素是 2，它是 $D_{34}=2$，于是组成新类 $G_9=\{G_3, G_4\}$。

（5）计算 G_9 与各类的距离，得到距离矩阵 $\boldsymbol{D}(2)$，见表 5-6。

表 5-6　$\boldsymbol{D}(2)$ 表

	G_9	G_7	G_8
G_7	6		
G_8	5	9	

（6）$\boldsymbol{D}(2)$ 中的最小元素是 5，它是 $D_{89}=5$，于是组成新类 $G_{10}=\{G_8, G_9\}$。

（7）计算 G_{10} 与各类的距离，得到距离矩阵 $\boldsymbol{D}(3)$，见表 5-7。

表 5-7　$\boldsymbol{D}(3)$ 表

	G_{10}
G_7	9

将 G_{10} 和 G_7 合并成一类，聚类过程结束，图 5-2 为用最长距离法聚类得到的聚类图。

用最长距离法得到的聚类结果与用最短距离法得到的聚类结果有所不同，但所分成的两大类是一致的。

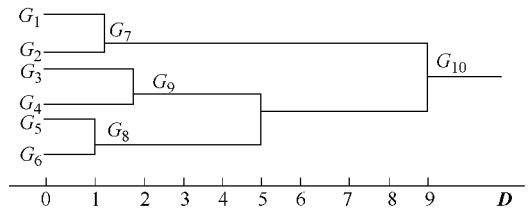


图 5-2 ［例 5-1］用最长距离法得到的聚类图

三、重心法

重心法是用一个类的重心来表示它处于多维空间中的位置，从物理意义上看，这是比较合理的。类与类之间的距离即可用重心之间的距离来表示。

设类 G_p 和类 G_q 所包含的样品数分别为 n_p 和 n_q。每个样品所测定的变量数是 m，则 G_p 和 G_q 的重心 $\overline{\boldsymbol{x}_p}$ 和 $\overline{\boldsymbol{x}_q}$ 的定义为：

$$\overline{\boldsymbol{x}_p}=\begin{pmatrix}\overline{x_{p1}}\\ \overline{x_{p2}}\\ \vdots\\ \overline{x_{pm}}\end{pmatrix} \qquad \overline{\boldsymbol{x}_q}=\begin{pmatrix}\overline{x_{q1}}\\ \overline{x_{q2}}\\ \vdots\\ \overline{x_{qm}}\end{pmatrix} \tag{5-18}$$

式（5-18）中 $\overline{\boldsymbol{x}_{pk}}$ 和 $\overline{\boldsymbol{x}_{qk}}$ 为：

$$\overline{\boldsymbol{x}_{pk}}=\frac{1}{n_p}\sum_{i\in G_p}\boldsymbol{x}_{ik} \tag{5-19}$$

$$\overline{\boldsymbol{x}_{qk}}=\frac{1}{n_q}\sum_{i\in G_q}\boldsymbol{x}_{ik} \tag{5-20}$$

则类与类之间的距离 D_{pq} 为：

$$D_{pq}=d_{\overline{x_p}\,\overline{x_q}} \tag{5-21}$$

若将类 G_p 和类 G_q 合并成新类 G_r，则显然 G_r 所含样品数 $n_r=n_p+n_q$，G_r 的重心向量 $\overline{\boldsymbol{x}_r}$ 可从 $\overline{\boldsymbol{x}_p}$ 和 $\overline{\boldsymbol{x}_q}$ 中求得：

$$\overline{\boldsymbol{x}_r}=\frac{1}{n_r}(n_p\cdot\overline{\boldsymbol{x}_p}+n_q\cdot\overline{\boldsymbol{x}_q}) \tag{5-22}$$

而新类 G_r 与其他各类 G_k 之间的距离 D_{rk} 的计算式为：

$$D_{rk}^2=\frac{n_p}{n_r}\cdot D_{pk}^2+\frac{n_q}{n_r}\cdot D_{qk}^2-\frac{n_pn_q}{n_r^2}\cdot D_{pq}^2 \tag{5-23}$$

【例 5-2】 用重心法将［例 5-1］中的六个样品聚类。

1. 计算各样品间的平方距离，得平方距离矩阵 $\boldsymbol{D}^2(0)$，见表 5-8，表 5-8 内各数据即为表 5-2 中各对应数据的平方。

2. $\boldsymbol{D}^2(0)$ 中最小数为 1，即 $\boldsymbol{D}_{12}^2(0)=\boldsymbol{D}_{56}^2(0)=1$。将 G_1 和 G_2 合并为 G_7，G_5 和 G_6 合并为 G_8。按式（5-23）计算 G_7 和 G_8 与其余各类的平方距离，得平方距离矩阵 $\boldsymbol{D}^2(1)$，见表 5-9。

表 5-8　$\boldsymbol{D}^2(0)$ 表

	G_1	G_2	G_3	G_4	G_5
G_2	1				
G_3	16	9			
G_4	36	25	4		
G_5	64	49	16	4	
G_6	81	64	25	9	1

表 5-9　$\boldsymbol{D}^2(1)$ 表

	G_7	G_3	G_4
G_3	12.25		
G_4	30.25	4	
G_8	64	20.25	6.25

表 5-10　$\boldsymbol{D}^2(2)$ 表

	G_7	G_9
G_9	20.25	
G_8	64	12.25

3. 对 $\boldsymbol{D}^2(1)$ 重复步骤 2，得到 $\boldsymbol{D}^2(2)$，见表 5-10，然后进一步重复运算得到 $\boldsymbol{D}^2(3)$，见表 5-11，至此所有样品都归为一类，聚类运算结束。

表 5-11　$\boldsymbol{D}^2(3)$ 表

	G_7
G_{10}	26.72

四、类平均法

用重心法进行系统聚类时，类与类之间的距离用重心来表示。在许多情况下，这种方法较好地反映了类与类之间的相近程度，但有时也会遇到一些问题。

从图 5-3 中，我们可以看到，图（a）中的 G_1 和 G_2 与图（b）中的 G_1 和 G_2 的相对位置是不同的，而按重心法两者的距离则完全相同，这是由于重心法只用一点来代表一类的位置所致，为了克服这一不足，有人建议将类 G_p 和类 G_q 之间的平方距离用各类元素两两之间的平方距离来表示，即：

$$D_{pq}^2=\frac{1}{n_p n_q}\sum_{\substack{i\in G_p\\ j\in G_q}} d_{ij}^2 \tag{5-24}$$

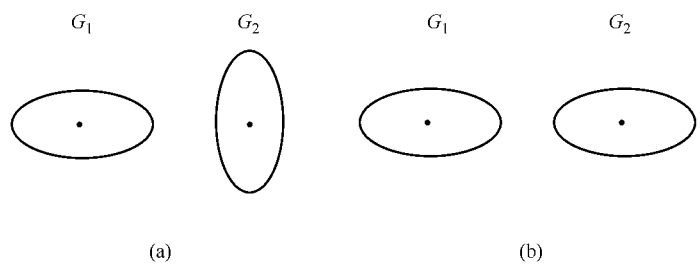


图 5-3　类 G_1 和类 G_2 不同的相对位置关系

G_p 和 G_q 组成新类 G_r 与其余类 G_k 的平方距离用式（5-25）计算：

$$D_{kr}^2=\frac{n_p}{n_r}D_{kp}^2+\frac{n_q}{n_r}D_{kq}^2 \tag{5-25}$$

类平均法克服了重心法的不足，是系统聚类法中较好的方法之一。

以上介绍的几种聚类方法，其并类的步骤是一样的，只是定义类与类之间的距离不同，从而得到不同的递推公式。其一般化的公式可表示如下：

$$D_{kr}^2=\alpha_p\cdot D_{kp}^2+\alpha_q\cdot D_{kq}^2+\beta D_{pq}^2+\gamma|D_{kp}^2-D_{kq}^2| \tag{5-26}$$

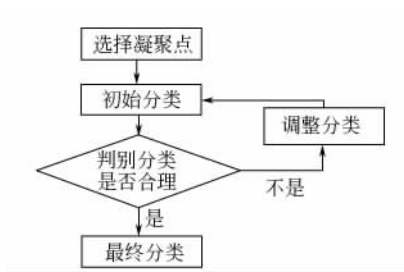
式（5-26）中的参数 α_p、α_q、β、γ 因聚类方法不同而取不同值，见表 5-12。如前所述，系统聚类法是目前最常用的聚类方法。但方法本身也存在一些不足，主要是采用系统分类法将样品并类结束后，分类结果便固定不变了。若样品进行变动或增加新的样品，聚类过程要做较大的改动。这说明方法的可变性较差。此外，由于分类过程需计算距离矩阵，当样品数量较大时计算工作量较大，用计算机计算时亦需占有较大的内存空间。为了增加聚类过程的灵活性，克服系统分类法的不足，发展了一个新的分类方法——逐步聚类法，下面加以介绍。

表 5-12 系统聚类法参数值表

聚类方法	α_p	α_q	β	γ
最短距离法	$\frac{1}{2}$	$\frac{1}{2}$	0	$-\frac{1}{2}$
最长距离法	$\frac{1}{2}$	$\frac{1}{2}$	0	$\frac{1}{2}$
重心法	$\frac{n_p}{n_r}$	$\frac{n_q}{n_r}$	$-\alpha_p\alpha_q$	0
类平均法	$\frac{n_p}{n_r}$	$\frac{n_q}{n_r}$	0	0

第三节 逐步聚类法

逐步聚类法的基本思想是在 n 个样品中选择一些有较好代表性的点，称为“凝聚点”，以此为基础对 n 个样品进行粗略的分类，然后不断修改凝聚点并让样品按某种原则向凝聚点聚集，最后得出最终分类。逐步聚类法的分类过程可简单地用下述框图表示。



一、凝聚点的选择和初始分类

如前所述，凝聚点是指一批有代表性的点，以凝聚点为中心形成类。选择凝聚点和确定初始分类的方法很多，下面选择几种主要的加以介绍。

（1）凭经验选择　选择凝聚点可以根据以往的环境调查的经验对样品所代表的区域进行粗略分类，并初步确定样品各类变量的取值范围，以此为基础可在粗略进行了分类的每一类中选择一个有代表性的样品作为凝聚点。

（2）人为随机选择　当对分类样品的特征和取值范围没有了解或了解甚少时，可人为地将样品分成 k 类，并计算每一类的重心，将这些重心作为凝聚点。

（3）密度法　先人为地选择两个正数 d_1 和 d_2，以每个样品为球心，以 d_1 为半径，落在这个球内的样品数（不包括球心的样品）叫做这个点的密度。显然，密度是样品点（即球心）和 d_1（球半径）的函数。

密度法选择凝聚点的过程是：首先选择具有最大密度的样品点作为第一凝聚点，然后选择次大密度的样品点并计算它与第一凝聚点的距离。若距离小于 d_2，则放弃该点作为凝聚点的选择；若距离大于 d_2，则选择该点作为第二凝聚点。用此方法按照样品的密度自大到小选下去，每一次和已选的任一凝聚点的距离小于 d_2 的样品取消，和已选的所有凝聚点的距离均大于 d_2 的样品作为新的凝聚点。d_1 和 d_2 的选择要适当，一般 $d_2>d_1$，通常取 $d_2=2d_1$。

为了说明用密度法选择凝聚点的过程，举一个简单的例子。

【例 5-3】　有 21 个样品，每个样品有两个观察值 x_1 和 x_2。用密度法选择该批样品的凝聚点。样品的观察数据如表 5-13 所示。

表 5-13　［例 5-3］中样品的观察值

样品序号	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
x_1	0	0	2	2	4	4	5	6	6	7	−4	−2	−3	−3	−5	1	0	0	−1	−1	−3
x_2	6	5	5	3	4	3	1	2	1	0	3	2	2	0	2	1	−1	−2	−1	−3	−5

任意两样品 i、j 之间的距离 d_{ij} 按式（5-27）进行计算：

$$d_{ij}=\sqrt{(x_{1i}-x_{1j})^2+(x_{2i}-x_{2j})^2} \tag{5-27}$$

计算了所有样品两两间的距离，选择判别半径 $d_1=2$，$d_2=4$，则可以得到各样品密度如表 5-14 所示。

表 5-14　［例 5-3］中样品密度

样品序号	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
密度	1	2	2	2	1	2	2	2	3	1	2	1	4	1	2	0	2	3	2	1	0

首先选择密度最大的13号样品作为第一凝聚点。密度其次的是9号和18号样品。按式（5-27），计算9号样品和18号样品与第一凝聚点13号样品的距离：

$$d_{9,13}=\sqrt{[6-(-3)]^2+(1-2)^2}=\sqrt{82}=9.06$$

$$d_{13,18}=\sqrt{(6-0)^2+[1-(-2)]^2}=\sqrt{45}=6.17$$

由于$d_{9,13}$和$d_{13,18}$均大于d_2，故选9号和18号样品作为第二、第三凝聚点。

密度为2的样品很多，我们按样品原次序号顺序来判别，密度为2的样品次序中第一为2号样品，计算2号样品与第一、二、三凝聚点的距离：

$$d_{2,9}=\sqrt{(0-6)^2+(5-1)^2}=\sqrt{52}=7.21$$

$$d_{2,13}=\sqrt{[0-(-3)]^2+(5-2)^2}=\sqrt{18}=4.24$$

$$d_{2,18}=\sqrt{(0-0)^2+[5-(-2)]^2}=\sqrt{49}=7$$

由于2号样品与第一、二、三凝聚点的距离均大于d_2，故选2号样品为第四凝聚点。计算其余样品与第一、二、三、四凝聚点的距离，若计算结果中，出现小于d_2的距离，则不选该点为凝聚点，反之则选该点为新的凝聚点。

按上述顺序进行计算判别，除21号样品被选作第五凝聚点外，其余均不被选作凝聚点。

确定了凝聚点之后，一般可按样品对各凝聚点的最近距离进行初始分类。有了初始分类的基础，可按一定的原则方法修改分类，最后得到最终的分类结果。

二、修改分类

对初始分类进行逐步修改的方法很多，这里仅讨论按批修改分类和逐点修改分类两种方法。

1. 按批修改法

按批修改法的步骤如下：

（1）选择凝聚点，并选择确定样品之间的距离定义；

（2）将所有样品按最近凝聚点归类；

（3）计算每一类的重心，将重心作为新的凝聚点；

（4）重复步骤（2）、（3）至所有新的凝聚点与老凝聚点重合为止。

以［例5-3］为例，说明按批修改法的过程。

（1）用密度法选择凝聚点。

取$d_1=\sqrt{5}$，$d_2=2\sqrt{5}$，得到三个凝聚点N_6（4，3）、N_{13}（−3，2）和N_{17}（0，−1）。

（2）样品间的距离用欧氏距离表示，计算各凝聚点和诸样品间的距离，见表

5-15。

表 5-15　样品和凝聚点之间的距离

	N_1	N_2	N_3	N_4	N_5	N_7	N_8	N_9	N_{10}	N_{11}	N_{12}	N_{14}	N_{15}	N_{16}	N_{18}	N_{19}	N_{20}	N_{21}
N_6	5	4.47	2.83	2	1	2.24	2.24	2.83	4.24	8	6.08	7.62	9.06	3.61	6.40	6.40	7.81	10.25
N_{13}	5	4.24	5.83	5.10	7.28	8.06	9	9.06	10.20	1.42	1	2	2	4.12	5	3.61	5.39	7
N_{17}	7	6	6.32	4.47	6.40	5.39	6.71	6.32	7.07	5.66	3.61	3.16	5.83	2.23	1	1	2.24	5

将所有样品按最近凝聚点归类，得到最初分类 $G_1^{(0)}$、$G_2^{(0)}$、$G_3^{(0)}$：

$G_1^{(0)}$ 包括：N_3、N_4、N_5、N_7、N_8、N_9、N_{10}、N_6

$G_2^{(0)}$ 包括：N_2、N_{11}、N_{12}、N_{14}、N_{15}、N_{13}

$G_3^{(0)}$ 包括：N_{16}、N_{18}、N_{19}、N_{20}、N_{21}、N_{17}

其中 N_1 因与凝聚点 N_6 和 N_{13} 的距离均为 5，暂不归类。

（3）计算各类的重心。

按式（5-18）类重心的定义，计算 $G_1^{(0)}$、$G_2^{(0)}$ 和 $G_3^{(0)}$ 的重心：

$$\boldsymbol{x}_{\overline{G_1^{(0)}}}=\begin{pmatrix}\overline{x_{11}}\\ x_{12}\end{pmatrix}=\begin{pmatrix}4.5\\ 2.375\end{pmatrix}$$

$$\boldsymbol{x}_{\overline{G_2^{(0)}}}=\begin{pmatrix}\overline{x_{21}}\\ x_{22}\end{pmatrix}=\begin{pmatrix}-2.83\\ 2.33\end{pmatrix}$$

$$\boldsymbol{x}_{\overline{G_3^{(0)}}}=\begin{pmatrix}\overline{x_{31}}\\ x_{32}\end{pmatrix}=\begin{pmatrix}-0.67\\ -1.83\end{pmatrix}$$

（4）以步骤（3）中计算所得的类重心为新的凝聚点，重复（2）、（3）两步，将样品重新分类，得到第一次迭代后的分类 $G_1^{(1)}$、$G_2^{(1)}$、$G_3^{(1)}$。

$G_1^{(1)}$ 包括：N_3、N_4、N_5、N_7、N_8、N_9、N_{10}、N_6

$G_2^{(1)}$ 包括：N_1、N_2、N_{11}、N_{12}、N_{14}、N_{15}、N_{13}

$G_3^{(1)}$ 包括：N_{16}、N_{18}、N_{19}、N_{20}、N_{21}、N_{17}

其各类的重心为：

$$\boldsymbol{x}_{\overline{G_1^{(1)}}}=\begin{pmatrix}4.5\\ 2.375\end{pmatrix}$$

$$\boldsymbol{x}_{\overline{G_2^{(1)}}}=\begin{pmatrix}-2.43\\ 2.06\end{pmatrix}$$

$$\boldsymbol{x}_{\overline{G_3^{(1)}}}=\begin{pmatrix}-0.67\\ -1.83\end{pmatrix}$$

（5）以第一次迭代后的类的重心为新的凝聚点，继续重复步骤（2）、（3）得第二次迭代后的分类 $G_1^{(2)}$、$G_2^{(2)}$、$G_3^{(2)}$。

$G_1^{(2)}$ 包括：N_3、N_4、N_5、N_7、N_8、N_9、N_{10}、N_6

$G_2^{(2)}$ 包括：N_1、N_2、N_{11}、N_{12}、N_{14}、N_{15}、N_{13}

$G_3^{(2)}$ 包括：N_{16}、N_{18}、N_{19}、N_{20}、N_{21}、N_{17}

各类的重心为：

$$\overline{\boldsymbol{x}_{G_1^{(2)}}}=\begin{pmatrix}4.5\\2.375\end{pmatrix}$$

$$\overline{\boldsymbol{x}_{G_2^{(2)}}}=\begin{pmatrix}-2.43\\2.06\end{pmatrix}$$

$$\overline{\boldsymbol{x}_{G_3^{(2)}}}=\begin{pmatrix}-0.67\\-1.83\end{pmatrix}$$

由于第二次迭代结果的重心与第一次迭代结果的重心重合，修改分类结束。

按批修改法进行逐步聚类计算工作量相对较小，聚类运算相对速度快，是一种简明实用的逐步聚类方法，按批修改法的缺点是其分类结果对最初凝聚点选择的依赖性较大，当最初凝聚点选择时带有一定主观因素时，其分类结果也就相应的带有一定的主观成分。

2. 逐点修改法

按批修改法是将样品全部归类后才修改凝聚点和初始分类，逐点修改法则是每进入一个样品就将其分类，同时不断修改凝聚点，逐点修改法的步骤如下。

（1）给出三个定数 K、C 和 R。K 表示分类数，C 表示凝聚点距离临界值，R 表示样品归类临界值，在 K、C、R 的选择时，一般要求 $R>C$。

（2）人为地选择 K 个凝聚点，并计算这 K 个凝聚点两两之间的距离。如果距离小于 C 则将相应的两个凝聚点合并，并用这两点的重心作为新的凝聚点。重复该步骤，直到所有凝聚点之间的距离均大于或等于 C 为止。

（3）将其余样品逐个输入，每进入一个样品计算该样品与所有凝聚点的距离。如果该样品与所有凝聚点的距离均大于 R，则样品作为新的凝聚点，否则将该样品归入最靠近它的凝聚点的那一类，并计算该类重心，以此新的重心作为凝聚点并计算它和其余各凝聚点的距离，若有小于 C 的，则进行合并，直至所有凝聚点之间的距离均大于或等于 C 为止。

（4）将 n 个样品重新逐个输入，按步骤（3）进行判别归类，若新的分类与前一次的分类结果相同，聚类过程结束，否则重复步骤（3）再进行判别归类。

以［例 5-3］为例，说明逐点分类法的过程。

（1）选取 $K=4$，$C=2.5$，$R=4.5$。

（2）取前 4 个样品 $N_1(0,6)$、$N_2(0,5)$、$N_3(2,5)$、$N_4(2,3)$ 作为凝聚点，点与点之间的距离用欧氏距离表示，它们之间的距离见表 5-16。

表 5-16

	N_1	N_2	N_3
N_2	1		
N_3	$\sqrt{5}$	2	
N_4	$\sqrt{13}$	$2\sqrt{2}$	2

由表 5-16 可见，最小距离 $d_{12}=1<C$。将 N_1、N_2 合并，记为 $G_1^{(0)}$ 并计算 $G_1^{(0)}$ 重心，得 $\boldsymbol{x}_{\overline{G_1^{(0)}}}=\begin{pmatrix}0\\5.5\end{pmatrix}$。然后计算该重心与凝聚点 N_3、N_4 的距离，见表 5-17。由表 5-17 可见最小距离 $d_{34}=2<C$，将 N_3 和 N_4 合并，记为 $G_2^{(0)}$，计算得其重心为 $\boldsymbol{x}_{\overline{G_2^{(0)}}}=\begin{pmatrix}2\\4\end{pmatrix}$。计算 $G_1^{(0)}$ 和 $G_2^{(0)}$ 之间的距离 $D_{12}^{(0)}=2.5=C$，故不再合并。则有新的凝聚点为 $G_1^{(0)}$ 和 $G_2^{(0)}$，其重心为 $\boldsymbol{x}_{\overline{G_1^{(0)}}}=\begin{pmatrix}0\\5.5\end{pmatrix}$ 和 $\boldsymbol{x}_{\overline{G_2^{(0)}}}=\begin{pmatrix}2\\4\end{pmatrix}$。

（3）将 N_5 输入。计算 N_5 与两个凝聚点的距离 $D_{15}=18.25$，$D_{25}=2$。由于 $D_{25}=2<R$，故 N_5 不作为新的凝聚点，将它归入 $G_2^{(0)}$，记为 $G_2^{(1)}$。计算 $G_2^{(1)}$ 的重心得 $\boldsymbol{x}_{\overline{G_2^{(1)}}}=\begin{pmatrix}2.67\\4.00\end{pmatrix}$。$\boldsymbol{x}_{\overline{G_2^{(1)}}}$ 与 $\boldsymbol{x}_{\overline{G_1^{(0)}}}$ 的距离 $D_{12}^{(1)}=3.06>C$，故这两类不合并。

表 5-17

	$G_1^{(0)}$	N_3
N_3	4.25	
N_4	10.25	2

重复该步骤，将其余样品一一输入，整个过程如表 5-18 所示。

表 5-18

输入样品	G_1		G_2		G_3		G_4	
	样品	重心	样品	重心	样品	重心	样品	重心
N_1,N_2,N_3,N_4	N_1,N_2	(0,5.5)	N_3,N_4	(2,4)				
N_5			N_5	(2.67,4)				
N_6			N_6	(3,3.75)				
N_7			N_7	(3.4,3.2)				
N_8			N_8	(3.83,3)				
N_9			N_9	(4.14,2.71)				
N_{10}			N_{10}	(4.5,2.38)				
N_{11}					N_{11}	(−4,3)		
N_{12}					N_{12}	(−3,2.5)		
N_{13}					N_{13}	(−3,2.33)		
N_{14}					N_{14}	(−3,1.75)		
N_{15}					N_{15}	(−3.4,1.8)		
N_{16}			N_{16}	(4.11,2.22)				
N_{17}					N_{17}	(−2.83,1.33)		
N_{18}					N_{18}	(−2.43,0.86)		
N_{19}					N_{19}	(−2.25,0.63)		
N_{20}					N_{20}	(−2.11,0.22)		
N_{21}							N_{21}	(−3,−5)

从表 5-18 可以看出，至输入 N_{10} 为止，只有两类，当输入 N_{11} 后，因它与两个凝聚点的距离大于 R，故 N_{11} 单独成一类 G_3，以后 N_{12}、N_{13}、N_{14}、N_{15} 也被归入 G_3，当 N_{16} 输入时，因与 G_2 重心的距离最近归入 G_2。随后 N_{17}、N_{18}、N_{19}、N_{20} 归入 G_3。最后输入 N_{21}，因它与三类的距离均大于 R，自成一类 G_4。

（4）将样品从 $N_1 \sim N_{21}$ 再输入一遍，这时 N_3 从 G_2 转入 G_1；N_{16} 从 G_2 转入 G_3，N_{20} 从 G_3 转入 G_4。得到的四类为：

G_1 包括：N_1、N_2、N_3；

G_2 包括：N_4、N_5、N_6、N_7、N_8、N_9、N_{10}；

G_3 包括：N_{11}、N_{12}、N_{13}、N_{14}、N_{15}、N_{16}、N_{17}、N_{18}、N_{19}；

G_4 包括：N_{20}、N_{21}。

（5）将样品从 $N_1 \sim N_{21}$ 重新输入一遍，这时只有 N_4 从 G_2 转到 G_1。

（6）将样品从 $N_1 \sim N_{21}$ 再输入一遍，这时样品聚类情况与前一次相同，聚类过程结束。得到的最终聚类结果为：

G_1 包括：N_1、N_2、N_3、N_4；

G_2 包括：N_5、N_6、N_7、N_8、N_9、N_{10}；

G_3 包括：N_{11}、N_{12}、N_{13}、N_{14}、N_{15}、N_{16}、N_{17}、N_{18}、N_{19}；

G_4 包括：N_{20}、N_{21}。

从上述逐点修改法聚类过程可以看出，聚类结果与定数 K、C、R 的取值有关。由于 K、C、R 的取值有一定的人为因素，因此在计算过程中，要注意结合实际问题的特征适当调整 K、C、R 的值。此外，逐点修改法的分类结果与样品的输入次序也有一定的关系。为使最终的分类结果更符合实际，需要根据问题的性质调整分类结果。

第四节　模糊聚类法

在自然科学的发展过程中，人们普遍接受这样一个原则：一个现象在能用定量的方法对其进行描述前，还不能认为对此现象有了彻底的了解。科学发展本身包含了深刻的定量化概念。然而，科学本身的深入发展又意味着研究对象的复杂化。一般而言，复杂化与精确化是难以兼容的，这就是说对极复杂的系统是难于用精确的方法进行描述的。为了解决复杂性与高度精确之间的深刻矛盾，人们必须做出这样的选择，或者将复杂的系统简化，使之能够用精确的方法进行描述，或者保持系统高度复杂性而降低描述现象的精确程度。将复杂的系统简化，使之能够用精确的方法进行描述是科学发展中所选择的主要方法。对复杂的系统，降低描述的精确程度通常被认为是学科发展不成熟的表现，从方法论上看亦发展缓慢。1965 年美国控制论专家查德（L. A. Zadeh）第一次提出了模糊集合论概念。

它为一门新的数学分支——模糊数学的产生提供了最初的思想，为描述复杂现象提供了新的概念和方法。

模糊性是普遍存在的，特别是在与人的感知辨别、决策过程有关的思维过程中存在大量的用传统的数学语言难于描述的复杂过程，模糊概念揭示了对这种复杂现象认识理解过程的实质。对复杂现象的精确描述不仅是极其困难的，并且在许多情况下也是多余的。正如一位乒乓球运动员在回球的一瞬间并不需要精确地知道乒乓球的方向和速度一样。

模糊数学作为数学的一个分支，目前发展还不是很成熟。但是，在对复杂系统的描述中，它的思想方法和概念开始突破传统数学方法难以逾越的障碍，使模糊数学在许多学科中受到普遍重视，在环境科学中也得到越来越多的应用。在这一节中将简单介绍模糊数学的基本概念和方法及在聚类分析中的应用。

一、模糊数学的基本概念

1. 集合和变量

集合和变量是数学集合论中最基本的概念。模糊数学是在普遍集合论基础上引申出来的。因此，在介绍模糊数学概念前需要了解集合和变量的基本概念。

(1) 集合　数学中集合的定义是：一些事物的全体称为一个集合，集合中的每一个事物称为集合中的元素。

集合是现代数学中最基本的概念之一。集合定义引出的概念是属于与包含。即在给定的论域U，假定u是集合$R(u)$的元素，则称u属于$R(u)$，记作$u\in R(u)$，$\in$表示属于；或者称$R(u)$包含u，记作$R(u)\ni u$，$\ni$表示包含。反之，假定u不是$R(u)$的元素，则称u不属于$R(u)$，记作$u\overline{\in}R(u)$，$\overline{\in}$表示不属于，或者称$R(u)$不包含u，记作$R(u)\overline{\ni}u$，$\overline{\ni}$表示不包含。在普通集合论中，由于$\in$和$\overline{\in}$在逻辑上是彼此否定的。因此，假定u是论域U上的任一个元素，$R(u)$是U上的一个子集，则在元素u和$R(u)$之间，要么$u\in R(u)$，要么$u\overline{\in}R(u)$，即$u\in R(u)$和$u\overline{\in}R(u)$不能都成立，也不能都不成立。

(2) 变量　集合论中变量的定义是：一个由变量名称X，论域U和论域上元素u的一个子集$R(X,u)$三个因素的组合$[X,U,R(X,u)]$称为变量。

论域U是指讨论问题的范围，它可以是有限的，也可以是无限的。元素u是指论域U上元素的单一名称，或称元素通名，而子集$R(X,u)$表达变量名称X加之于u值的一个限制。$R(X,u)$可简化为$R(X)$或$R(u)$。例如，考虑一个变量名称为“年纪”的变量。论域U可以取整数集合，子集$R(X,u)$取0～

100 间的整数，则该变量三因素可表示为：

X＝年纪

$U=\{0,1,2,3,4,\cdots\}$

$R(X,u)=\{0,1,2,3,\cdots,100\}$

2. 模糊变量和模糊子集

在普通集合论中，元素 u 和集合 $R(u)$ 之间只有属于和不属于两种关系。即要求论域上的任何一个元素或者绝对确定地属于子集 $R(u)$，或者绝对确定地不属于子集 $R(u)$，它表示的概念是边界确切的概念，而不是模棱两可的概念。但是，在现实生活中还存在许多边界不是严格确定的概念。例如“老年人”这一概念就是一个边界不严格确定的概念。对于一个 50 岁的人，很难绝对确切地判断他（她）属于或不属于“老年人”的范畴，而是处于符合与不符合之间，人们只能认为 50 岁的人在某种程度上属于“老年人”。这种不能确切地回答研究对象是否符合某一范围的概念称为模糊概念。模糊概念不能用普通集合来描述，而需要对普通集合的概念加以扩展才能表示，模糊变量和模糊子集正是在此基础上发展起来的。

（1）模糊变量　与普通变量类似，模糊变量也由三个因素组合成：变量名称 X，论域 U 和论域中的模糊子集 $R(X，u)$。模糊子集 $R(X，u)$ 中，u 是论域 U 上元素名称；$R(X，u)$ 表示变量名称 X 加之于 u 值的一个模糊限制，它可简记为 $R(X)$ 或 $R(u)$。

模糊变量的定义为：一个由变量名称 X，论域 U 和论域上元素 u 的一个模糊子集 $R(X，u)$ 三个因素的组合。

模糊变量与普通集合论中变量的区别在于加在论域中元素 u 的限制为一模糊限制。它不是确定地回答元素是否属于该子集，而是表示元素在“某种程度上”属于该子集。这里所说的“某种程度”在模糊数学中称为“隶属度”或“隶属函数”。有了“隶属度”的概念人们就能在逻辑上合理地处理模糊概念。

（2）模糊子集　模糊子集的定义为：论域 U 上的一个模糊子集 $R(X，u)$ 是指对任意的 $u\in U$，都有一个隶属程度 $\mu(0\leqslant\mu\leqslant1)$ 与之对应的子集。μ 称为模糊子集 $R(X，u)$ 的隶属函数，记作 $\mu(u)$。

当 $\mu(u)$ 仅取 0，1 两值时，元素 u 与模糊子集 $R(X，u)$ 之间的关系为“属于”（$\mu=1$）和“不属于”（$\mu=0$）两种，这时模糊子集 $R(X，u)$ 退化为边界确定的普通子集。因而可以认为普通子集是模糊子集的极端状况。由于模糊子集包含了隶属程度的概念，冲破了普通子集隶属程度只有 0 与 1 两种极端的状况，使模糊子集能够用来表达研究对象介于“属于”与“不属于”子集的中间状态的模糊概念。

在模糊子集中，变量的每一个数值的模糊限制是由一个隶属函数值所表征的。例如变量“年龄”的数值为：20、30、35。它们与模糊限制“年轻”之间相应的隶属函数值分别为：1、0.6 和 0.3，由此可以看出，“年轻”这个模糊概念可以用变量“年龄”值和与之对应的隶属函数来表示（见图 5-4）。

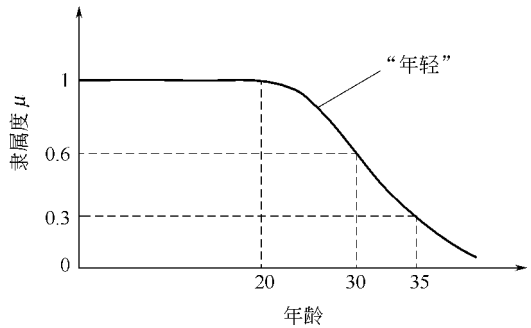


图 5-4　“年轻”的隶属函数

需要注意的一点是，隶属度概念和概率的概念是不同的。年龄 30 隶属于“年轻”这个模糊概念的程度是 0.6，它与年龄 30 的概率没有关系。隶属函数仅仅表示年龄 30 和人们关于“年轻”概念含义的匹配程度。

(3) 模糊集合的表示　在普通集合中，如一个有限集合 $R(u)=\{u_1, u_2, \cdots, u_n\}$ 可以用下式表示：

$$R(u)=u_1+u_2+\cdots+u_n \tag{5-28}$$

或

$$R(u)=\sum_{i=1}^{n}u_i \tag{5-29}$$

式 (5-28) 和式 (5-29) 中，“+”表示“联”，而不是一般四则运算中的加法。

将普通集合的表示式［式 (5-28)］推扩，对元素 u_1，u_2，…，u_n 其相应的隶属函数值为 μ_1，μ_2，…，μ_n 的模糊子集 $R(u)$ 可用下式表示：

$$R(u)=\mu_1u_1+\mu_2u_2+\cdots+\mu_nu_n \tag{5-30}$$

或：

$$R(u)=\sum_{i=1}^{n}\mu_iu_i \tag{5-31}$$

但是将普通集合的表示式［式 (5-28)］简单地推扩来表示模糊子集［即式 (5-30)］时有一个缺点：当隶属函数值 μ_i 和集合中元素值 u_i 都为数字时，μ_iu_i 的分量 μ_i 和 u_i 容易混淆，因此模糊数学的创导者查德建议使用分离符“/”，则式 (5-30) 和式 (5-31) 可以记为：

$$R(u)=\mu_1/u_1+\mu_2/u_2+\cdots+\mu_n/u_n \tag{5-32}$$

$$R(u)=\sum_{i=1}^{n}\mu_i/u_i \tag{5-33}$$

年龄 20、30、35 与模糊限制“年轻”相对应的隶属函数值为 1，0.6，0.3 的模糊子集 $Y(u)$ 可表示为：

$$Y(u)=1/20+0.6/30+0.3/35 \tag{5-34}$$

当模糊子集是连续集而不是分立集时，式（5-32）变为：

$$R(u)=\int_u \mu(u)/u \tag{5-35}$$

式（5-35）中，积分号表示项 $\mu(u)/u$ 的“联”。

上面介绍的模糊子集的表示法将元素值及与其匹配的隶属度结合起来表示。另一种模糊子集的表示法是直接用隶属函数来表示。例如代表“年老”与“年轻”两个模糊概念的模糊子集 $O(u)$ 和 $Y(u)$ 可直接用其隶属函数表示：

$$O(u)=\begin{cases}0 & (0<u<50)\\ \left[1+\left(\dfrac{u-50}{5}\right)^{-2}\right]^{-1} & (50<u<100)\end{cases} \tag{5-36}$$

$$Y(u)=\begin{cases}1 & (0<u<25)\\ \left[1+\left(\dfrac{u-25}{5}\right)^{2}\right]^{-1} & (25<u<100)\end{cases} \tag{5-37}$$

式（5-36）和式（5-37）中隶属函数的形式和结构可以根据对模糊概念的分析理解加以确定。

（4）模糊子集的水平集合　模糊子集的水平集合的定义是：若 $R(u)$ 是论域 U 的一个模糊子集，则 $R(u)$ 的一个 α 水平集合是 U 上所有在 $R(u)$ 中隶属函数值大于或者等于 α 的元素的组合，记作 $R_\alpha(u)$。α 是隶属函数值（0，1）之间的一个定数，通常称为门槛值。当 $\mu(u)\geqslant\alpha$ 时，元素 u 属于水平集合 $R_\alpha(u)$；反之，元素 u 不属于水平集合 $R_\alpha(u)$。由此可见，模糊子集 $R(u)$ 的一个 α 水平集合 $R_\alpha(u)$ 与元素 u 之间的关系只有“属于”或“不属于”，$R_\alpha(u)$ 有明确的边界。因此 α 水平集合 $R_\alpha(u)$ 是一个普通集合。$R_\alpha(u)$ 可以用式（5-38）表示：

$$R_\alpha(u)=\{u\,|\,\mu(u)\geqslant\alpha\} \tag{5-38}$$

【例 5-4】 设有模糊子集 $R(u)$ 的元素 2，1，7，6，9，其相应的隶属函数值为 0.1，0.3，0.5，0.9，1，求门槛值 α 为 0.1，0.3，0.4，0.8，1 时的水平集合。

由式（5-32）可将模糊子集 $R(u)$ 表示为：

$$R(u)=0.1/2+0.3/1+0.5/7+0.9/6+1/9$$

$\alpha=0.1$ 时，水平集合为

$$R_{0.1}(u)=\{u\,|\,\mu(u)\geqslant0.1\}=2+1+7+6+9$$

$\alpha=0.3$ 时，水平集合为

$$R_{0.3}(u)=\{u\,|\,\mu(u)\geqslant0.3\}=1+7+6+9$$

$\alpha=0.4$ 时，水平集合为

$$R_{0.4}(u)=\{u\,|\,\mu(u)\geqslant0.4\}=7+6+9$$

$\alpha=0.8$ 时，水平集合为

$$R_{0.8}(u)=\{u|\mu(u)\geqslant 0.8\}=6+9$$

$\alpha=1$ 时，水平集合为

$$R_1(u)=\{u|\mu(u)\geqslant 1\}=9$$

3. 模糊集合的基本运算

模糊集合的基本运算如下。

(1) 模糊子集的补　设 $A(u)$ 为论域 U 上的模糊子集。$A(u)$ 的隶属函数为 $\mu_A(u)$，模糊子集 $A(u)$ 的补的定义为：

$$A'(u)=\int_U^{[1-\mu_A(u)]/u} \tag{5-39}$$

式 (5-39) 中 $A'(u)$ 表示 $A(u)$ 的补。也可记作 $\neg A(u)$。

【例 5-5】 论域 $U=1+2+3+\cdots+10$ 上有模糊子集

$A(u)=0.5/3+0.3/5+1/6+0.2/8$，求 $A(u)$ 的补。

解　由模糊子集补的定义式 (5-39)，可有：

$$A'(u)=1/1+1/2+0.5/3+1/4+0.7/5+1/7+0.8/8+1/9+1/10$$

(2) 模糊子集的联　设 $A(u)$ 和 $B(u)$ 为论域 U 上的两个模糊子集，$A(u)$ 和 $B(u)$ 的隶属函数分别为 $\mu_A(u)$ 和 $\mu_B(u)$。模糊子集 $A(u)$ 和 $B(u)$ 联的定义为：

$$A(u)+B(u)=\int_U^{[\mu_A(u)\vee\mu_B(u)]/u} \tag{5-40}$$

式 (5-40) 中，$A(u)+B(u)$ 表示 $A(u)$ 和 $B(u)$ 的联，也可以记作 $A(u)\cup B(u)$。符号“$\vee$”表示 max，即取最大值的意思。

【例 5-6】 论域 $U=1+2+3+\cdots+10$ 上有模糊子集 $A(u)$ 和 $B(u)$：

$$A(u)=0.5/3+0.3/5+1/6+0.2/8$$

$$B(u)=0.2/3+0.3/4+0.7/6+1/9$$

求 $A(u)$ 和 $B(u)$ 的联。

由模糊子集联的定义式 (5-40)，可有：

$$\begin{aligned}A(u)+B(u)&=0.5/3\vee 0.2/3+0/4\vee 0.3/4+0.3/5\vee 0/5+\\&\quad 1/6\vee 0.7/6+0.2/8\vee 0/8+0/9\vee 1/9\\&=0.5/3+0.3/4+0.3/5+1/6+0.2/8+1/9\end{aligned}$$

(3) 模糊子集的交　设 $A(u)$ 和 $B(u)$ 为论域 U 上的两个模糊子集，$A(u)$ 和 $B(u)$ 的隶属函数分别为 $\mu_A(u)$ 和 $\mu_B(u)$。模糊子集 $A(u)$ 和 $B(u)$ 交的定义为：

$$A(u)\cap B(u)=\int_U^{[\mu_A(u)\wedge\mu_B(u)]/u} \tag{5-41}$$

式 (5-41) 中，$A(u)\cap B(u)$ 表示 $A(u)$ 和 $B(u)$ 的交。符号“$\wedge$”表示

min，即取最小值的意思。

【例 5-7】 论域 $U=1+2+3+\cdots+10$ 上有模糊子集 $A(u)$ 和 $B(u)$：

$$A(u)=0.5/3+0.3/5+1/6+0.2/8$$

$$B(u)=0.2/3+0.3/4+0.7/6+1/9$$

求 $A(u)$ 和 $B(u)$ 的交。

由模糊子集交的定义式（5-41）可有：

$$\begin{aligned}A(u)\cap B(u)=&0.5/3\wedge 0.2/3+0/4\wedge 0.3/4+0.3/5\wedge 0/5+\\&1/6\wedge 0.7/6+0.2/8\wedge 0/8+0/9\wedge 1/9\\=&0.2/3+0/4+0/5+0.7/6+0/8+0/9\\=&0.2/3+0.7/6\end{aligned}$$

（4）模糊子集的积 $A(u)$ 和 $B(u)$ 为论域 U 上的两个模糊子集，$A(u)$ 和 $B(u)$ 的隶属函数分别为 $\mu_A(u)$ 和 $\mu_B(u)$。模糊子集 $A(u)$ 和 $B(u)$ 积的定义为：

$$A(u)B(u)=\int_U^{\mu_A(u)\mu_B(u)/u} \tag{5-42}$$

式（5-42）中，$A(u)B(u)$ 表示 $A(u)$ 和 $B(u)$ 的积，作为积的概念的推扩，设有任意正数 k，模糊子集 $[A(u)]^k$ 可表示为：

$$[A(u)]^k=\int_U^{[\mu_A(u)]^k/u} \tag{5-43}$$

当 $k=2$ 时，则有模糊子集集中的定义式：

$$\mathrm{CON}[A(u)]=[A(u)]^2=\int_U^{[\mu_A(u)]^2/u} \tag{5-44}$$

当 $k=0.5$ 时，则有模糊子集扩张的定义式：

$$\mathrm{DIL}[A(u)]=[A(u)]^{0.5}=\int_U^{[\mu_A(u)]^{0.5}/u} \tag{5-45}$$

式（5-44）和式（5-45）中，CON 和 DIL 分别表示模糊子集“集中”和“扩张”的运算。

【例 5-8】 论域 $U=1+2+3+\cdots+10$ 上有模糊子集 $A(u)$ 和 $B(u)$：

$$A(u)=0.5/3+0.3/5+1/6+0.2/8$$

$$B(u)=0.2/3+0.3/4+0.7/6+1/9$$

求：$A(u)$ 和 $B(u)$ 的积、集中和扩张。

由于模糊子集积的定义式（5-42）有：

$$\begin{aligned}A(u)B(u)=&0.5\times(0.2/3)+0\times(0.3/4)+0.3\times(0/5)+\\&1\times(0.7/6)+0.2\times(0/8)+0\times(1/9)\\=&0.1/3+0.7/6\end{aligned}$$

由模糊子集集中的定义式（5-44），可有：

$$\mathrm{CON}[A(u)]=0.5^2/3+0.3^2/5+1^2/6+0.2^2/8$$
$$=0.25/3+0.09/5+1/6+0.04/8$$
$$\mathrm{CON}[B(u)]=0.2^2/3+0.3^2/4+0.7^2/6+1^2/9$$
$$=0.04/3+0.09/4+0.49/6+1/9$$

由模糊子集扩张的定义式（5-45），可有

$$\mathrm{DIL}[A(u)]=0.5^{0.5}/3+0.3^{0.5}/5+1^{0.5}/6+0.2^{0.5}/8$$
$$=0.7/3+0.54/5+1/6+0.45/8$$
$$\mathrm{DIL}[B(u)]=0.2^{0.5}/3+0.3^{0.5}/4+0.7^{0.5}/6+1^{0.5}/9$$
$$=0.45/3+0.54/4+0.84/6+1/9$$

（5）模糊子集的内积和外积　模糊子集内积和外积的定义为：设论域 U 上有模糊子集 $A(u)$ 和 $B(u)$，$A(u)$ 和 B(u) 内积和外积定义式如下：

$$A(u)\circ B(u)=\bigvee_{u\in U}[A(u)\wedge B(u)] \tag{5-46}$$

$$A(u)\odot B(u)=\bigwedge_{u\in U}[A(u)\vee B(u)] \tag{5-47}$$

式（5-46）和式（5-47）中，“$\circ$”表示两模糊子集内积运算符号；“$\odot$”表示两模糊子集外积运算符号。“$\vee$”和“$\wedge$”分别表示取上、下确界；对有限模糊集合，则为（2）、（3）中所介绍的表示取最大值与最小值的意思。

【例 5-9】　论域 $U=a+b+c+d$ 上有模糊子集 $A(u)$ 和 $B(u)$：

$$A(u)=0.2/a+0.4/b+0.3/c+0.8/d$$
$$B(u)=0.6/a+0.5/b+0.4/c+0.7/d$$

求：$A(u)$ 和 $B(u)$ 的内积与外积。

由内积与外积的定义式（5-46）和式（5-47），可有：

$$A(u)\circ B(u)=(0.2\wedge 0.6)\vee(0.4\wedge 0.5)\vee(0.3\wedge 0.4)\vee(0.8\wedge 0.7)$$
$$=0.2\vee 0.4\vee 0.3\vee 0.7$$
$$=0.7$$
$$A(u)\odot B(u)=(0.2\vee 0.6)\wedge(0.4\vee 0.5)\wedge(0.3\vee 0.4)\wedge(0.8\vee 0.7)$$
$$=0.6\wedge 0.5\wedge 0.4\wedge 0.8$$
$$=0.4$$

4. 隶属原则和择近原则

（1）隶属原则　模糊子集的边界不是绝对的，论域中的元素以不同的隶属度属于不同的模糊子集，这就需要有一个法则规定元素的归属。隶属原则就是解决此问题的判别原则。

隶属原则规定：

设 $R^{(1)}(u)$、$R^{(2)}(u)$、…、$R^{(n)}(u)$ 为论域上的 n 个模糊子集，u_0 是 U 中

的一个元素，若：

$$R^{(i)}(u_0)=\max[R^{(1)}(u_0),R^{(2)}(u_0),\cdots,R^{(n)}(u_0)]$$

则认为 u_0 相对隶属于 $R^{(i)}(u)$。

【例 5-10】 论域 $U=\{0,1,2,\cdots,100\}$ 上有“年老”与“年轻”两个模糊子集 $O(u)$ 和 $Y(u)$，其隶属函数分别由式（5-36）和式（5-37）定义。求年龄 35、45 和 55 属于 $O(u)$ 还是 $Y(u)$。

由式（5-36）和式（5-37），计算元素 35、45 和 55 属于模糊子集 $O(u)$ 和 $Y(u)$的隶属度。

$$O(35)=0$$

$$O(45)=0$$

$$O(55)=\left[1+\left(\frac{55-50}{5}\right)^{-2}\right]^{-1}=0.5$$

$$Y(35)=\left[1+\left(\frac{35-25}{5}\right)^{2}\right]^{-1}=0.2$$

$$Y(45)=\left[1+\left(\frac{45-25}{5}\right)^{2}\right]^{-1}=0.059$$

$$Y(55)=\left[1+\left(\frac{55-25}{5}\right)^{2}\right]^{-1}=0.027$$

$Y(35)>O(35)$，根据隶属原则，年龄 35 相对隶属于模糊子集 $Y(u)$。

$Y(45)>O(45)$，年龄 45 相对隶属于模糊子集 $Y(u)$。

$Y(55)<O(55)$，年龄 55 相对隶属于模糊子集 $O(u)$。

（2）择近原则　隶属原则解决了论域中元素相对于模糊子集的隶属关系。这是模糊判别中经常遇到的问题。除此之外，在模糊判别中还经常碰到论域 U 中若干模糊子集相互贴近程度的判别问题。这是择近原则所要回答的问题。择近原则的判别准则是贴近度。

贴近度的定义为：设 $A(u)$ 和 $B(u)$ 为论域 U 上的两个模糊子集，$A(u)$ 和 $B(u)$ 的贴近度以式（5-48）表示：

$$[A(u),B(u)]=\frac{1}{2}\{A(u)\circ B(u)+[1-A(u)\odot B(u)]\} \tag{5-48}$$

式（5-48）中$[A(u),B(u)]$ 表示模糊子集 $A(u)$ 和 $B(u)$ 的贴近度。$[A(u),B(u)]$的值越接近 1，$A(u)$ 和 $B(u)$ 越近似。

有了贴近度的概念就可以定义择近原则：

论域 U 上有 n 个模糊子集 $R^{(1)}(u)$，$R^{(2)}(u)$，…，$R^{(n)}(u)$ 及另一模糊子集 $B(u)$，若有 $1\leqslant i\leqslant n$ 使

$$[B(u),R^{(i)}(u)]=\max[B(u),R^{(j)}(u)]\quad j=1,2,3,\cdots,n \tag{5-49}$$

则称 $B(u)$ 与 $R^{(i)}(u)$ 最贴近，即 $B(u)$ 与 $R^{(n)}(u)$ 合并，匹配程度最高，

这就叫择近原则。

【例 5-11】 论域 $U=a+b+c+d$ 上有模糊子集 $A(u)$ 和 $B(u)$：

$$A(u)=0.2/a+0.4/b+0.3/c+0.8/d$$

$$B(u)=0.6/a+0.5/b+0.4/c+0.7/d$$

若论域 U 上另有一模糊子集 $C(u)$：

$$C(u)=0.5/a+0.4/b+0.5/c+0.6/d$$

问 $C(u)$ 并入 $A(u)$ 和 $B(u)$ 两子集中的哪一个更合理？

按式 (5-48)，分别求 $C(u)$ 和 $A(u)$、$B(u)$ 的贴近度：

$$[C(u),A(u)]=\frac{1}{2}\{C(u)\circ A(u)+[1-C(u)\odot A(u)]\}$$

$$=\frac{1}{2}\{(0.5\wedge 0.2)\vee(0.4\wedge 0.4)\vee(0.5\wedge 0.3)\vee(0.6\wedge 0.8)+$$
$$[1-(0.5\vee 0.2)\wedge(0.4\vee 0.4)\wedge(0.5\vee 0.3)\wedge(0.6\vee 0.8)]\}$$

$$=\frac{1}{2}[0.2\vee 0.4\vee 0.3\vee 0.6+(1-0.5\wedge 0.4\wedge 0.5\wedge 0.8)]$$

$$=\frac{1}{2}[0.6+(1-0.4)]=0.6$$

$$[C(u),B(u)]=\frac{1}{2}\{C(u)\circ B(u)+[1-C(u)\odot B(u)]\}$$

$$=\frac{1}{2}\{(0.5\wedge 0.6)\vee(0.4\wedge 0.5)\vee(0.5\wedge 0.4)\vee(0.6\wedge 0.7)+$$
$$[1-(0.5\vee 0.6)\wedge(0.4\vee 0.5)\wedge(0.5\vee 0.4)\wedge(0.6\vee 0.7)]\}$$

$$=\frac{1}{2}[0.5\vee 0.4\vee 0.4\vee 0.6+(1-0.6\wedge 0.5\wedge 0.5\wedge 0.7)]$$

$$=\frac{1}{2}[0.6+(1-0.5)]=0.55$$

由于 $[C(u),A(u)]>[C(u),B(u)]$，根据择近原则，$C(u)$ 并入 $A(u)$ 相对更合理。

二、模糊聚类分析

用模糊数学的方法进行聚类分析称为模糊聚类。由于环境科学中许多概念伴随模糊性，因此用模糊聚类的方法来处理环境科学中的许多聚类问题是十分合理的。模糊聚类分析的分类是建立在模糊关系基础上的，因此先简要地介绍模糊关系的基本概念。

1. 模糊关系基本概念

模糊关系的定义：所谓模糊集 $U(u)$ 到 $V(v)$ 的一个模糊关系 $R(u,v)$ 是

指$U(u)\times V(v)$的一个模糊子集，隶属度$\mu^{(R)}(u,v)$表示元素u与v具有关系$R(u,v)$的程度。

模糊关系定义中$U(u)\times V(v)$表示模糊集$U(u)$和$V(v)$中元素u和v之间的一种无约束的搭配，称为笛卡儿积集。在笛卡儿积集$U(u)\times V(v)$加上一个限制$R(u,v)$即表示了$U(u)\times V(v)$的一个模糊关系。$R(u,v)$是一个模糊子集，当$U(u)$和$V(v)$都是有限集合时，$R(u,v)$可用一个矩阵表示，这样的矩阵（元素是介于0，1之间的实数）称为模糊矩阵。

2. 模糊矩阵的合成运算

若$R(u,v)$是模糊集$U(u)$至$V(v)$的一个模糊关系，$S(v,w)$是模糊集$V(v)$至$W(w)$的一个模糊关系，则$R(u,v)$和$S(v,w)$的合成表示$U(u)$至$W(w)$的一个模糊关系，记为$R\circ S$，其运算式为：

$$R(u,v)\circ S(v,w)=\int_{U\times W}\bigvee\frac{\mu^{(R)}(u,v)\wedge\mu^{(S)}(v,w)}{(u,w)} \tag{5-50}$$

若$U(u)$、$V(v)$和$W(w)$是有限集合。$\boldsymbol{R}(u,v)$是$n\times m$维模糊矩阵，$\boldsymbol{S}(v,w)$是$m\times r$维模糊矩阵，则$\boldsymbol{R}(u,v)$和$\boldsymbol{S}(v,w)$的合成运算可表示为：

$$\boldsymbol{R}(u,v)\circ\boldsymbol{S}(v,w)=\bigvee_{j=1}^{m}\left[\mu^{(R)}(u,v)\wedge\mu^{(S)}(v,w)\right] \tag{5-51}$$

【例 5-12】 有模糊矩阵$\boldsymbol{R}=\begin{pmatrix}0.2 & 0.6\\ 0.4 & 0.8\\ 0.9 & 0.2\end{pmatrix}$，$\boldsymbol{S}=\begin{pmatrix}0.3 & 0.7\\ 0.1 & 0.4\end{pmatrix}$，求$\boldsymbol{R}$、$\boldsymbol{S}$的合成矩阵。

由式（5-51），可有$\boldsymbol{R}$、$\boldsymbol{S}$的合成矩阵：

$$\boldsymbol{R}\circ\boldsymbol{S}=\begin{pmatrix}0.2 & 0.6\\ 0.4 & 0.8\\ 0.9 & 0.2\end{pmatrix}\circ\begin{pmatrix}0.3 & 0.7\\ 0.1 & 0.4\end{pmatrix}$$

$$=\begin{pmatrix}(0.2\wedge0.3)\vee(0.6\wedge0.1) & (0.2\wedge0.7)\vee(0.6\wedge0.4)\\ (0.4\wedge0.3)\vee(0.8\wedge0.1) & (0.4\wedge0.7)\vee(0.8\wedge0.4)\\ (0.9\wedge0.3)\vee(0.2\wedge0.1) & (0.9\wedge0.7)\vee(0.2\wedge0.4)\end{pmatrix}$$

$$=\begin{pmatrix}0.2\vee0.1 & 0.2\vee0.4\\ 0.3\vee0.1 & 0.4\vee0.4\\ 0.3\vee0.1 & 0.7\vee0.2\end{pmatrix}=\begin{pmatrix}0.2 & 0.4\\ 0.3 & 0.4\\ 0.3 & 0.7\end{pmatrix}$$

由［例 5-12］中的运算可以看到，合成矩阵运算类似普通矩阵的乘法运算，所不同的是将普通矩阵乘法运算中的“×”改为“∧”，“+”改为“∨”。

3. 模糊聚类

以水体单元聚类为例，介绍模糊聚类的方法。

对于某个水体单元，假定它的水质以 n 个参数表示，则该水体水质可以一集合 $A=(a_1, a_2, \cdots, a_n)$ 表示。在环境质量的评价中，“环境质量”是一个模糊概念，对于单项污染参数 a_k，我们规定，隶属于“清洁水体”这一模糊概念的隶属函数如下：

$$\mu_{kj}=\begin{cases}1 & (0\leqslant a_{kj}\leqslant c_{ok})\\ \dfrac{c_{ok}}{a_{kj}} & (c_{ok}<a_{kj})\end{cases} \tag{5-52}$$

式（5-52）中，μ_{kj} 为第 j 个水体单元中第 k 项水质参数对于“清洁水体”概念的隶属函数值；a_{kj} 为第 j 个水体单元中第 k 项水质参数浓度值；c_{ok} 为地面水标准中第 k 项水质参数的浓度标准；k 为水体单元中水质参数的序数，$k=1,2,\cdots,n$。

则该水体单元水质可以用一模糊子集来表示：

$$A_j=(\mu_{1j},\mu_{2j},\cdots,\mu_{nj})$$

模糊子集 $A_i=(\mu_{1i},\mu_{2i},\cdots,\mu_{ni})$ 和 $A_j=(\mu_{1j},\mu_{2j},\cdots,\mu_{nj})$ 之间的模糊关系 $R(\mu_i, \mu_j)$ 的隶属函数选择如下形式：

$$r_{ij}=\frac{\sum_{k=1}^{n}\min(\mu_{ik},\mu_{jk})}{\sum_{k=1}^{n}\max(\mu_{ik},\mu_{jk})} \tag{5-53}$$

式（5-53）中，r_{ij} 为模糊子集 A_i 和 A_j 之间模糊关系 $R(\mu_i, \mu_j)$ 的隶属函数；i，j 为模糊子集 A 的序数，i，$j=1,2,\cdots,m$。

由式（5-53），可以计算得到表示模糊关系 $R(\mu_i, \mu_j)$ 的模糊矩阵 $\boldsymbol{R}$：

$$\boldsymbol{R}=\begin{pmatrix}r_{11} & r_{12} & \cdots & r_{1m}\\ r_{21} & r_{22} & \cdots & r_{2m}\\ \vdots & \vdots & \vdots & \vdots\\ r_{m1} & r_{m2} & \cdots & r_{mm}\end{pmatrix} \tag{5-54}$$

可以进行聚类分析的模糊矩阵 $\boldsymbol{R}$ 应具有如下三个性质：

（1）自反性：$r_{ii}=1$（$i=1,2,\cdots,m$）

（2）对称性：$r_{ij}=r_{ji}$

（3）传递性：$\boldsymbol{R}\circ\boldsymbol{R}\in\boldsymbol{R}$

由式（5-53）计算所得的模糊矩阵 $\boldsymbol{R}$［式（5-54）］已经满足了自反性和对

称性，但不一定满足传递性。对于不满足传递性的模糊矩阵 $\boldsymbol{R}$ 可以通过求传递闭包的方法加以改造，使其具有传递性，传递闭包的方法是根据模糊矩阵合成运算式（5-51）计算：

$$\boldsymbol{R}^2=\boldsymbol{R}\circ\boldsymbol{R} \tag{5-55}$$

由式（5-55）依次求出合成的模糊矩阵 $\boldsymbol{R}^2$、$\boldsymbol{R}^4$、…、$\boldsymbol{R}^{2n}$，当 $\boldsymbol{R}^{2n}=\boldsymbol{R}^{2(n+1)}$ 时，即 $r_{ij}^{2n}=r_{ji}^{2(n+1)}$ 时，则可认为模糊矩阵 $\boldsymbol{R}^{2n}$ 已具有传递性。

用取门槛值 $\alpha(0\leqslant\alpha\leqslant1)$ 的方法可以将模糊矩阵 $\boldsymbol{R}^{2n}$ 转换成为普通逻辑矩阵 $\boldsymbol{R}_\alpha$：

$$r_{ij}^1=\begin{cases}0 & (r_{ij}^{2n}<\alpha)\\ 1 & (r_{ij}^{2n}\geqslant\alpha)\end{cases} \tag{5-56}$$

式中，r_{ij}^1 为普通逻辑矩阵元素；r_{ij}^{2n} 为模糊矩阵 $\boldsymbol{R}^{2n}$ 的元素；$\boldsymbol{R}_\alpha$ 的意义为模糊矩阵 $\boldsymbol{R}^{2n}$ 的 α 水平集合［见式（5-38）］。当 α 的值由 1 下降到 0 时，所分的类由细变粗，逐渐归并为一类，形成一个动态聚类图，完成了聚类过程。

【例 5-13】 实测某水体得监测数据如表 5-19 所示。

表 5-19 ［例 5-13］中水体监测数据

水体单元	水质参数/(mg/L)				水体单元	水质参数/(mg/L)			
	COD	氨氮	酚	氰		COD	氨氮	酚	氰
1	10.6	10.5	0.16	0.24	4	40.6	7.3	0.07	0.10
2	33.4	4.4	0.02	0	5	73.8	11.5	0.14	0.05
3	12.4	9.0	0.12	0.20	6	94.7	13.2	0.48	0.06

用模糊聚类法对水体单元进行聚类。

聚类计算过程如下：

（1）查得有关水质参数的标准（即 c_{0k}）如下：COD 10mg/L；氨氮 0.5mg/L；酚 0.01mg/L；氰 0.05mg/L。由式（5-52），计算每个水体单元的隶属函数：

$A_1=(0.94,0.05,0.06,0.21)$

$A_2=(0.30,0.11,0.50,1)$

$A_3=(0.81,0.06,0.08,0.25)$

$A_4=(0.25,0.07,0.14,0.50)$

$A_5=(0.14,0.04,0.07,1)$

$A_6=(0.11,0.04,0.02,0.83)$

（2）根据式（5-53），计算得模糊矩阵 $\boldsymbol{R}$ 的元素：

$r_{11}=1$

$$r_{12}=\frac{0.30+0.05+0.06+0.21}{0.94+0.11+0.50+1}=0.24$$

$$r_{13}=\frac{0.81+0.05+0.06+0.21}{0.94+0.06+0.08+0.25}=0.85$$

$$r_{14}=\frac{0.25+0.05+0.06+0.21}{0.94+0.07+0.14+0.50}=0.35$$

$$r_{15}=\frac{0.14+0.04+0.06+0.21}{0.94+0.05+0.07+1}=0.22$$

$$r_{16}=\frac{0.11+0.04+0.02+0.21}{0.94+0.05+0.06+0.83}=0.20$$

$$r_{22}=1$$

$$r_{23}=\frac{0.30+0.06+0.08+0.25}{0.81+0.11+0.50+1}=0.29$$

$$r_{24}=\frac{0.25+0.07+0.14+0.50}{0.30+0.11+0.50+1}=0.50$$

$$r_{25}=\frac{0.14+0.04+0.07+1}{0.30+0.11+0.50+1}=0.65$$

$$r_{26}=\frac{0.11+0.04+0.02+0.83}{0.30+0.11+0.50+1}=0.52$$

$$r_{33}=1$$

$$r_{34}=\frac{0.25+0.06+0.08+0.25}{0.81+0.07+0.14+0.50}=0.42$$

$$r_{35}=\frac{0.14+0.04+0.07+0.25}{0.81+0.06+0.08+1}=0.26$$

$$r_{36}=\frac{0.11+0.04+0.02+0.25}{0.81+0.06+0.08+0.83}=0.24$$

$$r_{44}=1$$

$$r_{45}=\frac{0.14+0.04+0.07+0.50}{0.25+0.07+0.14+1}=0.51$$

$$r_{46}=\frac{0.11+0.04+0.02+0.50}{0.25+0.07+0.14+0.83}=0.52$$

$$r_{55}=1$$

$$r_{56}=\frac{0.11+0.04+0.02+0.83}{0.14+0.04+0.07+1}=0.80$$

$r_{66}=1$

模糊矩阵 $\boldsymbol{R}$ 可以表示为：

$$\boldsymbol{R}=\begin{pmatrix} 1 & 0.24 & 0.85 & 0.35 & 0.22 & 0.20 \\ 0.24 & 1 & 0.29 & 0.50 & 0.65 & 0.52 \\ 0.85 & 0.29 & 1 & 0.42 & 0.26 & 0.24 \\ 0.35 & 0.50 & 0.42 & 1 & 0.51 & 0.52 \\ 0.22 & 0.65 & 0.26 & 0.51 & 1 & 0.80 \\ 0.20 & 0.52 & 0.24 & 0.52 & 0.80 & 1 \end{pmatrix}$$

（3）改造矩阵 $\boldsymbol{R}$

$$\boldsymbol{R}^2=\boldsymbol{R}\circ\boldsymbol{R}=\begin{pmatrix} 1 & 0.35 & 0.85 & 0.42 & 0.35 & 0.35 \\ 0.35 & 1 & 0.42 & 0.52 & 0.65 & 0.65 \\ 0.85 & 0.42 & 1 & 0.42 & 0.42 & 0.42 \\ 0.42 & 0.52 & 0.42 & 1 & 0.52 & 0.52 \\ 0.35 & 0.65 & 0.42 & 0.52 & 1 & 0.80 \\ 0.35 & 0.65 & 0.42 & 0.52 & 0.80 & 1 \end{pmatrix}$$

$$\boldsymbol{R}^4=\boldsymbol{R}^2\circ\boldsymbol{R}^2=\begin{pmatrix} 1 & 0.42 & 0.85 & 0.42 & 0.42 & 0.42 \\ 0.42 & 1 & 0.42 & 0.52 & 0.65 & 0.65 \\ 0.85 & 0.42 & 1 & 0.42 & 0.42 & 0.42 \\ 0.42 & 0.52 & 0.42 & 1 & 0.52 & 0.52 \\ 0.42 & 0.65 & 0.42 & 0.52 & 1 & 0.80 \\ 0.42 & 0.65 & 0.42 & 0.52 & 0.80 & 1 \end{pmatrix}$$

$$\boldsymbol{R}^8=\boldsymbol{R}^4\circ\boldsymbol{R}^4=\begin{pmatrix} 1 & 0.42 & 0.85 & 0.42 & 0.42 & 0.42 \\ 0.42 & 1 & 0.42 & 0.52 & 0.65 & 0.65 \\ 0.85 & 0.42 & 1 & 0.42 & 0.42 & 0.42 \\ 0.42 & 0.52 & 0.42 & 1 & 0.52 & 0.52 \\ 0.42 & 0.65 & 0.42 & 0.52 & 1 & 0.80 \\ 0.42 & 0.65 & 0.42 & 0.52 & 0.80 & 1 \end{pmatrix}$$

由于 $\boldsymbol{R}^8=\boldsymbol{R}^4$，表明矩阵改造完毕。

（4）聚类　取门槛值 $\alpha(0\leqslant\alpha\leqslant1)$，模糊矩阵中元素按式（5-56）变为 0 或 1，即将模糊矩阵转换成了普通逻辑矩阵。

如取 $\alpha=0.80$，则得到矩阵：

$$\boldsymbol{R}_{\alpha=0.80}=\begin{pmatrix}1&0&1&0&0&0\\0&1&0&0&0&0\\1&0&1&0&0&0\\0&0&0&1&0&0\\0&0&0&0&1&1\\0&0&0&0&1&1\end{pmatrix} \tag{5-57}$$

由式（5-57），可将六个水体单元分成四类：

[1，3]，[2]，[4]，[5，6]

如取 $\alpha=0.60$，则得到矩阵：

$$\boldsymbol{R}_{\alpha=0.60}=\begin{pmatrix}1&0&1&0&0&0\\0&1&0&0&1&1\\1&0&1&0&0&0\\0&0&0&1&0&0\\0&1&0&0&1&1\\0&1&0&0&1&1\end{pmatrix} \tag{5-58}$$

由式（5-58），可将六个水体单元分成三类：

[1，3]，[2，5，6]，[4]

如取 $\alpha=0.50$，则得到矩阵：

$$\boldsymbol{R}_{\alpha=0.50}=\begin{pmatrix}1&0&1&0&0&0\\0&1&0&1&1&1\\1&0&1&0&0&0\\0&1&0&1&1&1\\0&1&0&1&1&1\\0&1&0&1&1&1\end{pmatrix} \tag{5-59}$$

由式（5-59），六个水体单元可分成两类：

[1，3]，[2，4，5，6]

如取 $\alpha=0.40$，则得到矩阵：

$$\boldsymbol{R}_{\alpha=0.40}=\begin{Bmatrix}1&1&1&1&1&1\\1&1&1&1&1&1\\1&1&1&1&1&1\\1&1&1&1&1&1\\1&1&1&1&1&1\\1&1&1&1&1&1\end{Bmatrix} \tag{5-60}$$

式（5-60）所表示的矩阵中，所有元素都为“1”，这样的矩阵称为“么”矩阵。此时，六个水体单元都归并为一类：

[1，2，3，4，5，6]

上述聚类过程可用一聚类图表示，见图 5-5，至此聚类过程完毕。

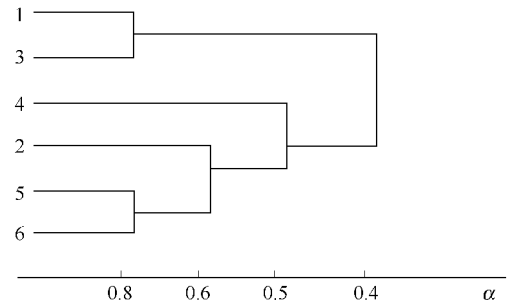


图 5-5 ［例 5-13］中六水体单元聚类图

第六章　时间序列分析基础

在研究分析环境科研和监测数据时，我们已经确认环境数据是反映随机现象的随机变量。但是在本章前，对环境数据进行的统计分析是将其作为在一定条件下的随机变量来进行的，即认为随机现象是在时间静止的情况下发生的。在这种条件下，分析环境数据所得到的环境质量特征是静态的。然而事实上环境污染现象是随时间推进而演变的过程，因此仅仅从静态的角度分析研究环境数据显然是不够的。为了从环境数据的统计分析中发掘出更多的信息，认识环境污染过程的动态特征，必须注意研究环境污染随时间变化的规律。随机变量依赖于变动参数（例如时间）的变化称为随机过程。环境污染过程是典型的随机过程，用随机过程的理论和方法来研究环境污染过程能够探索污染过程的动态特征，也是预测未来污染变化趋势的有用的方法。

随机过程理论在现代数学理论中占有重要位置，在实践上也有很广泛的应用。如信号传输过程中的白噪声、质点在液体或气体中的布朗运动、天气或气候的变化过程等。随着环境科学的发展，特别是随着环境监测技术和科研技术的发展，从大气、水质、噪声自动监测仪器得到了大量的监测数据，为污染过程的研究提供了基础，使随机过程的理论和方法在环境科学中得到了越来越广泛的应用。环境监测和科研数据的时间序列分析就是在随机过程理论基础上进行的。在实际的数据分析中，经常面对的是一些特定的随机过程，如马尔科夫过程、平稳随机过程、独立增量过程和维纳过程等。在本章中主要介绍平稳随机过程和马尔科夫过程在环境数据分析中的具体应用。

第一节　随机过程与时间序列的基本概念

一、随机过程的概念

自然界中，许多事物的变化过程遵循某种必然的变化规律，也就是说事物的变化过程可以用一个（或几个）时间 t 的确定函数来描述，这种变化过程称作确定性过程。例如地球绕太阳运转的运行轨迹，自由落体运动中落体的位置变化，声波在介质中的传播距离等，都是确定性过程。以介质中的声波传播为例，某一

介质中的声传播速度为 C，声波传播的距离有如下的关系式：

$$d=Ct \quad t\geqslant 0 \tag{6-1}$$

显然，这个关系式精确的描述了 $t\geqslant 0$ 时刻声波传播的距离。

自然界中还有另一类事物的变化过程，这类变化过程没有必然的规律，也就是说变化过程无法用时间 t 的确定函数来描述。或者事物变化全过程的某次观测结果是时间 t 的函数，但同一事物的变化过程独立地重复进行多次观测所得的结果并不相同。这类变化过程就是随机过程。环境污染过程就是一个典型的随机过程。如大气中污染物的浓度变化；排污口的排污量和浓度的变化；噪声污染源辐射强度的变化等。下面举一个具体的例子来说明随机过程的概念。

交通干线边上一测点测得的若干天 24 小时的噪声强度变化如图 6-1 所示。

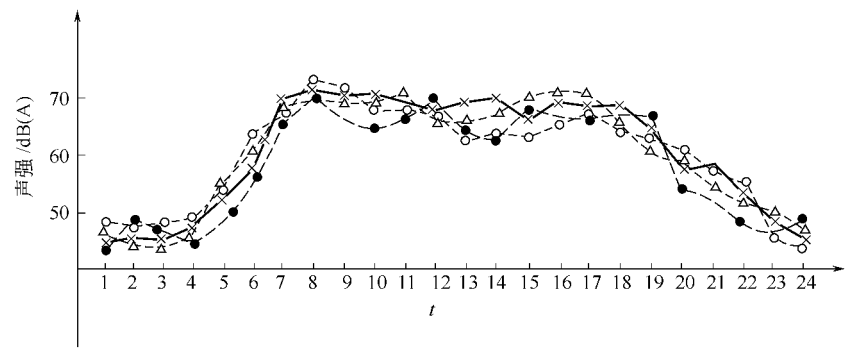


图 6-1　4 天 24 小时交通噪声强度变化

由图 6-1 可见，任何一天的交通噪声强度 24h 的变化都是无规律的、不可准确预测的。也就是说，任一时刻无法准确预测下一时刻的噪声强度。此外，同一测点的 4 天观测结果可以视为同一变化过程的 4 次独立观测结果（暂不考虑其他变化因素）。显然，同一时刻的 4 次重复观测结果完全不同，取值是随机的。

交通噪声强度受到道路车流量和车辆构成等多种随机因素的影响。事实上，交通噪声强度变化是一个弱平稳随机过程。

以上述概念为基础，我们引入随机过程的定义。

对于 $t\in T$，$X(t)$ 是依赖于时间 t 的一族随机变量，就称 $X(t)$ 是随机过程。或者对于每一个给定的 $t_i\in T$，$X(t_i)$ 都是随机变量，$X(t)$ 也称为随机过程。此处 T 代表规定此过程的所有时间点的集合，t 表示时间指标。

随机过程可以按照其状态分为连续型或离散型两类。

如果一个随机过程 $X(t)$ 对于一任意 $t_i\in T$，$X(t_i)$ 都是连续随机变量，我们就称这个随机过程为连续型随机过程。例如在地表水质量监测中，测量一段面的 COD 浓度，显然 COD 浓度的取值是在一范围内连续变化的，因而这一监测过程是连续型随机过程。

如果随机过程 $X(t)$ 对于任一 $t_i \in T$，$X(t_i)$ 都是离散型随机变量，那么就称这一随机过程为离散型随机过程。仍以地表水段面的 COD 测量为例子。如果以超标或不超标作为指标来测量，那么测量结果显然只有两种：超标和不超标，因而这样的测量过程是一个离散型随机过程。

随机过程还可以按照其时间参数的连续和离散进行分类。

如果随机过程 $X(t)$ 的参数集 T 是一有限或无限区间，则称 $X(t)$ 为连续参数随机过程，亦称随机函数（以下随机过程一词专指连续参数随机过程）。如果参数集 T 是离散的，是可列个数的集合，就称 $X(t)$ 为离散参数随机过程，又称作随机序列。例如图 6-1 的道路交通噪声强度变化过程，显然在 24 小时内参数集 T 是连续的，即 $T=\{t,0\leqslant t\leqslant 24\}$，因而道路交通噪声强度变化过程是连续参数随机过程。但实际测量交通噪声时，为了方便观测和适应数字化的要求，常常是每一小时观测一次，即参数集为 $T=\{1,2,3,\cdots,24\}$，显然道路交通噪声测量所获得的随机过程为一个随机序列：$\{X_1,X_2,X_3,\cdots,X_{24}\}$。在大多数情况下，只要数据采集合理，随机序列是可以近似描述随机过程的。

对一个固定的时刻 t 来说，随机过程表现为随机变量，通常称随机过程的任一固定时刻为随机过程的一个截口。如图 6-1 中 24 小时的任一个小时都可以作为随机过程的一个截口，截口上的任一个取值都是 t 时刻随机变量的一个样本。随机过程的任一次具体观测结果称为随机过程的一个现实。如图 6-1 是 4 天的道路交通噪声测量结果，即为随机过程的 4 个现实。观测的日子越多，显然随机过程的现实就越多。一个随机过程实际上是由无数个具有同一统计属性的现实组成的。读者在后面可以看到，有些随机过程只能获得数个或一个现实。随机过程的分析就只有在这有限个现实上进行。

随机过程的参数不一定是时间 t，还可以是别的参量。例如观测大气污染物浓度随高度变化时，可以把大气污染物浓度看作是高度 h 的随机函数。考虑污染物在河流中的扩散时，污染物的浓度可以看作是河流长度 l 的随机函数，即有 $X(l)$。一般将具有这种性质的随机过程广义地称为随机函数，在以后的分析讨论中，随机过程一词是专指参变量为时间 t 的随机函数。

二、随机过程的统计描述

根据定义，随机过程在任一固定时刻的状态是随机变量，因而可以利用随机变量的统计描述方法来描述随机过程的特性。

$X(t)$ 为一随机过程，对于任一 $t\in T$，即有 $X(t)$ 是一随机变量，它的分布函数是 t 的函数，有如下的形式：

$$F(x,t)=P\{X(t)\leqslant x\} \tag{6-2}$$

称作随机过程 $X(t)$ 的一维分布函数。若以下的积分存在：

$$F(x,t)=\int_{-\infty}^{x} f(x,t)\mathrm{d}x \tag{6-3}$$

则称 $f(x,t)$ 为随机过程 $X(t)$ 的一维概率密度函数。一维分布函数和概率密度函数描述了随机过程 $X(t)$ 在各个孤立时刻的统计特性，但不能反映随机过程在不同时刻所处状态之间的联系。

随机过程 $X(t)$ 在任意两时刻 t_1、t_2 所处的状态为 $X(t_1)$、$X(t_2)$。为了描述这两个状态之间的联系，此处引入随机变量 $X(t_1)$ 和 $X(t_2)$ 的联合分布函数，记作：

$$F(X_1,X_2;t_1,t_2)=P\{X(t_1)\leqslant x_1,X(t_2)\leqslant x_2\} \tag{6-4}$$

称为随机过程 $X(t)$ 的二维分布函数。

$$f(x_1,x_2;t_1,t_2)=\frac{\partial F(x_1,x_2;t_1,t_2)}{\partial x_1\ \partial x_2} \tag{6-5}$$

称为随机过程 $X(t)$ 的二维概率密度函数。

类似的，当 t 取 n 个时刻 t_1，t_2，…，t_n 时，随机过程 $X(t)$ 有 n 个状态：$X(t_1),X(t_2),\cdots,X(t_n)$。

n 个随机变量的联合分布函数为：

$$F(X_1,X_2,\cdots,X_n;t_1,t_2,\cdots,t_n)=P\{X(t_1)\leqslant x_1,X(t_2)\leqslant x_2,\cdots,X(t_n)\leqslant x_n\} \tag{6-6}$$

称为随机过程 $X(t)$ 的 n 维分布函数。

$$f(x_1,x_2,\cdots,x_n;t_1,t_2,\cdots,t_n)=\frac{\partial F(x_1,x_2,\cdots,x_n;t_1,t_2,\cdots,t_n)}{\partial x_1\ \partial x_2\cdots\partial x_n} \tag{6-7}$$

称为随机过程 $X(t)$ 的 n 维概率密度函数，当 n 足够大时，n 维分布函数能够较好地描述随机过程 $X(t)$ 的统计特性。但在实际问题中，给出随机过程 $X(t)$ 的 n 维分布函数非常复杂，甚至不可能，并且对大多数问题来说，也并不必要。实际上可以利用随机过程 $X(t)$ 的一些数字特征，如数学期望、方差、协方差函数、自相关函数等，描述 $X(t)$ 的基本统计特性。

随机过程 $X(t)$ 的数学期望就是这个过程的平均数，记作：

$$\mu(t)=E[X(t)]=\int_{-\infty}^{+\infty} xf(x,t)\mathrm{d}x \tag{6-8}$$

式中，$f(x,t)$ 是随机过程 $X(t)$ 的一维概率密度函数。$\mu(t)$ 又称作平均数函数，显然是时间 t 的函数。

随机过程的数学期望 $\mu(t)$ 是 $X(t)$ 的所有样本函数在时刻 t 的函数值的平均，表示了随机过程 $X(t)$ 在各个时刻的起伏中心，如图 6-2 的中间虚线所示。

随机过程的方差 $D(t)$ 的定义为：

$$D(t)=E[X(t)-\mu(t)]^2 \tag{6-9}$$

随机过程的均方差 $\sigma(t)$ 的定义为：

$$\sigma(t)=\sqrt{D(t)} \tag{6-10}$$

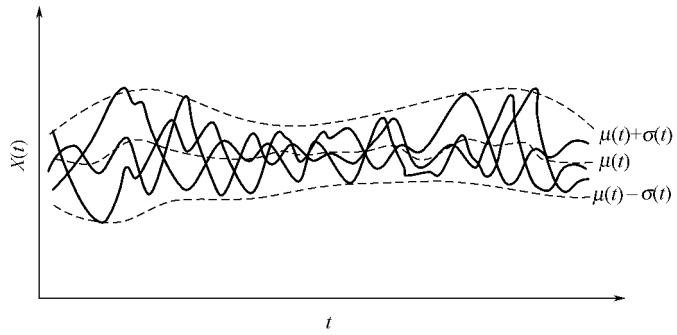


图 6-2

随机过程 $X(t)$ 的均方差 $\sigma(t)$ 反映了随机过程 $X(t)$ 任一时刻的所有可能取值相对于均值 $\mu(t)$ 的偏离程度，如图 6-2 的上下两端虚线所示。

数学期望和方差作为随机过程的两个主要数学特征，分别描述了随机过程的起伏中心和离散程度。但仅靠这两个统计特征，还不能较全面地描述随机过程。图 6-3 所示是两个数学期望和方差完全相同的随机过程。

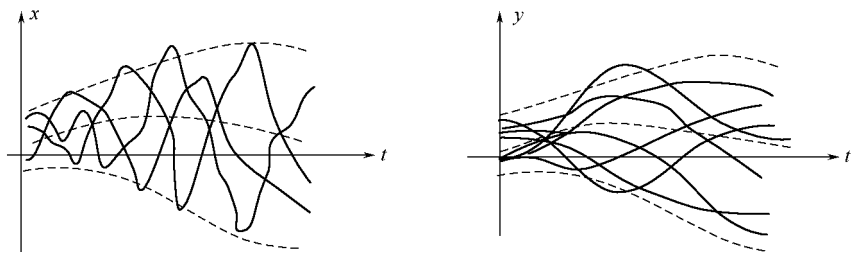


图 6-3　数学期望和方差完全相同的两个随机过程

由图 6-3 可见，两个随机过程还有其他显著不同的特征。$y(t)$ 的变化平滑、缓慢，而 $x(t)$ 的变化要起伏频繁得多、不规则得多。显然要描述随机过程的这类特征，还应引入称为自相关函数的随机过程统计特征值，记作 $R_X(t,\ t')$：

$$R_X(t,t')=E\{\{X(t)-E[X(t)]\}\{X(t')-E[X(t')]\}\}$$

$$=\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}[x_1(t)-\mu(t)][x_2(t')-\mu(t')]f(x_1,x_2,t,t')\mathrm{d}x_1\mathrm{d}x_2 \quad (6\text{-}11)$$

式中，$f(x_1,\ x_2,\ t,\ t')$ 为二维分布函数。

为了消除量纲和便于比较，通常使用自相关函数的标准化形式：

$$r_X(t,t')=\frac{R_X(t,t')}{\sigma(t)\sigma(t')} \quad (6\text{-}12)$$

$r_X(t,\ t')$ 又称作自相关系数。

自相关函数主要用来描述随机过程变化的不规则程度。它有以下几个性质。

(1) 对称性：$R_X(t,t')=R_X(t',t)$

(2) $t'=t$ 时，$R_X(t,t')=\sigma^2(t)$

(3) $r_X(t,t')$ 的取值范围为 $[-1,1]$

$r_X(t,t')=1$ 时为完全正相关；

$r_X(t,t')=0$ 时为完全不相关；

$r_X(t,t')=-1$ 时为完全负相关；

$r_X(t,t)\equiv 1$。

显然，随机过程的方差可以通过相关函数直接获得。一旦求出相关函数，就无需再单独计算方差。

环境科研和监测中，经常可以遇到相关函数分析的例子。如研究每天 24h 中每两个时刻的城市大气中 SO_2 的污染浓度的关系，就是一个随机过程的自相关函数问题。

在环境科研和监测中，也常遇到这样的问题，如研究某一时刻的污染源排放量和另一时刻的大气污染物浓度之间的联系，就必须引入互相关函数：

$$R_{XY}(t,t')=E\{\{X(t)-E[X(t)]\}\{Y(t')-E[Y(t')]\}\}$$

$$=\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}[x(t)-\mu_x(t)][y(t')-\mu_y(t')]f(x,y,t,t')\mathrm{d}x\mathrm{d}y \tag{6-13}$$

互相关系数：

$$r_{XY}(t,t')=\frac{R_{XY}(t,t')}{\sigma_x(t)\sigma_y(t')} \tag{6-14}$$

若两个随机过程是相互独立的，必然对任意的 t 和 t' 有：

$$r_{XY}(t,t')=0 \tag{6-15}$$

反过来不一定成立。

以下举一个随机过程的数字特征的计算实例。

【例 6-1】 随机过程 $X(t)$ 有如下的形式：

$X(t)=a\cos(\omega t+\theta)$ 式中，a 和 ω 是常数，θ 是在 $(0,2\pi)$ 的定义域上均匀分布的随机变量。求该随机变量的数学期望、方差和自相关函数。

显然该随机过程是一个随机相位正弦波。由定义概率密度应为：

$$f(\theta)=\begin{cases}\dfrac{1}{2}\pi & (0<\theta<2\pi)\\ 0 & (\text{其他})\end{cases}$$

因而随机过程的数学期望为：

$$\mu(t)=E[a\cos(\omega t+\theta)]=\int_0^{2\pi}a\cos(\omega t+\theta)\cdot\frac{1}{2\pi}\mathrm{d}\theta=0$$

自相关函数为：

$$R_X(t,t')=E\{[X(t)-\mu(t)][X(t')-\mu(t')]\}$$

$$=a^2\int_0^{2\pi}\cos(\omega t+\theta)\cos(\omega t'+\theta)\cdot\frac{1}{2\pi}\mathrm{d}\theta=\frac{a^2}{2}\cos[\omega(t-t')]=\frac{a^2}{2}\cos\omega\tau$$

式中　　$\tau=t-t'$

当 $t=t'$ 时，自相关函数即为随机过程的方差。

$$\sigma^2(t)=R_X(t,t')=\frac{a^2}{2}$$

相关分析技术无论是在硬件上还是软件上都很成熟。在环境监测与评价领域中应用也非常广。例如判别主要污染源的问题，就可以使用互相关技术。某一监测点附近地区受到 n 个污染源的同时污染，要鉴别污染该测点的主要污染源，最简便的方法是让附近的污染源分别排放，分别测量比较。由于实践中无法让工厂停工，污染源停止排放，这种方法不容易实行。此时可采用互相关技术进行分析。相关分析方法就是同步测量监测点和数个污染源的污染物浓度变化状况，然后将每一污染源附近的测量数据分别与监测点的测量数据做互相关分析，比较互相关函数来鉴别主要的污染源。

三、平稳随机过程

从是否“平稳”来对随机过程分类，常常会为研究分析随机过程带来许多方便。环境科研和监测中遇到的许多随机过程分析，一旦确认是平稳随机过程，问题就会大大简化。

如果随机过程的统计特性不随时间 t 而变化，就称该随机过程为平稳随机过程。严格的定义为：随机过程 $X(t)$ 的所有有限维分布函数不随时间而变化，即对任意给定的 n 和 τ，均有：

$$f_n(x_1,x_2,\cdots,x_n;t_1,t_2,\cdots,t_n)=f_n(x_1,x_2,\cdots,x_n;t_1+\tau,t_2+\tau,\cdots,t_n+\tau)\quad(6\text{-}16)$$

则称随机过程 $X(t)$ 为一平稳随机过程。显然，平稳随机过程的所有统计特性不随时间 t 的改变而改变，这是狭义平稳随机过程的定义，又称为强平稳随机过程。

在实际应用中，我们并不要求研究随机过程的所有统计特征是“平稳”的，而只要知道随机过程的几个主要的统计特征值是否“平稳”。因而需要引入平稳随机过程的另一个定义：

$$E[X(t)]=C\quad(6\text{-}17)$$

$$R_X(t,t')=R_X(t'-t)=R_X(\tau)\quad(6\text{-}18)$$

即随机过程 $X(t)$ 的任一时刻的数学期望为同一常量，自相关函数仅仅是时间间隔 τ 的函数。这是广义平稳随机过程的定义，又称为弱平稳随机过程。

根据定义，[例 6-1] 的随机相位正弦波的数学期望 $\mu(t)=0$，自相关函数 $R_X(t,t')=\frac{a}{2}\cos\omega\tau$，显然是个平稳随机过程。

在实际应用时，一个随机过程是否平稳固然可以通过概率统计方法来验证，但

更多的是通过分析问题的实质和直观现象，直接判断随机过程是否平稳。例如，在流量较高的道路两侧，在白天（8：00～16：00）8 小时中任一时刻的交通噪声强度符合正态分布，一般认为这一时段的道路交通噪声是平稳随机过程，如图 6-1 所示。

平稳随机过程的主要统计特征不随时间 t 改变，使得问题大大简化。但是如何来得到这些统计特征值呢？例如，如何求出随机过程的数学期望和自相关函数呢？按定义，如果有 n 个现实，那么随机过程的统计特征值为：

$$\mu(t)=\frac{1}{n}\sum_{j=1}^{n}x_j(t) \tag{6-19}$$

$$R(t,t')=\frac{1}{n-1}\sum_{j=1}^{n}[x_j(t)-\mu(t)][x_j(t')-\mu(t')] \tag{6-20}$$

即对任一 t 时刻的 n 个现实进行运算并求平均，就可得平稳过程的数学期望和相关函数估计值。又称为总体平均值。这种估值方法虽然较严格，但计算起来颇为烦琐。此外，在许多实际问题中，一般只能观测到一个现实，而不能获得多个现实。如某一地区的废水年排放总量等。我们现在的问题是，能否用一个现实在整个时间轴上的平均值来代替平稳随机过程的总体平均值。下面要探讨的是平稳随机过程的各态历经性。

首先看一下图 6-4 的例子，$X(t)$、$Y(t)$ 分别是两种类型的平稳随机过程。

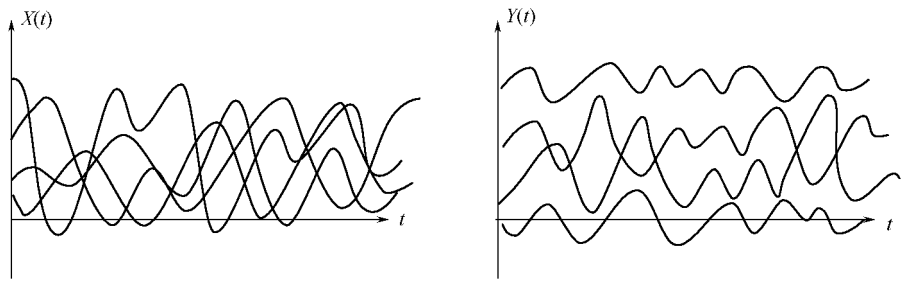


图 6-4　两种类型的平稳随机过程

由图可见，$X(t)$ 的每一个现实都随着时间 t 的增加围绕着数学期望上下波动，即沿着同一个水准在变动。而 $Y(t)$ 的每个现实都围绕着自己的一个特定水准在波动。显然对 $X(t)$ 来说，只要 t 充分长，几乎每个现实在它随时间的变化中都能经过每个可能的状态，也就是说 $X(t)$ 具有各态历经性。即，只要时间足够长，$X(t)$ 的每一个现实都能反映随机过程 $X(t)$ 的统计特征。

由于具有各态历经性的平稳随机过程的每一个现实都对随机过程的统计特性有代表性，因而在许多实际应用中常常用一个现实来得到随机过程的统计特征。例如要求出随机过程的数学期望，就可以用一个充分长时间内记录到的一个现实求平均值来近似，即：

$$E[X(t)]=\frac{1}{N}\sum_{i=1}^{N}x(t_i) \tag{6-21}$$

自相关函数为：

$$R(\tau)=\frac{1}{N-1}\sum_{i=1}^{N}[x(t_i)-\mu][x(t_i+\tau)-\mu] \tag{6-22}$$

一般来说，具有各态历经性质的平稳随机过程的自相关函数随着时间间隔 τ 逐渐增大而趋近于零，即：

$$\lim_{\tau\to\infty}R(\tau)=0 \tag{6-23}$$

【例 6-2】 某城市的大气监测点采集到的某个月 30 天的降尘数据如表 6-1 所示，该时段中降尘过程可近似作为平稳随机过程。试计算该市日降尘量的均值、方差和相关系数。

表 6-1 ［例 6-2］中大气监测点采集到的降尘数据/[t/(km² · d)]

日	浓度	日	浓度	日	浓度
1	2.0	11	0.6	21	1.1
2	2.6	12	0.6	22	2.0
3	2.2	13	1.2	23	1.8
4	1.4	14	0.7	24	2.7
5	1.5	15	1.0	25	2.8
6	2.3	16	1.1	26	2.1
7	2.7	17	1.4	27	2.2
8	1.6	18	1.6	28	3.0
9	1.6	19	1.3	29	2.0
10	0.8	20	2.1	30	1.6

解　显然作为平稳随机过程的降尘过程具有各态历经性。按定义，降尘过程的数学期望可用测得的一个现实求平均计算。

即均值为：$\bar{X}=\dfrac{\sum_{i=1}^{30}x(t_i)}{30}=1.72\text{t/(km}^2\cdot\text{d)}$

方差为：$\sigma^2=R(0)=\dfrac{1}{30-1}\sum_{i=1}^{30}[x(t_i)-\bar{X}]^2=0.458$

各项自相关系数可按下式计算：

$$r(\tau)=\frac{\dfrac{1}{30-1-\tau}\sum_{i=1}^{30-\tau}[x(t_i)-\bar{X}][x(t_i+\tau)-\bar{X}]}{\sigma^2}$$

可得到表 6-2 的数值。

表 6-2

τ	1	2	3	4	5	6	7	8	9
$r(\tau)$	0.626	0.453	0.330	0.335	0.183	0.036	−0.116	−0.265	−0.421

由表 6-2 得到的自相关系数示意图可一目了然，随着时间间隔增大，自相关系数急剧衰减。本例中的时间不够长，取值不够多，还不能明显地反映出 $\lim_{i=+\infty} R(\tau)=0$ 的规律（见图 6-5）。

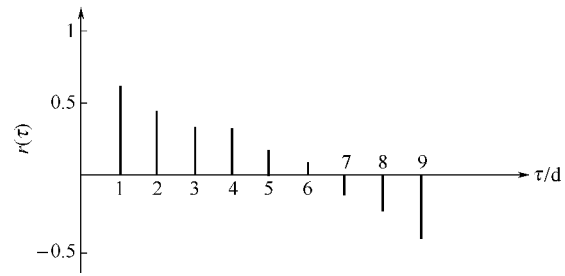


图 6-5

环境问题往往受到气象、水文等系统性因素的影响。因而严格说来，环境科研和监测中很难遇到真正的平稳随机过程。在实际处理环境污染问题时，应尽量选取合适的时段，采集恰当的时间序列，使随机过程满足或近似满足平稳条件的要求，从而运用平稳随机过程的研究处理方法和理论。

许多非平稳过程 $Y(t)$ 经过适当变换可用平稳过程 $X(t)$ 来描述，这类非平稳过程通常称作准平稳过程，最常见的准平稳过程有如下的三种类型：

$$Y(t)=X(t)+H(t) \tag{6-24}$$

$$Y(t)=X(t)H(t) \tag{6-25}$$

$$Y(t)=H_1(t)X(t)+H_2(t) \tag{6-26}$$

式中，$X(t)$ 为数学期望为零的平稳过程；$H(t)$ 为非随机的时间函数。可以证明，乘法模型和混合模型经过适当变换都可以简化为第一类的加法模型。下面讨论具有加法模型性质的非平稳时间序列。

显然，加法模型的数学期望为：

$$E[Y(t)]=E[X(t)+H(t)]=E[X(t)]+E[H(t)] \tag{6-27}$$

式中，$E[X(t)]=0$；$E[H(t)]=H(t)$即：$E[Y(t)]=H(t)$ (6-28)

式（6-27）和式（6-28）表明非随机的时间函数 $H(t)$ 就是加法模型的数学期望，因而识别准平稳过程的数学期望，就可使非平稳过程 $Y(t)$ 平稳化。

一般可以假设非随机时间函数 $H(t)$ 由两类函数叠加而成的。即：

$$H(t)=f(t)+T(t) \tag{6-29}$$

式中，$f(t)$ 为主值函数项，下一节中我们将看到 $f(t)$ 表述了时间序列的长期变化趋势；$T(t)$ 为周期函数项，表示影响时间序列的一些周期性因素。$f(t)$的识别可以利用下节介绍的趋势分析方法，$T(t)$ 的提取则可应用相关分析或谱分析等手段，此处不做详细介绍。

通常对非平稳随机过程做中心化处理，也可以实现平稳化，中心化的方法为：

$$x(t)=\frac{Y(t)-\mu(t)}{\sigma^2(t)} \tag{6-30}$$

式中，$\mu(t)$ 为非平稳过程 $Y(t)$ 的 t 时刻的数学期望；$\sigma^2(t)$ 为 t 时刻的方差，显然所产生的新的随机过程 $x(t)$ 的数学期望为零，方差为 1。

平稳随机过程的中心化在自回归等分析中也常常使用。应用中心化的平稳随机过程可以使问题大大简化。平稳随机过程 $x'(t)$ 的中心化方法为：

$$x(t)=\frac{x'(t)-\bar{X}}{\sigma^2} \tag{6-31}$$

式中，$\bar{X}$ 为平稳随机过程的数学期望；σ^2 为方差，显然中心化的平稳随机过程 $x(t)$ 的均值为零，方差为 1。

四、马尔科夫过程

随机过程的分析研究中，人们对不同时间的随机过程取值之间的关系特别予以重视，马尔科夫过程就是这种取值关系较为简单的一种随机过程。马尔科夫过程不同时间的取值关系是这样的，未来的状态仅依赖于现在的状态，而与现在状态是怎样从过去状态变化过来的无关。也就是说，这种随机过程的未来状态仅与已知的最后时刻的状态有关，而与更早的状态无关。

设随机过程 $X(t)$ 在时间 $t_1<t_2<\cdots<t_n$ 的取值为 x_1，x_2，…，x_n，马尔科夫过程必满足如下的概率关系：

$$P(x_n|x_{n-1},x_{n-2},\cdots,x_1)=P(x_n|x_{n-1}) \tag{6-32}$$

该式表明 t_n 时刻的随机过程取值仅与 t_{n-1} 时刻的取值有关，与更早时刻的取值无关。$P(x_n \mid x_{n-1})$ 是条件概率，又称为转移概率。

马尔科夫过程是无后效的随机过程。这表明这种过程的所有过去状态对未来状态的影响全部聚集在最后时刻的状态中。如果马尔科夫过程的状态取离散状态，时间也取离散时间，则这种离散的马尔科夫过程又称为马尔科夫链或简称为马氏链。

如果随机过程 $X(t)$ 共有 n 个状态，x_1，x_2，…，x_n，从状态 i 转移到状态 j 的转移概率为：

$$P(x_i|x_j)=p_{ij} \qquad (i,j=1,2,3,4,\cdots,n) \tag{6-33}$$

那么显然可以把所有的转移概率汇总在以下的转移矩阵中。由于 i，j 各有 n 个取值，因而矩阵显然也是 $n\times n$ 阶的：

$$[P_{ij}]=\begin{pmatrix} p_{11} & p_{12} & \cdots & p_{1n} \\ p_{21} & p_{22} & \cdots & p_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ p_{n1} & p_{n2} & \cdots & p_{mn} \end{pmatrix} \tag{6-34}$$

显然矩阵中每一行对应的是一个完备的事件，因而每一行的条件概率之和应为 1，即：

$$\sum_{i=1}^{n} p_{ij}=1 \tag{6-35}$$

矩阵中的任意一条件概率 p_{ij} 都必不为负值，即每一元素 $p_{ij}\geqslant 0$。

按照以上马尔科夫过程概率关系的定义，如果马尔科夫过程还满足以下的概率关系，就称之为齐次马尔科夫过程：

$$P(x_1|x_0)=P(x_2|x_1)=P(x_3|x_2)=\cdots=P(x_n|x_{n-1}) \tag{6-36}$$

表明转移概率只与状态的相对时段有关，而与在时间轴上的绝对位置无关。环境科研和监测中遇到的许多问题，都是与季节等因素有关。因而一般不是齐次马尔科夫过程。但在某一个不很长的时段中，可以近似认为满足齐次关系。

马尔科夫过程是未来状态仅取决于已知的最后时刻状态，但最后时刻却具有相对意义。例如要了解某一地区明年的大气污染平均水平，如果逐年的大气污染平均水平具有马尔科夫性质时，那么就意味着明年的大气污染平均水平仅与今年的现状有关，而与去年的污染水平无关，即今年的现状就是最后时刻的状态。但如果要了解后年的大气污染水平，显然明年的污染水平并不知道，已知的最后时刻状态仍然是今年的污染现状。这时要研究的未来状态（后年）与最后时刻状态(今年）就不是相隔一个时间步长，而是两个时间步长（两年）了。这种情况下，马尔科夫过程体现为两个时间步长后的状态仅与现状有关，而与更早的状态无关。所建立起来的马尔科夫矩阵是二阶矩阵，记作 $[p_{ij}(2)]$。

对于有 m 个步长的条件概率 $p_{ij}(m)$，显然可以把 m 步分解为 k 步和 $m-k$ 步 $(0<k<m)$。由马尔科夫过程的定义可知，从状态 x_i 经过 k 步到达状态 x_l 的转移概率为 $p_{il}(k)$，从状态度 x_l 经过 $m-k$ 步到达状态 x_j 的转移概率为 $p_{lj}(m-k)$，显然，$p_{il}(k)$ 和 $p_{lj}(m-k)$ 是独立无关的。因而应用概率乘法定理，从状态 x_i 经 m 步到达状态 x_j 的转移概率为：

$$p_{ij}(m)=\sum_{l=1}^{n} p_{il}(k)p_{lj}(m-k) \tag{6-37}$$

如令 $k=1$，则式（6-37）为：

$$p_{ij}(m)=\sum_{l=1}^{n} p_{il}p_{lj}(m-1) \tag{6-38}$$

式（6-38）可以用矩阵相乘的形式表达：

$$[p_{ij}(m)]=[P_{ij}][P_{ij}(m-1)] \tag{6-39}$$

式中，$[p_{ij}(m)]$ 为 m 个步长的转移矩阵，又称为 m 阶转移矩阵。

对于齐次马尔科夫过程，读者可证明，m 阶转移矩阵可以通过一阶转移矩阵的 m 次自乘而获得，即：

$$[p_{ij}(m)]=[P_{ij}]^m \tag{6-40}$$

表明只要有了一阶转移矩阵，就可以预测 m 个步长以后的未来状态（m 为

任意正整数，但 m 的取值必须使马氏过程满足齐次性条件）。

下一节中将具体讨论马尔科夫转移矩阵在时间序列的预测中的应用。这里仅举一个马尔科夫过程的简单例子。

【例 6-3】 如果城市的逐日降尘过程可以视为马尔科夫过程，以［例 6-2］的数据，按 0.5t/(km^2 · d) 为一档，试建立降尘过程的马尔科夫转移矩阵，并判断降尘过程是否具有马尔科夫性质。

解 根据表 6-1 的数据，可将数据分为：0.5～1.0、1.0～1.5、1.5～2.0、2.0～2.5、2.5～3.0 五个档次，即可由表建立一阶转移矩阵：

$$[p_{ij}]=\begin{bmatrix} \frac{3}{5} & \frac{2}{5} & 0 & 0 & 0 \\ \frac{1}{7} & \frac{2}{7} & \frac{2}{7} & \frac{2}{7} & 0 \\ \frac{1}{7} & \frac{1}{7} & \frac{3}{7} & 0 & \frac{2}{7} \\ 0 & \frac{2}{5} & 0 & \frac{1}{5} & \frac{2}{5} \\ 0 & 0 & \frac{2}{5} & \frac{2}{5} & \frac{1}{5} \end{bmatrix}$$

亦可建立二阶转移矩阵。

$$[p_{ij}(2)]=\begin{bmatrix} \frac{2}{5} & \frac{3}{5} & 0 & 0 & 0 \\ \frac{1}{7} & \frac{2}{7} & \frac{2}{7} & \frac{1}{7} & \frac{1}{7} \\ \frac{2}{6} & 0 & 0 & \frac{2}{6} & \frac{2}{6} \\ 0 & \frac{1}{5} & \frac{3}{5} & 0 & \frac{1}{5} \\ 0 & \frac{1}{5} & \frac{2}{5} & \frac{2}{5} & 0 \end{bmatrix}$$

根据多阶马尔科夫矩阵的定理，显然理论上二阶马尔科夫转移矩阵可由一阶转移矩阵自乘一次而获得，即有：

$$[p_{ij}(2)]=[p_{ij}(1)]^2=\begin{bmatrix} 0.417 & 0.354 & 0.114 & 0.114 & 0 \\ 0.167 & 0.294 & 0.204 & 0.139 & 0.196 \\ 0.167 & 0.312 & 0.339 & 0.155 & 0.180 \\ 0.057 & 0.194 & 0.274 & 0.314 & 0.160 \\ 0.057 & 0.217 & 0.251 & 0.160 & 0.314 \end{bmatrix}$$

比较实测数据获得的二阶转移矩阵和理论计算获得的二阶矩阵，可见差异很

明显。一般来说，样本数越多，马尔科夫过程的齐次性就越好，本例是 30 天逐日降尘时间序列，显然样本数不够多。

第二节　时间序列分析

环境科研和监测中遇到的许多问题都是连续型随机过程，但在实际分析处理这些问题时，为了简化问题，更为了满足数字化的要求，通常是以一定的时间间隔进行观测，所获得的是离散参数随机过程，也就是时间序列。如每 5 秒一次的道路交通噪声测量；大气中二氧化硫的隔日采样；地面水质监测中，一般河段的每年 6 次采样等，所获得的显然都是时间序列。

另外，有些数据只要是依时间顺序采集到的（无论是直接观测结果或是统计结果），一般都可以作为时间序列处理。

一、时间序列的趋势分析

环境科研和监测数据的趋势分析是环境数据资料分析的很重要的一部分。趋势分析一般能反映污染源的变化规律，评价环境污染控制管理措施的效益，并可做短期预报。

如前所述，一般随机过程 $Y(t)$ 通常可以表述为一个平稳随机过程 $X(t)$ 和一个非随机的时间函数 $H(t)$ 之和。$H(t)$ 反映的就是随机过程的总体变化趋势。趋势分析一般有以下几种方法。

1. 作图法

表 6-3 是某城市 18 年的二氧化硫日均值浓度。

表 6-3　某城市 18 年的二氧化硫日均值浓度/(mg/m^3)

年度	1970	1971	1972	1973	1974	1975	1976	1977	1978
浓度	0.121	0.142	0.165	0.184	0.170	0.196	0.233	0.210	0.254
年度	1979	1980	1981	1982	1983	1984	1985	1986	1987
浓度	0.306	0.271	0.346	0.315	0.280	0.430	0.324	0.307	0.504

可用上表数据做出二氧化硫浓度序列的时间演变曲线图，见图 6-6。

作图法分析时间序列的趋势，可以直接在图 6-6 的变化曲线上目估作出一条代表中心趋势的曲线或直线。这条线就代表该城市二氧化硫浓度的年变化趋势。也可以将时间按先后平分为两段，分别求两时间段上实测值的平均值，然后分别将该值点画在两时间段的中心时间上，连接两点为一直线，即为时间序列的变化趋势。本例中，1970～1978 年的二氧化硫平均值为 0.186mg/m^3，将该值在

1974 年的位置上找出。1979～1987 年的平均值为 0.342mg/m³，将该值在 1983 年的坐标上找出。连接两点，如图 6-6 中的直线所示，即为二氧化硫的年日均值变化趋势。

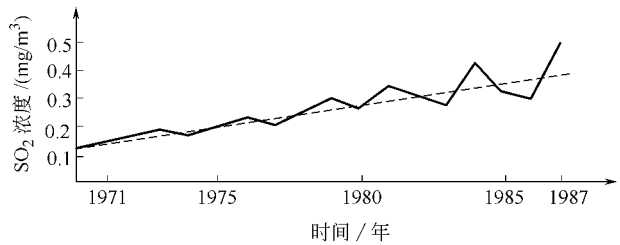


图 6-6　二氧化硫年度日均值浓度变化

这种趋势分析方法显然简单明了，但也是较粗糙的，在精度要求不很高的情况下可以使用。

2. 滑动平均法

滑动平均法又称为移动平均法，它是以一连串部分重叠的时间序列的平均值组合成新的序列的一种方法。对于涨落起伏较大的时间序列，重新组合的新序列较为平稳，可以反映时间序列的趋势值。

一般来说，若时间序列为：x_1，x_2，x_3，…，x_t，…，x_n

则从 t 时刻起取 m 项做平均有：

$$A\left(t+\frac{m-1}{2}\right)=\frac{1}{m}(x_t+x_{t+1}+\cdots+x_{t+m-1}) \tag{6-41}$$

式中，$A\left(t+\frac{m-1}{2}\right)$为 $t+\frac{m-1}{2}$时刻的平均值，显然 $t+1+\frac{m-1}{2}$时刻的滑动平均值为：

$$A\left(t+1+\frac{m-1}{2}\right)=\frac{1}{m}(x_{t+1}+x_{t+2}+\cdots+x_{t+m}) \tag{6-42}$$

当 t 取 1，2，…，$n-m+1$ 时，就可以得到整个序列的 m 项移动平均值，形成一个新的序列。显然，由定义 m 的取值以奇数最宜。

以表 6-3 的数据为例，做二氧化硫的三年滑动平均，三年滑动平均就是以每三年的数据（包括重叠部分）做平均，形成新的序列。例如 1980 年对应的滑动平均值为 1979～1981 年三年的平均值，即：

$$\overline{x_{1980}}=\frac{1}{3}(0.306+0.271+0.346)=0.308$$

依此，可以逐项计算出各年的滑动平均值，获得新的序列，如表 6-4 所示。

滑动平均后的序列作图如下（见图 6-7），与图 6-6 比较，显然曲线变平滑得多，趋势可以一目了然。

表 6-4 新的序列/(mg/m³)

年度	1970	1971	1972	1973	1974	1975	1976	1977	1978
滑动值		0.143	0.164	0.173	0.183	0.200	0.213	0.232	0.257
年度	1979	1980	1981	1982	1983	1984	1985	1986	1987
滑动值	0.277	0.308	0.311	0.314	0.342	0.345	0.354	0.378	

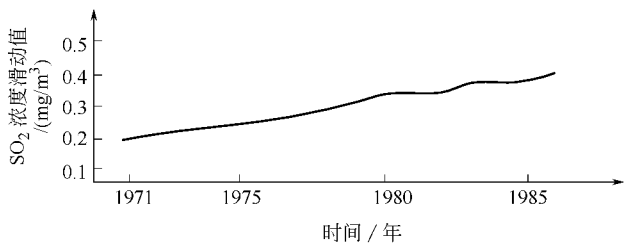


图 6-7 三年滑动平均二氧化硫变化趋势

一般来说，移动平均的时段越长，即滑动平均项 m 越大，时间序列的涨落起伏就被削弱得越多，滑动平均后的序列就越平滑，越能表现出趋势。对于有明显周期性的时间序列，滑动平均的时段最好取周期的整数倍，可以大大削减原序列的起伏，滤去序列中的周期性波动，突出时间序列的趋势。

3. 函数拟合法

一时间序列为 x_1，x_2，…，x_n，函数拟合法就是要找出一个时间的函数来表征这个时间序列的趋势。一般来说，序列总可以用一个多项式来表示：

$$A(t)=a_0+a_1t+a_2t^2+\cdots+a_mt^m \tag{6-43}$$

显然，用最小二乘法求得多项式函数的各项系数 a_0，a_1，…，a_m，即可完成多项式函数拟合。事实上，这就是一种多元回归的趋势分析方法。

在实际应用中，应根据实测数据的情况具体选取回归方程。显然取 $m=1$ 时为线性回归的情况，拟合的趋势为一直线。$m>1$ 时则趋势为一曲线。

函数拟合方法较作图法和滑动平均法复杂一些，但所获得的趋势是用函数形式来表征的，这对于进一步分析和预报是较便利的。因而在有些情况下，应采用这种拟合分析趋势的方法。

二、周期分析

随机过程讨论中，常用到“涨落起伏”来描述。其实“涨落起伏”就是许多大大小小的周期变化的迭加。在进行趋势分析时，我们总希望多滤掉一些周期性起伏，突出时间序列的变化趋势。但在有些分析中，例如分析一年四季大气污染

物浓度的变化规律，几个水期的水质变化规律等，我们非但不能消除周期性变化规律，反而要显示时间序列的这一特征，要进行周期分析。

先看一个周期变化的例子。图 6-8 是某测点测到的连续 30 天的氮氧化物浓度的变化。

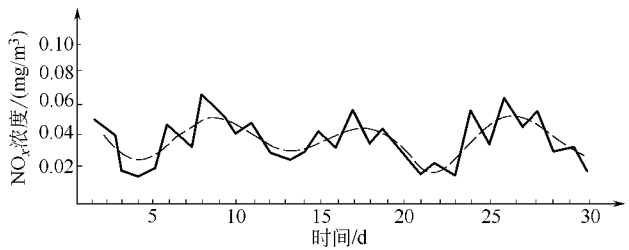


图 6-8　氮氧化物日均值浓度变化曲线

显然，从图中不容易发现明显的周期变化规律。但如果对该时间序列做滑动平均处理，就可以发现有一个以 9 天为周期的变化规律。当然，图 6-8 的曲线还包括许多周期更短的变化规律，要分析这些规律，就必须借助傅里叶变换技术，从频率域上获得时间序列的主要变化周期。

傅里叶级数理论表明，以任一周期 T 为区间的时间函数 $f(t)$，在满足狄氏条件时，总可以表述成一个基波及一系列高次谐波之和，即展开为如下的傅里叶级数：

$$f(t) = a_0 + \sum_{k=1}^{\infty}(a_k\cos\omega_k t + b_k\sin\omega_k t) \tag{6-44}$$

式中，$\omega_k = 2\pi k/T$ 称为第 k 个谐波的圆频率。在实际应用中，通常使用的参量是频率 f，与圆频率有如下关系：

$$f = \frac{\omega}{2\pi} \tag{6-45}$$

由三角函数的性质，显然傅里叶展开式中的各项系数 a_0，a_k，b_k 可由下式得到：

$$\begin{cases} a_0 = \dfrac{1}{T}\displaystyle\int_{-\frac{T}{2}}^{\frac{T}{2}} f(t)\,\mathrm{d}t \\ a_k = \dfrac{2}{T}\displaystyle\int_{-\frac{T}{2}}^{\frac{T}{2}} f(t)\cos\omega_k t\,\mathrm{d}t \\ b_k = \dfrac{2}{T}\displaystyle\int_{-\frac{T}{2}}^{\frac{T}{2}} f(t)\sin\omega_k t\,\mathrm{d}t \end{cases} \tag{6-46}$$

式中，a_0 在周期分析中称为直流成分，亦即时间 T 区间内 $f(t)$ 的均值。对于中心化的随机过程（即数学期望取零），显然有 $a_0 = 0$。频谱分析中最感兴趣的是 a_k 和 b_k。

利用欧拉公式，傅氏级数展开式也可以写成如下的形式：

$$f(t)=\sum_{k=-\infty}^{+\infty} c_k e^{i\omega_k t} \tag{6-47}$$

式中，c_k 称为复谱，有复数形式：

$$c_k=A_k e^{-i\theta_k} \qquad (c_0=a_0) \tag{6-48}$$

其中：

$$A_k=\frac{1}{2}\sqrt{a_k^2+b_k^2};\ \theta_k=-\mathrm{tg}^{-1}\left(\frac{b_k}{a_k}\right)$$

A_k 称为 $f(t)$ 的振幅谱，θ_k 称为相位谱。

在实际应用中，为了直观描述每个频率成分的比重，通常使用功率谱 p_k。功率谱的定义是复谱 c_k 的共轭乘积，即：

$$p_k=c_k\cdot c_k^*=\frac{1}{4}(a_k^2+b_k^2) \tag{6-49}$$

实际上就是振幅谱的平方。对于数学期望为零的随机过程 $f(t)$，读者很容易证明，所有频域上的功率谱之和正好等于 $f(t)$ 的方差 σ^2，因而功率谱又被称作方差谱。

对于时间序列 x_1，x_2，…，x_n，显然若取值间距为 τ，那么在 T 时间内共有 $N=T/\tau$ 个数值，傅里叶级数的系数为：

$$\begin{cases} a_0 = \dfrac{1}{N}\displaystyle\sum_{t=1}^{N} x_t \\ a_k = \dfrac{2}{N}\displaystyle\sum_{t=1}^{N} x_t\cos\omega_k t \\ b_k = \dfrac{2}{N}\displaystyle\sum_{t=1}^{N} x_t\sin\omega_k t \end{cases} \tag{6-50}$$

式中的 ω_k 此时为：

$$\omega_k=\frac{2\pi k}{N} \tag{6-51}$$

下面举一个时间序列的功率谱计算例子。

【例 6-4】 试对图 6-8 的时间序列做功率谱分析。图 6-8 对应的数据见表 6-5。

表 6-5 氮氧化物日均浓度变化表/(mg/m³)

日期	1	2	3	4	5	6	7	8	9	10
浓度	0.52	0.43	0.21	0.18	0.22	0.46	0.34	0.63	0.54	0.41
日期	11	12	13	14	15	16	17	18	19	20
浓度	0.48	0.32	0.28	0.34	0.48	0.37	0.59	0.34	0.44	0.34
日期	21	22	23	24	25	26	27	28	29	30
浓度	0.19	0.22	0.18	0.58	0.36	0.62	0.46	0.57	0.33	0.35

解　按时间序列频谱分析的方法，本例中显然有：

$$N=30 \qquad a_0=\frac{1}{N}\sum_{t=1}^{N}x_t=0.393$$

a_k，b_k，p_k 的计算结果见表 6-6。

表 6-6　计算结果

K	T/d	a_k	b_k	p_k
1	30	0.002	0.008	1.68×10^{-5}
2	15	0.020	−0.066	1.19×10^{-3}
3	10	0.009	−0.100	2.53×10^{-3}
4	7.5	0.004	0.069	1.18×10^{-3}
5	6	0.001	0.043	4.69×10^{-4}
6	5	0.002	0.020	1.05×10^{-4}
7	4.3	−0.031	0.009	2.65×10^{-4}
8	3.75	−0.024	0.032	4.06×10^{-4}

根据计算结果，可以得到频谱分析图（见图 6-9）。

所获得的是一组以 $\Delta\omega=1/N$ 为间隔的离散谱图，也就是说各次谐波的周期只能取 N/K。本例中周期长度只能取 30 天、15 天、10 天……。在许多情况下，实际时间序列存在的主要周期成分并不一定与给定的这些周期相重合，因而在 N 不太大的情况下（即时间序列不够长），漏掉一些主要周期成分的可能性很大，尤其是在低频段中。

图 6-9 表明氮氧化物日均浓度变化存在一个以 10 天为周期的主成分。这显然与我们对图 6-8 分析得到的 9 天为主周期的结论不一致。很明显，这是由于谐波分析中离散并给定的周期造成的。遇到这种情况，比较简单的方法是将序列长度图做调整，即稍稍改变 N 值，再做频谱分析。N 值改变的原则是使较明显的主要低频成分不漏掉。在上例中，如果取 $N=27$，读者可以验算，获得的谱图应为图 6-10。

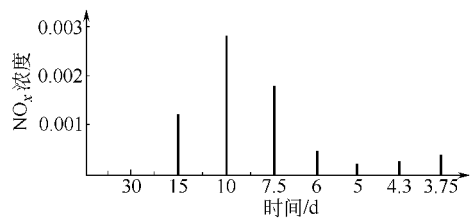


图 6-9　氮氧化物日均浓度变化谱图（一）

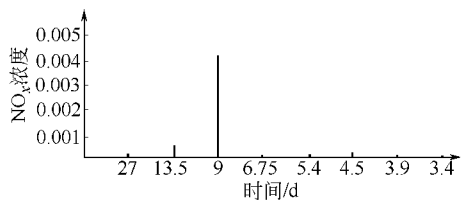


图 6-10　氮氧化物日均浓度变化谱图（二）

比较图 6-9 和图 6-10 的谱图可见，以 9 天为周期的主成分十分显著地突出来了。用这种方法避免低频主成分的漏失是有效的。

本例中只计算到 $K=8$，读者可以尝试一下 K 大于 8 的高频成分。一般来说 K 也不宜太大。后面我们将看到，$K>N/2$ 的高频成分就已经没有意义了。

读者可以发现，时间序列的傅氏频谱计算的计算量是非常大的，尤其是当序列很长的时候，即使借助于计算机，工作量还是较大，因而有必要简单介绍一种计算机使用的快速傅氏变换方法。

快速傅氏变换方法又称 FFT 技术，是由 Cooley 和 Tukey 于 1965 年提出的。FFT 技术就是综合傅氏变换和计算机运算的原理，将傅里叶系数的运算进行巧妙的组合，大大减少计算机的循环运算量，节省计算时间，FFT 技术早已被制成硬件或软件，读者可以根据具体情况选用，此处不做进一步的介绍。

事实上，功率谱与时间序列的相关函数是一对傅里叶变换对，也就是说互成傅里叶变换，因而在实际计算中，亦可直接用相关函数来分析频谱。如前所述，自相关函数包含了原时间序列的所有周期变化特征。

当时间 $T\to+\infty$（$N\to+\infty$）时，我们对随机过程的认识最全面。此时获得的频谱为连续谱。每个频率上对应的值称为功率谱密度。显然此时没有频率成分漏失，可知当 N 很大时，低频主成分不易漏失。美国环保局建议，时间序列的长度 T 应至少比感兴趣的最低频率成分的周期长 10 倍。例如我们感兴趣的是日、季、年的变化规律，那么至少应有 10 年的数据进行频谱分析。

周期分析在环境科研和监测中有很高的应用价值，主要有以下几个方面的应用。

1. 短期预报

掌握了时间序列的几个主要周期成分，一般就能大致判断未来近期的变化规律。如图 6-8 所示的氮氧化物浓度变化序列，已知其主周期是 9 天，那么显然 30 天后的日均值浓度变化规律可以根据这一周期来判断。当然在实际应用中除了利用主要周期外，还应考虑其他变化要素，共同建立起来的预报模式才具有一定的准确性。

2. 污染源鉴别

一般来说每一污染源都有各自的排放规律。从时域上说有各自的随机过程，从频域上来说就是有各自典型的排放周期图。周期分析鉴别主要污染源就是利用了这个性质。具体方法就是分别取得受污染点和各污染源的污染物浓度变化谱图，然后通过比较这些谱图的主周期成分来判断主要污染源。

例如图 6-11（a）是被污染点的污染物浓度变化谱图。显然在 ω_0 处有一峰值，为主周期。图 6-11（b）～（f）为被污染点附近 5 个主要污染源的污染物浓度变化谱图。比较可见，只有污染源（c）在 ω_0 处有明显有极大峰值，因而可以判断被污染点的主要污染源为（c）。这种污染源鉴别方法在变化周期较短的噪声、

电磁污染和振动等领域中应用尤为广泛。

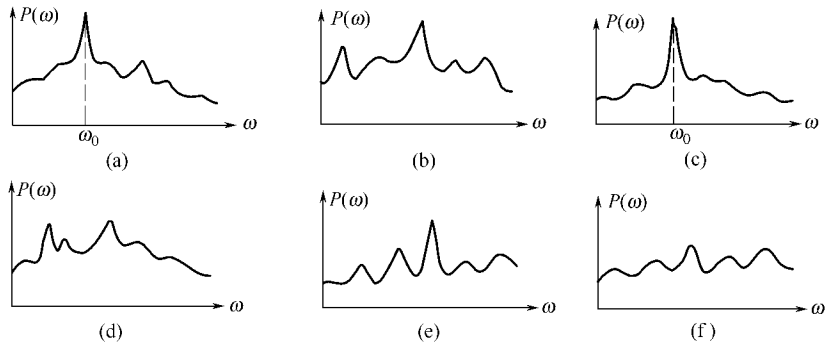


图 6-11　被污染点和其附近污染源的污染物浓度变化谱图

3. 环境污染过程分析

用谱分析来分析环境过程，可以获得更多的有用信息。例如图 6-12 是美国波多马克河口处测得的生物化学耗氧量（BOD）的频谱。由图可见，BOD 污染有两个主要期，分别为 12 小时和 24 小时。分析可知，两个主周期分别对应一天两次的潮汐和每天的下水流量。显然，如果上游某个企业废水排放具有不同的周期，那么下游的测点将在相应的频率上发现峰值，从而可以判断该企业是否产生了过量的污染。

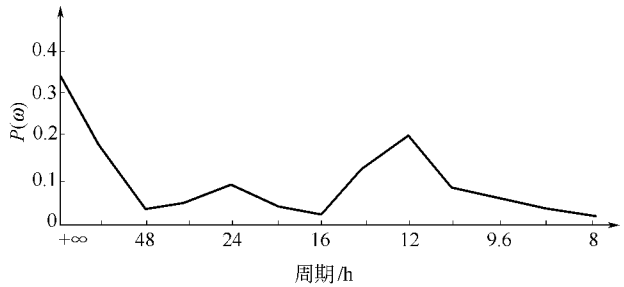


图 6-12　BOD 频谱

谱分析获得的主周期资料，可以据此对污染源采取相应的措施。这种措施也更有针对性、更有效。

谱分析中还可以得到一个重要的定理：采样定理。先看一下图 6-13 的例子。图（a）所示是一随机过程 $f(t)$，其频谱如图（b）所示，显然 $f(t)$ 是一连续型随机过程，存在一高频截止频率为 ω_0。在实际的环境问题中，大都使连续函数数量化，即按一定的时间间隔采样进行分析，使连续型随机过程变为时间序列。图 6-13（c）的 $X(t_i)$ 就是相应于 $f(t)$ 的时间序列，采样间隔为 T。

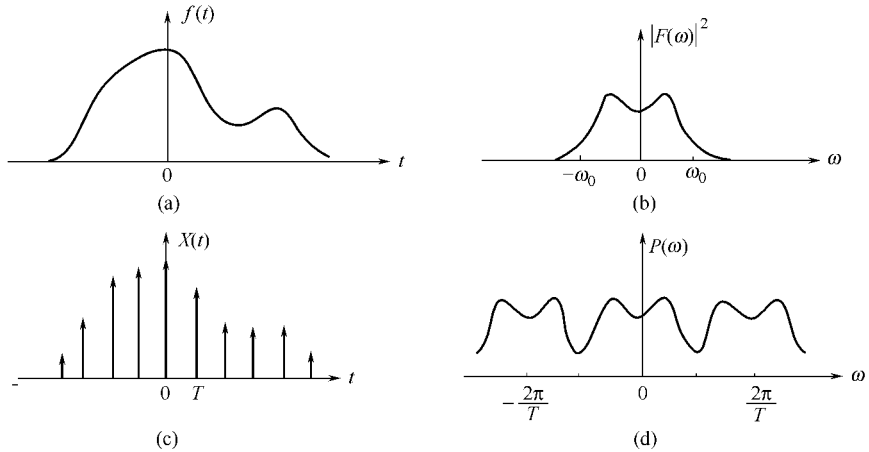


图 6-13

可以证明，图中的 $X(t_i)$ 可以表征为 $f(t)$ 与一个 $\delta(t-nT)$ 函数级数的乘积。即：

$$X(t_i)=f(t)\cdot\delta(t-nT)\quad(n=0,\pm1,\pm2,\cdots)\tag{6-52}$$

显然，$X(t_i)$ 的频谱即为 $f(t)$ 的频谱 $|F(\omega)|^2$ 和 $\delta(t-nT)$ 的频谱 $\delta(\omega-\omega_n)$ 的卷积。$\omega_n=2\pi n/T$（$n=0$，±1，±2，…）为采样频率。因而 $X(t_i)$ 的频谱 $P(\omega)$ 有图 6-13（d）的形式。为一连串相同的 $|F(\omega)|^2$ 的串接，频谱中心之间的间隔为 $2\pi/T$。

如果随机过程 $f(t)$ 的频谱存在一高频截止频率 ω_0，即 $f(t)$ 是一个有限带宽随机过程，大于 ω_0 频段的成分甚小，如图 6-13（b）所示。那么为了保证时间序列 $X(t_i)$ 的功率谱 $P(\omega)$ 的高频端 ω_0 频率上的成分不被重叠淹没，如图 6-13（d）所示，显然应选择这样一个采样周期 T，使得采样频率 $1/T>2f_0$，$f_0=\omega/2\pi$ 为高频截止频率。采样频率 $f=2f_0$，又称为奈奎斯特采样频率，它是保证采样后高频截止频率以内频段的信息能完整保留的最低采样频率。

显然，采样频率越高，所获得的信息量就越大、越完整。但实际的环境科研和监测工作中，采样频率高就意味着工作量大，因而在保证一定信息量的前提下，也不宜要求过高的采样频率。应该根据环境调查的具体情况，满足采样定理的要求，合理选择采样频率。

美国环保局规定，环境样本的采集中，采样频率应比随机过程的高频主频率高 3～4 倍。在实际的环境时间序列中，我们感兴趣的主频率可能不是高频截止频率，也就是说高于这个频率的频段上可能还有许多成分，不容忽略。因而适当增加采样频率可以确保足够的信息量。我们认为这一采样频率规定可以为我国的环境科技工作者借用。

【例 6-5】 试判断对图 6-8 的氮氧化物进行采样分析，应选择什么样的采样频率。

解 这种氮氧化物日均值浓度变化有一个以 9 天为周期的主频率，由图 6-10 可见，该主频率的峰值非常显著，因而可按该主频率选择采样周期。

按美国环保局的规定，显然应有：$f>(3\sim4)\ f_0$ 或 $T<T_0/(3\sim4)$（T_0 为主周期）。

即采样周期 T 应为 2～3 天（每 2～3 天采一次样）。

【例 6-6】 在图 6-12 的波多马克河口处 BOD 浓度测量的情况中，应选择的采样周期是多少？

解 显然该例中 BOD 污染有两个主周期，一个是 12 小时一次的潮汐变化，一个是 24 小时一次的下水变化，周期分别为 T_1、T_2。在选择采样周期时，按定理应考虑高频主频率的信息能否保证，即采样周期 T 为：$1/T\geqslant(3\sim4)\times 1/T_1$（$T_1=12$ 小时）

结果表明，BOD 污染的采样周期为 3～4 小时，即每天采样 6～8 次，即能获取潮汐变化的信息。

三、平稳时间序列的自回归模型及其应用

自回归模型是一种最常见的平稳时间序列的线性模型。主要用于揭示时间序列的内在联系、浓缩时间序列的信息、进行时间序列的描述和统计预报。

可以认为，平稳时间序列某一时刻 t 的状态可以由 t 时刻以前的所有时刻的状态来描述，因而可将时间序列的 t 时刻及 t 时刻过去时刻的状态建立一个线性回归方程。类似于前面介绍过的多元回归，但这种回归模型的变量为同一时间序列本身各不同时刻的取值，所以称为自回归方程。

实际模型建立过程中，通常取 t 时刻之前的有限个时刻，如 P 个时刻的状态来回归，可以建立如下的 P 阶自回归方程：

$$x(t)=\varphi_1 x(t-1)+\varphi_2 x(t-2)+\cdots+\varphi_p x(t-p)+a(t) \tag{6-53}$$

式中，φ_1，φ_2，…，φ_p 为 P 个常数。$a(t)$ 称为拟合残差，表示模拟值与实际值的偏差，一般是一个纯随机的白噪声序列。上式表明，t 时刻的 x 值可由 $(t-1)$，$(t-2)$，…，$(t-p)$ 个时刻的 x 值来描述。此处的 $X(t_i)$ 是经过中心化处理的时间序列，即均值为零，方差为 1。

现在的问题是如何选择回归方程的 P 个常数。显然，$\varphi_1,\varphi_2,\cdots,\varphi_p$ 的选择应使方程中的拟合偏差 $a(t)$ 成为一个白噪声序列。白噪声序列是一个数学期望为零、自相关函数为零、与任一随机过程完全不相关的随机过程，利用白噪声的这些特点和 $x(t)$ 为平稳时间序列这一事实，回归方程式（6-53）乘以 $x(t-1)$，然后再求数学期望，显然有：

$$r_1=\varphi_1+r_1\varphi_2+r_2\varphi_3+\cdots+r_{p-1}\varphi_p \tag{6-54}$$

式中，$r_1,r_2,\cdots,r_{p-1}$分别是$\tau=1,2,3,\cdots,p-1$时的平稳时间序列$x(t)$的自相关系数，类似的以$x(t-i)$乘回归方程式（$i=1,2,3,\cdots,p$），则可获得如下的P阶联立方程组：

$$\begin{cases}\varphi_1+r_1\varphi_2+r_2\varphi_3+\cdots+r_{p-1}\varphi_p=r_1\\ r_1\varphi_1+\varphi_2+r_1\varphi_3+\cdots+r_{p-2}\varphi_p=r_2\\ \vdots\\ r_{p-1}\varphi_1+r_{p-2}\varphi_2+\cdots+\varphi_p=r_p\end{cases} \tag{6-55}$$

这是一个P阶线性非齐次方程组，可表述为矩阵形式：

$$\boldsymbol{A}\boldsymbol{\psi}=\boldsymbol{R} \tag{6-56}$$

式中：

$$\boldsymbol{A}=\begin{bmatrix}1 & r_1 & r_2 & \cdots & r_{p-1}\\ r_1 & 1 & r_1 & \cdots & r_{p-2}\\ r_2 & r_1 & 1 & \cdots & r_{p-3}\\ \vdots & \vdots & \vdots & \vdots & \vdots\\ r_{p-1} & r_{p-2} & r_{p-3} & \cdots & 1\end{bmatrix} \tag{6-57}$$

为$P\times P$阶方阵，又称为系数矩阵。

$$\boldsymbol{\psi}=[\varphi_1\quad\varphi_2\quad\varphi_3\quad\varphi_4\quad\cdots\quad\varphi_p] \tag{6-58}$$

$$\boldsymbol{R}=[r_1\quad r_2\quad r_3\quad r_4\quad\cdots\quad r_p] \tag{6-59}$$

$\boldsymbol{\psi}$为自回归方程的系数阵，$\boldsymbol{R}$为相关系数阵，均为$P\times 1$阵，方程组有解：

$$\varphi_i=\frac{|\boldsymbol{D}_i|}{|\boldsymbol{A}|} \tag{6-60}$$

式中，$|\boldsymbol{A}|$是系数方阵的行列式值，$|\boldsymbol{D}_i|$是将系数方阵$\boldsymbol{A}$的第i行由R阵取代组成的新矩阵的行列式值，即：

$$\varphi_i=\begin{vmatrix}1 & r_1 & \cdots & r_1 & \cdots & r_{p-1}\\ r_1 & 1 & \cdots & r_2 & \cdots & r_{p-2}\\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots\\ r_{p-1} & r_{p-2} & \cdots & r_p & \cdots & 1\end{vmatrix}\Bigg/\begin{vmatrix}1 & r_1 & \cdots & r_{p-1}\\ r_1 & 1 & \cdots & r_{p-2}\\ \vdots & \vdots & \vdots & \vdots\\ r_{p-1} & r_{p-2} & \cdots & 1\end{vmatrix} \tag{6-61}$$

显然，通过时间序列的自相关系数就可求出P个φ值来，从而得到整个自回归方程的系数。

在实际应用中，自回归方程阶数P的选择是很重要的。当然，一般来说阶数越高，P值越大，对实际时间序列的拟合也就越好，但计算量也就大大增加。上一节介绍过，平稳随机过程的自相关系数$r(\tau)$一般随τ的增加逐渐变小，单调或波动衰减呈拖尾状趋于零。不同的时间序列自相关系数$r(\tau)$随τ增加而衰减的速率

是不一样的。因而可见，选择过大的 P 值也没有必要。根据实际的时间序列的自相关系数 $r(\tau)$ 的特征来选择适当的 P 值，就能保证一定的拟合精度。

前面说过，拟合残差 $a(t)$ 看作纯随机的白噪声序列。在自回归拟合中，应当把 $a(t)$ 作为随机修正量迭加入回归方程。$a(t)$ 的取值可有如下方法：在符合标准正态分布（$\mu=0$，$\sigma=1$）的随机变量中随机抽取一随机数 R，n 阶自回归方程的拟合残差 $a(t)$ 为：

$$a(t)=R\sqrt{1-(\varphi_1 r_1+\varphi_2 r_1+\cdots+\varphi_n r_n)} \tag{6-62}$$

下面举一个简单例子来说明自回归拟合的具体步骤和方法。

【例 6-7】 以［例 6-2］的城市 30 天降尘数据为例，对降尘过程作自回归拟合。

解　降尘过程的自相关系数 $r(\tau)$ 见表 6-7。

表 6-7　［例 6-8］中降尘过程的自相关系数

τ	1	2	3	4	5	6	7	8	9
$r(\tau)$	0.626	0.453	0.330	0.335	0.183	0.036	−0.116	−0.265	−0.421

由表 6-7 可见，降尘过程的时间序列的自相关系数以时差为两天之内最大，时间差大于两天时，记忆效应已经较弱。因而可以选取 $P=2$，即建立二阶自回归模型。

自回归方程为：$x(t)=\varphi_1 x(t-1)+\varphi_2 x(t-2)+a(t)$

自回归系数方程为：

$$\left.\begin{aligned}\varphi_1+r_1\varphi_2&=r_1\\ r_1\varphi_1+\varphi_2&=r_2\end{aligned}\right\}$$

本例中，由表 6-7 可知：

$$r_1=0.626$$
$$r_2=0.453$$

系数方程的解为：

$$\varphi_1=\begin{vmatrix}r_1 & r_1\\ r_2 & 1\end{vmatrix}\Big/\begin{vmatrix}1 & r_1\\ r_1 & 1\end{vmatrix}=\frac{r_1(1-r_2)}{1-r_1^2}=0.563$$

$$\varphi_2=\begin{vmatrix}1 & r_1\\ r_1 & r_2\end{vmatrix}\Big/\begin{vmatrix}1 & r_1\\ r_1 & 1\end{vmatrix}=\frac{r_2-r_1^2}{1-r_1^2}=0.1005$$

拟合残差为：

$$a(t)=R(t)\sqrt{1-r_1\varphi_1-r_2\varphi_2}=0.776R(t)$$

代入回归方程，即有：

$$x(t)=0.563x(t-1)+0.1005x(t-2)+0.776R(t)$$

注意，此处的 $x(t)$ 是经过中心化处理的时间序列，因而在对实际的降尘过程 $x'(t)$ 拟合时，应对方程做中心化的修正，即回归方程为：

$$\frac{x'(t)-\overline{X}}{\sigma^2(0)}=0.563\times\frac{x'(t-1)-\overline{X}}{\sigma^2(0)}+0.1005\times\frac{x'(t-2)-\overline{X}}{\sigma^2(0)}+0.776R(t)$$

式中，$\overline{X}=1.72$，为降尘过程的数学期望，$\sigma^2(0)=0.458$，为降尘过程的方差。修正后的降尘过程自回归方程为：

$$x'(t)=0.579+0.563x'(t-1)+0.1005x'(t-2)+0.355R(t)$$

式中，$R(t)$ 是满足标准正态分布的随机变量。上式就是降尘过程的二阶自回归模型。

$R(t)$ 的取值有一经验方法。首先在（0，1）中均匀分布的随机变量中随机抽取 12 个值（由计算机来实现，r_1，r_2，…，r_{12}），$R(t)$ 可由下式来计算：

$$R(t)=\sum_{i=1}^{12}r_i-6$$

即 12 个随机数相加再减 6。显然 $R(t)$ 的取值均匀分布在 0 的两侧。反复重复这一程序，就可以完成对每一个拟合状态的随机偏差修正。

根据已建立的二阶自回归模型和拟合残差修正的经验方法，可以获得模拟降尘过程时间序列（见表 6-8）。

显然表中只有 29 天的模拟数据。由二阶自回归模型的性质——当天的状态由前两天的状态决定，因而无法模拟第 1 天，第 2 天的降尘状态。第 31 天的状态是预测结果。

表 6-8 模拟降尘过程时间序列

t	3	4	5	6	7	8	9	10	11	12
$R(t)$	−0.096	−0.643	−0.587	−0.819	0.149	−1.361	0.589	0.504	−1.593	−0.307
$x'(t)$	2.2	1.9	1.4	1.3	2.1	1.9	2.0	1.8	0.7	0.9
t	13	14	15	16	17	18	19	20	21	22
$R(t)$	−0.689	0.444	−0.794	0.238	1.447	0.677	−1.189	−0.476	1.389	−0.097
$x'(t)$	0.7	1.5	0.8	1.3	1.8	1.7	1.2	1.3	2.4	1.4
t	23	24	25	26	27	28	29	30	31	
$R(t)$	0.521	−0.754	0.189	−0.202	−1.300	1.036	0.437	0.021	−0.025	
$x'(t)$	2.0	1.5	2.3	2.4	1.6	2.4	2.6	2.0	1.7	

将模拟计算结果与实测数据（见表 6-1）做比较，以检验二阶自回归模型的精度。图 6-14 是模拟数据与实测数据的比较，由图可见，模拟值与实际数据的趋势一致，但模拟值与实际数据的平均相对偏差为 0.341，显然偏差较大，有兴趣的读者可以对本例尝试建立三阶乃至更高阶的自回归模型。

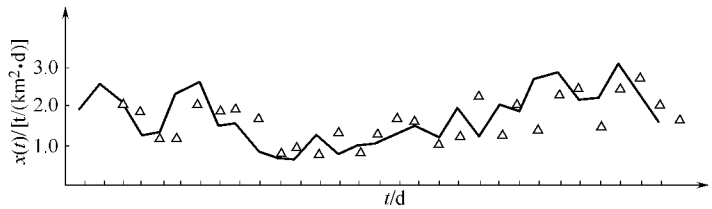


图 6-14　模拟与实测的比较

自回归模型的求解。回归方程的阶数越高就越复杂。因而在建模中，常常对完整的自回归方程进行因子筛选，选择一些相关较密切的时刻建立回归方程，这种方法又称为选点法。因子的筛选方法可以采用较严格的逐步回归方法，也可以根据自相关系数的大小粗略筛选，选出若干个相关系数最大的时刻组成自回归方程。例如图 6-15 是某个时间序列的自相关系数图。由图可见，γ_1，γ_6，γ_{12}三个自相关系数较大，若以这三个时间间隔建立自回归模型，即有：

$$x(t)=\varphi_1 x(t-1)+\varphi_6 x(t-6)+\varphi_{12} x(t-12) \tag{6-63}$$

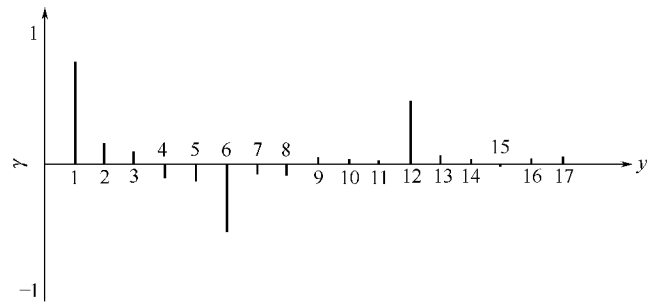


图 6-15　自相关系数变化图

系数的线性方程组为：

$$\left.\begin{aligned}\varphi_1+r_5\varphi_6+r_{11}\varphi_{12}&=r_1\\ r_5\varphi_1+\varphi_6+r_6\varphi_{12}&=r_6\\ r_{11}\varphi_1+r_6\varphi_6+\varphi_{12}&=r_{12}\end{aligned}\right\} \tag{6-64}$$

代入具体的 r_1，r_5，r_6，r_{11}，r_{12}值即可求出系数 φ_1，φ_6，φ_{12}，进而完成三阶自回归模型的建立。建模前进行因子筛选的选点法能够突出重点，减少计算量，实际应用很方便。

在使用自回归模型预报环境问题时，常常希望能预测未来多个时刻的状态。因而自回归方程应为：

$$x(t+k)=\varphi_1 x(t)+\varphi_2 x(t-1)+\cdots+\varphi_p x(t-p+1)$$

该式表明，未来 K 时刻的时间序列状态由过去 P 个时刻的状态描述。此处，预报依据时段为 P 个时刻，预报时效为 K 个时刻。这样的自回归模型对预报 $t+$

K 时刻的状态来说，必须跳过 $K-1$ 个不预报的时刻。自回归方程的系数线性方程组为：

$$\left.\begin{aligned}
&\varphi_1+r_1\varphi_2+r_2\varphi_3+\cdots+r_{p-1}\varphi_p=r_k\\
&r_1\varphi_1+\varphi_2+r_1\varphi_3+\cdots+r_{p-2}\varphi_p=r_{k+1}\\
&\qquad\vdots\\
&r_{p-1}\varphi_1+r_{p-2}\varphi_2+r_{p-3}\varphi_3+\cdots+\varphi_p=r_{k+p-1}
\end{aligned}\right\} \tag{6-65}$$

显然求解系数方程组即可完成 K 个时间步长的自回归模型的建立。

【例 6-8】 根据［例 6-2］的数据，对降尘过程做两个时间步长的二阶自回归拟合。

解 显然自回归方程的形式为：

$$x(t+1)=\varphi_1 x(t-1)+\varphi_2 x(t-2)$$

系数方程组为：

$$\varphi_1+r_1\varphi_2=r_2$$
$$r_1\varphi_1+\varphi_2=r_3$$

本例中，由表 6-2 可得：

$$r_1=0.626;r_2=0.453;r_3=0.330$$

代入系数方程组解得：

$$\varphi_1=\frac{r_2-r_1r_3}{1-r_1^2}=0.405$$

$$\varphi_2=\frac{r_3-r_1r_2}{1-r_1^2}=0.0763$$

因而自回归方程为：

$$x(t+1)=0.405x(t-1)+0.0763x(t-2)+a(t)$$

经过均值修正的两个时间步长的自回归方程为：

$$x'(t+1)=0.892+0.405x'(t-1)+0.0763x'(t-2)+0.386R(t)$$

读者不难验证，本例的两个时间步长二阶自回归模型的预报精度很差。本例除了要解释多步长自回归模型的建立方法外，也想表明自回归模型的拟合预报有其局限性。

自回归模型中，我们称 $a(t)$ 为拟合残差，认为它是一个白噪声序列。显然，$a(t)$ 越小，拟合结果就越好。因而在选择回归方程阶数和系数时，应使拟合残差 $a(t)$ 满足白噪声条件并尽可能的小。但在实际中，往往不能满足上述要求，也就是说无法使 $a(t)$ 成为真正的白噪声序列，甚至在 P 值取很大时也无法满足要求，导致了自回归模型的拟合偏差。

时间序列的线性模型还有滑动平均模型和自回归滑动平均模型，都能拟合和预报时间序列。由于建模比较复杂，本书不做介绍，有兴趣的读者可参阅有关

书籍。

四、马尔科夫转移矩阵模型

上一节我们介绍过马尔科夫过程及马尔科夫链，并介绍了马尔科夫转移矩阵的基本概念。这里要讨论马尔科夫转移矩阵在环境数据的分析拟合和预报中的应用。

自回归模型中，一阶自回归方程为：

$$x(t)=\varphi x(t-1)+a(t) \tag{6-66}$$

表明时间序列的全部历史状态对未来状态的影响都集中在最后的状态中，这显然就是马尔科夫链的情况。因而马尔科夫链只是自回归模型的一种特殊情况。根据已建立的马尔科夫转移矩阵来拟合预报时间序列的方法如下。

设马尔科夫链具有 n 个状态空间，一步长转移矩阵为：

$$[p_{ij}]=\begin{bmatrix} P_{11} & P_{12} & \cdots & P_{1n} \\ P_{21} & P_{22} & \cdots & P_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ P_{n1} & P_{n2} & \cdots & P_{mn} \end{bmatrix} \tag{6-67}$$

时间序列的初始状态概率分布为 $Q=(q_1, q_2, \cdots, q_n)$。令：

$$P_i^{(l)}=\sum_{j=1}^{l} p_{ij} \qquad q^{(l)}=\sum_{i=1}^{l} q_i$$

$P_i^{(l)}$ 是马尔科夫转移矩阵中第 i 行前 l 个元素之和。由于 $P_i^{(n)}=\sum_{j=1}^{n} p_{ij}=1$，因而 $P_i^{(l)}$ 的取值范围为：$0\leqslant P_i^{(l)}\leqslant 1$

$P_i^{(l)}$ 可以构成一个新的矩阵：

$$[P_i^l]=\begin{bmatrix} p_{11} & \sum_{j=1}^{2} p_{1j} & \cdots & \sum_{j=1}^{n} p_{1j} \\ p_{21} & \sum_{j=1}^{2} p_{2j} & \cdots & \sum_{j=1}^{n} p_{2j} \\ \vdots & \vdots & \vdots & \vdots \\ p_{n1} & \sum_{j=1}^{2} p_{nj} & \cdots & \sum_{j=1}^{n} p_{nj} \end{bmatrix}=\begin{bmatrix} p_1^{(1)} & p_1^{(2)} & \cdots & p_1^{(n)} \\ p_2^{(1)} & p_2^{(2)} & \cdots & p_2^{(n)} \\ \vdots & \vdots & \vdots & \vdots \\ p_n^{(1)} & p_n^{(2)} & \cdots & p_n^{(n)} \end{bmatrix} \tag{6-68}$$

马尔科夫链的模拟可以通过这个新产生的矩阵完成，具体步骤如下。

(1) 取在 (0,1) 数值区间上均匀分布的随机数 R_0，根据 R_0 来决定马尔科夫链的初始状态，如果：

$$q^{(l_0-1)}<R_0\leqslant q^{(l_0)}$$

那么就以第 l_0 个状态为时间序列的初始状态。

(2) 取在 (0,1) 上均匀分布的随机数 R_1，将第 l_0 行的元素和 R_1 进行比较，从而确定第一次转移后的状态，如果：

$$q_{l_0}^{(l_1-1)}<R_1\leqslant q_{l_0}^{(l_1)}$$

就认为时间序列从状态 l_0 转移到状态 l_1。

(3) 将 l_1 作为新的初始状态，重复步骤 (2)，可以使马尔科夫链从状态 l_1 转移到状态 l_2，如此反复进行下去，就可以得到一个新的状态序列：l_0，l_1，l_2，…，l_j，…，即得到所模拟的时间序列。

下面举一个例子说明马尔科夫转移矩阵模拟时间序列的具体做法。

【例 6-9】 利用［例 6-3］建立的马尔科夫转移矩阵，模拟某城市的逐日降尘的时间序列。

解 降尘通程的一步长马尔科夫转移矩阵为：

$$[p_{ij}]=\begin{bmatrix} \frac{3}{5} & \frac{2}{5} & 0 & 0 & 0 \\ \frac{1}{7} & \frac{2}{7} & \frac{2}{7} & \frac{2}{7} & 0 \\ \frac{1}{7} & \frac{1}{7} & \frac{3}{7} & 0 & \frac{2}{7} \\ 0 & \frac{2}{5} & 0 & \frac{1}{5} & \frac{2}{5} \\ 0 & 0 & \frac{2}{5} & \frac{2}{5} & \frac{1}{5} \end{bmatrix}$$

本例中时间序列的初始状态可随机选取。假设取得 $R_0=0.453$，根据［例 6-3］，本例中的 5 个状态分别对应的日降尘量范围为：(0.5～1.0)，(1.0～1.5)，(1.5～2.0)，(2.0～2.5)，(2.5～3.0)t/(km^2 · d)。

选取状态 3 为初始状态。一步长马尔科夫转移矩阵 $[p_{ij}]$ 组合成新的矩阵 $\boldsymbol{P}_i^{(l)}$，如下：

$$\boldsymbol{P}_i^{(l)}=\begin{bmatrix} 0.6 & 1 & 1 & 1 & 1 \\ 0.143 & 0.428 & 0.714 & 1 & 1 \\ 0.143 & 0.286 & 0.714 & 0.714 & 1 \\ 0 & 0.4 & 0.4 & 0.6 & 1 \\ 0 & 0 & 0.4 & 0.8 & 1 \end{bmatrix}$$

根据步骤 (2)，按顺序任意再抽取 29 个 (0，1) 区间内均匀分布的随机数，逐一比较确定转移到的下一个状态，即可得到表 6-9 的 30 天降尘的状态变化序列。

表 6-9　状态变化序列

随机数	0.453	0.838	0.528	0.974	0.975	0.049	0.499	0.654	0.263	0.149
状态	3	5	4	5	5	3	3	3	2	2
随机数	0.709	0.769	0.123	0.265	0.122	0.041	0.936	0.166	0.939	0.193
状态	3	5	3	2	1	1	2	2	4	2
随机数	0.295	0.321	0.492	0.772	0.663	0.839	0.193	0.073	0.297	0.348
状态	2	2	3	5	4	5	3	1	1	1

逐日状态变化序列所对应的具体降尘量应为表 6-9 确定的状态的下限加上状态范围与随机数 R 的乘积。

本例中，状态范围为 0.5t/(km^2 · d)，令状态下限为 $n_{i\min}$，因而逐日的降尘量 x_i 为：

$$x_i = n_{i\min} + 0.5 \times R_i$$

对 30 天的状态逐一计算，即可得表 6-10 的数据。

表 6-10　逐日降尘时间序列

日	1	2	3	4	5	6	7	8	9	10
降尘量	1.7	2.9	2.3	3.0	3.0	1.5	1.7	1.8	1.1	1.1
日	11	12	13	14	15	16	17	18	19	20
降尘量	1.9	2.9	1.6	1.1	0.6	0.5	1.5	1.1	2.4	1.1
日	21	22	23	24	25	26	27	28	29	30
降尘量	1.1	1.2	1.7	2.9	2.3	2.9	1.6	0.5	0.6	0.6

这就是利用马尔科夫转移矩阵模拟到的逐日降尘量的时间序列。

比较马尔科夫转移矩阵模拟的降尘过程和实测降尘过程的数学期望和方差有：

$$\overline{X_{实}} = 1.72\text{t/(km}^2 \cdot \text{d)} \quad \overline{X_{模}} = 1.67\text{t/(km}^2 \cdot \text{d)}$$

$$R_{(0)实} = 0.458 \qquad R_{(0)模} = 0.669$$

可见偏差还是很大的，正如前面指出的，时间序列越长、样本数越多，马尔科夫转移矩阵的模拟就越好。本例的样本数量显然是不够多的，只是为了较简明的说明马尔科夫转移矩阵的建立及其对时间序列的模拟和预报步骤和方法。在历史资料较丰富、样本数较多的情况下，使用马尔科夫转移矩阵进行模拟和预报可以收到令人满意的结果。

作者曾使用马尔科夫转移矩阵法对太阳辐射过程进行模拟，实验数据是 1982 年、1983 年两年在约旦的 Amman 测量的以小时为单位的太阳总辐射。

在实际模拟计算中，为方便起见，将水平面上的太阳总辐射数据变换成天气

晴朗指数 K 值进行模拟计算。天气晴朗指数 K 值的定义为：

$$K=\frac{H_h}{H_0}$$

式中，H_h 为地面水平面上太阳总辐射，H_0 为大气层外太阳总辐射。显然 H_0 是不受天气情况影响的量，可以通过计算得到。K 值的取值范围为：

$$0<K<1$$

将 K 值以 0.1 一档分档，共分为 10 个状态，对两年的实测数据分档进行统计，可建立如表 6-11 所示的马尔科夫转移矩阵。

表 6-11 马尔科夫转移矩阵

K 值 t 时刻	$t+1$ 时刻表 K 值				
	0～0.1	0.1～0.2	0.2～0.3	0.3～0.4	0.4～0.5
0～0.1	0.547	0.205	0.031	0.000	0.000
0.1～0.2	0.183	0.321	0.002	0.119	0.046
0.2～0.3	0.045	0.207	0.234	0.162	0.108
0.3～0.4	0.006	0.102	0.130	0.203	0.136
0.4～0.5	0.005	0.051	0.071	0.235	0.148
0.5～0.6	0.005	0.008	0.036	0.090	0.204
0.6～0.7	0.000	0.003	0.005	0.020	0.046
0.7～0.8	0.000	0.002	0.003	0.006	0.008
0.8～0.9	0.000	0.000	0.101	0.023	0.007
0.9～1.0	0.000	0.000	0.000	0.000	0.015
K 值 t 时刻	$t+1$ 时刻表 K 值				
	0.5～0.6	0.6～0.7	0.7～0.8	0.8～0.9	0.9～1.0
0～0.1	0.031	0.016	0.031	0.031	0.063
0.1～0.2	0.018	0.009	0.037	0.009	0.055
0.2～0.3	0.090	0.063	0.018	0.018	0.054
0.3～0.4	0.186	0.096	0.034	0.056	0.051
0.4～0.5	0.163	0.117	0.102	0.082	0.026
0.5～0.6	0.268	0.201	0.095	0.070	0.023
0.6～0.7	0.227	0.467	0.162	0.065	0.005
0.7～0.8	0.022	0.043	0.677	0.037	0.004
0.8～0.9	0.010	0.023	0.480	0.408	0.039
0.9～1.0	0.061	0.061	0.288	0.470	0.106

按前面介绍的马尔科夫转移矩阵模拟时间序列的方法，使用计算机产生的一系列（0，1）区间内均匀分布的随机数 R，将马尔科夫转移矩阵相应行的转移概率进行迭加比较，由此产生一系列状态转移序列，每个状态的取值下限加上 $0.1\times R$，即计算得相应的模拟 K 值，由此而得到太阳辐射 K 值的模拟时间序列。图 6-16 是模拟所得 K 值时间序列的概率密度函数（pdf）与实验数据的概率密度函数结果比较。

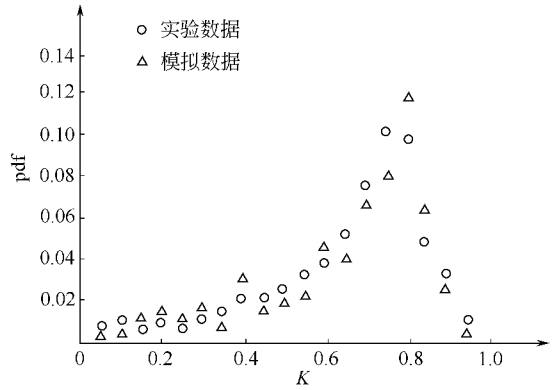


图 6-16　太阳辐射实验和模拟概率密度函数比较

由图可见，模拟数据概率密度函数与实验数据相比很接近，模拟结果还是较为满意的，说明马尔科夫转移矩阵是一种模拟时间序列可行的方法。

第七章　数据质量管理中的统计方法

在环境数据的产生过程中，特别是实验室分析测试工作中，通常用质量控制图的方法来控制与检验数据产生的过程中是否处于受控状态。过程处于统计控制中的标志是：来自过程的重复样本表现为来自一个稳定的概率分布的随机样本。控制图能依据一定的概率标准判定数据的质量指标是在偶然变异的范围内，还是出现了异常情况。一旦发现数据变异超出了偶然变异的范围，应及时研究其发生的原因，并采取措施消除它。数据产生过程中变异的发生有两类，一类是由“可指出的原因”引起的，如仪器的基准误差、试剂不纯、操作不当等，我们称数据的此种变异为系统误差。如果能够判断数据变异是由“可指出的原因”引起的，即系统误差，我们可以认为数据产生的过程处于脱控状态。我们可按照它的作用规律对它进行校正或设法消除，使数据产生的过程回到可以接受的变异水平。数据产生变异的另一类原因是由偶然变动的“一般原因”引起的，变异的绝对值和符号的变化时大时小、时正时负，以不可预定的方式变化着，我们称数据的此种变异为随机误差。随机误差是具有统计规律的误差，当数据测定的次数足够多时，数据变异的概率服从统计分布的规律。引起随机误差的因素是人们不可控制的，也不能通过调整数据产生的过程来修正。如果我们能够判断数据产生的变异仅仅是由“一般原因”引起的，即随机误差，我们可以认为数据产生的过程处于受控状态。在这种情况下，没有必要改变或调整数据产生的过程。

第一节　控制图的基本概念

控制图是判断数据变异是由“一般原因”（处于受控状态）引起的还是“可指出的原因”（处于脱控状态）引起的一种工具。在数据产生的过程中，一旦发现一个数据处于脱控状态，都应当对过程进行调整，使数据生产过程回到受控状态。如果我们研究的环境对象，用一个变量来描述，如河流水质中的某一种污染物浓度，则使用均值控制图，亦称 $\bar{x}$ 控制图。为了了解控制图的基本概念，我们先介绍 $\bar{x}$ 控制图的一些特征。

图 7-1 为 $\bar{x}$ 控制图的一般结构。控制图的中心线为处于控制状态过程的均值，垂直线表示要研究变量的测量尺度。从数据产生的过程中抽取一个样本 x_i，

计算出样本的均值 $\bar{x}$，将表示 $\bar{x}$ 值的数据点标在控制图上。图中 UCL 和 LCL 分别称为控制上限和控制下限。它对于决定过程是处于受控状态还是脱控状态是十分重要的。x_i 的值位于两个控制限之间时，表示数据产生的过程处于受控状态的概率很大。x_i 的值位于两个控制限之外，表示数据产生的过程处于脱控状态的概率很大。应当对过程进行调整。随着时间的推移，越来越多的数据点被标在控制图上。数据点的顺序是从左向右，同抽取样本的顺序相同。

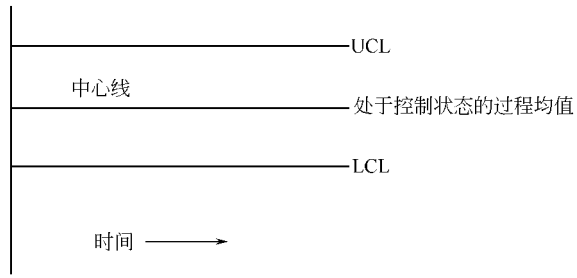


图 7-1 $\bar{x}$ 控制图的结构

除 $\bar{x}$ 控制图外，通常使用的其他控制图还有：检验样本中测量值全距的极差控制图（R 控制图），检验样本中有缺陷比率的 P 控制图和样本中有缺陷数量的 np 控制图。R 控制图、P 控制图和 np 控制图的基本结构和图 7-1 中 $\bar{x}$ 控制图的格式相似，其主要区别是使用的度量尺度不同。例如，在 P 控制图中，测量尺度是样本中有缺陷项目的比率，而不是样本的均值。本章中，我们主要介绍 $\bar{x}$ 控制图、R 控制图和 $\bar{x}$-R 控制图。

第二节 $\bar{x}$ 控制图

从正态总体 $N(\mu,\sigma^2)$ 中随机抽取容量为 n 的样本 x_i，样本均值 $\bar{x}$ 服从正态分布 $N\left(\mu,\frac{\sigma^2}{n}\right)$。由正态分布的知识知道，$\bar{x}$ 出现在区间（$\mu-2\sigma/\sqrt{n}$，$\mu+2\sigma/\sqrt{n}$）内的概率为 0.955，在此区间外的概率为 0.045；$\bar{x}$ 出现在区间（$\mu-3\sigma/\sqrt{n}$，$\mu+3\sigma/\sqrt{n}$）内的概率为 0.997，在此区间外的概率为 0.003。由于 0.003 是极小的概率，我们由此可判定，当 x_i 位于 $\bar{x}\pm3\sigma_{\bar{x}}$ 之间，可认为它处于受控状态。当 x_i 位于 $\bar{x}\pm3\sigma_{\bar{x}}$ 之外，可认为它处于脱控状态，即意味着可能出现了不正常情况。$\bar{x}$ 控制图的控制限可表示如下：

$$\mathrm{UCL}=\mu+3\sigma_{\bar{x}} \tag{7-1}$$

$$\mathrm{LCL}=\mu-3\sigma_{\bar{x}} \tag{7-2}$$

式中，$\sigma_{\bar{x}}=\dfrac{\sigma}{\sqrt{n}}$

在实际工作中，由于总体的均值和标准差一般是未知的，通常用样本的均值和标准差代替总体的均值和标准差来作控制图。则控制图的控制限可表示如下：

$$UCL=\bar{x}+3S_{\bar{x}} \tag{7-3}$$

$$LCL=\bar{x}-3S_{\bar{x}} \tag{7-4}$$

下面用［例 7-1］说明 $\bar{x}$ 控制图的实际应用。

【例 7-1】 用某种标准方法对含铜 0.250mg/L 的水质标准物做 20 次测定。测定结果如表 7-1 所示。

表 7-1 ［例 7-1］测定结果/(mg/L)

1	0.251	6	0.240	11	0.229	16	0.270
2	0.250	7	0.260	12	0.250	17	0.225
3	0.250	8	0.290	13	0.283	18	0.250
4	0.263	9	0.262	14	0.300	19	0.256
5	0.235	10	0.234	15	0.262	20	0.250

由表 7-1 中数据计算可得平均值 $\bar{x}=0.256$mg/L，标准差 $S_{\bar{x}}=0.020$mg/L。按图 7-1 $\bar{x}$ 控制图结构可得到如图 7-2 所示的控制图。

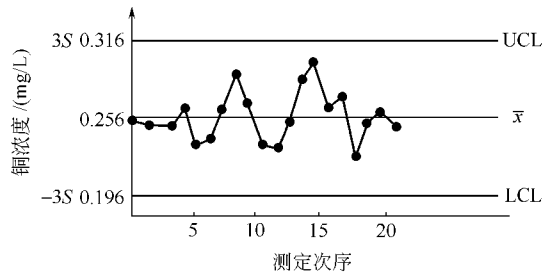


图 7-2　水中铜分析数据控制图

从控制图中可看出，测定结果都分布在上、下控制限之间，说明测定过程处于受控状态。

在实际工作中，我们还可以用极差 R 代替标准差来检验数据产生过程中的变异。其原因是因为极差容易计算。用极差 R 构造 $\bar{x}$ 控制图的控制上、下限，计算量较少。下面举例❶说明用极差 R 代替标准差构造 $\bar{x}$ 控制图的情况。

【例 7-2】 某公司为控制产品质量，每天从产品中抽取容量为 5 的随机样本，一共抽取了 20 个样本。得到样本的测定结果如表 7-2 所示。

❶ ［例 7-2］参考蒋子刚、顾雪梅编著的《分析测试中的数理统计与质量保证》。

表 7-2 ［例 7-2］测定结果

样本序号	观 测 值					样本均值 $\bar{x}_j$	样本极差 R_j
1	3.5065	3.5086	3.5144	3.5009	3.5030	3.5065	0.0135
2	3.4882	3.5085	3.4884	3.5250	3.5031	3.5026	0.0368
3	3.4897	3.4898	3.4995	3.5130	3.4969	3.4978	0.0233
4	3.5153	3.5120	3.4989	3.4900	3.4837	3.5000	0.0316
5	3.5059	3.5113	3.5011	3.4773	3.4801	3.4951	0.0340
6	3.4977	3.4961	3.5050	3.5014	3.5060	3.5012	0.0099
7	3.4910	3.4913	3.4976	3.4831	3.5044	3.4935	0.0213
8	3.4991	3.4853	3.4830	3.5083	3.5094	3.4970	0.0264
9	3.5099	3.5162	3.5228	3.4958	3.5004	3.5090	0.0270
10	3.4880	3.5015	3.5094	3.5102	3.5146	3.5047	0.0266
11	3.4881	3.4887	3.5141	3.5175	3.4863	3.4989	0.0312
12	3.5043	3.4867	3.4946	3.5018	3.4784	3.4932	0.0259
13	3.5043	3.4769	3.4944	3.5014	3.4904	3.4935	0.0274
14	3.5004	3.5030	3.5082	3.5045	3.5234	3.5079	0.0230
15	3.4846	3.4938	3.5065	3.5089	3.5011	3.4990	0.0243
16	3.5145	3.4832	3.5188	3.4935	3.4989	3.5018	0.0356
17	3.5004	3.5042	3.4954	3.5020	3.4889	3.4982	0.0153
18	3.4959	3.4823	3.4964	3.5082	3.4871	3.4940	0.0259
19	3.4878	3.4864	3.4960	3.5070	3.4984	3.4951	0.0206
20	3.4969	3.5144	3.5053	3.4985	3.4885	3.5007	0.0259

对每个容量都为 n 的 k 个样本，我们有样本的总体均值：

$$\bar{\bar{x}}=\frac{\bar{x}_1+\bar{x}_2+\cdots+\bar{x}_k}{k}=3.4995 \qquad (k=1,2,\cdots,20)$$

以及样本平均极差 $\overline{R}=\dfrac{R_1+R_2+\cdots+R_k}{k}=0.0253$

由式（7-1）和式（7-2），我们知道 $\bar{x}$ 控制图的控制上、下限为：

$$\mathrm{UCL}=\mu+3\frac{\sigma}{\sqrt{n}}$$

$$\mathrm{LCL}=\mu-3\frac{\sigma}{\sqrt{n}}$$

因此，为了构造 $\bar{x}$ 控制图的控制限，我们用极差来估计过程的标准差。现已证明，过程的标准差和极差之间有如下关系：

$$\hat{\sigma}=\frac{\overline{R}}{d_2}$$

上式中，d_2 是一个仅仅依赖于样本容量 n 的常数，由附表 10 可查阅。当样

本容量为 5 时，$d_2=2.326$，将 $\hat{\sigma}=\dfrac{\overline{R}}{d_2}$ 代入上、下控制限 UCL 和 LCL 计算式，我们得到：

$$\mathrm{UCL}=\overline{\overline{x}}+3\times\frac{\overline{R}}{d_2\sqrt{n}}=\overline{\overline{x}}+A_2\overline{R}$$

$$\mathrm{LCL}=\overline{\overline{x}}-3\times\frac{\overline{R}}{d_2\sqrt{n}}=\overline{\overline{x}}-A_2\overline{R}$$

式中，A_2 为一个仅依赖样本容量的常数，由附表 10 可查阅。当 $n=5$ 时，$A_2=0.577$，则我们有 $\bar{x}$ 控制图的上、下控制限为：

$$\begin{aligned}\mathrm{UCL}&=3.4995+0.577\times0.0253\\&=3.514\\\mathrm{LCL}&=3.4995-0.577\times0.0253\\&=3.485\end{aligned}$$

则我们制作得到该产品的 $\bar{x}$ 控制图如图 7-3 所示。

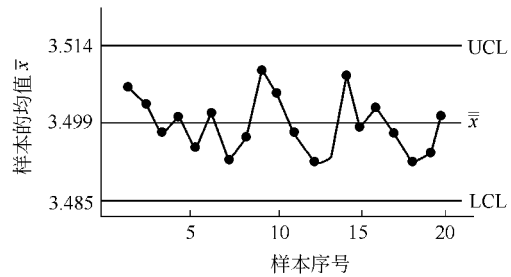


图 7-3　$\bar{x}$ 控制图

从［例 7-2］的控制图中，我们可以看到，20 个样本的均值都在控制限内，因此表明产品的生产过程处于受控状态。

第三节　*R* 控制图

本节我们讨论 R 控制图。虽然人们在质量控制过程中，经常使用 $\bar{x}$ 控制图，然而 R 控制图也常常被采用。其原因是，用样本极差 R 代替样本标准差 $S_{\bar{x}}$ 进行统计分析，可信程度差别不大。此外，极差 R 的计算简便也是一个优点。极差控制图可控制一个过程的变异。

为了构造 R 控制图，我们考虑用样本极差的平均值 $\overline{R}$ 来估计总体标准差。可以证明极差标准差的估计值 $\hat{\sigma}_R$ 为：

$$\hat{\sigma}_R = d_3 \frac{\overline{R}}{d_2} \tag{7-5}$$

式（7-5）中，d_2 和 d_3 是依赖于样本容量的常数，在附表 10 中可以查得。由此我们得到 R 控制图的上、下限计算式：

$$\mathrm{UCL} = \overline{R} + 3\hat{\sigma}_R = \overline{R} + 3d_3 \frac{\overline{R}}{d_2} \tag{7-6}$$

$$\mathrm{LCL} = \overline{R} - 3\hat{\sigma}_R = \overline{R} - 3d_3 \frac{\overline{R}}{d_2} \tag{7-7}$$

如果令：

$$D_4 = 1 + 3 \times \frac{d_3}{d_2} \tag{7-8}$$

$$D_3 = 1 - 3 \times \frac{d_3}{d_2} \tag{7-9}$$

将式（7-8）和式（7-9）代入式（7-6）和式（7-7），得到 R 控制图上、下限的表达式为：

$$\mathrm{UCL} = \overline{R} D_4 \tag{7-10}$$

$$\mathrm{LCL} = \overline{R} D_3 \tag{7-11}$$

式（7-10）和式（7-11）中的 D_4 和 D_3 也是依赖于样本容量的常数，亦可在附表 10 中查得。

【例 7-3】 用［例 7-2］中的数据制作 R 控制图。

查附表 10，当 $n=5$ 时，$D_3=0$，$D_4=2.114$

由式（7-10）和式（7-11），我们有：

$$\mathrm{UCL} = \overline{R} D_4 = 0.0253 \times 2.114 = 0.0534$$

$$\mathrm{LCL} = \overline{R} D_3 = 0.0253 \times 0 = 0$$

有了 R 控制图的 UCL 值和 LCL 值，可制作 R 控制图如图 7-4 所示。

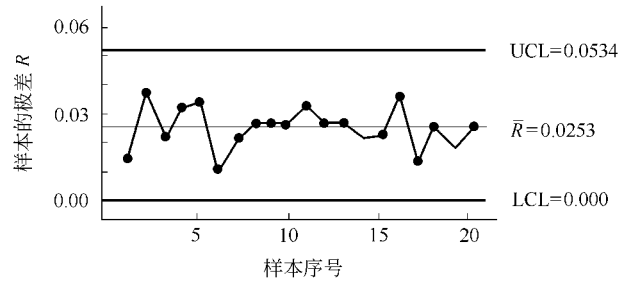


图 7-4 R 控制图

由于所有样本的极差都在控制限之内，因此可以认为过程处于受控状态。

第四节　$\bar{x}$-R 控制图

在实际使用中，有时还将 $\bar{x}$ 控制图和 R 控制图联合使用，即 $\bar{x}$-R 控制图。$\bar{x}$-R 控制图是最常用的过程统计管理图。它主要用于抽样频率较高并且过程比较稳定的过程。

下面举例[1]说明 $\bar{x}$-R 控制图的构造过程。

【例 7-4】 实验室分析某产品纯度得到数据如表 7-3 所示。

表 7-3　［例 7-4］产品纯度分析数据

时间序列	1	2	3	4	5	6	7	8	9	10
测定结果 1	99.7	99.7	99.4	99.5	99.7	99.6	99.3	99.6	99.5	99.7
测定结果 2	99.6	99.4	99.3	99.8	99.4	99.9	99.2	99.3	99.6	99.6
$\bar{x}$	99.65	99.55	99.35	99.65	99.55	99.75	99.25	99.45	99.55	99.65
R	0.1	0.3	0.1	0.3	0.3	0.3	0.1	0.3	0.1	0.1
时间序列	11	12	13	14	15	16	17	18	19	20
测定结果 1	99.6	99.5	99.7	99.7	99.4	99.6	99.2	99.7	99.4	99.5
测定结果 2	99.8	99.4	99.8	99.3	99.5	99.7	99.5	99.6	99.3	99.6
$\bar{x}$	99.7	99.45	99.75	99.5	99.45	99.65	99.35	99.65	99.35	99.55
R	0.2	0.1	0.1	0.4	0.1	0.1	0.3	0.1	0.1	0.1

其样本容量 $n=2$；样本数 $N=20$

由表 7-3 数据计算样本的总均值 $\bar{\bar{x}}$ 和极差均值 $\bar{R}$：

$$\bar{\bar{x}}=\frac{\sum \bar{x}_i}{N}=99.54$$

$$\bar{R}=\frac{\sum R_i}{N}=0.18$$

计算 $\bar{x}$ 控制图的上、下控制限：

$$\mathrm{UCL}_{\bar{x}}=\bar{\bar{x}}+A_2\bar{R}$$

$$\mathrm{LCL}_{\bar{x}}=\bar{\bar{x}}-A_2\bar{R}$$

查附表 10 可得，$A_2=1.880$，则有：

$$\mathrm{UCL}_{\bar{x}}=99.54+1.880\times0.18=99.88$$

$$\mathrm{LCL}_{\bar{x}}=99.54-1.880\times0.18=99.20$$

[1] 参考何祯等译《质量工具箱》。

计算 R 控制图的上、下控制限：

$$\mathrm{UCL}_R=\overline{R}D_4$$

$$\mathrm{LCL}_R=\overline{R}D_3$$

查附表 10 可得，$D_4=3.267$，$D_3=0$，则有：

$$\mathrm{UCL}_R=0.18\times3.267=0.59$$

$$\mathrm{LCL}_R=0.18\times0=0$$

由 $\mathrm{UCL}_{\bar{x}}$、$\mathrm{LCL}_{\bar{x}}$、UCL_R 和 LCL_R，我们可制得过程控制的 $\bar{x}$-R 控制图如图 7-5 所示。

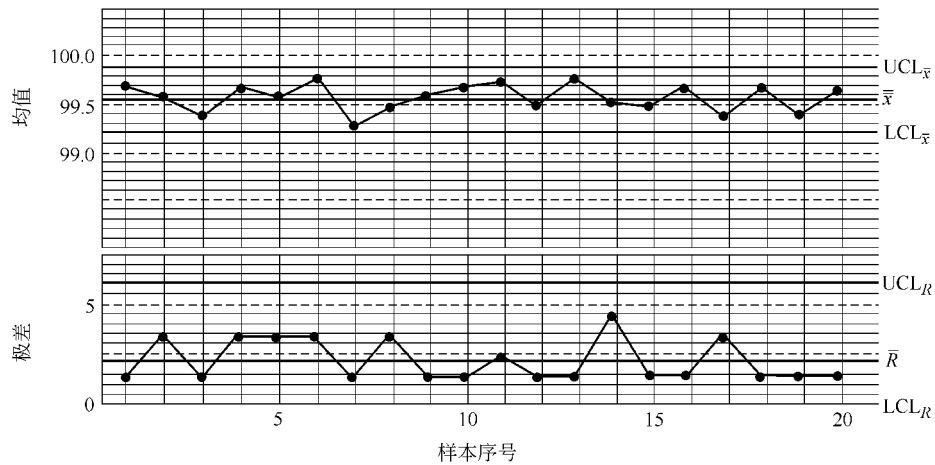


图 7-5 $\bar{x}$-R 控制图

由于所有数据点都落在 $\bar{x}$-R 控制图的控制限内，我们可以判断过程处于受控状态。

第五节 控制图的解释

根据控制图中数据点的位置和轨迹，我们可以用很小的错判概率来确定一个过程是否处于统计受控状态。控制图中，数据点的分布状况大部分是正常的，即符合正态分布规律，但也有不正常的分布状况。对一个统计过程控制来说，过程处于脱控状态的基本现象是数据点落在控制限之外。在这种情况下，应尽可能地分析查找原因，并及时进行纠正。

除了数据点处于控制限之外，可以认为过程处于脱控状态之外，在控制限之内的数据点轨迹，也可以显示出某些数据质量管理问题。下面我们来讨论过程脱控检验问题。

一、单点超出

如图 7-6 所示，我们可以看到，$\bar{x}$ 控制图中反映出 9 号数据点和 15 号数据点超出了上、下控制限。一般可认为这两个数据为可疑数据。为此，我们要对这两个数据的产生过程进行仔细分析。如果能够找出数据发生偏差的确切原因，我们可判定其为过程脱控，应对数据产生过程中的问题进行调整。如果在分析这两个数据产生的过程中，找不出确切的原因，虽然这两个数据点落在控制限外，我们还不能简单地判定其脱控。因为单个点超出控制限也有可能是随机因素引起的。在此种情况下，我们可以继续抽取样本，增加测定次数，并继续使用控制图。如果仍有偶尔单点超出控制限外，并分析不出确切原因，这就有可能是影响过程的因素较多，有些难以控制，因而随机波动较大。一般情况下，随着测定次数的增加，平均值波动变小，变化可能性不大；而标准差变小的可能性较大。一般而言，随着测定次数的增加，上控制限和下控制限的区间将逐步变窄。说明随着测定次数的增加，控制状态的可信度逐渐提高。

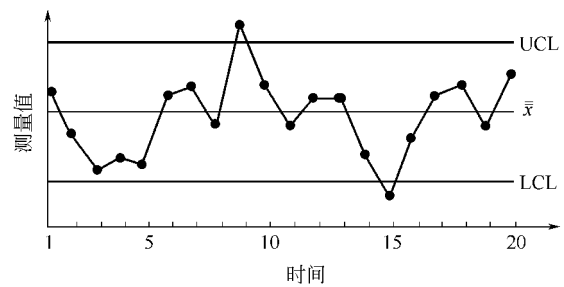


图 7-6　$\bar{x}$ 控制图

二、链分析

就一般情况而言，如果数据点都落在上、下控制限之间，我们可以认为数据产生的过程处于受控状态。但是，如果过多的数据点连续出现在中心的一侧，或数据点分布有明显上升或下降趋势，我们仍要注意，在这种情况下，数据产生的过程有可能失控。因为，数据点的单向趋势可能反映了数据产生过程中系统误差的存在。链分析方法是在此种情况下判别数据在产生过程中是否失控的一种实用方法。

在控制图中，将点连成线后，跨过中心线的次数称为链数。它反映了数据点在中心线两侧的分布情况。过多的点出现在中心线的同一侧，例如由 11 个数据点组成的链中有 10 个点在同一侧，说明过程可能已经失控。

图 7-7 为一数据产生过程的控制图。总共有 19 个数据点。

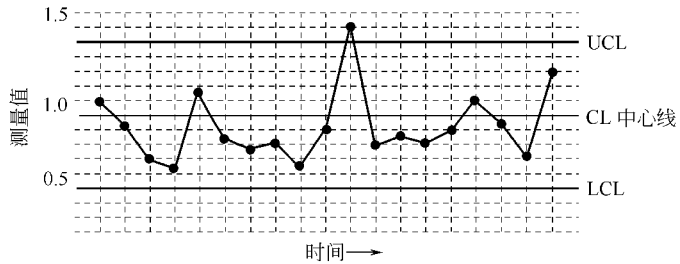


图 7-7 控制图例

折线穿越中心线的次数为 8，即链数为 8。查链数对照表（见表 7-4），总数为 19 时的合理链数范围为 5～14，可以认为控制图没有提供脱控信号。在计算链数时需要注意的是，总数中应不包括那些正好落在中心线上的数据点，并且，在计算链数时，若某个点正好在中心线上，其左右相邻的数据点都在中心线同侧，则不计为一个链数。链分析对于数据点在控制图上单向上升或下降的情况，也是可用的方法。

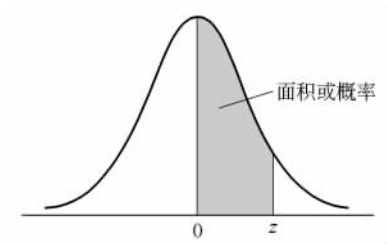
表 7-4 链数对照表

总 数	链数的合理范围	总 数	链数的合理范围	总 数	链数的合理范围
＜10	不能确定	28～29	9～20	48～49	17～32
10～11	3～8	30～31	10～21	50～59	18～33
12～13	3～10	32～33	11～22	60～69	22～39
14～15	4～11	34～35	11～25	70～79	26～45
16～17	5～12	36～37	12～25	80～89	31～50
18～19	5～14	38～39	13～26	90～99	35～56
20～21	6～15	40～41	14～27	100～109	39～62
22～23	7～16	42～43	15～28	110～119	44～67
24～25	8～17	44～45	15～30	120～129	48～73
26～27	8～19	46～47	16～31		

附表

附表 1　标准正态分布表

附表 1 中给出了在曲线下，平均值与大于平均值的 z 个标准差之间的面积。例如，对于 $z=1.25$，在曲线下，平均值与 z 之间的面积是 0.3944。

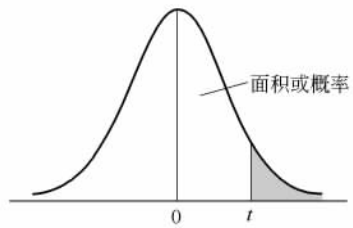


z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2518	0.2549
0.7	0.2580	0.2612	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.4940	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
2.6	0.4953	0.4955	0.4956	0.4957	0.4959	0.4960	0.4961	0.4962	0.4963	0.4964
2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.4970	0.4971	0.4972	0.4973	0.4974
2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.4980	0.4981
2.9	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986
3.0	0.4986	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.4990	0.4990

附表 2 相关系数的临界值 γ_α 表

f \ α	0.10	0.05	0.02	0.01	0.001	f \ α
1	0.98769	0.99692	0.999507	0.999877	0.9999938	1
2	0.90000	0.95000	0.98000	0.99000	0.99900	2
3	0.8054	0.8783	0.93433	0.95873	0.99116	3
4	0.7293	0.8114	0.8822	0.91720	0.97406	4
5	0.6694	0.7545	0.8329	0.8745	0.95074	5
6	0.6215	0.7067	0.7887	0.8343	0.92493	6
7	0.5822	0.6664	0.7498	0.7977	0.8982	7
8	0.5494	0.6319	0.7155	0.7646	0.8721	8
9	0.5214	0.6021	0.6851	0.7348	0.8471	9
10	0.4973	0.5760	0.6581	0.7079	0.8233	10
11	0.4762	0.5529	0.6339	0.6835	0.8010	11
12	0.4575	0.5324	0.6120	0.6614	0.7800	12
13	0.4409	0.5139	0.5923	0.6411	0.7603	13
14	0.4259	0.4973	0.5742	0.6226	0.7420	14
15	0.4124	0.4821	0.5577	0.6055	0.7246	15
16	0.4000	0.4683	0.5425	0.5897	0.7084	16
17	0.3887	0.4555	0.5285	0.5751	0.6932	17
18	0.3783	0.4438	0.5155	0.5614	0.6787	18
19	0.3687	0.4329	0.5034	0.5487	0.6652	19
20	0.3598	0.4227	0.4921	0.5368	0.6524	20
25	0.3233	0.3809	0.4451	0.4869	0.5974	25
30	0.2960	0.3494	0.4093	0.4487	0.5541	30
35	0.2746	0.3246	0.3810	0.4182	0.5189	35
40	0.2573	0.3044	0.3578	0.3932	0.4896	40
45	0.2428	0.2875	0.3384	0.3721	0.4648	45
50	0.2306	0.2732	0.3218	0.3541	0.4433	50
60	0.2108	0.2500	0.2948	0.3248	0.4078	60
70	0.1954	0.2319	0.2737	0.3017	0.3799	70
80	0.1829	0.2172	0.2565	0.2830	0.3568	80
90	0.1726	0.2050	0.2422	0.2673	0.3375	90
100	0.1638	0.1946	0.2301	0.2540	0.3211	100

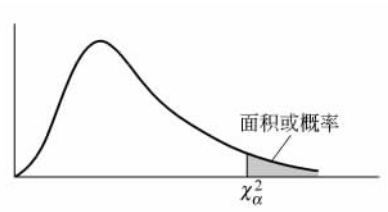
附表 3　t 分布表



附表 3 中的值是对于 t 分布上侧的一个面积或概率的 t 值。例如，当自由度为 10，上侧的面积为 0.05 时，$t_{0.05}=1.812$。

自由度	上侧面积				
	0.10	0.05	0.025	0.01	0.005
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.860	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.250
10	1.372	1.812	2.228	2.764	3.169
11	1.363	1.796	2.201	2.718	3.106
12	1.356	1.782	2.179	2.681	3.055
13	1.350	1.771	2.160	2.650	3.012
14	1.345	1.761	2.145	2.624	2.977
15	1.341	1.753	2.131	2.602	2.947
16	1.337	1.746	2.120	2.583	2.921
17	1.333	1.740	2.110	2.567	2.898
18	1.330	1.734	2.101	2.552	2.878
19	1.328	1.729	2.093	2.539	2.861
20	1.325	1.725	2.086	2.528	2.845
21	1.323	1.721	2.080	2.518	2.831
22	1.321	1.717	2.074	2.508	2.819
23	1.319	1.714	2.069	2.500	2.807
24	1.318	1.711	2.064	2.492	2.797
25	1.316	1.708	2.060	2.485	2.787
26	1.315	1.706	2.056	2.479	2.779
27	1.314	1.703	2.052	2.473	2.771
28	1.313	1.701	2.048	2.467	2.763
29	1.311	1.699	2.045	2.462	2.756
30	1.310	1.697	2.042	2.457	2.750
40	1.303	1.684	2.021	2.423	2.704
60	1.296	1.671	2.000	2.390	2.660
120	1.289	1.658	1.980	2.358	2.617
$+\infty$	1.282	1.645	1.960	2.326	2.576

附表 4　χ^2 分布表



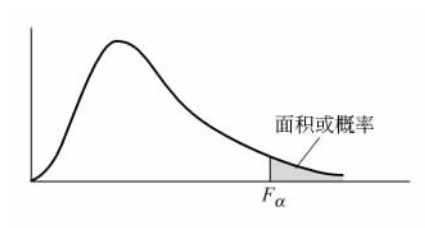
表中给出了 χ^2_α 值。其中 α 是 χ^2 分布上侧的面积或概率。例如，当自由度为 10，上侧的面积为 0.01 时，$\chi^2_{0.01}=23.2093$。

自由度	上侧面积									
	0.995	0.99	0.975	0.95	0.90	0.10	0.05	0.025	0.01	0.005
1	392704×10^{-10}	157088×10^{-9}	982069×10^{-9}	393214×10^{-8}	0.0157908	2.70554	3.84146	5.02389	6.63490	7.87944
2	0.0100251	0.0201007	0.0506356	0.102587	0.210720	4.60517	5.99147	7.37776	9.21034	10.5966
3	0.0717212	0.114832	0.215795	0.351846	0.584375	6.25139	7.81473	9.34840	11.3449	12.8381
4	0.206990	0.297110	0.484419	0.710721	1.063623	7.77944	9.48773	11.1433	13.2767	14.8602
5	0.411740	0.554300	0.831211	1.145476	1.61031	9.23635	11.0705	12.8325	15.0863	16.7496
6	0.675727	0.872085	1.237347	1.63539	2.20413	10.6446	12.5916	14.4494	16.8119	18.5476
7	0.989265	1.239043	1.68987	2.16735	2.83311	12.0170	14.0671	16.0128	18.4753	20.2777
8	1.344419	1.646482	2.17973	2.73264	3.48954	13.3616	15.5073	17.5346	20.0902	21.9550
9	1.734926	2.087912	2.70039	3.32511	4.16816	14.6837	16.9190	19.0228	21.6660	23.5893
10	2.15585	2.55821	3.24697	3.94030	4.86518	15.9871	18.3070	20.4831	23.2093	25.1882
11	2.60321	3.05347	3.81575	4.57481	5.57779	17.2750	19.6751	21.9200	24.7250	26.7569
12	3.07382	3.57056	4.40379	5.22603	6.30380	18.5494	21.0261	23.3367	26.2170	28.2995
13	3.56503	4.10691	5.00874	5.89186	7.04150	19.8119	22.3621	24.7356	27.6883	29.8194

续表

自由度	上侧面积									
	0.995	0.99	0.975	0.95	0.90	0.10	0.05	0.025	0.01	0.005
14	4.07468	4.66043	5.62872	6.57063	7.78953	21.0642	23.6848	26.1190	29.1413	31.3193
15	4.60094	5.22935	6.26214	7.26094	8.54675	22.3072	24.9958	27.4884	30.5779	32.8013
16	5.14224	5.81221	6.90766	7.96164	9.31223	23.5418	26.2962	28.8454	31.9999	34.2672
17	5.69724	6.40776	7.56418	8.67176	10.0852	24.7690	27.5871	30.1910	33.4087	35.7185
18	6.26481	7.01491	8.23075	9.39046	10.8649	25.9894	28.8693	31.5264	34.8053	37.1564
19	6.84398	7.63273	8.90655	10.1170	11.6509	27.2036	30.1435	32.8523	36.1908	38.5822
20	7.43386	8.26040	9.59083	10.8508	12.4426	28.4120	31.4104	34.1696	37.5662	39.9968
21	8.03366	8.89720	10.28293	11.5913	13.2396	29.6151	32.6705	35.4789	38.9321	41.4010
22	8.64272	9.54249	10.9823	12.3380	14.0415	30.8133	33.9244	36.7807	40.2894	42.7958
23	9.26042	10.19567	11.6885	13.0905	14.8479	32.0069	35.1725	38.0757	41.6384	44.1813
24	9.88623	10.8564	12.4011	13.8484	15.6587	33.1963	36.4151	39.3641	42.9798	45.5585
25	10.5197	11.5240	13.1197	14.6114	16.4734	34.3816	37.6525	40.6465	44.3141	46.9278
26	11.1603	12.1981	13.8439	15.3791	17.2919	35.5631	38.8852	41.9232	45.6417	48.2899
27	11.8076	12.8786	14.5733	16.1513	18.1138	36.7412	40.1133	43.1944	46.9630	49.6449
28	12.4613	13.5648	15.3079	16.9279	18.9392	37.9159	41.3372	44.4607	48.2782	50.9933
29	13.1211	14.2565	16.0471	17.7083	19.7677	39.0875	42.5569	45.7222	49.5879	52.3356
30	13.7867	14.9535	16.7908	18.4926	20.5992	40.2560	43.7729	46.9792	50.8922	53.6720
40	20.7065	22.1643	24.4331	26.5093	29.0505	51.8050	55.7585	59.3417	63.6907	66.7659
50	27.9907	29.7067	32.3574	34.7642	37.6886	63.1671	67.5048	71.4202	76.1539	79.4900
60	35.5346	37.4848	40.4817	43.1879	46.4589	74.3970	79.0819	83.2976	88.3794	91.9517
70	43.2752	45.4418	48.7576	51.7393	55.3290	85.5271	90.5312	95.0231	100.425	104.215
80	51.1720	53.5400	57.1532	60.3915	64.2778	96.5782	101.879	106.629	112.329	116.321
90	59.1963	61.7541	65.6466	69.1260	73.2912	107.565	113.145	118.136	124.116	128.299
100	67.3276	70.0648	74.2219	77.9295	82.3581	118.498	124.342	129.561	135.807	140.169

附表 5　F 分布表



表中给出了 F_α 值，其中 α 是 F 分布上侧的面积或概率。例如，当分子自由度为 12、分母自由度为 15、上侧的面积为 0.05 时，$F_{0.05}=2.48$。

$F_{0.05}$ 值表

分母自由度	分子自由度																		
	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	$+\infty$
1	161.4	199.5	215.7	224.6	230.2	234.0	236.8	238.9	240.5	241.9	243.9	245.9	248.0	249.1	250.1	251.1	252.2	253.3	254.3
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.41	19.43	19.45	19.45	19.46	19.47	19.48	19.49	19.50
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.70	8.66	8.64	8.62	8.59	8.57	8.55	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.86	5.80	5.77	5.75	5.72	5.69	5.66	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62	4.56	4.53	4.50	4.46	4.43	4.40	4.36
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.94	3.87	3.84	3.81	3.77	3.74	3.70	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.51	3.44	3.41	3.38	3.34	3.30	3.27	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.22	3.15	3.12	3.08	3.04	3.01	2.97	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.01	2.94	2.90	2.86	2.83	2.79	2.75	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.85	2.77	2.74	2.70	2.66	2.62	2.58	2.54
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.79	2.72	2.65	2.61	2.57	2.53	2.49	2.45	2.40
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.69	2.62	2.54	2.51	2.47	2.43	2.38	2.34	2.30

续表

$F_{0.05}$值表

分母自由度	分子自由度																		
	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	+∞
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.60	2.53	2.46	2.42	2.38	2.34	2.30	2.25	2.21
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.53	2.46	2.39	2.35	2.31	2.27	2.22	2.18	2.13
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.48	2.40	2.33	2.29	2.25	2.20	2.16	2.11	2.07
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.42	2.35	2.28	2.24	2.19	2.15	2.11	2.06	2.01
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.38	2.31	2.23	2.19	2.15	2.10	2.06	2.01	1.96
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.34	2.27	2.19	2.15	2.11	2.06	2.02	1.97	1.92
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.31	2.23	2.16	2.11	2.07	2.03	1.98	1.93	1.88
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.28	2.20	2.12	2.08	2.04	1.99	1.95	1.90	1.84
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.25	2.18	2.10	2.05	2.01	1.96	1.92	1.87	1.81
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.23	2.15	2.07	2.03	1.98	1.94	1.89	1.84	1.78
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27	2.20	2.13	2.05	2.01	1.96	1.91	1.86	1.81	1.76
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.18	2.11	2.03	1.98	1.94	1.89	1.84	1.79	1.73
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.16	2.09	2.01	1.96	1.92	1.87	1.82	1.77	1.71
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.15	2.07	1.99	1.95	1.90	1.85	1.80	1.75	1.69
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20	2.13	2.06	1.97	1.93	1.88	1.84	1.79	1.73	1.67
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19	2.12	2.04	1.96	1.91	1.87	1.82	1.77	1.71	1.65
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18	2.10	2.03	1.94	1.90	1.85	1.81	1.75	1.70	1.64
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.09	2.01	1.93	1.89	1.84	1.79	1.74	1.68	1.62
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.00	1.92	1.84	1.79	1.74	1.69	1.64	1.58	1.51
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.92	1.84	1.75	1.70	1.65	1.59	1.53	1.47	1.39
120	3.92	3.07	2.68	2.45	2.29	2.17	2.09	2.02	1.96	1.91	1.83	1.75	1.66	1.61	1.55	1.50	1.43	1.35	1.25
+∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83	1.75	1.67	1.57	1.52	1.46	1.39	1.32	1.22	1.00

续表

$F_{0.025}$值表

分母自由度	分子自由度																		
	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	＋∞
1	647.8	799.5	864.2	899.6	921.8	937.1	948.2	956.7	963.3	968.6	976.7	984.9	993.1	997.2	1001	1006	1010	1014	1018
2	38.51	39.00	39.17	39.25	39.30	39.33	39.36	39.37	39.39	39.40	39.41	39.43	39.45	39.46	39.46	39.47	39.48	39.49	39.50
3	17.44	16.04	15.44	15.10	14.88	14.73	14.62	14.54	14.47	14.42	14.34	14.25	14.17	14.12	14.08	14.04	13.99	13.95	13.90
4	12.22	10.65	9.98	9.60	9.36	9.20	9.07	8.98	8.90	8.84	8.75	8.66	8.56	8.51	8.46	8.41	8.36	8.31	8.26
5	10.01	8.43	7.76	7.39	7.15	6.98	6.85	6.67	6.68	6.62	6.52	6.43	6.33	6.28	6.23	6.18	6.12	6.07	6.02
6	8.81	7.26	6.60	6.23	5.99	5.82	5.70	5.60	5.52	5.46	5.37	5.27	5.17	5.12	5.07	5.01	4.96	4.90	4.85
7	8.07	6.54	5.89	5.52	5.29	5.21	4.99	4.90	4.82	4.76	4.67	4.57	4.47	4.42	4.36	4.31	4.25	4.20	4.14
8	7.57	6.06	5.42	5.05	4.82	4.65	4.53	4.43	4.36	4.30	4.20	4.10	4.00	3.95	3.89	3.84	3.78	3.73	3.67
9	7.21	5.71	5.08	4.72	4.48	4.32	4.20	4.10	4.03	3.96	3.87	3.77	3.67	3.61	3.56	3.51	3.45	3.39	3.33
10	6.94	5.46	4.83	4.47	4.24	4.07	3.95	3.85	3.78	3.72	3.62	3.52	3.42	3.37	3.31	3.26	3.20	3.14	3.08
11	6.72	5.26	4.63	4.28	4.04	3.88	3.76	3.66	3.59	3.53	3.43	3.33	3.23	3.17	3.12	3.06	3.00	2.94	2.88
12	6.55	5.10	4.47	4.12	3.89	3.73	3.61	3.51	3.44	3.37	3.28	3.18	3.07	3.02	2.96	2.91	2.85	2.79	2.72
13	6.41	4.97	4.35	4.00	3.77	3.60	3.48	3.39	3.31	3.25	3.15	3.05	2.95	2.89	2.84	2.78	2.72	2.66	2.60
14	6.30	4.86	4.24	3.89	3.66	3.50	3.38	3.29	3.21	3.15	3.05	2.95	2.84	2.79	2.73	2.67	2.61	2.55	2.49
15	6.20	4.77	4.15	3.80	3.58	3.41	3.29	3.20	3.12	3.06	2.96	2.86	2.76	2.70	2.64	2.59	2.52	2.46	2.40
16	6.12	4.69	4.08	3.73	3.50	3.34	3.22	3.12	3.05	2.99	2.89	2.79	2.68	2.63	2.57	2.51	2.45	2.38	2.32
17	6.04	4.62	4.01	3.66	3.44	3.28	3.16	3.06	2.98	2.92	2.82	2.72	2.62	2.56	2.50	2.44	2.38	2.32	2.25
18	5.98	4.56	3.95	3.61	3.38	3.22	3.10	3.01	2.93	2.87	2.77	2.67	2.56	2.50	2.44	2.38	2.32	2.26	2.19
19	5.92	4.51	3.90	3.56	3.33	3.17	3.05	2.96	2.88	2.82	2.72	2.62	2.51	2.45	2.39	2.33	2.27	2.20	2.13
20	5.87	4.46	3.86	3.51	3.29	3.13	3.01	2.91	2.84	2.77	2.68	2.57	2.46	2.41	2.35	2.29	2.22	2.16	2.09
21	5.83	4.42	3.82	3.48	3.25	3.09	2.97	2.87	2.80	2.73	2.64	2.53	2.42	2.37	2.31	2.25	2.18	2.11	2.04
22	5.79	4.38	3.78	3.44	3.22	3.05	2.93	2.84	2.76	2.70	2.60	2.50	2.39	2.33	2.27	2.21	2.14	2.08	2.00
23	5.75	4.35	3.75	3.41	3.18	3.02	2.90	2.81	2.73	2.67	2.57	2.47	2.36	2.30	2.24	2.18	2.11	2.04	1.97
24	5.72	4.32	3.72	3.38	3.15	2.99	2.87	2.78	2.70	2.64	2.54	2.44	2.33	2.27	2.21	2.15	2.08	2.01	1.94
25	5.69	4.29	3.69	3.35	3.13	2.97	2.85	2.75	2.68	2.61	2.51	2.41	2.30	2.24	2.18	2.12	2.05	1.98	1.91
26	5.66	4.27	3.67	3.33	3.10	2.94	2.82	2.73	2.65	2.59	2.49	2.39	2.28	2.22	2.16	2.09	2.03	1.95	1.88
27	5.63	4.24	3.65	3.31	3.08	2.92	2.80	2.71	2.63	2.57	2.47	2.36	2.25	2.19	2.13	2.07	2.00	1.93	1.85
28	5.61	4.22	3.63	3.29	3.06	2.90	2.78	2.69	2.61	2.55	2.45	2.34	2.23	2.17	2.11	2.05	1.98	1.91	1.83
29	5.59	4.20	3.61	3.27	3.04	2.88	2.76	2.67	2.59	2.53	2.43	2.32	2.21	2.15	2.09	2.03	1.96	1.89	1.81
30	5.57	4.18	3.59	3.25	3.03	2.87	2.75	2.65	2.57	2.51	2.41	2.31	2.20	2.14	2.07	2.01	1.94	1.87	1.79
40	5.42	4.05	3.46	3.13	2.90	2.74	2.62	2.53	2.45	2.39	2.29	2.18	2.07	2.01	1.94	1.88	1.80	1.72	1.64
60	5.29	3.93	3.34	3.01	2.79	2.63	2.51	2.41	2.33	2.27	2.17	2.06	1.94	1.88	1.82	1.74	1.67	1.58	1.48
120	5.15	3.80	3.23	2.89	2.67	2.52	2.39	2.30	2.22	2.16	2.05	1.94	1.82	1.76	1.69	1.61	1.53	1.43	1.31
＋∞	5.02	3.69	3.12	2.79	2.57	2.41	2.29	2.19	2.11	2.05	1.94	1.83	1.71	1.64	1.57	1.48	1.39	1.27	1.00

续表

$F_{0.01}$值表

分母自由度	分子自由度																		
	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	$+\infty$
1	4052	4999.5	5403	5625	5764	5859	5928	5982	6022	6056	6106	6157	6209	6235	6261	6287	6313	6339	6366
2	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37	99.39	99.40	99.42	99.43	99.45	99.46	99.47	99.47	99.48	99.49	99.50
3	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.35	27.23	27.05	26.87	26.69	26.60	26.50	26.41	26.32	26.22	26.13
4	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66	14.55	14.37	14.20	14.02	13.93	13.84	13.75	13.65	13.56	13.46
5	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16	10.05	9.89	9.72	9.55	9.47	9.38	9.29	9.20	9.11	9.06
6	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.72	7.56	7.40	7.31	7.23	7.14	7.06	6.97	6.88
7	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.47	6.31	6.16	6.07	5.99	5.91	5.82	5.74	5.65
8	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.67	5.52	5.36	5.28	5.20	5.12	5.03	4.95	4.86
9	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	5.11	4.96	4.81	4.73	4.65	4.57	4.48	4.40	4.31
10	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.71	4.56	4.41	4.33	4.25	4.17	4.08	4.00	3.91
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54	4.40	4.25	4.10	4.02	3.94	3.86	3.78	3.69	3.60
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	4.16	4.01	3.86	3.78	3.70	3.62	3.54	3.45	3.36
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10	3.96	3.82	3.66	3.59	3.51	3.43	3.34	3.25	3.17
14	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03	3.94	3.80	3.66	3.51	3.43	3.35	3.27	3.18	3.09	3.00
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80	3.67	3.52	3.37	3.29	3.21	3.13	3.05	2.96	2.87
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69	3.55	3.41	3.26	3.18	3.10	3.02	2.93	2.84	2.75
17	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.68	3.59	3.46	3.31	3.16	3.08	3.00	2.92	2.83	2.75	2.65
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51	3.37	3.23	3.08	3.00	2.92	2.84	2.75	2.66	2.57
19	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43	3.30	3.15	3.00	2.92	2.84	2.76	2.67	2.58	2.49
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	3.23	3.09	2.94	2.86	2.78	2.69	2.61	2.52	2.42
21	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40	3.31	3.17	3.03	2.88	2.80	2.72	2.64	2.55	2.46	2.36
22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26	3.12	2.98	2.83	2.75	2.67	2.58	2.50	2.40	2.31
23	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30	3.21	3.07	2.93	2.78	2.70	2.62	2.54	2.45	2.35	2.26
24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17	3.03	2.89	2.74	2.66	2.58	2.49	2.40	2.31	2.21
25	7.77	5.57	4.68	4.18	3.85	3.63	3.46	3.32	3.22	3.13	2.99	2.85	2.70	2.62	2.54	2.45	2.36	2.27	2.17
26	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18	3.09	2.96	2.81	2.66	2.58	2.50	2.42	2.33	2.23	2.13
27	7.68	5.49	4.60	4.11	3.78	3.56	3.39	3.26	3.15	3.06	2.93	2.78	2.63	2.55	2.47	2.38	2.29	2.20	2.10
28	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12	3.03	2.90	2.75	2.60	2.52	2.44	2.35	2.26	2.17	2.06
29	7.60	5.42	4.54	4.04	3.73	3.50	3.33	3.20	3.09	3.00	2.87	2.73	2.57	2.49	2.41	2.33	2.23	2.14	2.03
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98	2.84	2.70	2.55	2.47	2.39	2.30	2.21	2.11	2.01
40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80	2.66	2.52	2.37	2.29	2.20	2.11	2.02	1.92	1.80
60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63	2.50	2.35	2.20	2.12	2.03	1.94	1.84	1.73	1.60
120	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56	2.47	2.34	2.19	2.03	1.95	1.86	1.76	1.66	1.53	1.38
$+\infty$	6.63	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.41	2.32	2.18	2.04	1.88	1.79	1.70	1.59	1.47	1.32	1.00

附表 6　计算统计量 W 必需的系数 α_k（W）

k \ n	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1	0.7071	0.6872	0.6646	0.6431	0.6233	0.6052	0.5888	0.5739	0.5601	0.5475	0.5359	0.5251	0.5150	0.5056	0.4968	0.4886
2	—	0.1677	0.2413	0.2806	0.3031	0.3164	0.3244	0.3291	0.3315	0.3325	0.3325	0.3318	0.3306	0.3290	0.3273	0.3253
3	—	—	—	0.0875	0.1401	0.1743	0.1976	0.2141	0.2260	0.2347	0.2412	0.2460	0.2495	0.2521	0.2540	0.2553
4	—	—	—	—	—	0.0561	0.0947	0.1224	0.1429	0.1586	0.1707	0.1802	0.1878	0.1939	0.1988	0.2027
5	—	—	—	—	—	—	—	0.0399	0.0695	0.0922	0.1099	0.1240	0.1353	0.1447	0.1524	0.1587
6									—	0.0303	0.0539	0.0727	0.0880	0.1005	0.1109	0.1197
7									—	—	—	0.0240	0.0433	0.0593	0.0725	0.0837
8									—	—	—	—	—	0.0196	0.0359	0.0496
9									—	—	—	—	—	—	—	0.0163
10									—	—	—	—		—	—	—
11																
12																
13																
14																
15																
16																
17																
18																
19																
20																
21																
22																
23																
24																
25																

续表

k \ n	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34
1	0.4808	0.4734	0.4643	0.4590	0.4542	0.4493	0.4450	0.4407	0.4366	0.4328	0.4291	0.4254	0.4220	0.4188	0.4156	0.4127
2	0.3232	0.3211	0.3185	0.3156	0.3126	0.3098	0.3069	0.3043	0.3018	0.2992	0.2968	0.2944	0.2921	0.2898	0.2876	0.2854
3	0.2561	0.2565	0.2578	0.2571	0.2563	0.2554	0.2543	0.2533	0.2522	0.2510	0.2499	0.2487	0.2475	0.2463	0.2451	0.2439
4	0.2059	0.2085	0.2119	0.2131	0.2139	0.2145	0.2148	0.2151	0.2152	0.2151	0.2150	0.2148	0.2145	0.2141	0.2137	0.2132
5	0.1641	0.1686	0.1736	0.1764	0.1787	0.1807	0.1822	0.1836	0.1848	0.1857	0.1864	0.1870	0.1874	0.1878	0.1880	0.1882
6	0.1271	0.1334	0.1399	0.1443	0.1480	0.1512	0.1539	0.1563	0.1584	0.1601	0.1616	0.1630	0.1641	0.1651	0.1660	0.1667
7	0.0932	0.1013	0.1092	0.1150	0.1201	0.1245	0.1283	0.1316	0.1346	0.1372	0.1395	0.1415	0.1433	0.1449	0.1463	0.1475
8	0.0612	0.0711	0.0804	0.0878	0.0941	0.0997	0.1046	0.1089	0.1128	0.1162	0.1192	0.1219	0.1243	0.1265	0.1284	0.1301
9	0.0303	0.0422	0.0530	0.0618	0.0696	0.0764	0.0823	0.0876	0.0923	0.0965	0.1002	0.1036	0.1066	0.1093	0.1118	0.1140
10	—	0.0140	0.0263	0.0368	0.0459	0.0539	0.0610	0.0672	0.0728	0.0778	0.0822	0.0862	0.0899	0.0931	0.0961	0.0988
11			—	0.0122	0.0228	0.0321	0.0403	0.0476	0.0540	0.0598	0.0650	0.0667	0.0739	0.0777	0.0812	0.0844
12			—	—	—	0.0107	0.0200	0.0284	0.0358	0.0424	0.0483	0.0537	0.0585	0.0629	0.0669	0.0706
13			—	—	—	—	—	0.0094	0.0178	0.0253	0.0320	0.0381	0.0435	0.0485	0.0530	0.0572
14			—	—	—	—	—	—	—	0.0084	0.0159	0.0227	0.0289	0.0344	0.0395	0.0441
15			—	—	—	—	—	—	—	—	—	0.0076	0.0144	0.0206	0.0262	0.0314
16													—	0.0068	0.0131	0.0187
17													—	—	—	0.0062
18													—	—	—	—
19													—	—	—	—
20													—	—	—	—
21																
22																
23																
24																
25																

续表

k \ n	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50
1	0.4096	0.4068	0.4040	0.4015	0.3989	0.3964	0.3940	0.3917	0.3894	0.3872	0.3850	0.3830	0.3808	0.3789	0.3770	0.3751
2	0.2834	0.2813	0.2794	0.2774	0.2755	0.2737	0.2719	0.2701	0.2684	0.2667	0.2651	0.2635	0.2620	0.2604	0.2589	0.2574
3	0.2427	0.2415	0.2403	0.2391	0.2380	0.2368	0.2357	0.2345	0.2334	0.2323	0.2313	0.2302	0.2291	0.2281	0.2271	0.2260
4	0.2127	0.2121	0.2116	0.2110	0.2104	0.2098	0.2091	0.2085	0.2078	0.2072	0.2065	0.2058	0.2052	0.2045	0.2038	0.2032
5	0.1883	0.1883	0.1883	0.1881	0.1880	0.1878	0.1876	0.1874	0.1871	0.1868	0.1865	0.1862	0.1859	0.1855	0.1851	0.1847
6	0.1673	0.1678	0.1683	0.1686	0.1689	0.1691	0.1693	0.1694	0.1695	0.1695	0.1695	0.1695	0.1695	0.1693	0.1692	0.1691
7	0.1487	0.1496	0.1505	0.1513	0.1520	0.1526	0.1531	0.1535	0.1539	0.1542	0.1545	0.1548	0.1550	0.1551	0.1553	0.1554
8	0.1317	0.1331	0.1344	0.1356	0.1366	0.1376	0.1384	0.1392	0.1398	0.1405	0.1410	0.1415	0.1420	0.1423	0.1427	0.1430
9	0.1160	0.1179	0.1196	0.1211	0.1225	0.1237	0.1249	0.1259	0.1269	0.1278	0.1286	0.1293	0.1300	0.1306	0.1312	0.1317
10	0.1013	0.1036	0.1056	0.1075	0.1092	0.1108	0.1123	0.1136	0.1149	0.1160	0.1170	0.1180	0.1189	0.1197	0.1205	0.1212
11	0.0873	0.0900	0.0924	0.0947	0.0967	0.0986	0.1004	0.1020	0.1035	0.1049	0.1062	0.1073	0.1085	0.1095	0.1105	0.1113
12	0.0739	0.0770	0.0798	0.0824	0.0848	0.0870	0.0891	0.0909	0.0927	0.0943	0.0959	0.0972	0.0986	0.0998	0.1010	0.1020
13	0.0610	0.0645	0.0677	0.0706	0.0733	0.0759	0.0782	0.0804	0.0824	0.0842	0.0860	0.0876	0.0892	0.0906	0.0919	0.0932
14	0.0484	0.0523	0.0559	0.0592	0.0622	0.0651	0.0677	0.0701	0.0724	0.0745	0.0765	0.0783	0.0801	0.0817	0.0832	0.0846
15	0.0361	0.0404	0.0444	0.0481	0.0515	0.0546	0.0575	0.0602	0.0628	0.0651	0.0673	0.0694	0.0713	0.0731	0.0748	0.0764
16	0.0239	0.0287	0.0331	0.0372	0.0409	0.0444	0.0476	0.0506	0.0534	0.0560	0.0584	0.0607	0.0628	0.0648	0.0667	0.0685
17	0.0119	0.0172	0.0220	0.0264	0.0305	0.0343	0.0379	0.0411	0.0442	0.0471	0.0497	0.0522	0.0546	0.0568	0.0588	0.0608
18	—	0.0057	0.0110	0.0158	0.0203	0.0244	0.0283	0.0318	0.0352	0.0383	0.0412	0.0439	0.0465	0.0489	0.0511	0.0532
19	—	—	—	0.0053	0.0101	0.0146	0.0188	0.0227	0.0263	0.0296	0.0328	0.0357	0.0385	0.0411	0.0436	0.0459
20	—	—	—	—	—	0.0049	0.0094	0.0136	0.0175	0.0211	0.0245	0.0277	0.0307	0.0335	0.0361	0.0386
21							—	0.0045	0.0087	0.0126	0.0163	0.0197	0.0229	0.0259	0.0288	0.0314
22							—	—	—	0.0042	0.0081	0.0118	0.0153	0.0185	0.0215	0.0244
23							—	—	—	—	—	0.0039	0.0076	0.0111	0.0143	0.0174
24							—	—	—	—	—	—	—	0.0037	0.0071	0.0104
25							—	—	—	—	—	—	—	—	—	0.0035

附表 7　W 检验临界值表

n \ α	0.01	0.05	0.10
3	0.753	0.767	0.789
4	0.687	0.748	0.792
5	0.686	0.762	0.806
6	0.713	0.788	0.826
7	0.730	0.803	0.838
8	0.749	0.818	0.851
9	0.764	0.829	0.859
10	0.781	0.842	0.869
11	0.792	0.850	0.876
12	0.805	0.859	0.883
13	0.814	0.866	0.889
14	0.825	0.874	0.895
15	0.835	0.881	0.901
16	0.844	0.887	0.906
17	0.851	0.892	0.910
18	0.858	0.897	0.914
19	0.863	0.901	0.917
20	0.868	0.905	0.920
21	0.873	0.908	0.923
22	0.878	0.911	0.926
23	0.881	0.914	0.928
24	0.884	0.916	0.930
25	0.888	0.918	0.931
26	0.891	0.920	0.933
27	0.894	0.923	0.935
28	0.896	0.924	0.936
29	0.898	0.926	0.937
30	0.900	0.927	0.939
31	0.902	0.929	0.940
32	0.904	0.930	0.941
33	0.906	0.931	0.942
34	0.908	0.933	0.943
35	0.910	0.934	0.944
36	0.912	0.935	0.945
37	0.914	0.936	0.946
38	0.916	0.938	0.947
39	0.917	0.939	0.948
40	0.919	0.940	0.949
41	0.920	0.941	0.950
42	0.922	0.942	0.951
43	0.923	0.943	0.951
44	0.924	0.944	0.952
45	0.926	0.945	0.953
46	0.927	0.945	0.953
47	0.928	0.946	0.954
48	0.929	0.947	0.954
49	0.929	0.947	0.955
50	0.930	0.947	0.955

附表 8 符号检验中 γ 的临界值

n	双侧检验的 α				n	双侧检验的 α			
	0.01	0.05	0.10	0.25		0.01	0.05	0.10	0.25
	单侧检验的 α					单侧检验的 α			
	0.005	0.025	0.05	0.125		0.005	0.025	0.05	0.125
1	—	—	—	—	46	13	15	16	18
2	—	—	—	—	47	14	16	17	19
3	—	—	—	0	48	14	16	17	19
4	—	—	—	0	49	15	17	18	19
5	—	—	0	0	50	15	17	18	20
6	—	0	0	1	51	15	18	19	20
7	—	0	0	1	52	16	18	19	21
8	0	0	1	1	53	16	18	20	21
9	0	1	1	2	54	17	19	20	22
10	0	1	1	2	55	17	19	20	22
11	0	1	2	3	56	17	20	21	23
12	1	2	2	3	57	18	20	21	23
13	1	2	3	3	58	18	21	22	24
14	1	2	3	4	59	19	21	22	24
15	2	3	3	4	60	19	21	23	25
16	2	3	4	5	61	20	22	23	25
17	2	4	4	5	62	20	22	24	25
18	3	4	5	6	63	20	23	24	26
19	3	4	5	6	64	21	23	24	26
20	3	5	5	6	65	21	24	25	27
21	4	5	6	7	66	22	24	25	27
22	4	5	6	7	67	22	25	26	28
23	4	6	7	8	68	22	25	26	28
24	5	6	7	8	69	23	25	27	29
25	5	7	7	9	70	23	26	27	29
26	6	7	8	9	71	24	26	28	30
27	6	7	8	10	72	24	27	28	30
28	6	8	9	10	73	25	27	28	31
29	7	8	9	10	74	25	28	29	31
30	7	9	10	11	75	25	28	29	32
31	7	9	10	11	76	26	28	30	32
32	8	9	10	12	77	26	29	30	32
33	8	10	11	12	78	27	29	31	33
34	9	10	11	13	79	27	30	31	33
35	9	11	12	13	80	28	30	32	34
36	9	11	12	14	81	28	31	32	34
37	10	12	13	14	82	28	31	33	35
38	10	12	13	14	83	29	32	33	35
39	11	12	13	15	84	29	32	33	36
40	11	13	14	15	85	30	32	34	36
41	11	13	14	16	86	30	33	34	37
42	12	14	15	16	87	31	33	35	37
43	12	14	15	17	88	31	34	35	38
44	13	15	16	17	89	31	34	36	38
45	13	15	16	18	90	32	35	36	39

对于大于 90 的 n，γ 的近似值可取小于 $(n-1)/2-k\sqrt{n+1}$ 的最近的整数，对于 1%，5%，10%和 25%，k 的数值分别为 1.2879，0.9800，0.8224，0.5752。

附表 9　二样本秩和检验临界值表

$P(T \leqslant T_\alpha) \leqslant \alpha$，$n_1 < n_2$

$\alpha = 0.05$（下侧）

n_2 \ n_1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	—																			
2	—	—																		
3	—	—	6																	
4	—	—	6	11																
5	—	3	7	12	19															
6	—	3	8	13	20	28														
7	—	3	8	14	21	29	39													
8	—	4	9	15	23	31	41	51												
9	—	4	10	16	24	33	43	54	66											
10	—	4	10	17	26	35	45	56	69	82										
11	—	4	11	18	27	37	47	59	72	86	100									
12	—	5	11	19	28	38	49	62	75	89	104	120								
13	—	5	12	20	30	40	52	64	78	92	108	125	142							
14	—	6	13	21	31	42	54	67	81	96	112	129	147	166						
15	—	6	13	22	33	44	56	69	84	99	116	133	152	171	192					
16	—	6	14	24	34	46	58	72	87	103	120	138	156	176	197	219				
17	—	6	15	25	35	47	61	75	90	106	123	142	161	182	203	225	249			
18	—	7	15	26	37	49	63	77	93	110	127	146	166	187	208	231	255	280		
19	1	7	16	27	38	51	65	80	96	113	131	150	171	192	214	237	262	287	313	
20	1	7	17	28	40	53	67	83	99	117	135	155	175	197	220	243	268	294	320	348

$P(T \geqslant T_\alpha) \leqslant \alpha$，$n_1 < n_2$

$\alpha = 0.05$（上侧）

n_2 \ n_1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	—																			
2	—	—																		
3	—	—	15																	
4	—	—	18	25																
5	—	13	20	28	36															
6	—	15	22	31	40	50														
7	—	17	25	34	44	55	66													
8	—	18	27	37	47	59	71	85												
9	—	20	29	40	51	63	76	90	105											
10	—	22	32	43	54	67	81	96	111	128										
11	—	24	34	46	58	71	86	101	117	134	153									
12	—	25	37	49	62	76	91	106	123	141	160	180								
13	—	27	39	52	65	80	95	112	129	148	167	187	209							
14	—	28	41	55	69	84	100	117	135	154	174	195	217	240						
15	—	30	44	58	72	88	105	123	141	161	181	203	225	249	273					
16	—	32	46	60	76	92	110	128	147	167	188	210	234	258	283	309				
17	—	34	48	63	80	97	114	133	153	174	196	218	242	266	292	319	346			
18	—	35	51	66	83	101	119	139	159	180	203	226	250	275	302	329	357	386		
19	20	37	53	69	87	105	124	144	165	187	210	234	258	284	311	339	367	397	428	
20	21	39	55	72	90	109	129	149	171	193	217	241	267	293	320	349	378	408	440	472

$P(T \leqslant T_a) \leqslant \alpha$，$n_1 < n_2$

$\alpha = 0.025$(下侧)

n_2 \ n_1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	—																			
2	—	—																		
3	—	—	—																	
4	—	—	—	10																
5	—	—	6	11	17															
6	—	—	7	12	18	26														
7	—	—	7	13	20	27	36													
8	—	3	8	14	21	29	38	49												
9	—	3	8	14	22	31	40	51	62											
10	—	3	9	15	23	32	42	53	65	78										
11	—	3	9	16	24	34	44	55	68	81	96									
12	—	4	10	17	26	35	46	58	71	84	99	115								
13	—	4	10	18	27	37	48	60	73	88	103	119	136							
14	—	4	11	19	28	38	50	62	76	91	106	123	141	160						
15	—	4	11	20	29	40	52	65	79	94	110	127	145	164	184					
16	—	4	12	21	30	42	54	67	82	97	113	131	150	169	190	211				
17	—	5	12	21	32	43	56	70	84	100	117	135	154	174	195	217	240			
18	—	5	13	22	33	45	58	72	87	103	121	139	158	179	200	222	246	270		
19	—	5	13	23	34	46	60	74	90	107	124	143	163	183	205	228	252	277	303	
20	—	5	14	24	35	48	62	77	93	110	128	147	167	188	210	234	258	283	309	337

$P(T \geqslant T_a) \leqslant \alpha$，$n_1 < n_2$

$\alpha = 0.025$(上侧)

n_2 \ n_1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	—																			
2	—	—																		
3	—	—	—																	
4	—	—	—	26																
5	—	—	21	29	38															
6	—	—	23	32	42	52														
7	—	—	26	35	45	57	69													
8	—	19	28	38	49	61	74	87												
9	—	21	31	42	53	65	79	93	109											
10	—	23	33	45	57	70	84	99	115	132										
11	—	25	36	46	61	74	89	105	121	139	157									
12	—	26	38	51	64	79	94	110	127	146	165	185								
13	—	28	41	54	68	83	99	116	134	152	172	193	215							
14	—	30	43	57	72	88	104	122	140	159	180	201	223	246						
15	—	32	46	60	76	92	109	127	146	166	187	209	232	256	281					
16	—	34	48	63	80	96	114	133	152	173	195	217	240	265	290	317				
17	—	35	51	67	83	101	119	138	159	180	202	225	249	274	300	327	355			
18	—	37	53	70	87	105	124	144	165	187	209	233	258	283	310	338	366	396		
19	—	39	56	73	91	110	129	150	171	193	217	241	266	293	320	348	377	407	438	
20	—	41	58	76	95	114	134	155	177	200	224	249	275	302	330	358	388	419	451	483

$P(T \leqslant T_a) \leqslant \alpha$，$n_1 < n_2$

$\alpha = 0.01$（下侧）

n_2 \ n_1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	—																			
2	—	—																		
3	—	—	—																	
4	—	—	—	—																
5	—	—	—	10	16															
6	—	—	—	11	17	24														
7	—	—	6	11	18	25	34													
8	—	—	6	12	19	27	35	45												
9	—	—	7	13	20	28	37	47	59											
10	—	—	7	13	21	29	39	49	61	74										
11	—	—	7	14	22	30	40	51	63	77	91									
12	—	—	8	15	23	32	42	53	66	79	94	109								
13	—	3	8	15	24	33	44	56	68	82	97	113	130							
14	—	3	8	16	25	34	45	58	71	85	100	116	134	152						
15	—	3	9	17	26	36	47	60	73	88	103	120	138	156	176					
16	—	3	9	17	27	37	49	62	76	91	107	124	142	161	181	202				
17	—	3	10	18	28	39	51	64	78	93	110	127	146	165	186	207	230			
18	—	3	10	19	29	40	52	66	81	96	113	131	150	170	190	212	235	259		
19	—	4	10	19	30	41	54	68	83	99	116	134	154	174	195	218	241	265	291	
20	—	4	11	20	31	43	56	70	85	102	119	138	158	178	200	223	246	271	297	324

$P(T \geqslant T_a) \leqslant \alpha$，$n_1 < n_2$

$\alpha = 0.01$（上侧）

n_2 \ n_1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	—																			
2	—	—																		
3	—	—	—																	
4	—	—	—	—																
5	—	—	—	30	39															
6	—	—	—	33	43	54														
7	—	—	27	37	47	59	71													
8	—	—	30	40	51	63	77	91												
9	—	—	32	43	55	68	82	97	112											
10	—	—	35	47	59	73	87	103	119	136										
11	—	—	38	50	63	78	93	109	126	143	162									
12	—	—	40	53	67	82	98	115	132	151	170	191								
13	—	29	43	57	71	87	103	120	139	158	178	199	221							
14	—	31	46	60	75	92	109	126	145	165	186	208	230	254						
15	—	33	48	63	79	96	114	132	152	172	194	216	239	264	289					
16	—	35	51	67	83	101	119	138	158	179	201	224	248	273	299	326				
17	—	37	53	70	87	105	124	144	165	187	209	233	257	283	309	337	365			
18	—	39	56	73	91	110	130	150	171	194	217	241	266	292	320	348	377	407		
19	—	40	59	77	95	115	135	156	178	201	225	250	275	302	330	358	388	419	450	
20	—	42	61	80	99	119	140	162	185	208	233	258	284	312	340	369	400	431	463	496

附表 10 $\bar{x}$-R 控制图的系数

样本容量 n	d_2	A_2	d_3	D_3	D_4
2	1.128	1.880	0.853	0	3.267
3	1.693	1.023	0.888	0	2.574
4	2.059	0.729	0.880	0	2.282
5	2.326	0.577	0.864	0	2.114
6	2.534	0.483	0.848	0	2.004
7	2.704	0.419	0.833	0.076	1.924
8	2.847	0.373	0.820	0.136	1.864
9	2.970	0.337	0.808	0.184	1.816
10	3.078	0.308	0.797	0.223	1.777
11	3.173	0.285	0.787	0.256	1.744
12	3.258	0.266	0.778	0.283	1.717
13	3.336	0.249	0.770	0.307	1.693
14	3.407	0.235	0.763	0.328	1.672
15	3.472	0.223	0.756	0.347	1.653
16	3.532	0.212	0.750	0.363	1.637
17	3.588	0.203	0.744	0.378	1.622
18	3.640	0.194	0.739	0.391	1.608
19	3.689	0.187	0.734	0.403	1.597
20	3.735	0.180	0.729	0.415	1.585
21	3.778	0.173	0.724	0.425	1.575
22	3.819	0.167	0.720	0.434	1.566
23	3.858	0.162	0.716	0.443	1.557
24	3.895	0.157	0.712	0.451	1.548
25	3.931	0.153	0.708	0.459	1.541

参 考 文 献

1 邓勃. 数理统计方法在分析测试中的应用. 北京：化学工业出版社，1984

2 中山大学数学力学系. 概率论与数理统计. 北京：人民教育出版社，1980

3 M. 诺顿. 现代控制理论. 北京：科学出版社，1979

4 全浩. 环境管理与技术. 北京：中国环境科学出版社，1994

5 数学手册编写组. 数学手册. 北京：人民教育出版社，1979

6 南希. R. 泰戈. 质量工具箱. 北京：中国城市出版社，2004

7 陆雍森. 环境评价. 上海：同济大学出版社，1999

8 中国环境监测总站《环境水质监测质量保证手册》编写组. 环境水质监测质量保证手册. 北京：化学工业出版社，1994

9 吴鹏鸣等，环境空气监测质量保证手册. 北京：中国环境科学出版社，1989

10 蒋子刚，顾雪梅. 分析测试中的数理统计与质量保证. 上海：华东化工学院出版社，1991

11 高玉堂. 环境监测常用统计方法. 北京：原子能出版社，1981

12 玛·吉·纳特雷拉. 实验统计学. 上海：上海翻译出版社，1990

13 鞠正卫，陈文苗. 概率与数理统计. 哈尔滨：黑龙江科学技术出版社，1984

14 统计方法应用标准汇编小组. 统计方法应用国家标准汇编（2）. 北京：中国标准出版社，1990

15 托马斯 A. 威廉姆斯，丹尼斯 J. 斯威尼，戴维 R. 安德森. 商务与经济统计. 北京：机械工业出版社，2003