环境数据统计分析基础

程子峰 徐富春 编著

化学工业出版社
环境·能源出版中心

环境数据统计分析基础

程子峰 徐富春 编著

化学工业出版社
环境·能源出版中心
·北京·

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. 一北京：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。

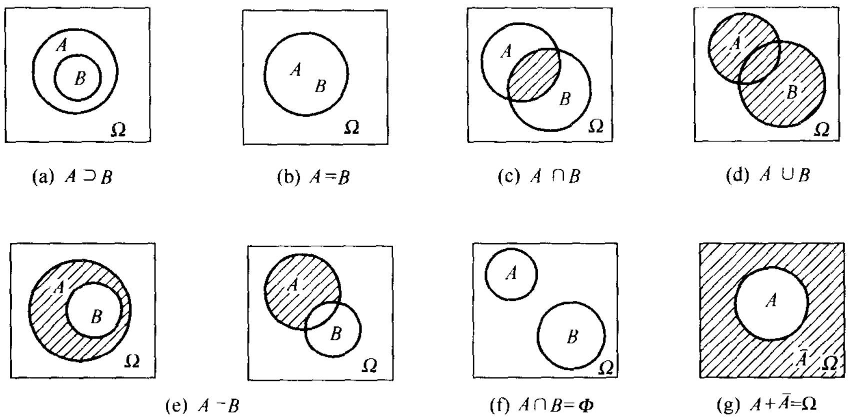

图 1-1 事件的基本关系

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：

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

（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) $A-B=A\cap\bar{B}$ 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：

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

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) $E(kX)=kE(X)$