新世纪高等学校教材

环境科学与工程系列教材

HUANJING TONGJI FENXI

北京师范大学环境学院 组编

环境统计分析

杨晓华 刘瑞民 曾 勇 编著

北京师范大学出版集团
BEIJING NORMAL UNIVERSITY PUBLISHING GROUP
北京师范大学出版社

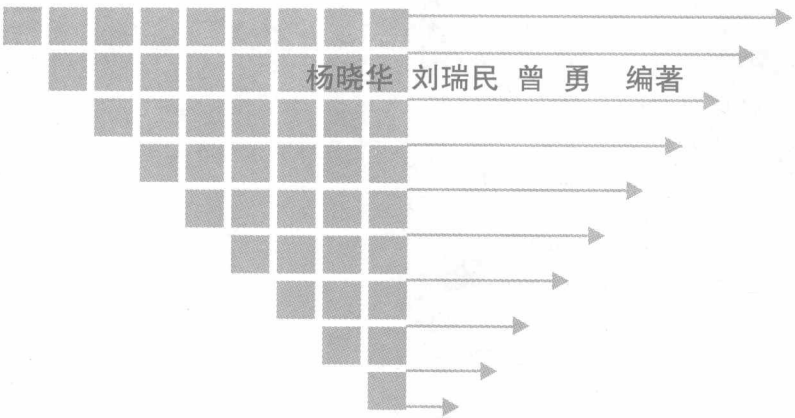
新世纪高等学校教材

环境科学与工程系列教材

北京师范大学环境学院 组编

环境统计分析

HUANJING TONGJI FENXI



杨晓华 刘瑞民 曾 勇 编著

北京师范大学出版集团
BEIJING NORMAL UNIVERSITY PUBLISHING GROUP
北京师范大学出版社

图书在版编目(CIP)数据

环境统计分析／杨晓华，刘瑞民，曾勇编著．—北京：北京师范大学出版社，2008.8
(环境科学与工程系列教材)
ISBN 978-7-303-09502-5

Ⅰ.环… Ⅱ.①杨… ②刘… ③曾… Ⅲ.环境统计－统计分析－高等学校－教材 Ⅳ.X11

中国版本图书馆 CIP 数据核字(2008)第 113056 号

出版发行：北京师范大学出版社 www. bnup. com. cn
北京新街口外大街 19 号
邮政编码：100875
印　　刷：北京新丰印刷厂
经　　销：全国新华书店
开　　本：170mm×230 mm
印　　张：20.25
字　　数：337 千字
印　　数：1～3 000 册
版　　次：2008 年 9 月第 1 版
印　　次：2008 年 9 月第 1 次印刷
定　　价：35.00 元

责任编辑：毛　佳　　装帧设计：高　霞
责任校对：李　菡　　责任印制：马鸿麟

前 言

环境统计分析是环境科学与环境工程的基础学科之一，是一门对环境系统不确定性问题进行数据处理、模型构建和分析的学科。环境系统，系指地球表面包括非生物、生物的各种环境因素及其相互关系的总和，是一个具有时、空、量、序变化的复杂巨系统。受人类活动、天文、气候和气象等众多因素的影响，环境系统中存在许多不确定性现象，并且有大量的数据需要进行统计分析和处理。环境的理论和实践对统计信息的需求急剧增加，对统计分析的理论和方法提出了更高的要求。在自然、社会与环境关系的基础上，用统计方法对环境问题予以量化描述和分析已成为环境研究的迫切需要。环境统计学的产生与发展使人们能够利用数理统计方法处理或解决环境中的不确定性问题，使其定量化，其中包括寻找变量之间的定量关系、从数据中发现环境趋势、探索环境系统变化规律。现代环境统计学一个很重要的标志就是模型技术的运用及量化分析。

全书分三大部分，共10章。其中，第1章属于基础篇，简要地介绍了环境统计分析的概率统计基础知识；第2～9章属于模型篇，阐述了环境一元线性回归分析、环境多元线性回归分析、环境系统聚类分析、环境模糊聚类分析、环境判别分析、环境主成分分析、环境因子分析、人工神经网络等方法、模型的原理，并给出了分析案例；第10章属于空间分析篇，介绍了环境空间统计分析的基本原理，

并给出了应用实例。全书的大多数例子都是用目前常用的统计分析语言 Matlab 编写实现的，是理论联系实际的经验总结，具有可操作性。本书适于做高等院校环境科学与环境工程专业的高年级本科生和研究生教材，对环境科学与环境工程、生态学、资源与管理、应用数学、地理科学等相关领域的学者和科研人员也有重要的参考价值。

本书第 1 章由杨晓华、曾勇执笔，第 2～9 章由杨晓华执笔，第 10 章由刘瑞民执笔，全书由杨晓华统稿。另外，尹心安参加了第 1 章的编写工作；王伟参加了第 3 章、第 4 章、第 10 章的编写工作；陈强、胡晓雪参加了第 5 章、第 6 章的编写工作；佘敦先参加了第 1 章、第 3 章、第 6 章、第 8 章的编写工作。2004 级、2005 级的博士研究生、2005 级的硕士研究生也提供了部分例题和习题。另外，习题答案均是用 Matlab 语言计算完成。

在本书的编写和出版过程中，北京师范大学环境学院院长杨志峰教授，副院长沈珍瑶、刘静玲教授，还有牛军峰、孙涛副教授以及北京师范大学出版社的胡廷兰、毛佳等同志对本书提出了许多宝贵意见。书中若干例题选自所列参考文献，在此一并表示感谢。由于我们的水平有限，书中错误在所难免，欢迎读者批评指正。

衷心感谢北京师范大学出版社给予的大力支持！

本书的完成得到国家重点基础研究发展规划项目（G2003CB415204）的资助，在此表示衷心的感谢！

编著者

2007 年 7 月

内 容 提 要

本书阐述了常用的环境统计分析方法，并给出了分析案例。首先简明扼要地介绍了环境统计分析的概率统计基础知识，又重点阐述了环境一元线性回归分析、环境多元线性回归分析、环境系统聚类分析、环境模糊聚类分析、环境判别分析、环境主成分分析和环境因子分析这些常用的环境统计分析模型；另外还给出了现代环境数据处理常用的人工神经网络方法和空间统计分析方法。对每一种方法，本书除了讲明基本原理外，还给出了大量的计算分析例题和案例。本书的部分例子是用目前实用的统计分析语言 Matlab 编写实现的，是理论联系实际的经验总结，具有实用性。本书适于做高等院校环境科学与环境工程专业的高年级本科生和研究生教材，对环境科学与环境工程、生态学、资源与管理、应用数学、地理科学等相关领域的学者和科研人员也有重要的参考价值。

目 录

第1章　概率统计基础

环境的理论和实践对统计信息的需求急剧增加，对统计分析的理论和方法提出了更高的要求。在自然、社会与环境关系的基础上，用统计方法对环境问题予以量化分析已成为环境科学工作者的迫切需要。环境统计学的产生与发展使人们能够利用数理统计方法处理或解决环境中的不确定性问题，使其定量化，其中包括寻找变量之间的定量关系、从数据中发现环境趋势、探索环境系统变化规律。为了能深刻理解和分析环境数据的数量特征和内在关系，需要我们首先掌握数理统计的基础知识。本章重点阐述环境统计分析的概率统计基础。

本章的主要内容是：

- 四种重要的概率分布；
- 随机向量的数字特征；
- 参数估计；
- 参数假设检验；
- 方差分析与试验设计初步。

1.1　四种重要的概率分布

在环境科学中，弄清统计分析对象的理论分布是关键的一环。土壤中的某些污染物、重金属的分布，大气中若干种微粒的浓度分布、监测值的误差分布等均服从正态分布或接近正态分布或取对数后服从正态分布。χ^2 分布、t 分布、F 分布是统计推断中经常碰到的另外三种分布。研究污染物在环境中的分布规律已是当前环境科学研究中重要的课题之一。

1.1.1　正态分布

市场上的食品很多是 1 kg 袋装，袋上标有“净含量 1 kg”的字样。但当用稍微精确一些的天平称那些食品的重量时，会发现有些可能会重些，有些可能会轻些，但都在 1 kg 左右。其中，多数离 1 kg 不远，离 1 kg 越近就越可能出现，离 1 kg 越远就越不可能。一般认为这种重量分布近似地服从正态分布(normal distribution)。近似地服从正态分布的变量很常见，如实验误差、商品的重量或

尺寸、某年龄人群的身高和体重等。在一定条件下，许多不是正态分布的样本均值在样本量很大时，也可用正态分布来近似。

若随机变量 X 的分布密度为：

$$f(x)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}} \qquad (-\infty<x<+\infty,\ \sigma>0) \tag{1.1}$$

则称 X 服从正态分布 $N(\mu, \sigma^2)$，简记为 $X\sim N(\mu, \sigma^2)$。其中，μ 为均值，σ 为标准差，σ^2 为方差（标准差的平方）。

正态分布的密度曲线是一个对称的、呈钟形的曲线(最高点在均值处)(图 1-1)。正态分布是一族分布，各种正态分布根据它们的均值和标准差不同而有区别。标准差为 1 的正态分布 $N(0, 1)$称为标准正态分布(standard normal distribution)。标准正态分布的密度函数与分布函数记为：

$$f(x)=\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}} \qquad (-\infty<x<+\infty) \tag{1.2}$$

$$\Phi(x)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x}e^{-\frac{y^2}{2}}dy \qquad (-\infty<x<+\infty) \tag{1.3}$$

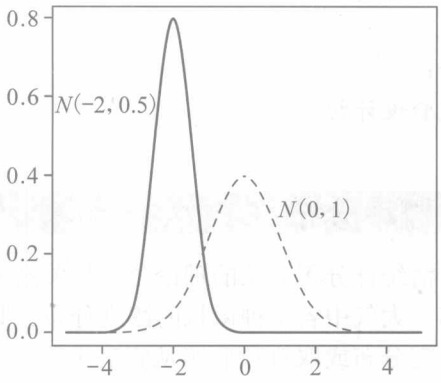


图 1-1　两条正态分布的密度曲线图
(左边是 $N(-2, 0.5)$分布，右边是 $N(0, 1)$分布)

在实际的生活中，我们经常会因为标准正态分布的优异特性而需要将一般的正态分布标准化，下面简单介绍一下正态分布的标准化过程。

设 $X\sim N(\mu, \sigma^2)$，作简单变换(减去其均值 μ，再除以标准差 σ)，则很容易得到随机变量 $Y=\frac{X-\mu}{\sigma}\sim N(0, 1)$。

因为：

$$E(Y)=E\left(\frac{X-\mu}{\sigma}\right)=\frac{1}{\sigma}[E(X)-\mu]=0$$

$$D(Y)=D\left(\frac{X-\mu}{\sigma}\right)=\frac{1}{\sigma^2}D(X)=1$$

这样就将一个普通的正态分布变成了一个标准的正态分布。

标准正态分布中还有一个十分重要的概念就是分位点。为了便于今后应用，对于标准正态随机变量，本书引入上侧分位点的定义(盛聚，1998)。

设 $X\sim N(0,\ 1)$，若 z_α 满足条件

$$P(X>z_\alpha)=\alpha \qquad (0<\alpha<1)$$

则称 z_α 为标准正态分布的上侧 α 分位点，如图 1-2 所示。

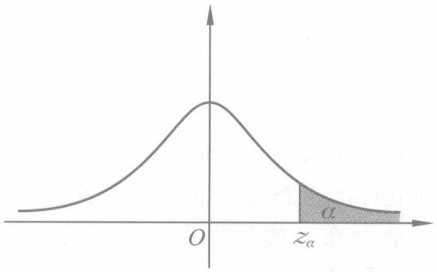


图 1-2　标准正态分布的上侧 α 分位点 z_α

例如，查附表 1 可知：$z_{0.01}=2.326\ 348$，$z_{0.05}=1.644\ 854$，$z_{0.10}=1.281\ 552$，$z_{0.154}=1.019\ 428$。

例 1.1　已知 $X\sim N(\mu,\ \sigma^2)$，求 X 在区间$(\mu-k\sigma,\ \mu+k\sigma)$的概率，这里 $k=1,\ 2,\ 3$。

解　$\forall a,\ b,\ 0<a<b$，有：

$$P(a<X<b)=\int_a^b \frac{1}{\sqrt{2\pi}\sigma}\mathrm{e}^{-\frac{1}{2\sigma^2}(x-\mu)^2}\,\mathrm{d}x$$

$$\xlongequal{t=\frac{x-\mu}{\sigma}}\int_{\frac{a-\mu}{\sigma}}^{\frac{b-\mu}{\sigma}}\frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{1}{2}t^2}\,\mathrm{d}t=\Phi\left(\frac{b-\mu}{\sigma}\right)-\Phi\left(\frac{a-\mu}{\sigma}\right)$$

这样在区间$(\mu-k\sigma,\ \mu+k\sigma)$的概率$(k=1,\ 2,\ 3)$为：

$$P(\mu-\sigma<X<\mu+\sigma)=\Phi(1)-\Phi(-1)=0.682\ 6$$

$$P(\mu-2\sigma<X<\mu+2\sigma)=\Phi(2)-\Phi(-2)=0.954\ 4$$

$$P(\mu-3\sigma<X<\mu+3\sigma)=\Phi(3)-\Phi(-3)=0.997\ 4$$

其中，$\Phi(-x)=1-\Phi(x)$。由此我们可以知道，属于正态分布的随机变量 X 之值，几乎都落在$(\mu-3\sigma,\ \mu+3\sigma)$区间里，落在该区间外的机会极少。

例 1.2　某地水体 COD 浓度 $X\sim N(5,\ 2^2)$，求 COD 浓度落在区间$(4,\ 8)$的

概率。

解 $\mu=5$，$\sigma=2$

$$\begin{aligned}P(4<X<8)&=\Phi\left(\frac{8-\mu}{\sigma}\right)-\Phi\left(\frac{4-\mu}{\sigma}\right)\\&=\Phi\left(\frac{8-5}{2}\right)-\Phi\left(\frac{4-5}{2}\right)\\&=\Phi(1.5)-\Phi(-0.5)\\&=0.933\ 2-0.308\ 5\\&=0.624\ 7\end{aligned}$$

1.1.2 χ^2 分布

一个由正态变量导出的分布是 χ^2 分布(chi-square distribution)。该分布在一些检验中会用到。n 个独立标准正态变量的平方和称为有 n 个自由度的 χ^2 分布，记为 $\chi^2(n)$。χ^2 分布为一族分布，成员由自由度区分。由于 χ^2 分布变量为正态变量的平方和，因此它不会取负值。

设 X_1，X_2，…，X_n 是取自标准正态总体 $N(0,1)$的容量为 n 的样本，那么 $\chi^2=\sum_{i=1}^{n}X_i^2$ 即为由正态分布导出的自由度为 n 的 $\chi^2(n)$分布。所谓自由度，就是指可以自由取值的数据的个数，或者指不受任何约束、可以自由变动的变量的个数。

对于任意一个 $\chi^2(n)$分布，它的概率密度函数为：

$$P(x)=\begin{cases}\dfrac{1}{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)}x^{\frac{n}{2}-1}\mathrm{e}^{-\frac{x}{2}} & (x>0)\\ 0 & (x\leqslant 0)\end{cases}\tag{1.4}$$

记为 $\chi^2\sim\chi^2(n)$，式中 n 为正整数，$\Gamma\left(\frac{n}{2}\right)$为 Γ 函数值，$\Gamma(z)=\int_0^{\infty}x^{z-1}\mathrm{e}^{-x}\mathrm{d}x$ $(z>0)$。

同正态分布类似，对于 χ^2 分布也有上侧 α 分位点。如果 $P(\chi^2>\chi_\alpha^2(n))=\alpha$，则称 $\chi_\alpha^2(n)$为上侧 α 分位点。对于不同的 α，n，上侧 α 分位点的值已制成表格（附表 3)，可以查到。例如对于 $\alpha=0.050$，$n=9$，查得 $\chi_{0.050}^2(9)=16.919$。但大部分书只给出到 $n=45$ 的上侧 α 分位点的值。费歇尔(R. A. Fisher)曾证明，当 n 充分大时，近似有

$$\chi_\alpha^2(n)\approx\frac{1}{2}(z_\alpha+\sqrt{2n-1})^2\tag{1.5}$$

其中，z_α 为标准正态分布的上侧 α 分位点。利用式(1.5)可以求当 $n>45$ 时，$\chi^2_\alpha(n)$分布的上侧 α 分位点的近似值。

例如，查附表1并计算，可得：

$$\chi^2_{0.010}(100)\approx\frac{1}{2}(2.326\ 348+\sqrt{199})^2\approx135.023\ 1$$

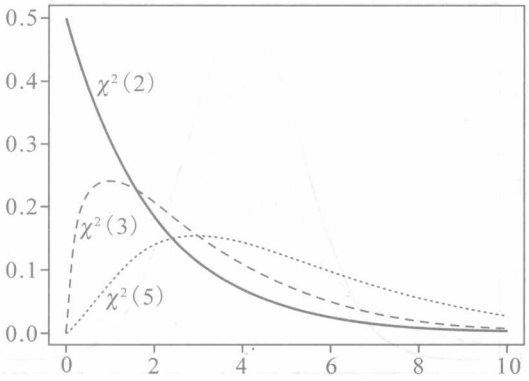


图1-3　自由度分别为2,3,5的 χ^2 分布密度曲线图

1.1.3　t 分布

设 $X\sim N(0,1)$，$Y\sim\chi^2(n)$，并且 X，Y 独立，则随机变量

$$t=\frac{X}{\sqrt{Y/n}}$$

服从自由度为 n 的 t 分布(t-distribution 或 student's t)，记为 $t\sim t(n)$。

对于任意一个 $t(n)$分布，它的概率密度函数为：

$$P(x)=\frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)\sqrt{n\pi}}\left(1+\frac{x^2}{n}\right)^{-\frac{n+1}{2}}\qquad(-\infty<x<+\infty)\tag{1.6}$$

式中，n 为正整数。

不同的样本量通过标准化所产生的 t 分布也不同，这样就形成一族分布。t 分布的分布曲线关于 $x=0$ 对称，它的密度曲线看上去有些像标准正态分布，但是中间瘦一些，而且尾巴长一些。当自由度 k 无限增大时，t 分布将趋近于标准正态分布 $N(0,1)$。

同样，类似于前面的两个分布，t 分布也有上侧 α 分位点的概念。