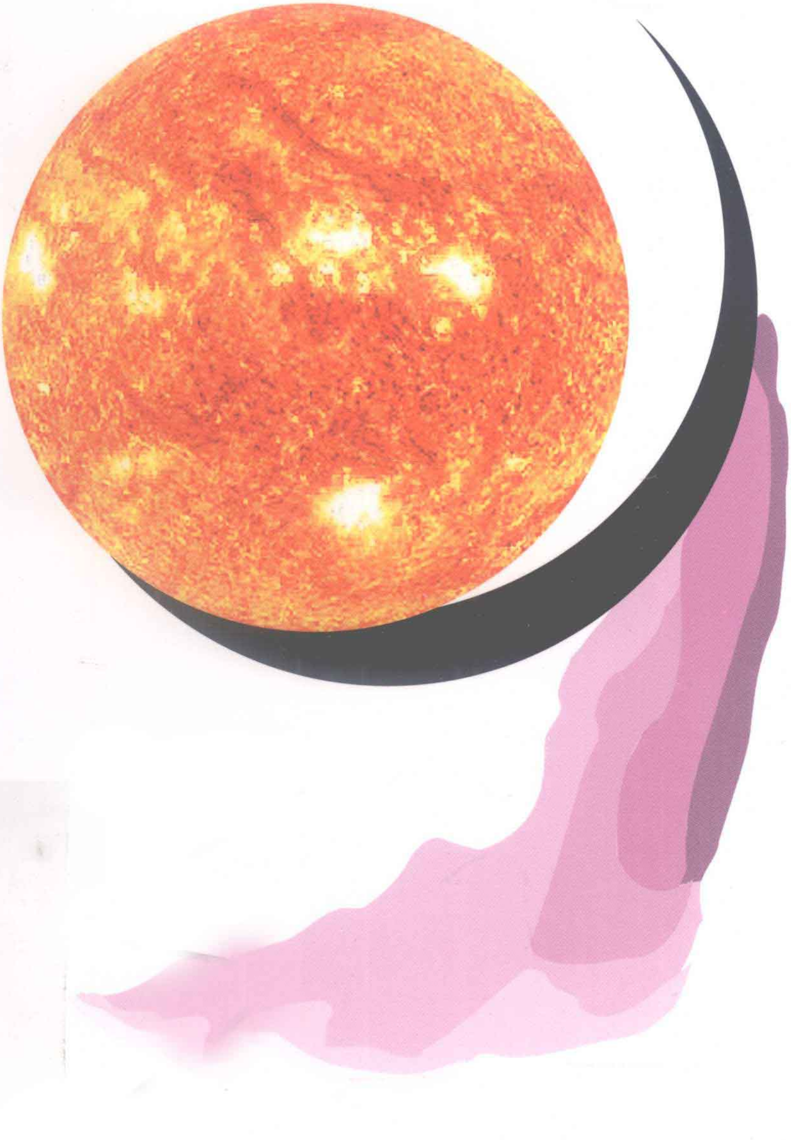
高等院校素质教育通选课教材

应用概率统计研究实例选讲

谢衷洁 著



北京大学出版社
PEKING UNIVERSITY PRESS

高等院校素质教育通选课教材

应用概率统计研究实例选讲

袁洁 著

北京大学出版社
PEKING UNIVERSITY PRESS

图书在版编目(CIP)数据

应用概率统计研究实例选讲/谢衷洁著. —北京：北京大学出版社，2011.8
(高等院校素质教育通选课教材)
ISBN 978-7-301-19439-3

Ⅰ. ①应…　Ⅱ. ①谢…　Ⅲ. ①概率论－高等学校－教材 ②数理统计－高等学校－教材　Ⅳ. ①O21

中国版本图书馆 CIP 数据核字(2011)第 172281 号

书　　　名：应用概率统计研究实例选讲
著作责任者：谢衷洁　著
责 任 编 辑：曾琬婷
标 准 书 号：ISBN 978-7-301-19439-3/C・0694
出 版 发 行：北京大学出版社
地　　　址：北京市海淀区成府路 205 号　100871
网　　　址：http://www.pup.cn　　　电子信箱：zpup@pup.pku.edu.cn
电　　　话：邮购部 62752015　发行部 62750672　理科编辑部 62767347　出版部 62754962
印　刷　者：涿州市星河印刷有限公司
经　销　者：新华书店
787mm×980mm　16 开本　21.5 印张　450 千字
2011 年 8 月第 1 版　　2011 年 8 月第 1 次印刷
定　　　价：45.00 元

内容简介

本书是一部有关概率统计如何应用于多学科实际问题的实证分析研究的教材。本书分为三部分:第一篇是补充一些时间序列分析、滤波与预报理论和方法等方面的基本知识,它们是读懂本书各个案例的理论基础。第二篇是实际案例分析,包括概率论的中心极限定理如何应用于我国通信卫星的交调分析,当年中国科学家如何基于我国的观测记录用统计方法发现天王星至少有6个以上的光环,随机过程的相互包含信息量如何解决一个激素分析的难题,滤波理论如何解决我国海洋重力勘探中的弱信号检测问题,高阶交叉过程如何检查出我国某些彩票中奖号码的问题,潜在周期分析如何发现智障儿童与正常儿童脑电波成分分布的差异,等等,它们都是非常有趣而真实的研究案例。从中读者可以看到其作者们在研究工作中是如何将概率统计的理论和方法广泛地应用于实际问题并成为解决各种问题的核心工具的。更重要的一点是,读者从各案例中将学习到“如何将一个实际问题转换成概率统计问题”,这也是数学联系实际的难点。第三篇中,不仅提供了大量真实的、涉及多学科领域的数据记录,而且作者根据当前研究中以及实际部门提出的问题,提出了一批很有科学意义的课题,以便于学生和其他读者做练习和研究之用。其中有的问题也是当今国际上许多人感兴趣的问题。

本书既可作为大学生和研究生的教材或教学参考书,也是广大应用概率统计工作者和相关领域的科研人员、工程师很有价值的参考书。

作 者 简 介

谢衷洁　北京大学数学科学学院教授、博士生导师。1959年毕业于北京大学数学专业概率论专门化，曾师从国际著名统计学家许宝騄(P. L. Hsu)和概率论学者江泽培(T. P. Chiang)。毕业后留校工作至今已五十多年。历任北京大学统计实验室主任、数理统计研究所副所长，中国概率统计学会理事兼副秘书长，时间序列分析专委会主任等职务。长期从事概率统计的教学和研究，取得了丰硕的研究成果。研究方向早期侧重于概率论，20世纪70年代后期转入时间序列分析及应用，如谱分析、数据建模、滤波与预报。90年代初他的研究组将小波分析引入到时间序列分析中取得了一批重要的学术成果。与日本早稻田大学、香港理工大学、奥地利 Kepler Linz 等高校都有过教学与科研的合作关系。他的应用研究涉及多个学科领域，如天文学、地球物理勘探、通信、生理学、内分泌学、金融工程、汇率及博彩业等。在国内外发表了近百篇学术论文，并在国内外的出版社出版了中、英、日三种语言的7部著作。教学和科研成果曾获国家自然科学奖(三等奖)、国家教学成果奖(二等奖)、国家教委科技进步奖(一等奖)、中国人民解放军军级科技进步奖，并在国外荣获第八届和第十四届维也纳控制论与系统研究大会颁发的"最优论文奖"。

前　言

应用概率统计实例选讲这门课程在北京大学开设已有二十多年的历史，是在江泽培教授倡议下开设的。其背景是：在20世纪80年代的历届全国概率统计会议上，老前辈王寿仁、魏宗舒、江泽培多位先生强烈呼吁中国概率统计工作者要多搞应用研究，要让概率统计在我国四化建设中发挥更大的作用。许多学者还指出：日本战后工业的复兴重要的一点要归功于在经济建设中广泛地运用了概率统计。当时来访的许多著名统计学家也指出，在西方大学中往往设有统计咨询课，在教师指导下让学生和实际部门的工作人员接触，为他们的问题提出看法和解决途径。当时我们认为设咨询课的条件还不具备，但可考虑把老师做过的实际问题的背景，如何将实际问题转化成数学问题，以及解决问题走过的弯路、失败和成功的原因等原原本本地拿到课堂上和学生们进行交流和探讨，并让学生自选题目做实习。在江先生、陈家鼎教授等同仁支持下，应用概率统计实例选讲课程就这样登场了。未曾想到这门课程自开设以来受到学生们的热烈欢迎，不仅在北京大学，在兄弟院校，甚至在日本早稻田大学等校都受到好评。于是世界科技出版社（World Scientific Press）决定出版该课所用的讲义，取名为 *Case Studies in Time Series Analysis*，并在魏宗舒先生帮助下，于1993年在国外出版。此书出版不久即受到国际著名统计学家 H. Tong，Guppta，Swelden 等人的好评（见影印版的封底）。

应该说该书主要反映的是作者在90年代以前从事的一些应用研究，近十多年间又解决了一些有趣的实际问题，而它们又有一定的学术深度，因此决定出版此新的中文版的教材。虽然原出版社曾多次提到希望出版修订版，但因考虑到国外原版书很贵，中国学生一般买不起，所以决定由北京大学出版社出版全新的中文版。

本课程原本是为北京大学概率统计系研究生开设的，基础自然是假定大学数学本科的主要课程是学过的。后来发现听课的学生中有许多是大学本科生，时间序列、测度论等还未学过或正在学。此外，有些教师对有关课程的内容讲得也很不全面，许多很有用的方法也不讲。于是，索性本课程一开始就对时间序列分析先作复习，这部分即相当于本书的第一篇。当然，实际问题的解决有时候是综合性的，因此在介绍各实际问题时，往往还要补充一些知识和工具。作为一门课程我们也不可能把基础讲得太宽了。例如近十年间，作者的研究组从事的一些课题，其所用的工具就大大超出了本书的范围，因而只能“忍痛割爱”了。

本课程的学习分为两个阶段，第一阶段是教师讲解基本知识和实例，时间约占课程的三分之二；第二阶段是利用剩下的时间让学生做练习，教师提供不同的真实数据和要求（有的数据可在本书的第三篇的附录中找到），其中有的课题要求对数据进行建模和对未来作预

报，有的要求寻找频谱特征，有的则希望证实一些学者的论断，等等。为了让做实习研究的读者对有关问题领域的知识有些了解，本书对每个问题都给了背景材料和建议的研究方向。最后，每个学生要按严格的格式写出论文作为本课程的答卷，教师则必须逐一地认真对学生的论文进行评改并给出成绩。此中困难的是：有许多学生积极性很高，他们有自己的问题和研究兴趣，于是自己收集和找来一些真实数据，打算自己运用课堂上学到的知识和方法去解决该问题，这时教师必须非常认真地对待这些学生。因为他们的问题可能是教师不熟悉的领域。一方面教师可能自己要学一些新东西，另一方面也要格外耐心地和他们讨论。他们的论文写得很不规范，方法也有不少缺点，甚至有错误，但是这些学生很可能在教师的帮助下做出新成果，并写出好论文。过去已有不少学生，他们有的还是本科生，其论文最后都发表在国内核心刊物上，这是很值得教师感到欣慰的事。

由于作者的知识面偏窄，能力很有限，本书的写作也比较匆忙，因而一定存在不少缺点和错误，欢迎读者给以批评指教，将来有机会再版时一定作补充和修正。

谢衷洁

2009年冬于北大数学学院

目　　录

第一篇　基础知识与方法简介

第二篇 实际案例研究分析

第三篇 数据与研究实习

第一篇　基础知识与方法简介

第一章　时间序列分析基础知识

§1　随机过程的定义及例子

在初等概率统计中，我们学过一元随机变量和多元随机变量及其分布；在统计学中，我们也学过从观测样本出发，如：

$$一元：x_1, x_2, \cdots, x_N; \tag{1.1.1}$$

$$n\ 元：x_1^{(k)}, x_2^{(k)}, \cdots, x_n^{(k)}\quad (k = 1, 2, \cdots, N), \tag{1.1.2}$$

去对研究对象 F(一元)或 $\boldsymbol{F}$(n 元)的性质，作一些估计、检验或其它统计推断．然而，在现实研究课题中，我们涉及的随机变量往往不是有限个，而是无穷多个，如$\{\xi_0, \xi_1, \xi_2, \cdots, \xi_n, \cdots\}$或$\{\cdots, \xi_{-k}, \cdots, \xi_{-1}, \xi_0, \xi_1, \cdots, \xi_n, \cdots\}$，或者随机变量足标是一个实数集，如$\{\xi_t, t \in (-\infty, +\infty)\}$，而且更困难的是：我们对研究对象的观测不可能像(1.1.2)那样对同一对象进行 N 次重复观测，而要求在只能获得对研究对象的一段观测，如$\{x_1, x_2, \cdots, x_N\}$(离散型)或$\{x_t, a \leqslant t \leqslant b\}$(连续型)的条件下，对总体进行统计推断或检测，这类问题的观测往往还是不可重复的．显然，这是初等概率统计中未学习过的一个新的领域．例如：

(1) 对某城市进行年平均温度的连续 N 年记录得到$\{x_t, t = 1, 2, \cdots, N\}$，问：明年 x_{N+1} 的预报值 $\hat{x}_{N+1}$ 最合理的是多少？

(2) 从杂乱的观测信号的一段记录$\{x_t, a \leqslant t \leqslant b\}$中，判断是否有信号 S_t 的成分？

(3) 对某天体发出的某类信号进行观测，得到记录$\{x_t, 0 \leqslant t \leqslant T\}$，问：该天体发出的信号有没有周期性规律？如果有，是单一的还是多个的周期？哪个最强，哪个次之？各个周期值是多少？

……

我们还可以在金融、通信中找到很多类似的问题．在本书第二篇中读者即可找到许多真实而有趣的实例．以上举的例子中，第一个是时间序列的预报问题，第二个是过程推断问题，第三个则属于随机过程的谱分析．这些课题的理论和方法，我们将逐一作简明介绍．

定义 1.1.1　设$(\Omega, \mathscr{F}, P)$是概率空间，T 是某个足标集，若对任意的 $t \in T$，$\xi_t(\omega)$ $(\omega \in \Omega)$是该概率空间上的一个随机变量，则这一族随机变量$\{\xi_t(\omega), t \in T\}$称为是**随机过程**．

以后，为简明起见，$\xi_t(\omega)$就省略为 ξ_t，其样本写成 x_t，而 T 主要是实数集 $\mathbf{R}$ 或直线上的一段区间$[a, b]$，或者序列 $t = 0, \pm 1, \pm 2, \cdots$，等等．

最简单的随机过程$\{\xi_t, t = 0, 1, 2, \cdots\}$是独立同分布的随机变量序列．例如，投掷硬币，令

$$\xi_t = \begin{cases} 1, & \text{如果出正面}, \\ 0, & \text{如果出反面}, \end{cases} \tag{1.1.3}$$

而一次次独立地投掷，则$\{\xi_t, t=0,1,2,\cdots\}$为一随机过程（以后离散足标情况下称为**随机序列**，且在不混淆的情况下，将随机过程或随机序列简记为ξ_t）.

另一重要的随机过程是实正态过程，其定义如下：设T是实数集，对任意$t\in T$，存在实值函数a_t，且对任意$s,t\in T$，存在二元实值函数$\sigma_{s,t}$，满足

(1) $\sigma_{s,t}=\sigma_{t,s}$; (1.1.4)

(2) 对任意$t_1,t_2,\cdots,t_n\in T$，矩阵

$$\boldsymbol{B}_n = (\sigma_{t_i,t_j})_{1\leqslant i,j\leqslant n} \tag{1.1.5}$$

是非负定矩阵，而$(\xi_{t_1},\cdots,\xi_{t_n})$的特征函数是

$$f_n(u_1,u_2,\cdots,u_n) = \exp\left\{ \mathrm{i}\sum_{k=1}^{n} a_{t_k}u_k - \frac{1}{2}\sum_{k=1}^{n}\sum_{l=1}^{n}\sigma_{t_k,t_l}u_k u_l \right\}, \tag{1.1.6}$$

则$\{\xi_t, t\in T\}$称为实的**正态过程**. 显见，a_t为均值函数，(1.1.5)为其协方差阵.

§2 宽平稳过程与严平稳过程

2.1 宽平稳过程

定义 1.1.2 设$\{\xi_t, t\in T\}$为实值随机过程，其期望和协方差皆存在，如果

(1) $E\xi_t\equiv a$对一切$t\in T$成立，a为常数；

(2) $E[(\xi_{t+\tau}-a)(\xi_t-a)]=R_\xi(\tau)$对一切$t,t+\tau\in T$成立，

则称$\{\xi_t, t\in T\}$为**宽平稳过程**，其中$R_\xi(\tau)$称为**协方差函数**.

通常称

$$B_\xi(\tau) = E(\xi_{t+\tau}\,\xi_t)$$

为**相关函数**. 显见，它和$R_\xi(\tau)$的关系为

$$R_\xi(\tau) = B_\xi(\tau) - |a|^2. \tag{1.1.7}$$

当$a=0$时，两者等价.

以后如不声明皆假定$a=0$，若不然可令

$$\eta_t = \xi_t - a, \quad t\in T,$$

则

$$E\eta_t = 0, \quad R_{\eta_t}(\tau) = B_{\eta_t}(\tau). \tag{1.1.8}$$

关于宽平稳序列，可举以下例子：

例 1.1.1（白噪声序列） 设$\{\xi_t, t=\mathbf{0},\pm\mathbf{1},\pm\mathbf{2},\cdots\}$是实的独立同分布序列，具有二阶矩，均值$a=\mathbf{0}$，则

$$E(\xi_{t+\tau}\xi_t)=\begin{cases}\sigma^2, & \tau=0,\\ 0, & \tau\neq 0.\end{cases} \tag{1.1.9}$$

显见,(**1.1.9**)式右边与 t 无关,故为宽平稳序列.

例 1.1.2(随机振荡过程) 设 ξ 为零均值、具有二阶矩的随机变量: $E\xi=0, E\xi^2=\sigma^2$; θ 为服从均匀分布 $U[0,2\pi]$ 的随机相位,则

$$\zeta_t=\xi\cos(\omega t+\theta),\quad t\in T \tag{1.1.10}$$

是宽平稳过程,其中 ξ 与 θ 独立,ω 为圆频率(常数).

读者可自行证明:

$$E\zeta_t=0,$$

$$E(\zeta_{t+\tau}\zeta_t)=\frac{\sigma^2}{2}\cos\omega\tau,\quad t,\tau\in T, \tag{1.1.11}$$

其中用到 $E[\cos(C+\theta)]=0$,C 为常数,$\theta\sim U[0,2\pi]$.

以上结果可推广到

$$\zeta_t=\sum_{k=1}^{N}\xi_k\cos(\omega_k t+\theta_k),\quad t\in T$$

的情形. 也就是说,如果 $\{\xi_k\}$ 相互独立,零均值,方差为 $\{\sigma_k^2\}$,而 $\{\theta_k\}$ 为与 $\{\xi_k\}$ 独立并自身相互独立同服从分布 $U[0,2\pi]$ 的随机变量,则 $\{\zeta_t,t\in T\}$ 是宽平稳过程.

宽平稳性是建立在随机过程的一、二阶矩"时不变"的基础上的,即均值、方差和协方差不随时间 t 的推移而改变. 还有另一类平稳性是建立在分布律上的,相应的随机过程称为严平稳过程.

2.2 严平稳过程

定义 1.1.3 设 $\{\xi_t,t\in T\}$ 是实值平稳过程,若对任意的 $n>0$,取 $t_1,t_2,\cdots,t_n\in T,\tau_1,\tau_2,\cdots,\tau_n\in T$, $\{\xi_{t_1},\xi_{t_2},\cdots,\xi_{t_n}\}$ 和 $\{\xi_{t_1+\tau_1},\xi_{t_2+\tau_2},\cdots,\xi_{t_n+\tau_n}\}$ 具有相同的分布函数

$$F_{t_1,t_2,\cdots,t_n}(x_1,x_2,\cdots,x_n)=F_{t_1+\tau_1,t_2+\tau_2,\cdots,t_n+\tau_n}(x_1,x_2,\cdots,x_n), \tag{1.1.12}$$

则称 $\{\xi_t,t\in T\}$ 是**严平稳过程**.

最简单的严平稳过程就是独立同分布序列 $\{\xi_t,t=0,\pm1,\cdots\}$.

关于"宽"与"严"的性质有以下几点值得读者注意:

(1) 严平稳过程不意味着一定是宽平稳过程. 因为在定义 1.1.3 中并未要求随机变量 ξ_t 具有一、二阶矩. 大家知道 Cauchy 分布就没有期望和方差.

(2) 宽平稳过程未必是严平稳过程. 因为一、二阶矩对时间推移的不变性并不能得到分布函数对时间推移的不变性. 读者可自己构造一个过程来说明这一问题.

(3) 严平稳过程若具有二阶矩,则它必然是宽平稳的. 因为一、二阶矩可理解为都是建立在对分布的积分基础上的,而分布对时间推移的不变性就使得均值、方差、协方差也不变.

(4) 对正态过程而言,“宽”与“严”等价.因为正态分布只用到一、二阶的参数.

以后我们主要研究实的宽平稳过程(或序列).为简便起见,以后就将宽平稳过程简称为平稳过程,而且皆假定均值 $a=0$,除非有声明.

关于相关函数 $R_\xi(\tau)$,它具有以下常用性质:

(1) $R_\xi(0)=\sigma^2\geqslant 0$; (1.1.13)

(2) $|R_\xi(\tau)|\leqslant R_\xi(0)$; (1.1.14)

(3) 对 $t_1,t_2\cdots,t_n\in T$,矩阵$(R(t_i-t_j))_{1\leqslant i,j\leqslant n}$是非负定矩阵;

(4) 当 $T=\mathbf{R}$ 或 $T=[0,+\infty)$时,$R_\xi(\tau)$是连续函数.

值得注意的是:$R_\xi(\tau)$连续并不意味着随机过程$\{\xi_t,t\in T\}$的观测曲线 $x_t(t\in T)$是连续函数.

例 1.1.3(随机电报信号) 设 $T=(-\infty,+\infty)$,ξ_t 只取“1”和“0”两种状态(见图 1.1.1):

$$\xi_t\sim\begin{pmatrix}1 & 0\\ 1/2 & 1/2\end{pmatrix},\quad t\in T. \tag{1.1.15}$$

在 Δ 时间间隔内,状态可能发生变化的次数 $\mu(\Delta)$服从 Poisson 分布:

$$P\{\mu(\Delta)=k\}=\frac{(\lambda_0\Delta)^k}{k!}\mathrm{e}^{-\lambda_0\Delta},\quad k=0,1,2,\cdots, \tag{1.1.16}$$

其中 $\Delta,\lambda_0>0$,ξ_t 与 $\mu(\Delta)$是独立的(即任意 Δ 间隔内波形变化的次数不受 t 时刻随机过程所处状态的影响).可以算出(参见谢(1990)):

(1) $E(\xi_t)=\dfrac{1}{2}$;

(2) $R_\xi(\tau)=\dfrac{1}{4}\mathrm{e}^{-2\lambda_0|\tau|}$ (见图 1.1.2).

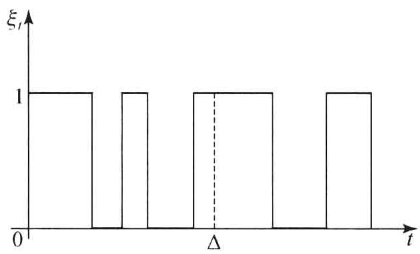


图 1.1.1 随机电报信号的记录

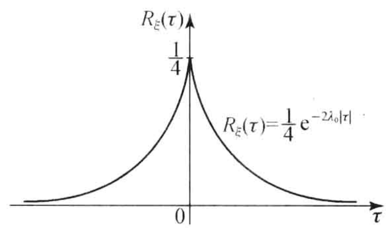


图 1.1.2 随机电报信号的相关函数

由相关函数 $R_\xi(\tau)$与 t 无关知道$\{\xi_t,t\in T\}$是平稳的,而且由图 1.1.2 知道 $R_\xi(\tau)$是连续函数,但电报信号 ξ_t 显然是间断函数.

§3 平稳过程的谱函数与谱密度

定义 1.1.2 不难推广到复平稳过程.