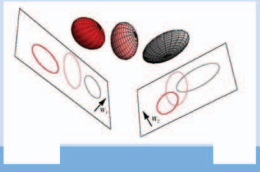
河北省重点学科应用统计学丛书

概率统计模型与优化

GAILÜ TONGJI MOXING YU YOUHUA



赵秀恒 梁建英 张良勇 王倩影 著

河北科学技术出版社

河北省重点学科应用统计学丛书

概率统计模型与优化

GAILÜ TONGJI MOXING YU YOUHUA

赵秀恒 梁建英 张良勇 王倩影 著

河北科学技术出版社

图书在版编目（CIP）数据

概率统计模型与优化 / 赵秀恒等著. —— 石家庄 ：河北科学技术出版社，2015.6

ISBN 978-7-5375-7682-6

Ⅰ. ①概… Ⅱ. ①赵… Ⅲ. ①概率统计-统计模型 Ⅳ. ①O211

中国版本图书馆 CIP 数据核字（2015）第 163459 号

概率统计模型与优化

赵秀恒 梁建英 张良勇 王倩影 著

出版发行	河北科学技术出版社
地　　址	石家庄市友谊北大街 330 号（邮编：050061）
印　　刷	石家庄燕赵创新印刷有限公司
开　　本	787×1092　1/16
印　　张	16.25
字　　数	290 千字
版　　次	2015 年 6 月第 1 版
	2015 年 6 月第 1 次印刷
定　　价	50.00 元

内 容 简 介

统计学能得到迅速的发展，主要原因就是它可与其他学科融合，根据实际问题的需要，不断探索新的数据分析方法，逐渐形成新的理论。在经济、社会、人口、医学等众多研究领域中，人们通过各种方式收集数据，然后对数据进行统计分析，利用分析结果指导社会实践。

随着经济、人口、医学、生物、金融、环境等学科的发展，我们所面临的数据也越来越多样化、复杂化，随之而来对相应统计模型与方法的需求也越来越紧迫。所以，本书从几个侧面，给出了一些统计模型及研究方法。

该书内容主要涉及精算数学中随机利率下的寿命统计模型，包括随机利率下的离散型精算统计模型、连续型精算统计模型、全离散型精算统计模型、全连续型精算统计模型；精算数学中的风险统计模型，包括基于常数利息力下连续型再保险的风险模型、基于整值时间序列数据的风险模型；微阵列数据下的统计模型，从理论和应用两个方面阐述了微阵列数据下的多重假设检验的一些理论与方法；研究了一些借助计算机来处理随机数据的方法，包括机器学习与模式识别，使计算更具有智能化；利用贝叶斯方法研究不确定性模型以及效用理论的决策问题；最后一章介绍了马尔科夫链模型及其应用。

目 录

第 1 章　绪　论

1.1　精算统计模型

保险精算学是依据经济学的理论知识和基本原理，利用现代数学理论知识和方法，对各种保险经济活动未来的财务风险进行分析、估价和管理的一门综合性的应用科学. 因此，随着保险业的发展，保险精算学的研究受到越来越多学者关注，可详见李秀芳，曾庆五（2005）；卢仿允，曾庆五（2001）.

寿险业作为保险业中的重要行业之一，对经济的发展有着重要的影响. 因此，对寿险业中存在的风险的预测是至关重要的. 寿险业的风险性主要受费用率、死亡率、利率的影响. 对于费用率和死亡率，保险公司可以通过严格的核保核赔对费用率进行合理的预测，以及通过不同的调查方式对死亡率进行评估预测. 利率受社会环境、政府政策、经济变化、自然灾害等多因素的影响，要对利率未来的变化进行合理的预测是比较艰难的. 传统寿险精算模型是基于固定利率条件下进行研究，这与现实中利率的随机性存在很大的偏差. 因此，降低利率不确定性所带来的风险，更好的方法就是将利率随机化进行建模. 本章是在寿险精算理论学习及借鉴国内外学者研究的基础上，构建一些随机利率下的统计模型.

1.1.1　利息理论

（1）利息的度量

期初本金为 1 元，利率为 i，单利条件下的积累函数可表示为

$$a(t)=1+it,\ t=0,1,2,\cdots \tag{1.1}$$

贴现函数可表示为

$$d=\frac{1}{1+it} \tag{1.2}$$

期初本金为 1 元，复利条件下的积累函数可表示为

$$a(t)=(1+i)^t, t=0,1,2,\cdots \tag{1.3}$$

贴现函数可表示为

$$d=\frac{1}{(1+i)^t} \tag{1.4}$$

即 $v=\frac{1}{1+i}$，称 v 为贴现因子.

下文所涉及的利息在不说明的情况下，都指复利的情况.

(2) 年金的现值

期初给付定期年金的现值的计算模型为

$$\ddot{a}_{\overline{n}|}=v+v^2+\cdots v^n=\frac{1-v^n}{d} \tag{1.5}$$

期末给付定期年金的现值的计算模型为

$$a_{\overline{n}|}=1+v+v^2+\cdots v^{n-1}=\frac{1-v^n}{i} \tag{1.6}$$

1.1.2 生存模型

一个人的寿命是从刚刚出生到死亡瞬间的时间所跨越的长度，它是不能预先确定的，其在概率上称之为随机变量，将其记为 X；另外，记 x 岁的人为 (x)，则 (x) 剩余寿命也是随机变量，记为 T.

一个刚出生的婴儿在 x 岁之前死亡的概率为

$$F(x)=P(X\leqslant x), x\geqslant 0 \tag{1.7}$$

称 $F(x)$ 为寿命分布函数.

在精算学中，我们常用另一个函数来描述寿命的分布，这个函数定义为

$$S(x)=P(X>x), x\geqslant 0 \tag{1.8}$$

称为生存函数，表示 0 岁的人活过 x 岁的概率. 显然，生存函数与寿命分布函数有如下关系

$$S(x)=1-F(x), x\geqslant 0 \tag{1.9}$$

接下来考虑 (x) 的余寿 T，余寿 T 的分布函数为

$$\begin{aligned} F_T(t)&=P(T\leqslant t)=P(x<X\leqslant x+t\,|\,X>x)\\ &=\frac{F(x+t)-F(x)}{1-F(x)}=\frac{S(x)-S(x+t)}{S(x)} \end{aligned} \tag{1.10}$$

余寿 T 的概率密度函数为

$$f_T(t)=F'_T(t)=-\frac{S'(x+t)}{S(x)} \tag{1.11}$$

下面来介绍精算学中用国际通用符号来表示的关于 T 的各种概率，表示如下

$$ {}_tq_x = P(T \leqslant t) = \frac{S(x) - S(x+t)}{S(x)}, t \geqslant 0 \tag{1.12} $$

$$ {}_tp_x = P(T > t) = 1 - {}_tq_x = \frac{S(x+t)}{S(x)}, t \geqslant 0 \tag{1.13} $$

$$ {}_{t|u}q_x = {}_tp_x - {}_{u+t}q_x = \frac{S(x+t) - S(x+t+u)}{S(x)} = {}_tp_x \cdot {}_uq_{x+t}, t \geqslant 0, u \geqslant 0 \tag{1.14} $$

在精算学中，寿命分布除了用生存函数 $S(x)$ 来表示外，还习惯于用生存函数的相对变化率 $-\dfrac{S'(x)}{S(x)}$ 来表示，称之为死力，记为 μ_x，则

$$ f_T(t) = F'_T(t) = -\frac{S'(x+t)}{S(x)} = \frac{S(x+t)}{S(x)}\left[-\frac{S'(x+t)}{S(x)}\right] = {}_tp_x\mu_{x+t} \tag{1.15} $$

1.1.3　传统寿险精算理论

假设 (x) 进行投保，给付保险金额为 1 元时，签单时保险金给付现值随机变量记为 Z，余寿 T 的取整余寿记为 K.

文章所涉及的保险金额不特殊说明的情况下，都假定为 1 元.

(1) 趸缴纯保费

趸缴纯保费的相关概念：

用 x 表示投保年龄，b_t 表示保险金给付函数，v_t 表示折现函数，t 为从签单到死亡的时间长度，T 为被保险人的余寿随机变量，$K = [T]$ 表示取整余寿随机变量. 定义现值函数 $z_t = b_tv_t$，表示未来保险金给付在签单时的现值，定义现值函数的期望为趸缴纯保费，即一次性缴清的纯保费.

(2) 趸缴纯保费精算模型

对于死亡即付寿险的趸缴纯保费：

(a) n 年定期保险的趸缴纯保费

$$ \overline{A}_{x:\overline{n}|} = E_T(Z) = \int_0^n v^t \cdot f_T(t)dt = \int_0^n v^t \cdot {}_tp_x \cdot \mu_{x+t}dt \tag{1.16} $$

(b) 终身寿险的趸缴纯保费

$$ \overline{A}_x = E_T(Z) = \int_0^\infty v^t \cdot f_T(t)dt = \int_0^\infty v^t \cdot {}_tp_x \cdot \mu_{x+t}dt \tag{1.17} $$

(c) n 年期两全保险的趸缴纯保费

$$\bar{A}_{x:\overline{n}|} = E_T(Z) = \int_0^n v^t \cdot {}_tp_x \cdot \mu_{x+t}dt + v^n \cdot {}_np_x \tag{1.18}$$

对于死亡年末给付的寿险的趸缴纯保费：

(a) n 年定期保险的趸缴纯保费

$$A^1_{x:\overline{n}|} = E_T(Z) = \sum_{k=0}^{n-1} v^{k+1} \cdot {}_kp_xq_{x+k} \tag{1.19}$$

(b) 终身寿险的趸缴纯保费

$$A_x = E_T(Z) = \sum_{k=0}^{\infty} v^{k+1} \cdot {}_kp_xq_{x+k} \tag{1.20}$$

(c) n 年期两全保险的趸缴纯保费

$$A_{x:\overline{n}|} = E_T(Z) = \sum_{k=0}^{n-1} v^{k+1} \cdot {}_kp_xq_{x+k}dt + v^n \cdot {}_np_x \tag{1.21}$$

(3) 生存年金

生存年金是指在已知某人生存的条件下，按预先约定的金额以连续方式或以一定的周期进行一系列的给付的保险，且每次年金给付必须以年金受领人生存为条件. 一旦年金受领人死亡，给付便立即停止. 生存年金精算模型分为.

对连续给付型生存年金的精算现值的计算模型：

(a) n 年定期生存年金

$$\bar{a}_{x:\overline{n}|} = \int_0^n v^t \cdot {}_tp_xdt \tag{1.22}$$

(b) 终身生存年金

$$\bar{a}_x = \int_0^{\infty} v^t \cdot {}_tp_xdt \tag{1.23}$$

对离散给付型生存年金的精算现值的计算模型：

(a) 期初给付生存年金

(i) n 年定期生存年金

$$\ddot{a}_{x:\overline{n}|} = \sum_{k=0}^{n-1} v^k \cdot {}_kp_x \tag{1.24}$$

(ii) 终身生存年金

$$\ddot{a}_x = \sum_{k=0}^{\infty} v^k \cdot {}_kp_x \tag{1.25}$$

(b) 期末给付生存年金

(i) n 年定期生存年金

$$a_{x:\overline{n}|}=\sum_{k=1}^{n}v^{k}\cdot{}_{k}p_{x} \tag{1.26}$$

(ii) 终身生存年金

$$a_{x}=\sum_{k=0}^{\infty}v^{k}\cdot{}_{k}p_{x} \tag{1.27}$$

(4) 年缴纯保费

由于趸缴纯保费的方式要求投保人一次缴纳数目很大的保费，实为一般收入的投保人难以负担. 因此，在实际业务中，绝大多数的寿险业务采用分期缴费的方式，按年缴的纯保费，称为年缴纯保费. 年缴纯保费的计算模型分为.

全连续型年缴纯保费模型：

(a) n 年定期的年缴纯保费

$$\overline{p}(\overline{A}^{1}_{x:\overline{n}|})=\frac{\overline{A}^{1}_{x:\overline{n}|}}{\overline{a}_{x:\overline{n}|}} \tag{1.28}$$

(b) 终身寿险的年缴纯保费

$$\overline{p}(\overline{A}_{x})=\frac{\overline{A}_{x}}{\overline{a}_{x}} \tag{1.29}$$

(c) n 年期两全保险的趸缴纯保费

$$\overline{p}(\overline{A}_{x:\overline{n}|})=\frac{\overline{A}_{x:\overline{n}|}}{\overline{a}_{x:\overline{n}|}} \tag{1.30}$$

全离散型年缴纯保费模型：

(a) n 年定期的年缴纯保费

$$p(A^{1}_{x:\overline{n}|})=\frac{A^{1}_{x:\overline{n}|}}{\ddot{a}_{x:\overline{n}|}} \tag{1.31}$$

(b) 终身寿险的年缴纯保费

$$p(A_{x})=\frac{A_{x}}{\ddot{a}_{x}} \tag{1.32}$$

(c) n 年期两全保险的趸缴纯保费

$$p(A_{x:\overline{n}|})=\frac{A_{x:\overline{n}|}}{\ddot{a}_{x:\overline{n}|}} \tag{1.33}$$

半连续型年缴纯保费模型：

(a) n 年定期的年缴纯保费

$$p(\overline{A}^{1}_{x:\overline{n}|})=\frac{\overline{A}^{1}_{x:\overline{n}|}}{\ddot{a}_{x:\overline{n}|}} \tag{1.34}$$

(b) 终身寿险的年缴纯保费

$$p(\overline{A}_{x})=\frac{\overline{A}_{x}}{\ddot{a}_{x}} \tag{1.35}$$

(c) n 年期两全保险的趸缴纯保费

$$p(\overline{A}_{x:\overline{n}|})=\frac{\overline{A}_{x:\overline{n}|}}{\ddot{a}_{x:\overline{n}|}} \tag{1.36}$$

1.2 风险统计模型

风险理论已成为概率论和数理统计的一个非常活跃的分支，是当今精算界和数学界研究的热门话题，它主要运用统计方法，用数学模型定量描述保险经营过程并对其相关数字特征进行研究，进而定量分析保险公司的破产风险，为保险公司提供早期的预警系统，以提高保险业的经营管理能力和自身的竞争力.

风险理论已经发展了很长一段时间，最早对其作出贡献的有 Edmund Halley 和 Daniel Bernoulli，前者构造了世界上第一张生命表，后者提出了以极大效用原理作为决策法则的思想.

到了 20 世纪，瑞典精算师 Flip Lundberg 于 1903 年完成的博士论文《Approximerad Framstallning av Sannolikhets funkionen》中提出了一类重要的随机过程—Poisson 过程，并指出齐次 Poisson 过程是保险公司负债数据的一个主要模型，也就是所谓的经典风险模型. 但是 Flip Lundberg 的研究并没有以严格的现代数学为基础. 在随后的时间里，以 Harald Cram′er 为首的瑞典学派将 Flip Lundberg 的工作建立在严格的数学基础之上，并建立了风险理论与随机过程理论之间的关系见 Cram′er (1930，1945，1954，1955). Lundberg 和 Cram′er 的工作现在已经被公认为研究经典风险理论的基础.

1.2.1 两个基本的风险模型

首先给出本章的基本假设：令 (Ω,F,P) 为一个完备的概率空间，文中模型的所有随机变量和随机过程均是定义在该概率空间上的.

在经典模型中，均是从盈余入手对模型进行研究和讨论的，用 U 表示. 若不考虑除了保费和理赔之外的影响盈余的因素，如利率、附加费、分红及

再保险等等，则保险公司在某时刻的盈余由初始资金，加上保费收入，再除去索赔支出构成.

(1) 离散时间经典风险模型

时间为离散的，于是我们用 1，2，…，n 来表示时刻，则记保险公司在时刻 n 的盈余为

$$U(n)=u+cn-S(n) \tag{1.37}$$

其中：$S(n)=\sum_{k=1}^{N(n)}Y_k$ ，

假设：

①初始盈余 $U(0)=u\geqslant 0$.

②常数 $c>0$ 表示保险公司单位时间征收的保险费率.

③索赔额序列 $\{Y_k,k\geqslant 1\}$ 为独立同分布的非负随机变量序列，且 $E(Y_1)=\mu,var(Y_1)=\sigma^2$.

④索赔次数序列 $\{N(n),n\geqslant 0\}$ 为服从参数 λ 的齐次 Poisson 过程，则有：

$$E[N(n)]=Var[N(n)]=\lambda n.$$

定义保险公司的破产时刻为

$$N_u=\inf\{n:U(n)<0\}$$

保险公司的破产概率定义为

$$\Psi(u)=P\{\exists n>0,U(n)<0\,|\,U(0)=u\}=P\{N_u<+\infty\,|\,U(0)=u\}$$

(2) 连续时间经典风险模型

定义：$\{X_n,n\geqslant 1\}$ 为索赔时间间隔序列，且 $T_n=\sum_{k=1}^{n}X_k$ 为第 n 次索赔时刻，并约定 $T_0=0$.

$U(t)$ 表示保险公司在时刻 t 的盈余，以常数保费率 $c>0$ 收取保费，则保险公司在时刻 t 的盈余可以表示为

$$U(t)=u+ct-S(t) \tag{1.38}$$

其中：$S(t)$，$t\geqslant 0$ 表示到时刻 t 的索赔总额，且

$$S(t)=\sum_{i=1}^{N(t)}Y_i$$

假定：

① 初始盈余 $U(0)=u\geqslant 0$.

② $\{N(t),t\geqslant 0\}$ 是强度为 λ 的齐次 Poisson 过程，表示到时刻 t 的总索赔

次数，则 $E[N(t)]=Var[N(t)]=\lambda t$.

③ 索赔额序列 $\{Y_k,k\geqslant 1\}$ 和索赔额时间间隔序列 $\{X_n,n\geqslant 1\}$ 均为独立同分布的非负随机变量序列，其分布函数分别为 $G(x)=P\{X\leqslant x\}$ 和 $F(y)=P\{Y\leqslant y\}$，且有 $G(0)=F(0)=0,E(Y_1)=\mu,Var(Y_1)=\sigma^2$.

④ $\{X_i,i\geqslant 1\}$、$\{Y_k,k\geqslant 1\}$ 与 $\{N(t),t\geqslant 0\}$ 均相互独立.

索赔总额过程 $\{S(t),t\geqslant 0\}$ 为一个复合泊松过程.

同样的，保险公司的破产时刻和破产概率分别定义为

$$T_u=\inf\{t:U(t)<0\},\Psi(u)=P\{T_u<+\infty\}=P\{\bigcup_{t\geqslant 0}U(t)<0\}$$

在实际的经营过程中，保险公司为了保证自己运作上的安全，通常要求

$$E[ct-S(t)]=ct-E(S(t))=(c-\lambda\mu)t>0$$

令 $\theta=\frac{c}{\lambda\mu}-1$，即需满足 $\theta>0$，θ 称为相对安全负载（safety loading），且 $\theta\in(0,1)$.

1.2.2 时间序列数据

按时间次序排列的随机变量序列

$$X_1,X_2,\cdots \tag{1.39}$$

称为时间序列. 如果用

$$x_1,x_2,\cdots,x_N \tag{1.40}$$

分别表示随机变量 $X_1,X_2,\cdots,X_N$ 的观测值，就称（1.40）是时间序列（1.39）的 N 个观测样本. 这里 N 是观测样本的个数. 如果用

$$x_1,x_2,\cdots \tag{1.41}$$

表示 $X_1,X_2,\cdots$ 的依次观测值，就称（1.41）是（1.39）的一次实现或一条轨道.

时间序列数据是一种复杂的数据对象，在社会生活中的各个领域广泛存在着大量的时间序列数据有待进一步的分析和处理. 时间序列分析的主要任务就是根据观测数据的特点为数据建立尽可能合理的统计模型，然后利用模型的统计特性去解释数据的统计规律，以期达到控制或预报的目的.

（1）传统的时间序列模型

传统的时间序列模型处理的都是连续数据，例如某地区的月平均气温、股票日交易价格等. ARMA 模型是时间序列分析中应用最为广泛的一类模型. 然而，在近十几年里 ARCH 模型取得了极为迅速的发展，已被广泛地用

于验证金融理论中的规律描述以及金融市场的预测和决策. ARCH 模型作为一种度量金融时间序列数据波动性的有效工具，并应用于与波动性有关广泛研究领域，包括政策研究、理论命题检验、季节性分析等方面. 采用 ARCH 模型来模拟波动性，将会对期货交易制度设计，风险控制制度设计和投资组合风险管理策略研究，提供一个更为广阔的研究空间，几类模型的数学描述如下：

AR(p) 模型（Autoregressive Model）

如果 $\{\varepsilon_t\}$ 是白噪声 WN（0，σ^2），实数 a_1，a_2，…，a_p（$a_p \neq 0$）使得多项式 $A(z)$ 的零点都在单位圆外

$$A(z)=1-\sum_{j=1}^{p}a_j z^j \neq 0, |z| \leqslant 1$$

就称 p 阶差分方程

$$X_t = \sum_{j=1}^{p}a_j X_{t-j} + \varepsilon_t, t \in Z$$

是一个 p 阶自回归模型.

MA(q) 模型（Moving Average Model）

如果 $\{\varepsilon_t\}$ 是白噪声 WN（0，σ^2），实数 $b_1, b_2, \cdots, b_q (b_q \neq 0)$ 使得

$$B(z)=1+\sum_{j=1}^{q}b_j z^j \neq 0, |z| < 1$$

就称

$$X_t = \varepsilon_t + \sum_{j=1}^{q}b_j \varepsilon_{t-j}, t \in Z$$

是一个 q 阶滑动平均模型.

ARCH 模型（Autoregressive conditional heteroskedasticity model）：

$$X_t = \beta_0 + \beta_1 X_{t-1} + \cdots + \beta_p X_{t-p} + u_t$$

$$\sigma_t^2 = Eu_t^2 = \alpha_0 + \alpha_1 u_{t-1}^2 + \cdots + \alpha_q u_{t-q}^2$$

（2）整值时间序列数据

离散取非负整数值的时间序列数据（计数数据）在实际生活中相当普遍，比如：每月的失业人数、一个医院的每天住院病人数、一个车站的日客流量、未支付的信用分期付款（在信用评分中有用）、事故或者事故索赔次数（在确定保费时用）、住房抵押贷款中的预付费人数（在定价住房抵押贷款证券时用），因此统计学者对整值时间序列的研究越来越感兴趣，已成为统计学者的

研究热点问题.

整数值时间序列广泛地、大量地存在，而且彼此之间各有差异. 刻画这类数据的模型起源于 20 世纪 80 年代，由 Al－Osh and Alzaid (1987) 提出了整数值时间模型 (INAR 模型). Weiss, C. H. (2008) 综述了 2008 年之前整数值时间序列的研究成果. Kim and Zhang (2008), Zhang and Wang (2010) 提出了符号稀疏算子，基于此算子的 INAR 模型可以处理负整数值和负相关的数据，特别在处理非平稳的整数值数据时发挥了重要的作用. 近年来，整数值自回归模型再次引起人们的兴趣，Drost and van den Akker 等 (2009a, 2009b) 分别在 JRSS B 和 Bernolli 等统计学高档次杂志发表了最新研究成果. 上述模型属于稀疏算子 (thinning operator) 模型，即利用稀疏算子来构建模型. 这类模型的缺点是结构的状态空间模型，它常用 Poisson 分布来构造模型. 一个具有重大影响的模型是 Ferland (2006) 等提出的整数值广义自回归条件异方差 (Generalized Autoregressive Conditional Heteroscedastic, GARCH) 模型，它可以处理整数值数据中的异方差性，即假设整数值数据的条件分布是 Poisson 分布，条件均值 (称强度过程) 是历史观察值和历史强度的线性函数. 由于这个模型是处理金融数据非常成功的 GARCH 模型的整数值推广，从而一经出现就引起了广泛关注，并引发了一些相关研究. Fokianos and Rahbek (2009) 考虑了模型的遍历性和基于似然的参数估计量，Fokianos and Fried (2010) 研究了干预 (intervention) 效应问题，Weiss (2009) 建立了自协方差函数的递推关系，Weiss (2010a, 2010b) 给出边际分布和高阶矩，Neumann (2011) 给出了模型的混合性质和遍历性.

INAR 模型：

$$X_t=\alpha\circ X_{t-1}+\varepsilon_t,\ 0<\alpha<1,\ t=1,\ 2,\ \cdots$$

其中 $\alpha\circ X_{t-1}=\sum_{i=1}^{X_{t-1}}B_i$，$X_0$ 为非负随机变量，$\{\varepsilon_t\}$ 是 i. i. d. 取非负整数值的随机变量列，$\{B_i\}$ 是 i. i. d. 贝努力随机变量列，$P(B=1)=\alpha$.

INMA 模型：

$$X_t=\alpha\circ\varepsilon_{t-1}+\varepsilon_t,0<\alpha<1,t=1,2,\cdots$$

其中 $\alpha\circ\varepsilon_{t-1}=\sum_{i=1}^{\varepsilon_{t-1}}B_i$，$X_0$ 为非负随机变量，$\{\varepsilon_t\}$ 是 i. i. d. 取非负整数值的随机变量列，$\{B_i\}$ 是 i. i. d. 贝努力随机变量列，$P(B=1)=\alpha$.

1.3 微阵列数据

近年来，基于“微阵列”的新技术大量涌现、不断迅速发展，并逐步成为基因表达谱分析、新的生物学标志物发现、疾病机制的研究、疾病诊断、药物筛选等强有力的工具之一（Giltnane and Rimm，2004）. 这一类技术包括CDNA微阵列、高密度寡核苷酸阵列、蛋白阵列、组织阵列以及组织化学阵列等. 由于这些技术均基于高通量方法，很快就能产生庞大的高维生物学数据（Schena，et al. 1996）. 同时，随着分子生物学相关学科的迅猛发展，除了人类基因序列（Donham，et al. 1999），越来越多的动植物、微生物的全基因组序列也得以测定，基因序列数据正在迅猛增长（Abbott，1998）. 这些微阵列数据不同于传统的医学资料，具有其自身的特殊性，因而给传统的统计学分析带来了前所未有的挑战.

1.3.1 微阵列数据的统计方法

微阵列数据一般表示为 $m\times n$ 的矩阵形式，m 表示基因数，n 表示生物样本，通常只有几个或者几十个. 表1—1给出了一个关于前列腺癌的微阵列实验数据（Singh，et al.，2002），包括50个正常对照和52个前列腺癌患者6 033个基因的表达水平，以6 033×102的矩阵排列，即 $m=6\,033$，$n=102$. 研究者一个主要目的是，哪些基因在正常者和前列腺癌患者中的表达水平不同？需要识别出这些基因以进一步分析.

微阵列实验及其他高通量检测技术的兴起，无疑将成为本世纪的主流，微阵列实验主要的优势在于能同时大量地、全面性地检测上万个基因的表达量，通过基因芯片，可在短时间内找出可能受疾病影响的基因，作为早期诊断的生物标记. 然而，由于这一类技术的高度自动化、规模化及微型化的特性，使得他们所生成的数据量非常庞大且数据形态比一般实验数据更加复杂. 因此，传统统计分析方法已经不能适应. 与此同时，统计学家提出了非常多的新的统计理论和方法来分析微阵列实验数据，这些方法广泛地被生物学研究人员所使用. 微阵列数据常见的统计分析有以下几点：