高等学校本科生公共课教材

统计模拟及其R实现

肖枝洪 朱强 编著

WUHAN UNIVERSITY PRESS
武汉大学出版社

高等学校本科生公共课教材

统计模拟及其 R 实现

肖枝洪　朱　强　编著

武 汉 大 学 出 版 社

图书在版编目(CIP)数据

统计模拟及其R实现/肖枝洪,朱强编著.—武汉:武汉大学出版社,2010.4

高等学校本科生公共课教材

ISBN 978-7-307-07657-0

Ⅰ.统… Ⅱ.①肖… ②朱… Ⅲ.统计—模拟实验—高等学校—教材 Ⅳ.C8-33

中国版本图书馆CIP数据核字(2010)第043735号

责任编辑:杨 华　　责任校对:黄添生　　版式设计:马 佳

出版发行:**武汉大学出版社**　(430072 武昌 珞珈山)

(电子邮件:cbs22@whu.edu.cn 网址:www.wdp.com.cn)

印刷:湖北省京山德兴印务有限公司

开本:720×1000 1/16　印张:16.25　字数:288千字　插页:1

版次:2010年4月第1版　2010年4月第1次印刷

ISBN 978-7-307-07657-0/C·257　定价:26.00元

内 容 提 要

本书系统地介绍了统计模拟的一些实用方法和技术,同时也介绍了 R 语言及其编程方法. 在对条件期望、条件方差、Poisson 过程和 Markov 链的基本知识进行简单介绍之后,介绍了如何利用计算机产生随机数以及如何利用这些随机数产生任意分布的随机变量、随机过程等知识; 介绍了一些分析统计数据的方法和技术,如 Bootstrap、模拟精度改进技术等; 介绍了如何利用统计模拟来判断所选的随机模型是否拟合实际的数据; 介绍了处理缺失数据的 EM 算法和进行 Bayesian 统计推断的 MCMC 算法及一些新发展起来的统计模拟技术; 最后介绍了动态模型的模拟. 本书对每一章节中的例子,都给出了用 R 语言编写的模拟程序.

本书可作为统计学、计算数学与应用数学、保险学与管理学、精算学、工程技术等专业本科生教材或其他专业研究生教材,也可供相关专业人士参考.

前　言

在当今信息时代,统计知识得到越来越多的应用,各种统计方法层出不穷.在统计领域中,统计计算技术应运而生,而且发展非常迅速。它一方面使得许多统计方法得到广泛应用,另一方面使得有些难以在理论上进行论证的问题通过模拟得到证实.

2005 年以前,我国出版的有关统计模拟的著作还比较少,一般只是在理论上对统计计算方法进行介绍,而且使用 R 软件给出具体计算程序的更少。然而不过四五年,有关统计模拟和对 R 软件使用进行介绍的书籍开始大量出版.而我们适时出版这部书,也正是适应时代的需要.同时,本书也是省级教改项目“适应新时期人才培养的需要,构建概率统计类课程体系”和校级研究生创新工程项目“基于复杂数据的统计计算与模拟”的一项研究成果.提高大学生以及研究生进行数据挖掘、建立统计模型的能力是本项目改革的重点.因此,本书在写作上特别注意以下 6 个显著的特点:

(1) 统计计算方法全面,脉络分明:分别介绍了逆变量法、筛选法、条件期望法、分层抽样法、重要抽样法、EM 算法、MCMC 法等.既考虑了离散型随机变量的模拟,又考虑了连续型随机变量的模拟,同时还考虑了随机过程的模拟.不仅考虑了一维随机变量的模拟,而且考虑多维随机向量的模拟.

(2) 统计计算方法细腻详实,所编写的程序可以通过跟踪验证其准确性.

(3) 密切结合实际问题,例如经济中的期权策略实施、维修问题、排队问题等.

(4) 每章内容由浅入深,浅显易懂,但又能给人以更多的启示.同时还配有若干练习,帮助读者加强理解与巩固相关的知识.

(5) 本书第六章每提出一个模拟方法,都要和其他的方法相比较,尽量使算法更有效.通篇构成一个有机的整体,读者阅读此书,会感到逻辑性强,问题提出的背景清晰.

(6) 本书运用 R 语言将许多例子进行了编程,以供读者借鉴,使读者感觉有种工具在手的感觉,可以增强自信心,克服畏难心理.

本书的内容和练习取舍了 Sheldon M. Ross 的 *Simulation* 一书的内容和练

习,R 软件的介绍借鉴了李东风的《统计软件教程》,在此表示衷心的感谢. 本书的出版得到了武汉大学出版社、华中农业大学教务处和研究生处的大力支持,在此表示衷心的感谢! 本书适合大、中专院校从事统计或计算机方面的工作或学习的教师和本科生、研究生阅读. 章节的安排、内容的组织以及程序的编写,都是由肖枝洪和朱强两人共同商榷而定. 虽然我们在编写的过程中作了很大努力,力求准确,但是由于水平有限,所著之书可能存在许多错误,敬请读者批评指正.

编　者

2010 年 1 月于狮子山

目　　录

第 1 章
预备知识

我们设想读者已经具备了初等概率统计的基础知识,因此本章仅对初等概率统计中介绍得较少而在本书以后有关章节所需要的内容进行介绍.

1.1 矩母函数与生成函数

有时直接求若干独立随机变量和的分布很不方便,下面引进矩母函数和生成函数来解决此问题. 矩母函数和生成函数在第六章介绍重要抽样法时也会用到.

定义 1.1 设随机变量 X 的密度函数为 $p(x)$,称 $E[\exp(tX)]$ 为 X 的矩母函数,记为 $g_X(t)$,或简记为 $g(t)$,即

$$g(t) = E[\exp(tX)] = \int \exp(tx)\, p(x)\, \mathrm{d}x.$$

注: 矩母函数 $g(t)$ 与 X 的分布函数相互唯一确定,通过矩母函数可计算 X 的各阶矩:

$$E[X^n] = g^{(n)}(0),$$

式中,$g^{(n)}(t)$ 指函数 $g(t)$ 关于 t 的 n 阶导数. 若随机变量 X 与 Y 相互独立,则 $g_{X+Y}(t) = g_X(t)\,g_Y(t)$.

我们将一些常用分布的矩母函数列表如表 1–1.

表 1–1 一些常用分布的矩母函数

分布名称	密度函数	矩母函数
$U(a,b)$	$p(x) = \dfrac{1}{b-a},\ a < x < b$	$\dfrac{1}{t(b-a)}(\mathrm{e}^{bt} - \mathrm{e}^{at})$
$E(\lambda)$	$p(x) = \lambda \mathrm{e}^{-\lambda x},\ x > 0$	$\dfrac{\lambda}{\lambda - t},\ t < \lambda$

续表

分布名称	密度函数	矩母函数
$N(\mu,\sigma^2)$	$p(x)=\dfrac{1}{\sqrt{2\pi}\sigma}\exp\left(-\dfrac{(x-\mu)^2}{2\sigma^2}\right)$	$\exp(\mu t+\frac{1}{2}\sigma^2t^2)$
$\mathrm{Gamma}(\alpha,\lambda)$	$p(x;\alpha,\lambda)=\dfrac{\lambda^\alpha}{\Gamma(\alpha)}x^{\alpha-1}e^{\lambda x},\ x>0$	$\left(\dfrac{\lambda}{\lambda-t}\right)^\alpha,\ t<\lambda$

注: $\chi^2(n)$ 实质上为 $\mathrm{Gamma}\left(\frac{1}{2},\frac{n}{2}\right)$.

定义 1.2 若 X 为离散型随机变量,称 $E(s^X)$ 为其概率生成函数,记为 $\phi_X(s)$,或简记为 $\phi(s)$. 若 $P\{X=k\}=p_k,k=0,1,2,\cdots$,则

$$\phi_X(s)=\sum_{k=0}^{\infty}p_ks^k.$$

易知

$$p_0=\phi_X(0);$$

$$p_k=\frac{1}{k!}\frac{d^k}{\mathrm{d}s^k}\phi_X(s)\Big|_{s=0},\quad k=1,2,\cdots.$$

注: 生成函数 $\phi(s)$ 与 X 的分布函数相互唯一确定. 通过生成函数可计算 X 的各阶矩. 例如:

$$E[X]=\phi'(1),\quad E[X^2]=\phi''(1)+\phi'(1).$$

若随机变量 X 与 Y 相互独立,则 $\phi_{X+Y}(s)=\phi_X(s)\phi_Y(s)$. 我们将一些常用分布的生成函数列表如表 1–2。

表 1–2 **一些常用分布的生成函数**

分布名称	概率函数	生成函数
0–1 分布	$p(x)=p^x(1-p)^{1-x},x=0,1$	$1+(s-1)p$
二项分布 $B(n,p)$	$p(x)=C_n^xp^x(1-p)^{n-x},x=0,1,\cdots,n$	$(1+(s-1)p)^n$
几何分布 $G(p)$	$p(x)=p(1-p)^x,x=0,1,\cdots$	$\dfrac{p}{1-s(1-p)}$
Poisson 分布 $P(\lambda)$	$p(x)=\dfrac{\lambda^x}{x!}e^{-\lambda},x=0,1,\cdots$	$e^{\lambda(1-s)}$

1.2　条件期望和条件方差

1. 条件分布

如果 X 和 Y 是离散型随机变量,且 $P\{X=x,Y=y\}$ 为 X 和 Y 的联合概率函数,则可以计算在已知 $X=x$ 的条件下 Y 的条件分布,即若 X 的边缘概率函数 $p_X(x)>0$,则条件概率函数 $P\{Y=y \mid X=x\}=\dfrac{P\{Y=y,X=x\}}{P\{X=x\}}$. 如果 X 和 Y 是连续型随机变量,且 $p(x,y)$ 为 X 和 Y 的联合分布密度函数,则可以计算在已知 $X=x$ 的条件下 Y 的条件分布密度函数,即若 X 的边缘密度函数 $p_X(x)>0$,则条件分布密度函数为:

$$p(y \mid x)=\frac{p(x,y)}{p_X(x)}.$$

【例 1.1】　令 X 和 Y 的联合密度函数如下:

$$p(x,y)=\begin{cases} x+y, & 0 \leqslant x \leqslant 1, 0 \leqslant y \leqslant 1, \\ 0, & \text{否则}. \end{cases}$$

试求 $P\left\{X<\dfrac{1}{4} \mid Y=\dfrac{1}{3}\right\}$ 的值.

解　易知 $p_Y(y)=y+\dfrac{1}{2}$, 故

$$p(x \mid y)=\frac{p(x,y)}{p_Y(y)}=\frac{x+y}{y+\frac{1}{2}}.$$

所以,

$$P\left\{X<\frac{1}{4} \mid Y=\frac{1}{3}\right\}=\int_0^{\frac{1}{4}} p\left(x \mid \frac{1}{3}\right) \mathrm{d}x=\int_0^{\frac{1}{4}} \frac{x+\frac{1}{3}}{\frac{1}{3}+\frac{1}{2}} \mathrm{d}x=\frac{11}{80}.$$

2. 条件数学期望

由于条件数学期望在统计模拟中有着重要的地位,我们在此给出其定义.

如果 X 和 Y 是具有联合概率函数的离散随机变量,则在给定 $X=x$ 的条件下 Y 的条件期望 $E[Y \mid X=x]$ 为

$$E[Y \mid X=x]=\sum_y y P\{Y=y \mid X=x\}=\frac{\sum_y y P\{X=x,Y=y\}}{P\{X=x\}}.$$

即,给定 $X=x$ 的条件下 Y 的条件期望是以给定 $X=x$ 的条件下 Y 取值 y 的条件概率为权的加权平均值.

类似地,如果 X 和 Y 是具有联合密度函数 $f(x,y)$ 的连续随机变量,则在给定 $X=x$ 的条件下 Y 的条件期望 $E[Y\mid X=x]$ 为

$$E[Y\mid X=x]=\frac{\int y f(x,y)\,\mathrm{d}y}{\int f(x,y)\,\mathrm{d}y}.$$

由定义知 $E[Y\mid X]$ 为 X 的函数,且这个函数在 $X=x$ 时取值为 $E[Y\mid X=x]$. 同时 $E[Y\mid X]$ 它自身也是一个随机变量. 下述性质是非常有用的.

命题 1.1

$$E[E[Y\mid X]]=E[Y]. \tag{1.2.1}$$

如果 X 是一个离散型随机变量,则方程(1.2.1)可化为

$$E[Y]=\sum_x E[Y\mid X=x]P\{X=x\};$$

而如果 X 是一个具有密度函数 g 的连续型随机变量,则方程(1.2.1)可化为

$$E[Y]=\int E[Y\mid X=x]g(x)\,\mathrm{d}x.$$

我们在 X 和 Y 是离散型随机变量情形下给出命题 1.1 的证明.

证
$$\begin{aligned}\sum_x E[Y\mid X=x]P\{X=x\} &= \sum_x\sum_y yP\{Y=y\mid X=x\}P\{X=x\}\\ &= \sum_x\sum_y yP\{X=x,Y=y\}\\ &= \sum_y y\sum_x P\{X=x,Y=y\}\\ &= \sum_y yP\{Y=y\}\\ &= E[Y].\end{aligned}$$

【例 1.2】 随机变量 (X,Y) 的联合分布律如下表:

Y \ X	1	2	3	$p_{\cdot j}$
1	$\frac{2}{27}$	$\frac{4}{27}$	$\frac{1}{27}$	$\frac{7}{27}$
2	$\frac{5}{27}$	$\frac{7}{27}$	$\frac{3}{27}$	$\frac{15}{27}$
3	$\frac{1}{27}$	$\frac{2}{27}$	$\frac{2}{27}$	$\frac{5}{27}$
$p_{i\cdot}$	$\frac{8}{27}$	$\frac{13}{27}$	$\frac{6}{27}$	

试求 $E[X|Y]$ 的分布律, EX 及 $E(E[X|Y])$.

解　当 $Y=1$ 时,有

$$P\{X=1|Y=1\}=\frac{P\{X=1,Y=1\}}{P\{Y=1\}}=\frac{2}{7},$$

同理, $P\{X=2|Y=1\}=\frac{4}{7}$, $P\{X=3|Y=1\}=\frac{1}{7}$. 故

$$E[X|Y=1]=\sum_{i=1}^{3}iP\{X=i|Y=1\}=\frac{13}{7}.$$

类似地有

$$E[X|Y=2]=\sum_{i=1}^{3}iP\{X=i|Y=2\}=\frac{28}{15}.$$

$$E[X|Y=3]=\sum_{i=1}^{3}iP\{X=i|Y=3\}=\frac{11}{5}.$$

又 $P\{E[X|Y]=E[X|Y=j]\}=P\{Y=j\}$, $j=1,2,3$. 故 $E[X|Y]$ 的分布律如下:

$E[X\|Y=j]$	$\frac{13}{7}$	$\frac{28}{15}$	$\frac{11}{5}$
$P\{E[X\|Y]=E[X\|Y=j]\}=P\{Y=j\}$	$\frac{7}{27}$	$\frac{15}{27}$	$\frac{5}{27}$

$$E(E[X|Y])=\frac{13}{7}\times\frac{7}{27}+\frac{28}{15}\times\frac{15}{27}+\frac{11}{5}\times\frac{5}{27}=\frac{52}{27},$$

而

$$EX=\sum_{i=1}^{3}iP\{X=i\}=1\times\frac{8}{27}+2\times\frac{13}{27}+3\times\frac{6}{27}=\frac{52}{27},$$

即

$$EX=E(E[X|Y])=\frac{52}{27}.$$

3. 条件方差

给定 $X=x$ 的条件下 Y 的条件方差如下:

$$\mathrm{Var}(Y|X)=E[(Y-E[Y|X])^2|X].$$

同样, $\mathrm{Var}(Y|X)$ 也是 X 的函数,且这个函数在 $X=x$ 时取值为 $\mathrm{Var}(Y|X=x)$. 化简得到

$$\mathrm{Var}(Y|X)=E[Y^2|X]-(E[Y|X])^2.$$

对上式两边同时取期望得到:

$$E[\mathrm{Var}(Y|X)] = E[E[Y^2|X]] - E[(E[Y|X])^2]$$
$$= E[Y^2] - E[(E[Y|X])^2]. \qquad (1.2.2)$$

又因为 $E[(E[Y|X])] = E[Y]$，我们有

$$\mathrm{Var}(E[Y|X]) = E[(E[Y|X])^2] - (E[Y])^2. \qquad (1.2.3)$$

将式(1.2.2)与式(1.2.3)相加得到下面的等式——著名的**条件方差公式**:

$$\mathrm{Var}(Y) = E[\mathrm{Var}(Y|X)] + \mathrm{Var}(E[Y|X]).$$

【例 1.3】 从某大学中任意挑选一个学院，然后从此学院任意挑选 n 个学生. 令 X 表示这些学生中来自武汉市的人数. 令 Q 表示该学院来自于武汉市人数所占的比例，因为学院之间比例不同，所以 Q 也为随机变量. 给定 $Q=q$，则 $X \sim B(n,q)$，从而，$E[X|Q=q] = nq$，$\mathrm{Var}(X|Q=q) = nq(1-q)$. 假设随机变量 $Q \sim U(0,1)$，上述方式建立的模型称为**分层模型**，记为

$$Q \sim U(0,1),$$
$$X|Q=q \sim B(n,q).$$

则

$$E[X] = EE[X|Q] = E[nQ] = \frac{n}{2},$$
$$\mathrm{Var}(X) = E[\mathrm{Var}(X|Q)] + \mathrm{Var}(E[X|Q])$$
$$= E[nQ(1-Q)] + \mathrm{Var}(nQ) = \frac{n}{6} + \frac{n^2}{12}.$$

1.3 随机过程简介

以前我们学过的概率统计中随机变量序列基本上是独立同分布的. 本书中还将考虑相依的随机变量序列. 例如，日气温将形成以时间为序的随机变量序列，而且一天的气温明显地与前一天的气温不是独立的.

随机过程 $\{X_t, t \in T\}$ 是一个随机变量集合，**状态空间** S 是随机变量 X_t 取值的集合，集合 T 称为**指标集**，指标集可以是离散的，例如 $T=\{0,1,2,\cdots,\}$，也可以是连续的，例如 $T=[0,\infty)$，这取决于应用的需要. 对于指标集 T 为离散情形时的随机过程有时也称为随机变量序列.

【例 1.4】 (天气变化问题) 令状态集 $S=\{$晴，多云$\}$，指标集 $T=\{0,1,\cdots\}$，一个典型的序列为

晴，晴，多云，多云，晴，$\cdots$.

将上述序列用 X_t 依次记录下来，则 $\{X_t, t \in T\}$ 就是一个具有离散的状态空间 S 和离散的指标集 T 的随机过程.

下面,我们将讨论一类常用的随机过程.

1. Poisson **过程**

假设事件在 $[0,t]$ 上任意时刻发生且令 $N(t)$ 表示在时间 $[0,t]$ 上事件发生的个数. 如果

(1) $N(0)=0$;

(2) $N(t)$ 是独立增量过程,即任取 $0<t_1<t_2<\cdots<t_n$,

$$N(t_1),N(t_2)-N(t_1),\cdots,N(t_n)-N(t_{n-1})$$

相互独立;

(3) 对任何 $t>0,s\geqslant 0$, 增量 $N(s+t)-N(s)$ 服从参数为 λt 的 Poisson 分布,即

$$P\{N(s+t)-N(s)=k\}=\frac{(\lambda t)^k e^{-\lambda t}}{k!};$$

则称这些事件构成一个具有速率 $\lambda(\lambda>0)$ 的 Poisson 过程.

条件(1) 描述了过程开始于时刻 0. 条件(2),这个独立增量假设阐述了到时刻 t 发生的事件数($N(t)$)与在时刻 t 和 $t+s$ 之间发生的事件数($N(t+s)-N(t)$)相互独立. 条件(3),这个平稳增量假设阐述了 $N(t+s)-N(t)$ 的概率分布对任何 t 都是相同的.

对于一个 Poisson 过程,令 X_1 表示第一个事件来到的时刻. 进而对 $n>1$,令 X_n 表示第 $n-1$ 个事件和第 n 个事件之间的间隔时间. 序列 $\{X_n,n=1,2,\cdots\}$ 称为间隔时间序列. 例如,如果 $X_1=1,X_2=4$, 则 Poisson 过程的第一个事件将发生在时刻 1 和第二个事件将发生在时刻 4 之间(如图 1-1).

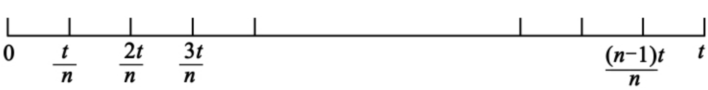

图 1-1　间隔时间

现在我们来确定 X_n 的分布. 首先注意到事件 $\{X_1>t\}$ 发生当且仅当 Poisson 过程在区间 $[0,t]$ 上没有一个事件发生,从而

$$P\{X_1>t\}=P\{N(t)=0\}=e^{-\lambda t}.$$

因此 X_1 是均值为 $\dfrac{1}{\lambda}$ 的指数分布. 为了获得 X_2 的分布,注意到

$$\begin{aligned}P\{X_2>t\mid X_1=s\}&=P\{\text{在}(s,s+t]\text{上没有事件发生}\mid X_1=s\}\\&=P\{\text{在}(s,s+t]\text{上没有事件发生}\}\\&=e^{-\lambda t}.\end{aligned}$$

其中最后的两个方程成立是因为独立增量性. 由前述讨论知, X_2 也是均值为 $\frac{1}{\lambda}$ 的指数随机变量,进而 X_2 和 X_1 相互独立. 重复这样的讨论,我们得到命题 1.2.

命题 1.2 设 $X_1, X_2, \cdots$ 为 Poisson 过程中事件发生的间隔时间序列,则 $X_1, X_2, \cdots$ 是独立同分布的随机变量序列,且共同分布为具有参数 λ 的指数分布.

令 $S_n = \sum_{i=1}^{n} X_i$ 表示第 n 个事件发生的时刻. 根据事实: S_n 不大于 t 当且仅当到时间 t 时至少有 n 个事件发生,我们看到

$$P\{S_n \leqslant t\} = P\{N(t) \geqslant n\} = \sum_{j=n}^{\infty} \mathrm{e}^{-\lambda t} \frac{(\lambda t)^j}{j!}.$$

易知上式左边是 S_n 的分布函数. 对它进行微分得到 S_n 的密度函数 $p_n(t)$ 如下:

$$\begin{aligned} p_n(t) &= \sum_{j=n}^{\infty} j\lambda \mathrm{e}^{-\lambda t} \frac{(\lambda t)^{j-1}}{j!} - \sum_{j=n}^{\infty} \lambda \mathrm{e}^{-\lambda t} \frac{(\lambda t)^j}{j!} \\ &= \sum_{j=n}^{\infty} \lambda \mathrm{e}^{-\lambda t} \frac{(\lambda t)^{j-1}}{(j-1)!} - \sum_{j=n}^{\infty} \lambda \mathrm{e}^{-\lambda t} \frac{(\lambda t)^j}{j!} \\ &= \lambda \mathrm{e}^{-\lambda t} \frac{(\lambda t)^{n-1}}{(n-1)!}. \end{aligned}$$

定理 1.1 设 $\{N(t), t \geqslant 0\}$ 是参数为 λ 的 Poisson 过程,则在给定 $N(t) = n$ 的条件下,事件的到达时刻 $(S_1, S_2, \cdots, S_n)$ 的联合分布密度为

$$f(s_1, s_2, \cdots, s_n \mid N(t) = n) = \frac{n!}{t^n}, \quad 0 < s_1 < s_2 < \cdots < s_n \leqslant t. \tag{1.3.1}$$

证明 给定 $N(t) = n$, 设 n 个事件的到达时刻分别为 $(S_1, S_2, \cdots, S_n)$. 对充分小的增量 $\Delta s_i, i = 1, 2, \cdots, n$, 事件 $N(t) = n$ 和在 $[s_i, s_i + \Delta s_i), i = 1, 2, \cdots, n$ 中恰发生一起事件而在 $[0, s_1), [s_1 + \Delta s_1, s_2), \cdots, [s_{n-1} + \Delta s_{n-1}, s_n), [s_n + \Delta s_n, t]$ 中没有事件发生等价.

$$\begin{aligned} & f(s_1, s_2, \cdots, s_n \mid N(t) = n) \Delta s_1 \Delta s_2 \cdots \Delta s_n \\ =\ & P\{s_i \leqslant S_i < s_i + \Delta s_i, i = 1, 2, \cdots, n \mid N(t) = n\} + o(\Delta s_1 \Delta s_2 \cdots \Delta s_n) \\ =\ & \frac{P\{s_i \leqslant S_i < s_i + \Delta s_i, i = 1, 2, \cdots, n, N(t) = n\}}{P\{N(t) = n\}} \\ & + o(\Delta s_1 \Delta s_2 \cdots \Delta s_n), \end{aligned} \tag{1.3.2}$$

而