21世纪保险精算系列教材

Actuarial Science

21世纪保险精算系列教材
精算师考试用书
中国人民大学风险管理与精算中心主编

Actuarial Models

精算模型（第二版）

肖争艳　编著

中国人民大学出版社

21世纪保险精算系列教材

Actuarial Science

21世纪保险精算系列教材
精算师考试用书
中国人民大学风险管理与精算中心主编

Actuarial Models

精算模型（第二版）

肖争艳　编著

中国人民大学出版社
· 北京 ·

图书在版编目（CIP）数据

精算模型/肖争艳编著. —2版. —北京：中国人民大学出版社，2015.7
21世纪保险精算系列教材
ISBN 978-7-300-21540-2

Ⅰ.①精… Ⅱ.①肖… Ⅲ.①保险-计算方法-教材 Ⅳ.①F840.4

中国版本图书馆CIP数据核字（2015）第148321号

21世纪保险精算系列教材
精算模型（第二版）
肖争艳 编著
Jingsuan Moxing

出版发行	中国人民大学出版社		
社　　址	北京中关村大街31号	**邮政编码**	100080
电　　话	010－62511242（总编室）		010－62511770（质管部）
	010－82501766（邮购部）		010－62514148（门市部）
	010－62515195（发行公司）		010－62515275（盗版举报）
网　　址	http://www.crup.com.cn		
	http://www.ttrnet.com（人大教研网）		
经　　销	新华书店		
印　　刷	北京鑫丰华彩印有限公司	**版　　次**	2013年1月第1版
规　　格	185 mm×260 mm　16开本		2015年7月第2版
印　　张	20.75 插页1	**印　　次**	2015年7月第1次印刷
字　　数	432 000	**定　　价**	36.00元

总　序

自1775年英国公平人寿最早将运用数学工具为产品定价的专门人员命名为精算师以来，精算师职业在国际上已有200多年的发展历史。这一职业最早在人寿和养老金业务中发挥作用，之后逐步向非寿险、社会保障等领域扩展。20世纪以后，精算师的职业进一步延伸到银行、投资、公司财务、金融工程等领域。精算师职业领域的扩展与精算职业组织的发展和精算教育水平的提高密切相关。1848年后欧美一些国家陆续成立的精算师协会以及国际精算师协会，为提高全球精算教育标准做出了贡献。例如，国际精算师协会早在1998年就公布了初级精算教育标准，要求2005年后加入国际精算师协会的成员在精算教育标准上符合国际教育标准。2007年，国际精算师协会再次公布了重新修订的初级精算教育标准及教育大纲。国际上著名的精算师职业组织，包括北美寿险精算师协会、北美非寿险精算师协会、英国精算师协会等，也从2000年后陆续对其精算教育标准和精算师考试体系进行改革，强调精算学与统计学、金融学、投资学、会计学、经济学等学科的融合，强调精算学科培养复合型风险管理人才的目标。

我国精算教育和精算师职业发展起步较晚，1992年后才陆续引入北美寿险精算师考试、英国精算师考试、日本精算师考试、北美非寿险精算师考试等，2000年后，中国精算师考试体系逐步建立起来。目前，中国精算师考试的考点已增加到15个。2006年12月，民政部批准中国精算师协会正式筹备成立。中国精算师协会的成立，必将进一步推动中国精算教育和精算师职业的发展，也迫切要求对当前的精算教育体系和精算师考试体系进行必要的改革，以尽快向国际精算师协会发布的精算教育标准看齐。

中国人民大学统计学院是国内较早开展风险管理与精算教育的大学学院之一。1992年就开始招收风险管理与精算专业方向的硕士研究生，1993年开始招收该方向的本科生，1996年招收了该专业方向的第一批博士研究生。2004年，经教育部批准备案，统计学院设立了独立的风险管理与精算学硕士学位点和博士

学位点，标志着在风险管理与精算人才培养上，形成了学士、硕士、博士多层次、专业化的人才培养教育体系。其专业课程设置完全与国际接轨，涵盖了北美、英国和中国精算师初级课程考试的基本内容，教学大纲紧跟国际精算师协会公布的精算教育指南，同时根据学科发展的国际趋势，每年重新修订课程和教学大纲。在研究方面，设立了中国人民大学风险管理与精算中心。多年来，在寿险风险管理与精算、非寿险特别是汽车保险风险管理与精算、养老金、社会保障等领域取得了很多有影响的成果，进一步促进了风险管理与精算教育的发展。为适应我国精算教育改革与发展的需要，并与国际精算师协会的精算教育标准接轨，中国人民大学风险管理与精算中心精心组织编写了一套精算学系列教材，分两个阶段完成。第一阶段涵盖精算师考试初级课程的全部专业课内容，包括《金融数学》、《风险理论》、《寿险精算学》、《非寿险精算学》、《精算中常用的统计模型》5 本教材和配套的学习辅导书，共 10 本。第二阶段涵盖精算师考试高级课程的全部内容，分寿险、非寿险、养老金、健康保险、社会保障、投资等不同系列。这套教材一方面可以满足各高校精算专业的教学需求，另一方面也可以作为参加各类精算师资格考试学员的学习参考资料，同时，也可以作为对精算学科有兴趣的同仁了解和学习精算的参考书。

这套教材的特点，一是在内容上涵盖了北美寿险、北美非寿险、英国、中国精算师考试最新的内容，同时紧跟国际精算师协会提出的精算教育标准，涵盖了国际精算教育大纲的基本内容；二是为了便于读者自学和教师讲授，我们为每本教材编写了学习辅导书，辅导书中包括学习要点、教材习题解答和一部分补充练习题及其解答等；三是在写法上，力求把精算学的数理理论与实务结合起来，注意精算学背后的实践意义，努力从实际意义上解释各种数学关系。

本套教材凝结了中国人民大学风险管理与精算中心全体教师的心血，特别是王晓军、孟生旺、黄向阳、王燕、肖争艳、肖宇谷等老师，他们为本套教材的编写付出了极大的艰辛，统计学院部分硕士研究生和本科生对辅导书中的习题解答进行了验证，感谢他们为本套教材做出的贡献，同时也感谢中国人民大学出版社的编辑们为本套教材的出版付出的辛勤劳动。

袁卫

前　言

保险的基本职能是分散风险，因此如何定义和测量风险是保险精算学的重要内容。精算模型是使用统计学和数学等研究工具，对保险赔付损失以及资金流入和流出一个保险系统的过程进行定量分析的学科。精算模型通过建立相关风险模型来研究保险风险的性质，并为现实的保险经营进行有效的风险分析和控制提供技术支持。本书旨在向读者阐述精算建模的过程，即如何从实际数据出发建立一个合适的精算模型。

编写本书的目的之一是满足精算专业的学生参加精算师资格考试的需求，所以在编写过程中参考了北美寿险精算师协会（SOA）的考试课程“Exam C-Construction and Evaluation of Actuarial Models”和中国精算师考试“A3 精算模型”的考试大纲，在内容的取舍上基本与 SOA 的考试大纲相符。全书大体可以分为三部分。第一部分是风险模型，介绍随机变量和风险度量的基本知识，讨论短期内单个保单的理赔额分布和保单组合的总理赔额分布，以及保险公司在长期内的破产概率。第二部分是模型的估计和选择，根据保险数据的特点，详细阐述精算建模的两种方法：经验模型法和参数模型法。第三部分是信度理论和随机模拟，研究如何合理利用先验信息和索赔经验对个体的未来损失进行预测，并介绍随机模拟技术及其在精算模型中的应用。

随着保险业的发展，市场竞争日趋激烈，精算师资格考试的内容也在不断更新，因此本书第一版的内容已经难以满足该门考试的要求。本书在第一版的基础上进行了如下几个方面的修改：增加了关于随机变量的尾部特征和大样本数据下死亡率的估计的内容，删除了多变量参数模型，大幅修改了随机模拟内容，补充了对应内容的习题，更正了原书中的错误。

阅读本书的读者需要具有概率统计学和高等数学的基本知识，还需要掌握使用 MATLAB，R 和 Excel 等统计软件进行计算的能力。但本书并不要求读者具有很好的保险知识背景，本书在首次出现保险术语之处都给出了定义。因此本书

也适合数学和统计专业的学生自学。为了方便读者自学，本书还设计了一定数量的例题和习题，给出了所有习题的解答过程。

本书也可作为精算模型和风险理论的本科课程教材，建议在大学本科三年级或四年级安排一个学期每周 3 个学时的教学。对于每周只有 2 个学时的课程（即 2 学分课程），可删减随机变量基础知识、长期聚合风险模型、信度理论等教学内容。本书提供了第 2～9 章的教学课件，欢迎采用本书的教师对这些课件进行修改，以适用于个人教学。

本书在编写过程中得到了许多同志的大力支持和帮助，凝结了大家的劳动成果。为本书修订做出贡献的有高荣、唐玥、邱子真、李铭章和孙文昭等，在此表示特别感谢。感谢中国人民大学统计学院应用统计专业（风险管理与精算方向）2010 级、2011 级、2012 级本科班学生，他们对本书提出了许多宝贵的修改意见，并仔细核对全书的例子和习题解答，使本书的内容得以完善。

在本书的修订过程中，虽然我们力求完善，但由于作者水平有限，书中的错误和疏漏之处在所难免，欢迎读者提出宝贵意见，作者将不胜感激。

本书的课件、习题的解答过程和答案可登录中国人民大学出版社工商管理分社的网站 www. rdjg. com. cn 查阅或下载。

肖争艳

目 录

第 1 章 Chapter 1 随机变量的基本知识

精算模型总是试图描述未来可能发生的随机性损失。损失的随机性有三层含义：损失发生的次数是随机的；损失发生的时间是随机的；损失的大小是随机的。对一份特定的保单而言，这三种随机性至少满足一种。而在概率论里，这三种随机性都能用相应的随机变量来刻画。为了后面叙述的方便，本章将对本书经常用到的随机变量的基本知识进行简要的概述。

1.1 概率空间、随机变量及分布函数

随机试验是概率论中的基本概念。事件可以看作随机试验的一种结果。试验的结果事先不能准确地预言，但具有三种特征：

(1) 可以在相同的条件下重复进行。

(2) 每次试验的结果不止一个，但预先知道试验所有可能的结果。

(3) 每次试验前不能确定哪个结果会出现。

记随机试验的基本结果为 ω，称为样本点。随机试验所有可能的结果组成的集合称为试验的样本空间，记为 Ω。Ω 中的样本点也称为基本事件，样本空间 Ω 称为必然事件，空集 $\varnothing$ 为不可能事件。Ω 的子集 A 由基本事件组成，通常称为事件。在实际问题中，人们一般不会对样本空间的所有子集感兴趣，而是关注某些事件及其发生的可能性大小。我们可用下面的概念来描述上述问题。

定义 1.1 设 Ω 是一个样本空间（或任意一集合），F 是 Ω 中某些子集组成的集合。如果满足：

(1) $\Omega \in F$；

(2) 若 $A \in F$，则 $A^c \in F$；

(3) 若 $A_n \in F(n=1, 2, \cdots)$，则

$$\bigcup_{n=1}^{\infty} A_n \in F$$

则称 F 为 σ 代数，(Ω, F) 为可测空间，F 中的元素称为事件。

如果 F 是事件的 σ 代数，则 $\varnothing \in F$；当 $A_n \in F(n \geqslant 1)$，$\bigcap_{n=1}^{\infty} A_n \in F$。

定义 1.2 设（Ω，F）是可测空间，$P(\cdot)$ 是定义在 F 上的实值函数。如果：

(1) 任意 $A \in F$，$P(A) \geqslant 0$；

(2) $P(\Omega)=1$；

(3) 对两两不相容的事件 A_1，A_2，…，A_n，…（即 $A_i \cap A_j = \varnothing$，$i \neq j$），有

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$$

则称 P 是（Ω，F）上的概率，（Ω，F，P）是概率空间，$P(A)$ 称为事件 A 的概率。

事件 A 的概率可理解为对事件 A 发生的可能性的度量。由定义可知，事件的概率具有如下性质：

(1) $P(\varnothing)=0$；

(2) $0 \leqslant P(A) \leqslant 1$；

(3) 若 $A \in F$，$B \in F$，且 $A \subset B$，则 $P(A) \leqslant P(B)$；

(4) 若 $A \in F$，$B \in F$，则 $P(A \cup B) = P(A) + P(B) - P(A \cap B)$；

(5) 若 $A_n \in F$（$n \geqslant 1$），则 $P\left(\bigcup_{i=1}^{\infty} A_i\right) \leqslant \sum_{i=1}^{\infty} P(A_i)$；

(6) 若 $A_n \in F$（$n \geqslant 1$），且 $A_n \subset A_{n+1}$（$n \geqslant 1$），则 $P\left(\bigcup_{n=1}^{\infty} A_n\right) = \lim_{n \to \infty} P(A_n)$；

(7) 若 $A_n \in F$（$n \geqslant 1$），且 $A_{n+1} \subset A_n$（$n \geqslant 1$），则 $P\left(\bigcap_{n=1}^{\infty} A_n\right) = \lim_{n \to \infty} P(A_n)$。

定义 1.3 设（Ω，F，P）是概率空间，X 是定义在 Ω 上的取值于实数域 R 的函数。如果对任意实数 $x \in R$，$(\omega | X(\omega) \leqslant x) \in F$，则称 $X(\omega)$ 是 F 上的随机变量，并称 $F_X(x) = P(X \leqslant x)$ 为 X 的累积分布函数，有时简称为分布函数。

分布函数 $F(x)$ 满足下列性质：

(1) $0 \leqslant F(x) \leqslant 1$；

(2) $F(x)$ 递增；

(3) $F(x)$ 右连续，即 $\lim_{x \to x_0^+} F(x) = F(x_0)$；

(4) $\lim_{x \to -\infty} F(x) = 0$，$\lim_{x \to +\infty} F(x) = 1$。

随机变量可以用来描述现实中我们感兴趣但又无法预先知道的事件，如被保险人的损失次数、索赔额和保险公司的盈余额。在概率论中，常用的随机变量有两种类型：离散型随机变量和连续型随机变量。离散型随机变量常用来描述被保险人的理赔次数的分布。离散型随机变量的概率特性除了用分布函数表示外，还可以用分布律来描述（$p_k = P(N=k)$，$k=0$，1，2，…，n）。例如，某投保汽

车险的驾驶员在一年之内发生索赔次数 N 的分布如下：

$$p_0=P(N=0)=0.8,\quad p_1=P(N=1)=0.1,\quad p_2=P(N=2)=0.075$$
$$p_3=P(N=3)=0.025,\quad p_k=P(N=k)=0,\quad k\geqslant 4$$

常见的离散分布有泊松分布、负二项分布、几何分布、二项分布等。我们将在本书附录 4 中列出这些分布的性质。

连续型随机变量常用来描述损失额的大小，连续分布的概率特征还可以用分布密度函数来刻画。

定义 1.4　设 (Ω, F, P) 是概率空间，X 是随机变量，$F_X(x)$ 是其分布函数，如果存在函数 $f(x)$，使得

$$F(x)=\int_{-\infty}^{x} f(t)\,\mathrm{d}t$$

则 $f(x)$ 称为连续型随机变量 X 的分布密度函数。

用来描述损失大小的随机变量 X 的分布密度函数通常具有下列特征：

(1) 非负性。损失额应该都是非负的，因此 $P(x\geqslant 0)=1$ 且 $x\leqslant 0$ 时，$f(x)=0$。

(2) 损失额应该是连续变化的，因而 $f(x)$ 是连续的。

(3) 损失额小的保险事故发生的可能性较大，损失额大的保险事故发生的可能性较小，但不可以忽略。从直观上说，损失额分布密度函数的尾部较厚（见图 1—1）。

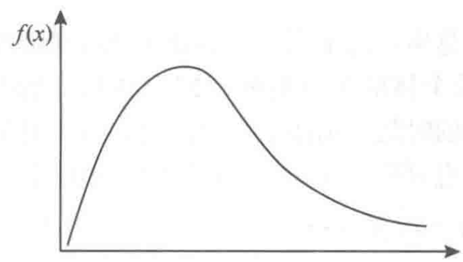

图 1—1　典型的损失额分布密度图

常用于描述损失的分布有指数分布、伽玛分布、对数正态分布、威布尔分布、帕累托分布等。本书附录 3 将列出一些常见连续分布的性质。

在实际问题中，对于某些随机试验的结果有时需要用多个随机变量来描述。例如，为了研究保险公司在一段时期内总理赔额的性质，不仅需要知道每次理赔额的大小，而且要知道这段时期内的理赔次数。随机向量 $\boldsymbol{X}=(X_1, X_2, \cdots, X_n)$ 的分布不仅与随机变量 $X_1, X_2, \cdots, X_n$ 的分布有关，还依赖于 $X_1, X_2, \cdots, X_n$ 之间的相互关系。

定义 1.5　对于 n 维随机向量 $\boldsymbol{X}=(X_1, X_2, \cdots, X_n)$，它的 n 维分布函数（或联合分布函数）定义为：

$$F(x_1,x_2,\cdots,x_n)=P(X_1\leqslant x_1,X_2\leqslant x_2,\cdots,X_n\leqslant x_n)$$

如果 $f(x_1, x_2, \cdots, x_n)=\dfrac{\partial^n F}{\partial x_1 \partial x_2 \cdots \partial x_n}$ 对所有（x_1，x_2，…，x_n）$\in R^n$ 都存在，则称 $\boldsymbol{X}=(X_1, X_2, \cdots, X_n)$ 是连续型随机向量，函数 $f(x_1, x_2, \cdots, x_n)$ 是 $\boldsymbol{X}=(X_1, X_2, \cdots, X_n)$ 的联合密度函数，并且

$$F(x_1,x_2,\cdots,x_n)=\int_{-\infty}^{x_1}\int_{-\infty}^{x_2}\cdots\int_{-\infty}^{x_n} f(t_1,t_2,\cdots,t_n)\mathrm{d}t_1\mathrm{d}t_2\cdots\mathrm{d}t_n$$

如果 $\boldsymbol{X}=(X_1, X_2, \cdots, X_n)$ 的所有可能取值是有限个 n 维向量或可列无穷个 n 维向量，则称 $\boldsymbol{X}=(X_1, X_2, \cdots, X_n)$ 是离散型随机向量，称 $P(X_1=x_1, X_2=x_2, \cdots, X_n=x_n)$ 为 n 维随机向量 $\boldsymbol{X}=(X_1, X_2, \cdots, X_n)$ 的联合分布列。

定义 1.6 设 $F(x_1, x_2, \cdots, x_n)$ 是 $\boldsymbol{X}=(X_1, X_2, \cdots, X_n)$ 的联合分布函数，则 X_1，X_2，…，X_n 的边际分布定义为：

$$\begin{aligned}&F_{k_1,k_2,\cdots,k_n}(x_{k_1},x_{k_2},\cdots,x_{k_n})\\&=F(\infty,\cdots,\infty,x_{k_1},\infty,\cdots,\infty,x_{k_2},\infty,\cdots,\infty,x_{k_n},\infty,\cdots,\infty)\end{aligned}$$

特别地，X_k 的边际分布为：

$$F_{X_k}(x_k)=F(\infty,\cdots,\infty,x_k,\infty,\cdots,\infty)$$

1.2 生存函数与危险率函数

在寿险精算中，通常用生存函数和危险率函数来描述个体的生存时间。设随机变量 X 表示个体的生存时间，即个体从初始时刻开始直至死亡、发生疾病、失效或者违约的时间。描述生存时间统计特征的基本函数是生存函数，它反映被观察个体在任意时刻 x（$x\geqslant 0$）仍然生存的概率，我们将其定义为：

$$S(x)=P(X>x) \tag{1—2—1}$$

当 X 为连续型随机变量时，生存函数与累积分布函数互补，即 $S(x)=1-F(x)$，这里 $F(x)=P(X\leqslant x)$，同时，生存函数也是概率密度函数 $f(x)$ 的积分，即 $S(x)=P(X>x)=\int_x^{\infty} f(y)\mathrm{d}y$，因此

$$f(x)=-\frac{\mathrm{d}S(x)}{\mathrm{d}x} \tag{1—2—2}$$

$S(x)$ 的图形叫做生存曲线，陡峭的生存曲线表示较低的生存概率或较短的生存时间，平缓的生存曲线表示较高的生存概率或较长的生存时间。

【例 1—1】

生存时间随机变量服从指数分布，其生存函数为 $S(t)=\mathrm{e}^{-\frac{1}{\theta}t}$（$t\geqslant 0$，$\theta>0$），图 1—2 为指数生存曲线。

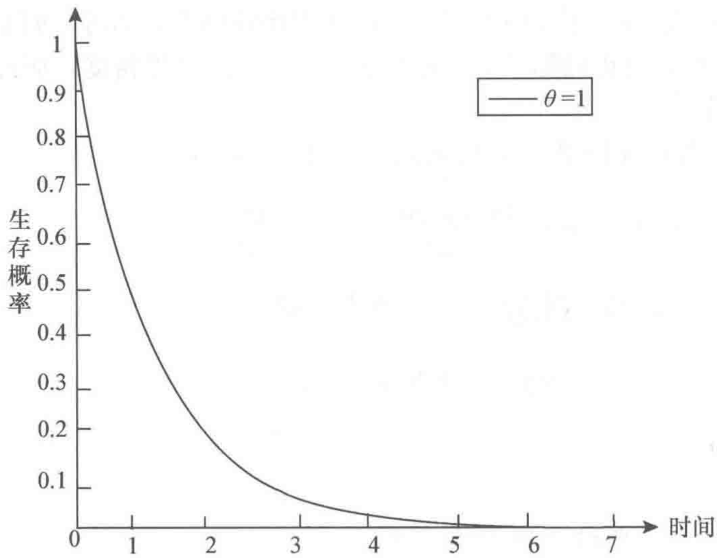

图 1—2　指数生存曲线

当 X 为离散型随机变量时，假设其概率分布函数 $p(x_j)=P(X=x_j)(j=1, 2, \cdots, n)$，其中 $x_1<x_2<\cdots<x_n$，那么 X 的生存函数为：

$$S(x)=P(X>x)=\sum_{x_j>x}p(x_j) \tag{1—2—3}$$

【例 1—2】

假设生存时间 X 服从离散均匀分布，概率分布函数为：

$$p(x_j)=P(X=j)=\frac{1}{2},\quad j=1,2$$

其生存函数为：

$$S(x)=P(X>x)=\sum_{x_j>x}p(x_j)=\begin{cases}1, & 0\leqslant x<1\\ 0.5, & 1\leqslant x<2\\ 0, & x\geqslant 2\end{cases}$$

危险率函数是生存分析中的另一个基本函数，它描述被观察个体在某时刻存活的条件下，在以后的单位时间内死亡的（条件）概率。危险率函数又称条件瞬时死亡率、死亡密度。在人口学中，它被称为死亡力，在可靠性研究中称为条件失效率。

危险率函数的定义为：

$$h(x)=\frac{f(x)}{S(x)} \tag{1—2—4}$$

显然，$h(x)$ 是在生存到时刻 x 的条件下的死亡密度。

我们注意到，式（1—2—2）和式（1—2—4）都是对个体在 x 时刻死亡密度

的瞬时度量，但 $f(x)$ 只需个体在初始时刻生存即可，而 $h(x)$ 需要个体在时刻 x 生存，这也是称 $f(x)$ 为时刻 x 死亡的无条件密度，而称 $h(x)$ 为条件密度的原因。

将式（1—2—2）代入式（1—2—4），有

$$h(x)=-\frac{\mathrm{d}S(x)/S(x)}{\mathrm{d}x}=-\frac{\mathrm{d}\ln S(x)}{\mathrm{d}x} \tag{1—2—5}$$

式（1—2—5）两边从 0 到 x 积分，得

$$\int_0^x h(y)\mathrm{d}y=-\ln S(x) \tag{1—2—6}$$

所以

$$S(x)=\mathrm{e}^{-\int_0^x h(y)\mathrm{d}y} \tag{1—2—7}$$

在实际应用中，还经常用到累积危险率函数，记为 $H(x)$，其定义为：

$$H(x)=\int_0^x h(y)\mathrm{d}y=-\ln S(x) \tag{1—2—8}$$

因此

$$S(x)=\mathrm{e}^{-H(x)} \tag{1—2—9}$$

与生存函数相比，危险率函数通常可以反映死亡（失效）机制更为详细的信息。因此，在概括性地描述生存数据时，危险率函数通常占据主导地位。

1.3 随机变量的数字特征

随机变量的概率分布（分布函数或密度函数）包含了随机变量的全部信息。但是在很多情况下，人们并不需要获得一个随机变量的全部信息，只需要了解它的某些特征值，如均值、最大值、最小值、与均值的离散程度、分布的对称情况。这些信息通常可以由随机变量的数字特征来反映。

定义 1.7 （1）如果 $\sum_{j=1}^{\infty}|x_j|^k p_j<\infty$，取值于 $\{x_j\}$ 的离散型随机变量 X 的 k 阶原点矩定义为：

$$\mu_k=\sum_j x_j^k P(X=x_j)$$

（2）连续型随机变量 X 的 k 阶原点矩定义为：

$$\mu_k=E(X^k)=\int_{-\infty}^{\infty}x^k\mathrm{d}F(x)=\int_{-\infty}^{\infty}x^k f(x)\mathrm{d}x$$

如果$\int_{-\infty}^{\infty}|x|^k \mathrm{d}F(x)<\infty$，则 $F(x)$ 为分布函数。

特别地，$\mu_1=E(X)$ 称为 X 的数学期望或均值，用来描述平均水平。在精算模型中，设每张保单的实际赔付额为 X，则均值 $E(X)$ 通常称为纯保费，它是保费定价的基础。

定义 1.8　(1) 如果 $\sum_{j=1}^{\infty}|x_j-\mu_1|^k p_j<\infty$，取值于 $\{x_j\}$ 的离散型随机变量 X 的 k 阶中心矩定义为：

$$\mu'_k=\sum_j (x_j-\mu_1)^k P(X=x_j)$$

(2) 连续型随机变量 X 的 k 阶中心矩定义为：

$$\mu'_k=E((X-\mu_1)^k)=\int_{-\infty}^{\infty}(x-\mu_1)^k \mathrm{d}F(x)=\int_{-\infty}^{\infty}(x-\mu_1)^k f(x)\mathrm{d}x$$

特别地，当 $k=2$ 时，μ'_2 称为 X 的方差，记为 $Var(X)$，通常用 σ^2 来表示。计算方差还有一个常用的公式：

$$\mu'_2=E(X-\mu_1)^2=E(X^2)-\mu_1^2$$

方差的平方根称为 X 的标准差，记为 $Std(X)$，通常用 σ 来表示。它与随机变量 X 具有相同的量纲。

除了均值和方差外，还经常使用以下指标来描述随机变量的分布特征：

(1) 变异系数 (coefficient of variation)：$r=\sigma/\mu_1$。

(2) 偏度：$r_3=\mu'_3/\sigma^3$。

(3) 峰度：$k=\mu'_4/\sigma^4$。

方差和标准差、变异系数都可以用来描述分布的离散程度。方差和标准差以均值为中心计算分布的离散程度，它们的优点是直观、容易理解。但是，一方面，其数值的大小取决于随机变量 X 本身水平的高低，也就是说与 X 的均值大小有关。X 的绝对水平高，离散程度的测度值自然也大；绝对水平低，离散程度的测度值自然也小。另一方面，采用不同计量单位计量的随机变量，其离散程度的测度值也不同。因此，用方差和标准差无法比较平均水平不同或计量单位不同的不同组别的随机变量与均值的离散程度。

变异系数则从相对的角度来观察差异和离散程度。由于变异系数等于标准差除以均值，分子和分母具有相同的量纲，因此消除了量纲的影响。在比较两个不同分布的差异程度时，变异系数比标准差更适合。变异系数越大，说明分布的离散程度越大；变异系数越小，说明分布的离散程度越小。

偏度和峰度这两个指标用来描述分布形状。偏度是对分布偏斜方向和程度的测度。当分布对称时，离差三次方后正负离差可以相互抵消，因此偏度 $r_3=0$。当分布不对称时，正负离差不能抵消，就形成了正的或负的偏态。若 $r_3>0$，说明正向偏差值较大，可以判断分布函数正偏或者右偏。若 $r_3<0$，说明负离差数

值较大，可判断为负偏或者左偏。r_3 的绝对值越大，说明偏斜程度越高。

峰度是对分布密度尖峰程度的测度。峰度的值通常是与标准正态分布相比较而言的。可以证明，如果随机变量服从标准正态分布，则峰度值为 3。若 $k>3$，说明分布比正态分布更尖，为尖峰分布；若 $k<3$，说明分布比正态分布更平，为平峰分布。

对于二维随机变量（X，Y），除了讨论 X 和 Y 的数学期望和方差，还需要讨论 X 和 Y 之间的相互关系的数字特征。

定义 1.9 称

$$E(X-E(X))E(Y-E(Y))$$

为随机变量 X 和 Y 的协方差，记为 $Cov(X, Y)$。称

$$\rho_{XY}=\frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}$$

为随机变量 X 和 Y 的相关系数。

从定义 1.9 可以推出协方差和相关系数具有如下简单性质：

(1) $Var(X+Y)=Var(X)+Var(Y)-2Cov(X,Y)$；

(2) $Cov(X,Y)=E(XY)-E(X)E(Y)$；

(3) $|\rho_{XY}|\leqslant 1$；

(4) $|\rho_{XY}|=1$ 的充要条件是，存在常数 a，b 使得 $P(Y=a+bX)=1$。

从性质（4）可以看出，ρ_{XY} 是一个可以用来表示 X，Y 之间线性关系紧密程度的量，当 $|\rho_{XY}|$ 较大时，我们通常说 X，Y 线性相关程度较高；当 $|\rho_{XY}|$ 较小时，我们通常说 X，Y 线性相关程度较低；当 $\rho_{XY}=0$ 时，则称 X 和 Y 不相关。

1.4 随机变量的矩母函数和母函数

定义 1.10 设随机变量 X 的分布函数为 $F(x)$，其特征函数、矩母函数和母函数分别定义为：

(1) 特征函数：$\varphi_X(t)=E(e^{itX})=\int_{-\infty}^{+\infty}e^{itx}dF(x)=\int_{-\infty}^{\infty}e^{itx}f(x)dx$（或者 $=\sum_k e^{itx_k}P(X=x_k)$）。

(2) 矩母函数：$M_X(t)=E(e^{tX})=\int_{-\infty}^{+\infty}e^{tx}dF(x)=\int_{-\infty}^{\infty}e^{tx}f(x)dx$（或者 $=\sum_k e^{tx_k}P(X=x_k)$）。

(3) 母函数：$P_X(t)=E(t^X)=\int_{-\infty}^{+\infty}t^x dF(x)=\int_{-\infty}^{\infty}t^x f(x)dx$（或者 $=\sum_k t^{x_k}P(X=x_k)$）。