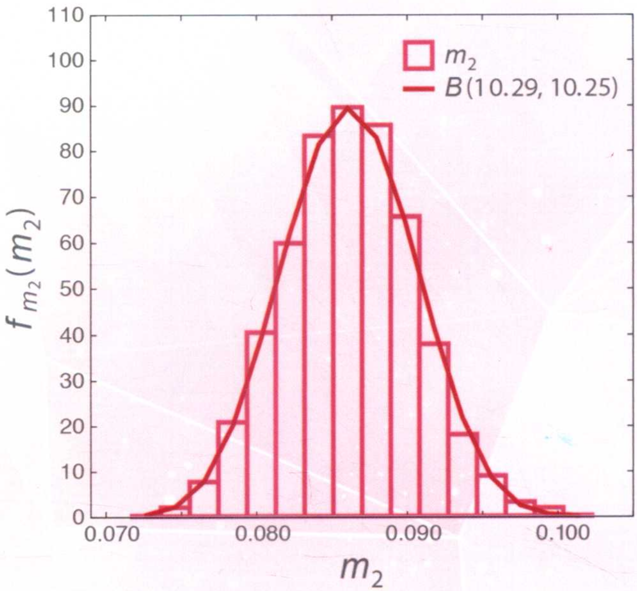

不确定性量化导论

INTRODUCTION TO UNCERTAINTY QUANTIFICATION

王鹏 修东滨 著

科学出版社

不确定性量化导论

王　鹏　修东滨　著

科学出版社
北　京

内 容 简 介

本书共六章，向读者较为全面地介绍了不确定性量化这一交叉研究领域的基本概念、常用方法和最新研究进展。第一章是全书的绪论；第二章回顾了概率与统计基础知识；第三章描述了随机系统的构建模拟；第四章论述了 PDF/CDF 方法；第五章阐述了当下最为常用的参数不确定性量化方法——广义多项式混沌法；第六章则基于数据同化这一概念向读者介绍了最新的模型不确定性量化方法。由于后三章的内容相对独立，读者可以选择单章节阅读。

本书是不确定性量化研究的基础参考书，可用作本领域高年级本科生或者研究生的入门教材。

图书在版编目（CIP）数据

不确定性量化导论/王鹏，修东滨著. —北京：科学出版社，2019. 1
ISBN 978-7-03-059472-3

Ⅰ. ①不… Ⅱ. ①王… ②修… Ⅲ. ①不确定系统-量化-研究
Ⅳ. ①N94

中国版本图书馆 CIP 数据核字（2018）第 255453 号

责任编辑：许 健／责任校对：谭宏宇
责任印制：黄晓鸣／封面设计：殷 靓

科学出版社 出版
北京东黄城根北街 16 号
邮政编码：100717
http://www.sciencep.com

苏州市越洋印刷有限公司 印刷
科学出版社发行 各地新华书店经销
*
2019 年 1 月第 一 版 开本：B5（720 × 1000）
2019 年 1 月第一次印刷 印张：6
字数：120 000

定价：80.00 元

（如有印装质量问题，我社负责调换）

前言

PREFACE

不确定性普遍存在于我们的世界之中。无论是基础科学研究，还是复杂工程系统设计，乃至我们的日常生活，都会受到大量不确定性因素的影响。这些不确定性可能来自于事物本身的随机性 (aleatory)，也可能来源于信息的缺失和误差。随着计算机仿真技术的提升和大数据时代的到来，科学计算与数据信息已经成为许多领域了解复杂系统的重要工具，其中不确定性的量化研究也成为当下最为活跃的新兴交叉课题之一。

本书旨在对不确定性量化 (uncertainty quantification，简称 UQ) 这门新兴学科做简要介绍，希望读者在阅读后能对不确定性的基础概念以及当下主要的不确定性量化方法有所了解。同时，笔者也想在此强调不确定性量化研究的初衷来源于人们对不确定性系统状态的预测，所以研究目标与研究方法同样重要，需要多学科的知识。我们希望读者在阅读过程中也适当回到各自所要研究问题的原点，并通过与不同学科背景的学者交流，最终根据目标系统，选择与开发最为合适的不确定性量化方法。此外，人工智能、大数据等新兴信息技术的迅速发展，对不确定性量化的发展也起到了促进作用。

本书可能无法涵盖当下所有的最新研究成果，笔者期待在以后的再版中补全，更希望能通过本书抛砖引玉，激发读者对不确定性研究的兴趣。

王　鹏

2018 年 5 月 30 日

目录 CONTENTS

第一章 绪论

不确定性普遍存在于我们的现实世界之中，它的量化研究对于基础科学原理的探索、工程产业升级乃至人类社会的稳定发展都具有重大意义。本章旨在向读者简要介绍不确定性量化 (uncertainty quantification, UQ) 这门新兴交叉学科。

1.1 研究背景

不确定性广泛存在于自然世界、工程系统与我们的社会生活之中。在微观物理世界，电子等微小粒子会受到环境中电场、磁场等各类噪声的扰动，呈现出多元反应，科学家通过对这些随机扰动进行量化分析，可以预测粒子的微观变化，进而探究与控制大量粒子的宏观物理特性。在现代电网控制管理中，风、光等可再生能源在时空尺度上往往呈现出强烈的波动性，严重影响了整个系统的稳定性，电力工程师们通过对此类不确定性的量化与精准的预测，可以协助电网系统的供需平衡，实现电网的智能化。在现代社会生活之中，互联网的普及与大数据时代给人们带来了前所未有的海量信息，如何过滤、减少其中的虚假信息 (噪声)，探知社会现象背后的规律，避免谣言的传播，也是人类社会生活共同面临的核心问题。

人类对不确定性的认知经历了漫长的过程。在距今约 2500 年前的春秋时期，伟大的东方思想家老子在《道德经》中提出 “知不知，尚矣；不知知，病也”，在认知层面上强调了不确定性的重要性。无独有偶，西方哲学史中最重要的哲学家之一大卫 · 休谟 (David Hume) 也强调“对不确定性的认知是人类知识的起点”。伴随着现代科学的发展，20 世纪著名科学家诺伯特 · 维纳 (Norbert Wiener)、范坎彭 (N.

G. van Kampen) 也先后在科学层面上指出大部分系统充满了不确定性 [1, 2]。

1.2 发展状况

近年来伴随着计算机性能的飞速提升，科学计算已经成为许多领域了解复杂系统的主要工具，国际学术界与工程界涌现出“不确定性量化”这一新兴交叉研究领域。它以数学与统计方法为基础，涵盖概率、微分方程、逼近论、图论与网络理论、遍历理论、测度论、随机过程、时间序列、贝叶斯分析、重要性抽样、非参数技术、多元分析等多个研究方向，运用计算机实现数学算法，对现实复杂系统进行模拟、预测。因此，不确定性量化研究不仅需要数学与统计方法的创新，也需要计算机技术的进步，更需要具体应用研究领域专业知识的发展。美国国家研究委员会在 2013 年出版的《2025 年的数学科学》中也强调了“不确定性量化是一门交叉性极强的综合性学科”[3]。

通过近年来的发展，不确定性量化已成为国际学术界最为活跃的前沿交叉学科之一。鉴于传统的单一学科会议与期刊已无法满足其学术交流，美国 Begell 出版集团、美国工业与应用数学学会 (Soceity of Industrial and Applied Mathematics, SIAM) 和美国统计学会 (American Statistical Association, ASA) 先后于 2011 年、2014 年创办了针对不确定性量化研究的专业学术期刊：*International Journal for Uncertainty Quantification* 与 *SIAM/ASA Journal for Uncertainty Quantification*；美国工业与应用数学学会于 2012 年也开创了两年一届的“SIAM Uncetainty Quantification”专业会议，其 2018 年 4 月的第四届年会吸引了全球约 800 名专业人士参会。目前，美国能源部 (Department of Energy，DOE) 的六个国家实验室 (洛斯阿拉莫斯、桑迪亚、劳伦斯伯克利、橡树岭、阿贡、太平洋西北) 也成立了跨专业、跨部门的不确定性量化团队。

鉴于不确定性量化对国家科技实力和经济发展的重大影响，美国政府非常重视这一学科的基础研究。美国能源部、国防部先进研究项目局 (Defense Advanced Research Projects Agency，DARPA)、国家自然科学基金委员会 (National Science Foundation，NSF)、国家核安全局 (National Nuclear Safety Administrcian，NNSA) 近八年来先后投入巨资 (表 1.1)，在大气污染防治、环境保护、核材料埋存、电路设计、飞机研发、智能电网等领域开展了基础研究。

表 1.1 美国政府对不确定性量化研究的部分资助列表

资助单位	能源部	自然科学基金委员会	国家核安全局	国防部先进研究项目局
项目名称	SciDAC	SAMSI	PSAAP	EQUiP
资助金额	\$10 000 000	\$10 000 000	\$20 000 000	\$30 000 000

美国产业界也对不确定性量化的应用研究异常重视。在航空发动机这一高精尖领域，制造商为了减少研发成本，往往需要通过数值运算来模拟产品或部件的真实运作过程，进而实现其优化设计。例如劳斯莱斯 250-C20R 涡轮轴发动机，其叶片形状需要九个几何参数来进行量化。但是，昂贵且漫长的数值实验使得工程师们无法通过穷举的方式，对这些几何参数的各种组合进行性能分析。如何通过有限的样本，实现叶片性能的最优化设计，是劳斯莱斯 (Rolls-Royce) 不确定性量化团队的核心目标 [4] (图 1.1)。在石油勘探过程中，人们往往对地下状况一无所知，动辄百万美元的取样成本迫使雪佛龙 (Chevron)、美孚 (Mobil) 等石油公司的研发部门必须通过有限的点信息对整个区域的地质情况进行精准的建模 (图 1.2)。

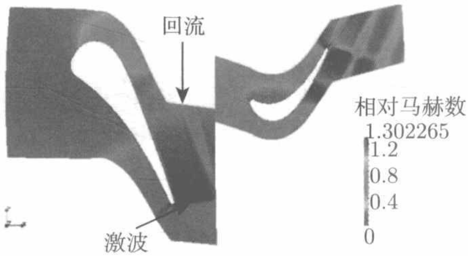

图 1.1 飞机发动机叶片设计的数值模拟 [4]

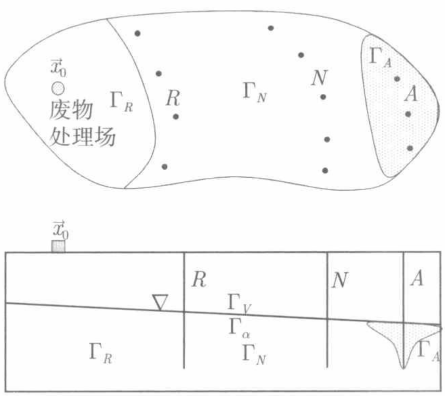

图 1.2 地质勘探中的区域建模 [5]

在具体研究中，不确定性量化主要考虑系统参数、模型与计算这三种不确定因素对系统状态的影响 (图 1.3)。

1) 参数不确定性　系统状态往往受到诸多参数的影响，而这些参数所呈现出的时空波动性 (异质性)，需要大量样本才能准确描述。鉴于取样技术、取样成本等实际操作因素所限，大多数情况下我们只有少量的参数样本数据。即使在大数据时代，随着取样技术的进步和数据成本的降低，人们得到的海量数据在测量、传输、读取过程中，依然会或多或少地受到各类随机因素的干扰，而这些干扰的叠加与放大势必会对最终系统状态造成影响。所以，我们可以减少参数 (数据) 不确定性，但是无法完全消除它。

2) 模型不确定性　数学模型是对现实规律的近似。但是，鉴于人类认知的局限性与差异性，同一规律或者物理现象往往会有不同的理解、不同的简化与不同的模型，而现实规律永远存在，它并不随人类的认知变化而改变。所以，如何选取与组合不同的数学模型，对现实系统状态进行有效的预测，是模型不确定性量化的目标。

3) 计算不确定性　在确认模型并获取相关参数后，人们可以通过数学模型的计算来模拟现实系统。由于大多数模型是连续的数学方程，在实际数值计算中，需要离散化，将微分运算转为差值运算、积分预算转为加值预算，这些离散过程势必会引入误差不确定性，进而影响系统状态的数值拟合。

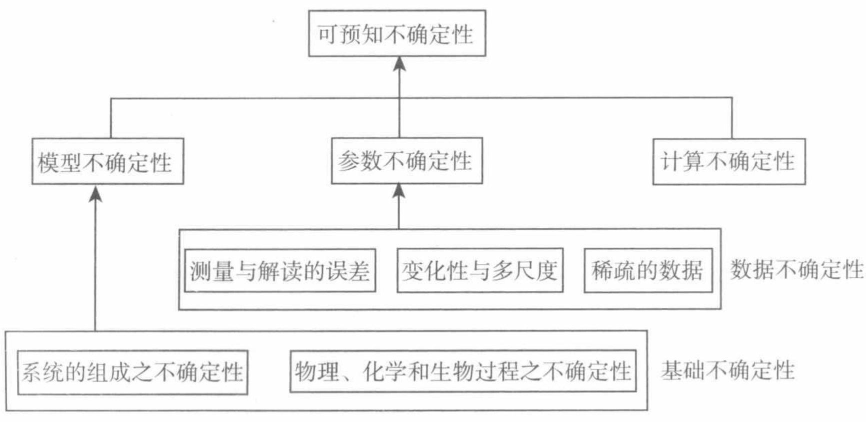

图 1.3　不确定性量化的研究对象

举例来说，在航空发动机的工业设计中，工程师们需要对气体流动、叶片转动、

燃料燃烧等过程进行建模，从而对产品的实际性能进行分析。在数值模拟中，气体压力、流速、叶片材料强度与熔点、燃料密度等大量参数会引入参数不确定性，流体力学、固体力学、燃烧学等多组数学方程也有各自不同的简化 (模型不确定性) 与数值算法的近似 (计算不确定性)。这些不确定性因素相互叠加、相互放大，会严重影响设计人员对整个系统实际性能的考量。虽然当下计算机科学发展迅猛，人工智能技术也呈现爆发式增长，但是很多基础科学问题，例如描述流体运动的纳维斯 · 斯托克斯 (Navier Stokes) 方程，在提出百年之后，其复杂性依然很难随着计算机硬件条件的提升而得到完全解决。

当前不确定性量化的研究热点是参数不确定性。在实际研究中，科研工作者通常将不确定参数等同于随机变量或随机过程，$Z(\mathbf{x},t) \equiv Z(\mathbf{x},t;\omega)$，即参数不仅在时空间 $(\mathbf{x},t)$ 变化，同时也在概率空间 ω 变化。而含有这些随机参数的原确定系统随之变为随机系统，它的解是原系统 (状态) 输出的统计信息 (概率密度函数或概率分布函数)。相对于经典随机微分方程，上述系统的随机输入不再是维纳过程、泊松过程等理想化过程，而来源于实际测量数据，呈现出一定的时空关联性 [6–9]。因此，随机微积分等经典的随机分析数学方法并不直接适用于此类随机系统。

现有的参数不确定性量化方法可以分为统计型 (如蒙特卡罗) 和随机数学型 (如随机有限元、统计矩微分方程、摄动法、PDF/CDF 方法等) 两大种。

(1) 基于蒙特卡罗方法

根据系统参数的人们概率分布，产生一组相互独立的数据作为系统输入；通过数值模拟，可以得到相应的输出，进而提取这个集合的统计信息，例如系统的均值和方差。蒙特卡罗模拟 (Monte Carlo simulation，MCS) 是一种对原系统进行重复性的数值模拟，将相应的输出结果进行统计的方法。它可以考虑高维度随机变量，简单易用。但是，其收敛比较慢，需要大量的输出样本，进而提高了计算成本，同时，它无法提供系统状态的随机变化规律。基于上述原因，蒙塔卡罗方法也往往被冠以 “简单粗暴” 的标签。近年来此类方法的技术发展主要针对收敛速度的提升，例如拉丁超立方采样 (参见文献 [10, 11])、准蒙特卡洛采样 (参见文献 [12-14]) 等方法。遗憾的是，这些新的改进也带来了额外的限制条件，影响了它们的适用性。

(2) 扰动法

扰动法 (perturbation methods) 将随机参数在其均值附近进行有限的泰勒展开，是最受欢迎的非抽样类方法。由于方程系统的二阶项后会变得较为复杂，大多数情况的展开会截至二阶项。此类方法虽然已被广泛应用在诸多工程领域 [15−17]，但是无法考虑过多的不确定参数和系统状态 (即输入和输出的总维度通常小于 10)。

(3) 算子法

算子法 (operator-based methods) 包含诺伊曼展开 [18, 19] 和加权积分 [20, 21]，都是基于系统控制方程随机算子的方法。此类方法类似于扰动法，无法考虑过高的随机维度 (随机输入与输出数量)，且强烈依赖于控制方程的算子，故较为适用于稳态问题。

(4) 统计矩方程

统计矩方程 (moment equations) 的核心目标是计算系统状态的统计矩。通过对原系统方程进行随机平均，可以推导出系统状态各阶统计矩的确定方程。但是在推导过程中，低阶统计矩方程的求解往往需要高阶统计矩的信息，进而引入了闭包问题，必须根据具体的问题来选取特殊的近似方法来解决。

(5) PDF/CDF 方法

PDF/CDF 方法起源于统计物理 [22−24]，通过引入系统状态的精细概率密度函数 (PDF) 或者精细概率分布函数 (CDF)，推导出系统状态的 PDF/CDF 方程，在流体力学中得到了广泛应用 [25]。近年来随着数值算法框架的改进，人们也可以通过计算系统状态的精细概率密度函数或精细概率分布函数方程 [26−30]，以统计的形式求解系统状态的分布信息。

(6) 广义多项式混沌法

广义多项式混沌法 (generalized polynomial chaos，gPC) [31] 是经典多项式混沌方法的泛化 [32]，也是应用最广泛和使用最多的一种方法 [33]。它的核心是将随机解表示为随机输入的正交多项式，在本质上是将随机空间以扰动的形式变现出来，具有很好的收敛速度。

综上所述，蒙特卡罗方法作为一种简单易行的方法被欧洲学者广为推崇；PDF/CDF 方法作为新兴的量化框架正在逐渐被广大科研工作者接受；广义多项式混沌法是当下美国学界最为活跃的研究课题，也是本书的重点。虽然广义混沌多项式方法的快速发展使得本书无法涵盖所有方向与应用，但是笔者希望通过对这一领域重点文献的总结，帮助读者了解行业内的热门领域，为读者在今后科研中的文献检索等工作奠定基础。

1.3 本书结构

本书面对的读者应具有线性代数、微分方程和概率的基础知识。笔者旨在通过简单介绍不确定性量化这一新兴交叉学科，帮助读者迅速掌握其理论基础。本书的组织形式适用于国内大学本科高年级或研究生一个学期的学习，章节具体结构如下：第 2 章回顾了概率与统计基础知识；第 3 章介绍随机系统的构建模拟，如随机输入信息 (参数化) 及所得到的随机输出信息；第 4 章介绍 PDF/CDF 方法；第 5 章介绍广义多项式混沌法，及其主要数值实现 —— 随机伽辽金法和随机配点法的基本理论要点；第 6 章简单介绍数据同化。

第二章 概率与统计基础知识

本章将回顾随机计算所需的概率与统计基础知识。

2.1 单元随机变量

不确定性广泛存在于世界之中，使得很多事件结果呈现出随机化。在概率上，我们用数字来表示每一个可能发生事件的结果。例如扔骰子的结果 X 可以由 $X(\omega) \in \{1,2,3,4,5,6\}$ 来表示。其中，ω 是随机事件 (扔骰子) 的可能结果，存在于全部可能结果空间 Ω，而随机变量 $X = X(\omega)$ 是一个定义在 Ω 空间上的真值函数。为了描述某个随机事件，我们首先需要将 Ω 空间中的子集归类 $\mathcal{F}$，也称之为 $\sigma-$域或 $\sigma-$代数。

定义2.1 一个 $\sigma-$ 域的 $\mathcal{F}$ 是 Ω 子集的集合，且满足以下条件:

1) $\mathcal{F}$ 不是空集: $\emptyset \in \mathcal{F}$ 且 $\Omega \in \mathcal{F}$;

2) 如果 $A \in \mathcal{F}$，那么 $A^c \in \mathcal{F}$;

3) 如果 $A_1, A_2, \cdots, \in \mathcal{F}$，那么

$$\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}, \quad \bigcap_{i=1}^{\infty} A_i \in \mathcal{F}$$

回到扔骰子的例子中，当我们考虑某特定结果 (如点数为 3) 时，不仅事件 $\{\omega : X(\omega) = 3\}$ 属于 $\mathcal{F}$; $\{\omega : 2 < X(\omega) < 5\}$, $\{\omega : X(\omega) \geqslant 5\}$, $\{\omega : X(\omega) \leqslant 2\}$ 以及更多的相关事件都属于 $\mathcal{F}$。

根据定义，我们可以看到最小的 $\sigma-$ 域是 $\mathcal{F}_1 = \{\emptyset, \Omega\}$，最大 $\sigma-$ 域 (也称为 Ω 的幂集) 是 $\mathcal{F}_2 = 2^{\Omega} \triangleq \{A : A \subset \Omega\}$，其中 $A \neq \emptyset, A \neq \Omega$。对于特定的子集 $\mathcal{C}$,

也存在着一个最小的 $\sigma-$ 域，$\sigma(\mathcal{C})$，即 $\mathcal{C}$ 的生成 $\sigma-$域。这样的话，我们可以得到 $\mathcal{F}_1=\sigma(\{\varnothing\})$，$\mathcal{F}_2=\sigma(\mathcal{F}_2)$。

综上所述，$\mathcal{F}$ 的基本操作如 $\bigcup$、$\bigcap$，$\mathcal{C}$ 所带来的结果不应超出 $\mathcal{F}$ 的范畴，这也是 $\sigma-$ 域 $\mathcal{F}$ 的直观意义。

2.1.1　概率

概率是辅助我们度量事件发生可能的重要工具。例如，在上述扔骰子事件中，如果每次投放骰子的情况相同，那么在无穷次重复试验后，我们会发现每个骰子面的可能性 (概率) 皆为 1/6，即

$$P(\{\omega:X(\omega)=1\})=\cdots=P(\{\omega:X(\omega)=6\})=\frac{1}{6}$$

概率论中的大数定律也告诉我们：在试验不变的条件下，特定随机事件发生的频率将随着重复试验次数的增加而接近于它的真实概率。

定义2.2 (概率空间)　概率空间由样本空间 Ω、$\sigma-$域 $\mathcal{F}\subset 2^{\Omega}$ 和概率测度 P 共同构成，通常用 $(\Omega,\mathcal{F},P)$ 来表示，且满足下列条件：

1) $0\leqslant P(A)\leqslant 1,\forall A\in\mathcal{F}$;

2) $P(\Omega)=1$;

3) 对于事件 $A_1,A_2,\cdots\in\mathcal{F}$ 且 $A_i\bigcap A_j=\varnothing,\forall i\neq j$,

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

4) 对于事件 $A,B\in\mathcal{F}$

$$P(A\cup B)=P(A)+P(B)-P(A\cap B)$$
$$P(A^c=1-P(A)),\quad P(\Omega)=1,\quad P(\varnothing)=0$$

2.1.2　分布

首先引入随机事件的概率分布函数 F_X 和概率分布 P_X 的概念。

定义2.3 (分布函数)　随机事件 X 的分布函数 F_X 是以下概率的集合：

$$F_X(x)=P(X\leqslant x)=P(\{w:X(w)\leqslant x\}),x\in\mathbb{R}\tag{2.1}$$

通过上述定义，不仅可以计算系统结果在某个区间 $(a,b]$ 上的概率：$P\{w : a < x(w) \leqslant b\} = F_X(b) - F_X(a)$；也可以得到某特定数值的概率：$P(X = x) = F_X(x) - \lim_{\epsilon\to 0} F_X(x-\epsilon)$。

对于结果空间中某些事件的集合 $B \subset \mathbb{R}$，我们可以通过计算得到其发生概率。

定义2.4 (分布) 随机事件 X 的分布 P_X 是某些事件集 $\omega : X(\omega) \in B$ 发生概率的集合

$$P_X(B) = P(X \in B) = P(\{\omega : X(\omega \in B)\})$$

其中，$\mathbb{R}$ 的合适子集也叫做博雷尔集 (Borel set)，也称为博雷尔 $\sigma-$ 域：$\sigma(\{(a,b] : -\infty < a < b < \infty\})$。

定义2.5 (离散随机变量) 当随机事件的概率分布函数出现跳跃时，

$$F_X(x) = \sum_{x_k \leqslant x} p_k, \quad x \in \mathbb{R} \tag{2.2}$$

我们将相应的随机变量称之为离散随机变量。

离散随机变量的取值是有限的 $x_1, x_2, \cdots$，其概率可以表示为：$p_k = P(X = x_k)$。下边是两种常见的离散分布，它们的概率密度分布如图 2.1 所示。

1) 离散平均分布 $\mathcal{U}(a,b)$:

$$P(X = k) = \frac{1}{b-a+1}, \quad k \in \{a, a+1, \cdots, b-1, b\}$$

2) 泊松分布 $P(\lambda)$, $\lambda > 0$:

$$P(X = k) = e^{-\lambda}\frac{\lambda^k}{k!}, \quad k = 0, 1, \cdots$$

定义2.6 (连续随机变量) 当随机事件的概率分布函数没有任何间断时，

$$\lim_{\epsilon\to 0} F_X(x+\epsilon) = F_X(x), \quad \forall x \tag{2.3}$$

我们将相应的随机变量称之为联系随机变量。

连续随机变量在结果空间上任意一点的概率为 0，且大多数连续分布可以用相应的概率密度 f_X 来描述：

$$F_X(x) = \int_{-\infty}^{x} f_X(y)\mathrm{d}y, \quad x \in \mathbb{R} \tag{2.4}$$

其中 f_X 满足以下性质：

$$f_X(x) \geqslant 0, \quad \int_{-\infty}^{\infty} f_X(y)\mathrm{d}y = 1, \quad \forall x \in \mathbb{R}$$

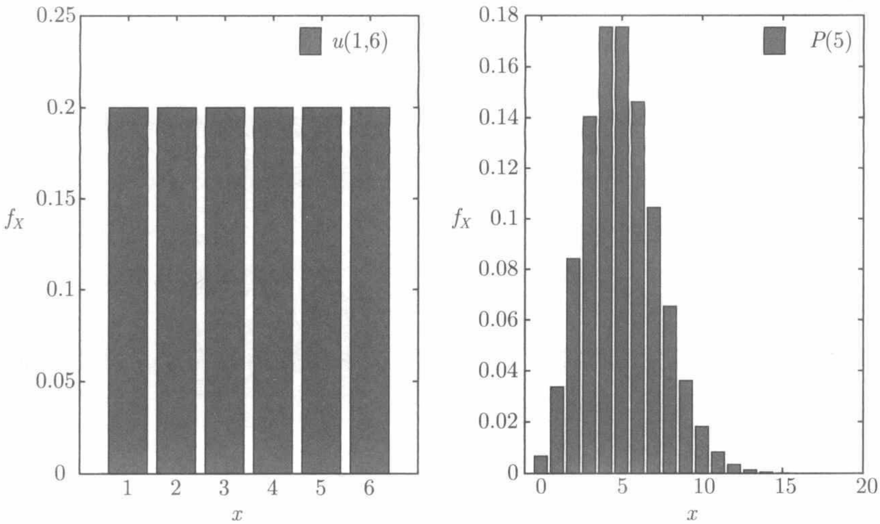

图 2.1　离散平均分布 $X \sim \mathcal{U}(1,6)$ 与泊松分布的概率密度函数 $X \sim P(5)$

下边我们介绍两种常见的连续分布，它们的概率分布函数如图 2.2 所示。

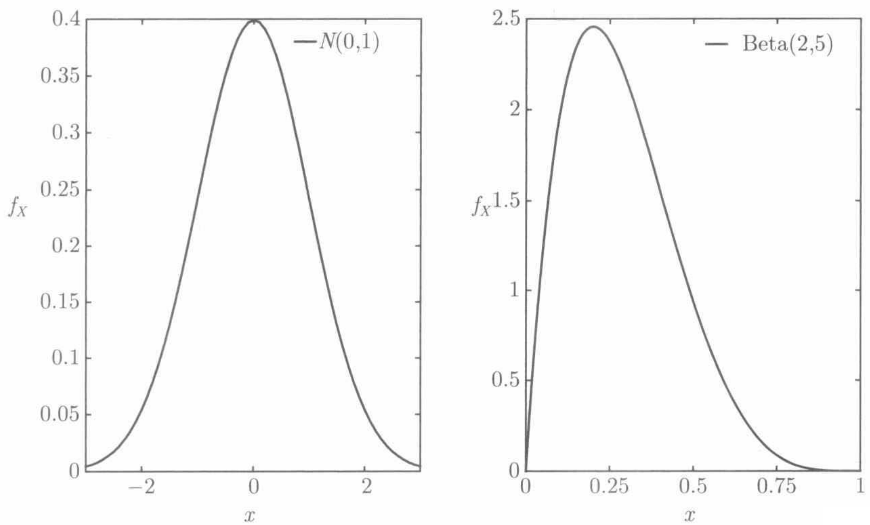

图 2.2　标注正态分布 $X \sim \mathcal{N}(0,1)$ 与贝塔分布的概率密度函数 $X \sim \mathrm{Beta}(2,5)$

1) 正态分布 (高斯分布)$X \sim \mathcal{N}(\mu, \sigma^2)$, $\mu \in \mathbb{R}$, $\sigma^2 > 0$，是极为重要的概率分布